

# ETHOS-Push

## An Approach toward Empathetic Orchestration of Scrutable Push



**Kieran Fraser**

Supervisor: Prof. Owen Conlan

School of Computer Science and Statistics  
University of Dublin, Trinity College

This dissertation is submitted for the degree of  
*Doctor of Philosophy*



## Declaration

I declare that this thesis has not been submitted as an exercise for a degree at this or any other university and it is entirely my own work, unless where otherwise indicated.

I agree to deposit this thesis in the University's open access institutional repository or allow the Library to do so on my behalf, subject to Irish Copyright Legislation and Trinity College Library conditions of use and acknowledgement.

---

Kieran Fraser  
November 2021





## Acknowledgements

The work completed in this Thesis would not have been possible without the generous support and guidance I have received from my family, friends and colleagues over the past number of years.

I would first like to thank my supervisor, Prof. Owen Conlan, for his unwavering support, advice and encouragement during this period. I would also like to thank my fellow students and colleagues within the ADAPT Centre and those within the wider Trinity community, with whom I have had the pleasure of sharing this challenging, but rewarding, journey - your friendship and camaraderie has been a great source of joy.

Finally, I would like to thank my family. In particular my parents, Barry and Edel, and my aunt, Therese, for every opportunity they have provided to me. Your encouragement, reassurance and support throughout this time has been invaluable.



## Abstract

There was a time, not so long ago, when our physical world would have seemed unfathomably large. But thanks, in part, to the meteoric rise of technology and its ability to connect far reaching parts of the universe, it has since diminished in size. The space and silence that was once infinite is now fast filling up with a noisy clamour borne of the virtual world, which is quickly expanding and, unless tempered, threatens to reduce the size of our physical world further, until it exists solely behind a softly glowing screen.

Mobile devices are ubiquitous in the modern world and rarely leave our side. Subsequently, interruptions are more frequent and attention increasingly difficult to protect. Push-notifications are an extremely valuable tool for disseminating useful and timely information, but their current use lacks the intelligence to tactfully adapt to the transient physical and cognitive contextual needs of subscribers, to whom they are delivered. Those pushing notifications resort to increasingly persuasive strategies to entice engagement and, at times, do so without empathy for the human to whom they are pushing. Therefore, subscribers of this technology require tools for protection, awareness and empowerment to retain ownership of their attention and make informed engagement decisions.

This thesis presents ETHOS-Push, an approach for creating, delivering and presenting notifications which are empathetic to subscribers and promote scrutability. The requirements of the ETHOS-Push approach were informed through a rigorous review of literature and observational studies of push-notification engagements in-the-wild. A sustainable, data-driven approach toward intelligent delivery was evaluated; well-defined principles of persuasion were used to engineer features for scrutability; and finally, overtly enticing text was moderated at the point of creation to reduce the exploitative influence of notifications employing information-gap strategies.



# Table of contents

|                                                        |           |
|--------------------------------------------------------|-----------|
| List of figures                                        | xv        |
| List of tables                                         | xxi       |
| Abbreviations                                          | xxvii     |
| <b>1 Introduction</b>                                  | <b>1</b>  |
| 1.1 Overview . . . . .                                 | 1         |
| 1.2 Problem Definition and Objectives . . . . .        | 3         |
| 1.3 Motivation . . . . .                               | 5         |
| 1.4 Research Questions . . . . .                       | 6         |
| 1.5 Thesis Contributions . . . . .                     | 7         |
| 1.6 Research Methodology . . . . .                     | 8         |
| 1.7 Thesis Structure . . . . .                         | 10        |
| 1.8 Publications . . . . .                             | 11        |
| 1.8.1 Conference Papers . . . . .                      | 11        |
| 1.8.2 Workshop Papers . . . . .                        | 12        |
| <b>2 Background</b>                                    | <b>13</b> |
| 2.1 Overview . . . . .                                 | 13        |
| 2.2 Methodology . . . . .                              | 13        |
| 2.3 How Are Notifications Used? . . . . .              | 19        |
| 2.4 What Impact Do Notifications Have? . . . . .       | 23        |
| 2.4.1 Interruptions . . . . .                          | 23        |
| 2.4.2 Attention . . . . .                              | 27        |
| 2.5 How Is Notification Technology Evolving? . . . . . | 29        |
| 2.6 Generative Methods . . . . .                       | 38        |
| 2.6.1 Synthea . . . . .                                | 40        |

|          |                                                                                      |           |
|----------|--------------------------------------------------------------------------------------|-----------|
| 2.6.2    | Synthetic Data Vault . . . . .                                                       | 41        |
| 2.6.3    | Synthpop . . . . .                                                                   | 43        |
| 2.7      | Limitations within Literature . . . . .                                              | 44        |
| 2.8      | Summary . . . . .                                                                    | 46        |
| <b>3</b> | <b>Understanding Notification Engagement In-the-wild</b>                             | <b>47</b> |
| 3.1      | Overview . . . . .                                                                   | 47        |
| 3.2      | Native Mobile Push . . . . .                                                         | 48        |
| 3.2.1    | Data Collection . . . . .                                                            | 48        |
| 3.2.1.1  | Background-Sensing . . . . .                                                         | 50        |
| 3.2.1.2  | Experience Sampling Method . . . . .                                                 | 52        |
| 3.2.2    | Participant Recruitment . . . . .                                                    | 53        |
| 3.2.3    | Privacy & Ethics . . . . .                                                           | 53        |
| 3.2.4    | Dataset . . . . .                                                                    | 54        |
| 3.2.5    | Context and Notification Engagement . . . . .                                        | 55        |
| 3.2.5.1  | The Role of Sender Relationship on Notification<br>Engagement . . . . .              | 56        |
| 3.2.5.2  | The Importance of Pushing with the Right App .                                       | 61        |
| 3.2.5.3  | Environmental factors . . . . .                                                      | 67        |
| 3.2.5.4  | Updated Notifications . . . . .                                                      | 72        |
| 3.2.5.5  | Changing Contexts . . . . .                                                          | 75        |
| 3.2.5.6  | Key Findings . . . . .                                                               | 81        |
| 3.2.6    | Smartphone Engagement Etiquette . . . . .                                            | 82        |
| 3.2.6.1  | Perceived Behaviour . . . . .                                                        | 82        |
| 3.2.6.2  | Smartphone Use in Company . . . . .                                                  | 84        |
| 3.2.6.3  | Hope, Smartphone Usage & Notifications . . . . .                                     | 86        |
| 3.2.6.4  | Key Findings . . . . .                                                               | 91        |
| 3.3      | Web Push . . . . .                                                                   | 92        |
| 3.3.1    | Data Collection . . . . .                                                            | 92        |
| 3.3.1.1  | Pre-study questionnaire . . . . .                                                    | 95        |
| 3.3.1.2  | Mid-study Questionnaire . . . . .                                                    | 96        |
| 3.3.1.3  | Post-study Questionnaire . . . . .                                                   | 97        |
| 3.3.2    | Participant Recruitment . . . . .                                                    | 97        |
| 3.3.3    | Privacy & Ethics . . . . .                                                           | 98        |
| 3.3.4    | Dataset . . . . .                                                                    | 99        |
| 3.3.5    | Understanding Participant Characteristics & Attitudes to-<br>ward Web Push . . . . . | 99        |

|          |                                                                   |            |
|----------|-------------------------------------------------------------------|------------|
| 3.4      | Limitations . . . . .                                             | 103        |
| 3.4.1    | Technological Limitations . . . . .                               | 103        |
| 3.4.2    | Experimental Design Limitations . . . . .                         | 104        |
| 3.4.3    | Experimental Execution Limitations . . . . .                      | 105        |
| <b>4</b> | <b>Generation &amp; Simulation of Push-Notifications</b>          | <b>107</b> |
| 4.1      | Overview . . . . .                                                | 107        |
| 4.2      | Predicting Notification Engagement Actions . . . . .              | 108        |
| 4.2.1    | Notification Data & Features . . . . .                            | 109        |
| 4.2.2    | Predicting Notification Engagement Action . . . . .               | 110        |
| 4.3      | Generating Synthetic Notification-Context Pairs . . . . .         | 114        |
| 4.3.1    | Generative Adversarial Networks . . . . .                         | 114        |
| 4.3.2    | Evaluation Metrics . . . . .                                      | 117        |
| 4.3.3    | Implementation, Synthesis Results & Discussion . . . . .          | 123        |
| 4.3.4    | Gym-push . . . . .                                                | 131        |
| 4.4      | Evaluating Synthetic Notifications In-the-wild . . . . .          | 131        |
| 4.4.1    | Experiment Approach . . . . .                                     | 132        |
| 4.4.2    | Results . . . . .                                                 | 133        |
| 4.4.2.1  | Synthetic Data In-the-wild . . . . .                              | 134        |
| 4.4.2.2  | Delivery Agents In-the-wild . . . . .                             | 137        |
| 4.5      | Limitations . . . . .                                             | 141        |
| 4.6      | Summary . . . . .                                                 | 142        |
| <b>5</b> | <b>Scrutable Push-notifications</b>                               | <b>143</b> |
| 5.1      | Overview . . . . .                                                | 143        |
| 5.2      | Engineering Persuasive & Scrutable Features of Push-notifications | 144        |
| 5.2.1    | Dataset . . . . .                                                 | 145        |
| 5.2.2    | Feature Engineering . . . . .                                     | 146        |
| 5.2.2.1  | Authority . . . . .                                               | 146        |
| 5.2.2.2  | Scarcity . . . . .                                                | 148        |
| 5.2.2.3  | Liking . . . . .                                                  | 149        |
| 5.2.2.4  | Social Proof . . . . .                                            | 150        |
| 5.2.2.5  | Commitment & Consistency . . . . .                                | 152        |
| 5.2.2.6  | Reciprocity . . . . .                                             | 153        |
| 5.2.3    | Total Persuasion . . . . .                                        | 155        |
| 5.2.4    | Offline Evaluation of Persuasive & Scrutable Features . . .       | 155        |
| 5.3      | Evaluating Persuasive & Scrutable Push In-the-wild . . . . .      | 159        |

|          |                                                                                            |            |
|----------|--------------------------------------------------------------------------------------------|------------|
| 5.3.1    | Experiment Approach . . . . .                                                              | 159        |
| 5.3.1.1  | Persuasive Feature Definition . . . . .                                                    | 160        |
| 5.3.1.2  | The SCRUPLE . . . . .                                                                      | 160        |
| 5.3.1.3  | Adaptive Presentation Templates . . . . .                                                  | 162        |
| 5.3.2    | Results . . . . .                                                                          | 165        |
| 5.3.2.1  | Persuasive Notification Engagement In-the-wild . . . . .                                   | 165        |
| 5.3.2.2  | Impact of Adaptive Presentation of Push-notifications . . . . .                            | 168        |
| 5.4      | Limitations . . . . .                                                                      | 169        |
| 5.5      | Summary . . . . .                                                                          | 170        |
| <b>6</b> | <b>Exploring the Effects of Enticing Notification Text Content</b>                         | <b>173</b> |
| 6.1      | Overview . . . . .                                                                         | 173        |
| 6.2      | Enticing Notification Text In-the-wild . . . . .                                           | 175        |
| 6.2.1    | WeAreUs Text Feature . . . . .                                                             | 175        |
| 6.2.2    | Extracting Notification Enticement . . . . .                                               | 178        |
| 6.2.3    | Notification Features Paired with Enticing Text . . . . .                                  | 179        |
| 6.2.4    | Notification Engagement Actions Toward Enticing Text . . . . .                             | 186        |
| 6.3      | Training a NMS with Notification Text . . . . .                                            | 186        |
| 6.4      | Enticing vs Summarised Notification Text . . . . .                                         | 189        |
| 6.4.1    | Procuring Notifications & Summarising Text . . . . .                                       | 190        |
| 6.4.2    | Exploring Enticing vs Summarised Text . . . . .                                            | 192        |
| 6.4.3    | Influence of Enticing vs Summarised Notification Titles<br>using Simulated Users . . . . . | 195        |
| 6.5      | Conditional Notification Text Generation . . . . .                                         | 196        |
| 6.5.1    | Language Model Background . . . . .                                                        | 197        |
| 6.5.2    | Language Model Implementation . . . . .                                                    | 198        |
| 6.5.3    | Enticing Notification Text Analysis . . . . .                                              | 199        |
| 6.5.4    | Influence of Conditionally Generated Notification Titles on<br>Simulated Users . . . . .   | 203        |
| 6.6      | Evaluation of Autonomous Notification Curation<br>In-the-wild . . . . .                    | 205        |
| 6.6.1    | Experiment Approach . . . . .                                                              | 206        |
| 6.6.2    | Enticement and Impact on Engagement . . . . .                                              | 208        |
| 6.6.3    | Participant Assessment of Curation Methods . . . . .                                       | 210        |
| 6.6.3.1  | Diversity . . . . .                                                                        | 211        |
| 6.6.3.2  | Informative . . . . .                                                                      | 211        |
| 6.6.3.3  | Credibility . . . . .                                                                      | 212        |



|          |                                                                     |            |
|----------|---------------------------------------------------------------------|------------|
| 6.7      | Limitations . . . . .                                               | 214        |
| 6.8      | Summary . . . . .                                                   | 214        |
| <b>7</b> | <b>Conclusions</b>                                                  | <b>217</b> |
| 7.1      | Summary of Contributions . . . . .                                  | 218        |
| 7.1.1    | Research Question 1 . . . . .                                       | 218        |
| 7.1.2    | Research Question 2 . . . . .                                       | 219        |
| 7.1.3    | Research Question 3 . . . . .                                       | 220        |
| 7.1.4    | Research Question 4 . . . . .                                       | 220        |
| 7.1.5    | Public Resources . . . . .                                          | 221        |
| 7.2      | Discussion . . . . .                                                | 221        |
| 7.3      | Future Directions . . . . .                                         | 223        |
| 7.4      | Coding Effort . . . . .                                             | 224        |
| 7.5      | Final Remarks . . . . .                                             | 226        |
|          | <b>References</b>                                                   | <b>227</b> |
|          | <b>Appendix A Preliminary: Exploring Open Data for Optimal Push</b> | <b>249</b> |
| A.1      | Overview . . . . .                                                  | 249        |
| A.2      | Open-source Notifications . . . . .                                 | 251        |
| A.2.0.1  | Nodobo . . . . .                                                    | 252        |
| A.2.0.2  | Worldwide Mobile App User Behavior . . . . .                        | 252        |
| A.2.0.3  | Device Analyzer . . . . .                                           | 252        |
| A.2.0.4  | CrowdSignals.io . . . . .                                           | 253        |
| A.2.0.5  | Mobile Data Challenge (MDC) . . . . .                               | 253        |
| A.2.0.6  | Friends & Family (F&F) . . . . .                                    | 253        |
| A.2.0.7  | Additional Update . . . . .                                         | 254        |
| A.3      | Synthetic & Augmented Notifications . . . . .                       | 254        |
| A.3.1    | Challenge . . . . .                                                 | 255        |
| A.3.2    | Proposed Solution . . . . .                                         | 255        |
| A.3.3    | Engineering Notification Features . . . . .                         | 257        |
| A.3.3.1  | Notification Date & Time . . . . .                                  | 257        |
| A.3.3.2  | Notification Sender & Ranking . . . . .                             | 257        |
| A.3.3.3  | Notification Subject & Ranking . . . . .                            | 259        |
| A.3.3.4  | Notification Application & Ranking . . . . .                        | 260        |
| A.3.3.5  | User Events/Activities . . . . .                                    | 261        |
| A.4      | Evaluating a Notification Management System . . . . .               | 264        |
| A.4.1    | NAbs Notification Management System . . . . .                       | 264        |

---

|                   |                                                          |            |
|-------------------|----------------------------------------------------------|------------|
| A.4.2             | Parameter Optimisation . . . . .                         | 267        |
| A.4.2.1           | Tuning FKB parameters . . . . .                          | 269        |
| A.4.2.2           | Tuning FMF parameters . . . . .                          | 269        |
| A.5               | Results . . . . .                                        | 269        |
| A.6               | Limitations . . . . .                                    | 270        |
| A.7               | Summary . . . . .                                        | 271        |
| <b>Appendix B</b> | <b>Supplementary Material for Chapter 3</b>              | <b>273</b> |
| B.1               | Smartphone Usage and Notification Engagement . . . . .   | 273        |
| B.1.1             | Participant Age, Apps and Notifications . . . . .        | 274        |
| B.1.2             | Notification Topics & Places of Engagement . . . . .     | 274        |
| B.1.3             | Notification Receptivity & Device Interactions . . . . . | 277        |
| B.1.4             | Elapsed Engagement Time & App Significance . . . . .     | 278        |
| B.1.5             | Key Findings . . . . .                                   | 282        |
| B.1.6             | The Role of the Sender . . . . .                         | 284        |
| B.1.7             | Pushing with the Right App . . . . .                     | 288        |
| B.2               | Additional Figures & Tables . . . . .                    | 290        |
| <b>Appendix C</b> | <b>Supplementary Material for Chapter 5</b>              | <b>305</b> |
| C.1               | SCRUPLE Persuasive Feature Definitions . . . . .         | 305        |

# List of figures

|      |                                                                                                                                                             |    |
|------|-------------------------------------------------------------------------------------------------------------------------------------------------------------|----|
| 2.1  | Analysis of co-authorship and keyword co-occurrence within the post-screened corpus of literature referencing push-notification research 2014-2020. . . . . | 17 |
| 2.2  | Review method for identifying push-notification literature. . . . .                                                                                         | 18 |
| 3.1  | Example of a geofence implemented using the Android SDK (image from Android's Documentation [And20c]) . . . . .                                             | 51 |
| 3.2  | Notifications recorded per day across all participants during the WeAreUs study period . . . . .                                                            | 55 |
| 3.3  | Method used to identify a significant contact at moments of notification delivery . . . . .                                                                 | 56 |
| 3.4  | Engagement behaviour of notifications associated with significant and insignificant contacts inferred at moments of notification delivery                   | 58 |
| 3.5  | Method of identifying app significance . . . . .                                                                                                            | 63 |
| 3.6  | Notification engagement behaviours across significant and insignificant apps . . . . .                                                                      | 63 |
| 3.7  | Distribution of notification category and priority level across significant and insignificant apps . . . . .                                                | 66 |
| 3.8  | Place types at moments of notification delivery and action . . . . .                                                                                        | 68 |
| 3.9  | Ambient noise at moments of notification delivery and action . . . . .                                                                                      | 71 |
| 3.10 | Ambient light at moments of notification delivery and action . . . . .                                                                                      | 72 |
| 3.11 | Method of tracking updated notification content implemented in WeAreUs app . . . . .                                                                        | 73 |
| 3.12 | Notification behaviour perception (1:VeryPoor; 2:Poor; 3:Neutral; 4:Good; 5:VeryGood) over notifications with differing levels of updates                   | 75 |
| 3.13 | The relationship between multiple notification updates and notification behaviours . . . . .                                                                | 76 |

|      |                                                                                                                                                                                                                                   |     |
|------|-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-----|
| 3.14 | Changes in battery level between moments of notification delivery and action . . . . .                                                                                                                                            | 79  |
| 3.15 | Changes in day between moments of notification delivery and action                                                                                                                                                                | 80  |
| 3.16 | Changes in ambient light category between moments of notification delivery and action . . . . .                                                                                                                                   | 81  |
| 3.17 | Scores of participant behaviour when acting on notifications of differing types (1:VeryPoor; 2:Poor; 3:Neutral; 4:Good; 5:VeryGood)                                                                                               | 83  |
| 3.18 | Scores of participant behaviour when acting on notifications in differing contexts (1:VeryPoor; 2:Poor; 3:Neutral; 4:Good; 5:VeryGood)                                                                                            | 84  |
| 3.19 | Scores for necessity of smartphone use in different contexts (1:VeryUnnecessary; 2:Unnecessary; 3:Neutral; 4:Necessary; 5:VeryNecessary)                                                                                          | 86  |
| 3.20 | Average State Hope responses over days of the WeAreUs Study . .                                                                                                                                                                   | 88  |
| 3.21 | Average State Hope scores per participant during the WeAreUs Study                                                                                                                                                                | 88  |
| 3.22 | Average State Hope scores recorded with different company present                                                                                                                                                                 | 91  |
| 3.23 | Pushd study design . . . . .                                                                                                                                                                                                      | 92  |
| 3.24 | The 3 distinct phases of the Pushd study evaluating different components of the ETHOS-Push framework . . . . .                                                                                                                    | 93  |
| 3.25 | The creation, presentation and delivery combinations for each phase of the Pushd study. . . . .                                                                                                                                   | 95  |
| 3.26 | Results from Mindful and Attention Awareness Scale (MAAS) and the Curiosity & Exploration Inventory (CEI-II) responses. . . . .                                                                                                   | 101 |
| 3.27 | Participant sentiment toward a) subscribing to web push-notifications b) sharing data for automated decision making; Participant identification of enticing text c) aggregated responses d) responses by participant age. . . . . | 102 |
| 3.28 | Most popular a) topic and b) website subscriptions . . . . .                                                                                                                                                                      | 103 |
| 4.1  | Algorithm performance predicting notification engagement actions                                                                                                                                                                  | 111 |
| 4.2  | Permutation feature importance when predicting notification engagement actions . . . . .                                                                                                                                          | 113 |
| 4.3  | Training a Generative Adversarial Network . . . . .                                                                                                                                                                               | 115 |
| 4.4  | Training a Conditional Generative Adversarial Network . . . . .                                                                                                                                                                   | 116 |
| 4.5  | WGAN-GP generator and critic loss convergence whilst training. .                                                                                                                                                                  | 125 |
| 4.6  | Comparison of synthetic data quality scores achieved by a selection of generative models (lowest scores indicate better synthetic data quality). . . . .                                                                          | 127 |

|      |                                                                                                                                                          |     |
|------|----------------------------------------------------------------------------------------------------------------------------------------------------------|-----|
| 4.7  | Comparison of synthetic data utility scores achieved by a selection of generative models (higher scores indicate better synthetic data quality). . . . . | 128 |
| 4.8  | Comparison of synthetic data privacy scores achieved by a selection of generative models. . . . .                                                        | 129 |
| 4.9  | Intelligent agents tasked with managing notification delivery. . . .                                                                                     | 132 |
| 4.10 | Mutual information between real notification engagement features (left) and synthetic notification engagement features (right) . . . .                   | 135 |
| 4.11 | Visualisations shown to participants when discerning real from synthetic data . . . . .                                                                  | 136 |
| 5.2  | F1 scores of selected algorithms when predicting notification action using differing feature data . . . . .                                              | 157 |
| 5.3  | Comparison of CTR of users using differing notification data sets. . . .                                                                                 | 158 |
| 5.4  | Screenshots of the SCRUPLE. . . . .                                                                                                                      | 162 |
| 5.5  | Notification presentation examples. . . . .                                                                                                              | 163 |
| 5.6  | Notification presentation examples. . . . .                                                                                                              | 163 |
| 5.7  | Distribution of persuasion within Web Push notifications. . . . .                                                                                        | 166 |
| 5.8  | Participant feedback on use of the SCRUPLE tool for enlightened push-notification engagement. . . . .                                                    | 168 |
| 6.1  | A word cloud illustrating the most commonly featured words found in the privacy-preserving text field collected during the WeAreUs study . . . . .       | 176 |
| 6.2  | Analysis of the privacy-preserving text feature from the WeAreUs notification data set . . . . .                                                         | 177 |
| 6.3  | Notification enticement classifier accuracy and loss . . . . .                                                                                           | 178 |
| 6.4  | WeAreUs notification text classified as neutral (left) or enticing (right) . . . . .                                                                     | 179 |
| 6.5  | Analysis of enticing notification text found in the WeAreUs data set                                                                                     | 179 |
| 6.6  | Feature importance rankings when predicting notification actions                                                                                         | 187 |
| 6.7  | Average performance metrics of DQN agents trained on notification text feature . . . . .                                                                 | 188 |
| 6.8  | Agent CTR for enticing vs neutral notifications . . . . .                                                                                                | 189 |
| 6.9  | Example of Google Alert notification feed . . . . .                                                                                                      | 190 |
| 6.10 | Word clouds of Alert data set. Left: Title text; Right: Summary text                                                                                     | 191 |

|      |                                                                                                                                                                                 |     |
|------|---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-----|
| 6.11 | Wordclouds of Alert notifications classified as enticing. Left: summarisation of the content of enticing notifications. Right: the title of the enticing notification . . . . . | 192 |
| 6.12 | Top bigrams and trigrams extracted from enticing notification texts of the Alert data set . . . . .                                                                             | 193 |
| 6.13 | Analysis of enticing notification text found in the Alert data set . . . . .                                                                                                    | 195 |
| 6.14 | DQN notification CTRs for enticing titles and summarised enticing notifications . . . . .                                                                                       | 196 |
| 6.15 | Fine-tuning GPT-2 on Alert notifications . . . . .                                                                                                                              | 198 |
| 6.16 | Inferred enticement distribution of enticing and neutral notifications generated by GPT-2 . . . . .                                                                             | 200 |
| 6.17 | Word clouds of GPT-2 data set. Left: conditioned as neutral; Right: conditioned as enticing . . . . .                                                                           | 201 |
| 6.18 | Top trigrams extracted from neutral and enticing notification titles generated by GPT-2 . . . . .                                                                               | 201 |
| 6.19 | Distribution of enticement over differing title topics generated by GPT-2 . . . . .                                                                                             | 203 |
| 6.20 | Sentiment analysis of GPT-2 generated notification titles . . . . .                                                                                                             | 204 |
| 6.21 | DQN notification CTRs for enticing and neutral GPT-2 generated notification titles . . . . .                                                                                    | 205 |
| 6.22 | Phase 1 of the Pushd web push-notification study . . . . .                                                                                                                      | 206 |
| 6.23 | Phase 1 of the Pushd web push-notification study . . . . .                                                                                                                      | 207 |
| 6.24 | Likert scores (5-point scale) of participants measuring the diversity of notifications from each curation method. . . . .                                                       | 211 |
| 6.25 | Likert scores (5-point scale) of participants measuring the informativeness of notifications from each curation method. . . . .                                                 | 212 |
| 6.26 | Credibility scores (1-10) of notifications from each curation method. . . . .                                                                                                   | 213 |
| 6.27 | Believability scores (1-4) of notifications from each curation method. . . . .                                                                                                  | 214 |
| A.1  | Timings of notification delivery in the synthetic data set drawn from real-world events . . . . .                                                                               | 258 |
| A.2  | Frequency of differing subject and sender types present in synthetic notifications . . . . .                                                                                    | 259 |
| A.3  | Frequency of differing apps present in synthetic notifications and differing event types occurring per participant . . . . .                                                    | 261 |
| A.4  | Web app created for exploring the synthetic notification data set . . . . .                                                                                                     | 263 |

|      |                                                                                                                                             |     |
|------|---------------------------------------------------------------------------------------------------------------------------------------------|-----|
| A.5  | Fuzzy membership function inferring importance of the relationship between notification recipient and sender . . . . .                      | 265 |
| A.6  | Preferred delivery distribution of work and social notifications during working hours (assumed 9am to 5pm) . . . . .                        | 266 |
| A.7  | Particle Swarm Optimisation Algorithm . . . . .                                                                                             | 268 |
| A.8  | Performance comparison between using original parameters and parameters learned using synthetic data set and PSO. . . . .                   | 270 |
| B.1  | Number of apps installed and frequency of notification categories received for participants of different age . . . . .                      | 273 |
| B.2  | Average number of notifications received per day and frequency of notification topics received per participant . . . . .                    | 275 |
| B.3  | Notifications delivered and engaged with over differing place types                                                                         | 276 |
| B.4  | Changes of place type between notification delivery and notification action . . . . .                                                       | 276 |
| B.5  | Number of device interactions per day (single participant) and the average unlock count per day (all participants) . . . . .                | 277 |
| B.6  | Elapsed time between notification engagement points of interest: seen, decision & response. . . . .                                         | 280 |
| B.7  | Participant app usage metrics . . . . .                                                                                                     | 280 |
| B.8  | Average relevance scores for notifications received by participants (1:VeryIrrelevant; 2:Irrelevant; 3:Neutral; 4:Relevant; 5:VeryRelevant) | 281 |
| B.9  | Frequency of LED colours implemented in notifications pushed by different apps . . . . .                                                    | 283 |
| B.10 | Frequency of vibration patterns implemented by different apps when pushing notifications . . . . .                                          | 284 |
| B.11 | Screenshots of the WeAreUs mobile application used for capturing notification engagement and mobile usage data . . . . .                    | 292 |





# List of tables

|     |                                                                                                                                                                                |    |
|-----|--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|----|
| 2.1 | Selection criteria and method for systematic review of push-notification technology . . . . .                                                                                  | 15 |
| 2.2 | Result of systematic review of push-notification literature . . . . .                                                                                                          | 16 |
| 2.3 | Literature describing <i>in-the-wild</i> studies of mobile notifications, context-aware computing and interruptibility 2017-2020. . . . .                                      | 30 |
| 2.4 | Summary of Notification features referenced in literature. . . . .                                                                                                             | 36 |
| 2.5 | Summary of Context features referenced in literature. . . . .                                                                                                                  | 37 |
| 2.6 | Summary of Usage features referenced in literature. . . . .                                                                                                                    | 38 |
| 3.1 | Engagement times for notifications with differing associated contact significance . . . . .                                                                                    | 59 |
| 3.2 | Observed frequencies and percentages of notifications responded to on different days with respect to contact significance . . . . .                                            | 61 |
| 3.3 | Results of Mann-Whitney U tests run to determine if there were differences between significant and insignificant apps on a number of continuous notification features. . . . . | 65 |
| 3.4 | Results of chi-square tests of homogeneity run to determine if there were differences in probabilities between apps inferred as significant and insignificant. . . . .         | 65 |
| 3.5 | Crosstabulation of app significance and notification category (significance accepted at $p < .001$ ). . . . .                                                                  | 66 |
| 3.6 | Place types in which notifications were received and the subsequent action taken . . . . .                                                                                     | 68 |
| 3.7 | Place types in which notifications were acted on (opened/dismissed)                                                                                                            | 69 |
| 3.8 | Results of Kendall's tau-b correlation for determining relationships between ambient light and notification engagement features . . . .                                        | 71 |

|      |                                                                                                                                                                                                                                                |     |
|------|------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-----|
| 3.9  | Identifying differences between opened and dismissed notification context states using test of two proportions (* denotes significant results; (F) indicates Fisher's Exact Test was used due to insufficient sample size)                     | 77  |
| 3.10 | The State Hope Scale presented to participants at moments they unlocked their smartphone                                                                                                                                                       | 87  |
| 3.11 | Results of Spearman's rank-order correlation for identifying relationships between average State Hope and notification engagement behaviours. ( del. = moment of delivery; act. = moment of action                                             | 90  |
| 3.12 | Final participant group allocations during each phase of Pushd study.                                                                                                                                                                          | 100 |
| 4.1  | Evaluation metrics, identified in literature and implemented in industry, which quantify Quality, Utility and Privacy of synthetic data.                                                                                                       | 118 |
| 4.2  | Differing model implementation hyperparameters used when generating synthetic push-notifications                                                                                                                                               | 123 |
| 4.3  | WGAN-GP model utility, quality and privacy scores for varying batch size ( $b$ ) and latent dimension ( $\ell$ ) (Top 10 at maximising utility of synth data). Best values achieved emboldened (max or min depending on metric).               | 124 |
| 4.4  | WGAN-GP model utility, quality and privacy scores for varying batch size ( $b$ ) and latent dimension ( $\ell$ ) (Top 10 at minimising quality loss between real and synth). Best values achieved emboldened (max or min depending on metric). | 125 |
| 4.5  | WGAN-GP model utility, quality and privacy scores for varying number of training epochs. Best values achieved emboldened (max or min depending on metric).                                                                                     | 125 |
| 4.6  | Utility, quality and privacy scores of WGAN-GP model trained using varying amount of real samples as seed. Best values achieved emboldened (max or min depending on metric).                                                                   | 126 |
| 4.7  | Quality, Utility and Privacy metrics of synthetic data generated during experiment.                                                                                                                                                            | 135 |
| 4.8  | The odds ratio of correctly distinguishing real notification engagement data.                                                                                                                                                                  | 138 |
| 4.9  | Chi squared goodness-of-fit test evaluating proportion of participants correctly distinguishing real from synthetic data.                                                                                                                      | 138 |

|      |                                                                                                                                                               |     |
|------|---------------------------------------------------------------------------------------------------------------------------------------------------------------|-----|
| 4.10 | Web push-notification engagement during deployment of intelligent delivery agents. . . . .                                                                    | 139 |
| 4.11 | Change in Net Promoter Score when intelligent agents deployed. .                                                                                              | 140 |
| 4.12 | Evaluating agents trained on synthetic vs real notification engagement data. . . . .                                                                          | 140 |
| 5.1  | Reciprocity scores for increasing elapsed time between app engagement and notification delivery using decay rate of 0.02. . . . .                             | 155 |
| 5.2  | Results of Spearman's rank-order correlation between mean feature scores and mean participant CTR. . . . .                                                    | 166 |
| 5.3  | Results of Friedman test to determine differences in participant CTRs for differing levels of persuasive notification. . . . .                                | 167 |
| 5.4  | Results of Wilcoxon signed-rank test to determine differences in participant CTRs when differing notification representations presented. . . . .              | 170 |
| 5.5  | Results of Kruskal-Wallis H test to determine differences in CTRs or NPS scores between groups presented with differing notification representations. . . . . | 170 |
| 6.1  | Clickbait and Non-clickbait examples . . . . .                                                                                                                | 178 |
| 6.2  | Observed frequencies of notifications, with and without enticing text, delivered through differing apps . . . . .                                             | 180 |
| 6.3  | Observed frequencies of notifications delivered, with and without enticing text, and their inferred topic . . . . .                                           | 182 |
| 6.4  | Observed frequencies of delivered notifications, with and without enticing text, and their assigned categories . . . . .                                      | 182 |
| 6.5  | Observed frequencies of delivered notifications, with and without enticing text, and their assigned LED colours . . . . .                                     | 183 |
| 6.6  | Observed frequencies of delivered notifications, with and without enticing text, and their assigned priority level . . . . .                                  | 184 |
| 6.7  | Observed frequencies of delivered notifications, with and without enticing text, and their assigned vibration pattern . . . . .                               | 185 |
| 6.8  | Enticing and neutral notification examples from the Alert data set                                                                                            | 191 |
| 6.9  | A sample of notification titles from topics generating largest proportion of enticing notifications . . . . .                                                 | 194 |
| 6.10 | The four released model sizes of GPT-2. Table source: [Rad+]. . .                                                                                             | 198 |

|                                                                                                                                                                    |     |
|--------------------------------------------------------------------------------------------------------------------------------------------------------------------|-----|
| 6.11 Enticing and neutral notification title examples extracted from the GPT-2 generated notification data set . . . . .                                           | 199 |
| A.1 Features recorded in the Friends & Family data set . . . . .                                                                                                   | 254 |
| A.2 Features used for design and evaluation of existing NMS . . . . .                                                                                              | 256 |
| A.3 F&F data features used for creating NMS compatible features . . . . .                                                                                          | 257 |
| A.4 Mapping of app categories to notification subjects. . . . .                                                                                                    | 260 |
| A.5 Mapping of event types to assumed relevant times. . . . .                                                                                                      | 262 |
| B.1 Observed frequencies and percentages of notifications responded to at different times of day with respect to contact significance . . . . .                    | 285 |
| B.2 Observed frequencies and percentages of notification categories with respect to contact significance . . . . .                                                 | 286 |
| B.3 Observed frequencies and percentages of notification topics with respect to contact significance . . . . .                                                     | 287 |
| B.4 Observed frequencies and percentages of notification <b>app</b> with respect to contact significance . . . . .                                                 | 288 |
| B.5 Crosstabulation of app significance and LED colour of notification (significance accepted at $p < .007$ ). . . . .                                             | 289 |
| B.6 Crosstabulation of app significance and vibration patterns of notification (significance accepted at $p < .007$ ). . . . .                                     | 290 |
| B.7 Crosstabulation of app significance and priority level of notification (significance accepted at $p < .008$ ). . . . .                                         | 291 |
| B.8 A list of notification engagement features recorded by the WeAreUs app . . . . .                                                                               | 291 |
| B.9 A list of aggregated smartphone usage metrics recorded by the WeAreUs app at hourly intervals . . . . .                                                        | 293 |
| B.10 A list of qualitative and quantitative features collected through the ESM questionnaire presented at moments of notification action . . . . .                 | 293 |
| B.11 A list of qualitative and quantitative features collected through the ESM questionnaire presented at moments participants unlocked their smartphone . . . . . | 294 |
| B.12 A list of demographic features gathered from participants when signing-up to the app/study . . . . .                                                          | 294 |
| B.13 Observed frequencies and percentages of notifications responded to at different place types with respect to contact significance . . . . .                    | 295 |

---

|                                                                                                                          |     |
|--------------------------------------------------------------------------------------------------------------------------|-----|
| B.14 Crosstabulation of app significance and notification topic (signifi-<br>cance accepted at $p < .005$ ). . . . .     | 296 |
| B.15 Pre-study questionnaire presented to participants of the Pushd study.                                               | 297 |
| B.16 The Mindful Attention Awareness Scale (MAAS) presented as part<br>of the Pushd pre-study questionnaire. . . . .     | 298 |
| B.17 The Curiosity and Exploration Inventory (CEI-II) presented as<br>part of the Pushd pre-study questionnaire. . . . . | 299 |
| B.18 Mid-study questionnaire presented to participants of the Pushd<br>study. . . . .                                    | 300 |
| B.19 Mid-study questionnaire presented to participants of the Pushd<br>study. . . . .                                    | 301 |
| B.20 Post-study questionnaire presented to participants of the Pushd<br>study. . . . .                                   | 302 |



# Abbreviations

|                   |                                            |
|-------------------|--------------------------------------------|
| <b>AI</b>         | Artificial Intelligence                    |
| <b>CTR</b>        | Click-through Rate                         |
| <b>DCR</b>        | Distance to Closest Record                 |
| <b>ESM</b>        | Experience Sampling Method                 |
| <b>ETHOS-Push</b> | EmpaTHetic Orchestration of Scrutable Push |
| <b>FIS</b>        | Feature Importance Similarity              |
| <b>GAN</b>        | Generative Adversarial Network             |
| <b>HR</b>         | Hitting Rate                               |
| <b>IQR</b>        | Interquartile Range                        |
| <b>KPI</b>        | Key Performance Indicator                  |
| <b>MAE</b>        | Mean Absolute Error                        |
| <b>MLP</b>        | Multilayer Perceptron                      |
| <b>ML</b>         | Machine Learning                           |
| <b>MSE</b>        | Mean Squared Error                         |
| <b>NLP</b>        | Natural Language Processing                |
| <b>NMS</b>        | Notification Management System             |
| <b>NM</b>         | Notification Management                    |
| <b>NPS</b>        | Net Promoter Score                         |

|                |                                                                  |
|----------------|------------------------------------------------------------------|
| <b>PC</b>      | Predictive Capability                                            |
| <b>RL</b>      | Reinforcement Learning                                           |
| <b>SCRUPLE</b> | SCRUtable Push LEdger                                            |
| <b>SRA</b>     | Synthetic Ranking Agreement                                      |
| <b>WGAN-GP</b> | Wasserstein Generative Adversarial Network with Gradient Penalty |
| <b>WGAN</b>    | Wasserstein Generative Adversarial Network                       |



# Chapter 1

## Introduction

### 1.1 Overview

The world has, over the past 12 months, undergone a significant societal, cultural and digital shift. The COVID-19 pandemic has challenged and transformed many aspects of our lives. Due to the very nature of the virus and the necessity for the population at large to distance themselves socially from one another to reduce its transmission and destructive influence, adoption and use of technology has not only flourished but become indispensable to daily life. Whilst the events of the last 12 months have certainly accelerated adoption of digital media across many lagging domains, the ubiquity of technology has been prevalent for some time already - a key example: the mobile phone. In developed countries, 91% claim ownership of a mobile device, followed by 90% in developing countries [Del17]. The number of unique mobile users grew from 4.53 billion in 2015 to 5.22 billion in 2020 [Kem21]. The opportunities and access to digital media provided by the mobile device means, as a society, we are more connected and capable of consuming vast amounts of content now, more than ever before.

This ubiquitous connectivity proved a lifeline during the numerous "lock-downs" that tested the reserve of many during the global pandemic. The technology facilitated the continuation of education [Dha20], work [Leo20], healthcare [WC] and social interaction [MM20b] with minimal disruption whilst the world adapted to a new socially distanced and remote way of life - elements of which may remain for years to come, particularly as we adapt and reap the benefits of technological

enhancements. A key feature in the early fight against the virus were mobile push-notifications. *Exposure Notifications* [Goo20b] were used to quickly and privately alert those who came into contact with a positive COVID case so they could act early in breaking the chain of transmission. This was a prime and powerful example of how contextual mobile data, BLE proximity sensors, could be harnessed and paired with push-notification technology to provide a timely intervention for the safety and well-being of the mobile user.

Push technology has itself become increasingly ubiquitous in recent years and can now be implemented not just in mobile devices, but in any web browser, through a smart watch, Internet of Things (IoT) enabled device, even in plants. An MIT team, using plant nanobionics, programmed spinach plants to transmit an email notification when certain chemicals were detected [Won+17]. The proof-of-concept design demonstrated that plants enabled with push technology could alert us to early signs of pollution or explosives and support early interventions, but it also indicates how we could soon be inundated with notifications pushed by organisms other than humans. Similarly, with the increased ubiquity of push-notification capabilities, the opportunities for businesses, brands and app owners to reach users with new information, offers and media rich experiences has also expanded. 66.7% of mobile users opt in to receive push-notifications from apps installed on their device [Par21a], although this differs wildly depending on the OS (Android opt-in rate: 91%; iOS opt-in rate: 44%) [Bon20]. Use of notifications increase engagement for brands and apps, but also satisfy the needs of the subscribers interacting with media online who, when surveyed, expressed reasons for engagement such as: *stay up-to-date with news and current events; find funny or entertaining content; fill up spare time; stay in touch with what my friends are doing; and make sure I don't miss out on anything ('FOMO')*.

The engagements users have with their mobile device and their incoming push-notifications have the potential to tell a rich story of the transient needs, priorities and aspirations of the user in varying context and cognitive states. Applications, in the past, had the ability to harness this information and throttle the influential impact of push-notifications to drive behaviour, such as in the case of Exposure Notifications. However, mobile users in recent years, thanks in part to the introduction of new privacy laws and campaigns against harvesting data for exploitative purposes, have become more cognizant of the value of their data which they impart to devices and the subsequent apps and services which utilise it. As such, the landscape for data-driven approaches for intelligent technologies

is changing, with focus turning toward consent, privacy, trust, transparency, explainability and data minimisation.

## 1.2 Problem Definition and Objectives

According to the *Digital 2021: Global Overview Report* [Kem21], the number of unique smartphone, internet and social media users rose by 1.8, 7.3 and 13.2 percent, respectively, from January 2020 to January 2021. 96% of internet users aged 16 to 64 own a smartphone, 64.4% a laptop or desktop computer, 34.3% a tablet and 23.3% a smart watch - devices all capable of pushing notifications. This also coincided with growth in the number of app downloads during 2020: 142.9 billion, an increase of 23.7% over 2019. Subsequently, time spent interacting with mobile apps also increased, with users spending on average over 4 hours a day on their mobile device - 10 to 19 hours a month on social media platforms such as *Facebook*, *WhatsApp* or *Instagram* and an average of 7 to 23 hours a month on streaming platforms such as *YouTube* and *Netflix*. Despite the expanding audience and increased appetite for digital media, the massive influx of apps available and smart devices capable of jostling for engagement has meant attention has never been a scarcer commodity. Unfortunately, a consequence of this is a war for attention, in which empathy for the human - to which attention belongs - is often disregarded in favour of metrics such as *impressions*, *likes*, *subscribes* and *follows*.

*When there is contention for attention, those who seek it turn to the most reliable attention getters: sex, hierarchy, calamity, and so forth.* [DB01]

Perhaps chief, in the arsenal of tools for grabbing attention, is the push-notification. Now ubiquitous across technologies (mobile, web, IoT), a push has the power to interrupt and grab attention at every moment of our life, due in part to the penetration of mobile technology and our attachment to it. Left unchecked, this has the potential to wreak havoc on our productivity [KPD16], relationships, mind and overall health and well-being [PH18]. Notifications pushed at opportune moments have the ability to inform, engage and provide real value to the subscriber. However, that value must be weighed against the cost of interruption. The context in which a push-notification is delivered is important when weighing its value versus its cost.

Research has shown that incoming notifications distract the user severely from their current tasks, regardless of whether the notification is given immediate attention [Tho+14; Lee+14b; Mar+15; War+17]. A significantly high influx of notifications has also been shown to resonate negatively with those receiving them [Sah+14]. Push notifications from various social media sites, email accounts and other mobile apps, compete continuously for precedence. The problem with this scenario is that it occurs in real-time, with no awareness or empathy for the consumer of the information [Oul+12; Smi+14]. Extreme or aggressive tactics are therefore resorted to by those pushing notifications to "win" the attention of the user [Liu+19]. Research pertaining to notification management and interruptibility attempts to explore end-user interactions with notifications pushed at users through various devices and attempts to identify receptive moments for messages to be delivered [CDM15; MHM17; HK; KS; Oko+15; Qin+14; HI05].

Many of these approaches for managing notifications are data-driven which can be difficult to sustain in the current climate where consent and regulation, both positive progressions for end-users of technology, are difficult waters to navigate when gathering granular, sensitive and big data to be used for autonomous decision making, such as during Notification Management. Many research studies invest time developing ad-hoc mobile applications for gathering mobile notification data due to the limited number of open source data sets available which contain the unique features required to test hypotheses [Meh+15b; KS; Oko+15; Qin+14]. There is also a lack of standard features, methods of collection, units of measurement and benchmark data sets with which to consistently build upon [Chu+15]. In particular, with respect to data minimisation, awareness of the contexts in which certain methods or features are significant and others not would aid researchers and app developers process only what is necessary, thus reducing the amount of data harvested from notification subscribers, protecting their sensitive information from potential exposure.

Consequently, the work proposed in this thesis seeks to find approaches to intelligent push-notification orchestration which is imbued with empathy for the subscriber such that: they are empowered to understand how notifications persuade; their data enhances more sustainable data-driven systems; and the data-driven systems are conditioned to better protect their best interests with regard to push-notifications. In light of this, the research objectives of this work are:

- **RO1:** Perform a state-of-the-art review of push-notification, interruption management and context-aware computing literature
- **RO2:** Understand influence and engagement with respect to push-notification arrival in varying contexts
- **RO3:** Design, implement and evaluate an approach toward sustainable production of push-notification engagement data
- **RO4:** Design, implement and evaluate an empathetic approach to creation, delivery, presentation and scrutability of push-notifications

## 1.3 Motivation

The first key motivation behind this work was to contribute an approach to push-notification management which adopts greater intelligence and empathy for the notification subscriber and harnesses the tools and techniques usually leveraged by those pushing notifications to better empower subscribers to understand and manage interruptions caused by their incoming notifications. Recent revelations by key industry leaders in the field of Human Computer Interaction, such as Jeff Seibert (Twitter) and Bailey Richardson (Instagram), have shone light upon the exploitative nature of technology which we use on a daily basis. Tristan Harris, an ex-Google product manager, pioneered the Time Well Spent movement after his internal presentation at Google regarding the destructive implications of nudging users to spend more time on their platform failed to incite real change in how technology, such as push notifications, were being used - exploitatively. The movement challenged the big tech conglomerates to adapt to a changing culture in which users would no longer be satisfied with being constantly steered or nudged (via notification) toward actions regardless of subsequent negative impacts on their health and well-being. Similarly, a recent Netflix documentary "The Social Dilemma" received critical acclaim for its examination and narrative around the influential depth and control exhibited by technology such as push-notifications and the role they play in addiction, anxiety and the spread of disinformation. Therefore this work aimed to contribute toward Humane Technology by inserting empathy at key points of the push-notification life-cycle: creation, presentation, delivery and reflection.

Another key motivator for this work was to contribute toward methods which could help sustain data-driven approaches to Notification Management whilst minimising the risks associated with exposing sensitive information, such as notification engagements, to intelligent NMSs systems, whose computational resources generally reside outside of subscriber control (in the cloud). Technology users, for good reason, have become increasingly wary of granting consent to intelligent systems which harvest data to support autonomous decisions. Additionally, due to the new General Data Protection Rules, there are also tighter restrictions around the processing of data which must be minimally fit for purpose. The new regulations also add further barriers for freely sharing data sets, such as notification-context pairs, which could be used for standardisation, collaboration, transfer learning and, of particular relevance to academia, comparative evaluation.

## 1.4 Research Questions

As the availability of user attention decreases and numbers of those seeking it grow progressively, increasingly persuasive and exploitative methods are leveraged to try and seize it, sometimes unbeknownst to end-users, using the medium of push-notifications. In order to facilitate empathy-driven approaches to protect subscribers, sustainable methods which utilise minimal user data need to be implemented.

Therefore, the primary thesis of this dissertation is that attributes describing notification engagement behaviour, derived and generated using minimal/synthetic data, can be used to autonomously orchestrate the tasks of empathetic push-notification creation, delivery and presentation whilst facilitating scrutability.

In support of this thesis, the following 4 research questions were proposed to investigate, through empirical studies, the design, implementation and evaluation of the ETHOS-Push approach:

- **RQ1:** How do subscribers act toward, perceive and become influenced by push-notifications in-the-wild?
- **RQ2:** Can a generative approach to push-notifications help facilitate intelligent notification management?

- **RQ3:** Do persuasive principles applied to push-notifications enhance scrutability?
- **RQ4:** Can autonomous moderation, at the point of notification creation, reduce overtly influential notifications?

## 1.5 Thesis Contributions

The work outlined in this thesis strives to contribute toward the body of research which is challenging modern technology, and specifically push-notification technology, to be and do better by the global population of users with which it engages. A mixed methods research methodology was adopted to inform a data-driven approach toward the sustainable, empathetic and scrutable orchestration of push-notifications. As such, one major contribution and two minor contributions are borne from this work.

The major contribution of this work is ETHOS-Push, an innovative approach toward EmpaTHetic Orchestration of Scrutable Push-notifications. The approach focuses on four significant points within the life-cycle of a push-notification (creation, delivery, presentation and reflection) and applies intelligence at these points to facilitate an empathetic experience for subscribers receiving push-notifications. An *empathetic* experience, in the context of this work, is one which satisfies the following requirements:

- **R1:** the notification content does not overtly entice engagement (intervention at creation)
- **R2:** past subscriber engagement data is used minimally to inform intelligent systems e.g. for optimal delivery (intervention at delivery)
- **R3:** notification presentation is personalised and adaptive to varying contexts (intervention at presentation)
- **R4:** notification influence is scrutable and understandable facilitating awareness (intervention at reflection)

The ETHOS-Push approach advances the state-of-the-art through the development, implementation and evaluation of each requirement described above.

Moreover, the experiments discussed in this thesis showed that overtly influential notification content had an impact on notification engagement and that, through identification and moderation at the point of notification creation, that same influence could be placated, protecting subscribers from being constantly enticed to engage. Experiments also demonstrated that NMSs, which are normally dependent on large amounts of subscriber data, could be assuaged with synthetic data generated to imitate real notification engagements, albeit with a slight performance loss for the sake of increased privacy. An application of a conditional generative model was also shown to identify preferred moments for contextual delivery of notifications which improved engagement whilst facilitating greater control. Experiments which examined persuasive principles with respect to push-notification engagement identified new features which contributed to the lure of certain nudges, and a tool developed to facilitate greater awareness of such persuasion, the SCRutable Push LEgger (SCRUPLE), was shown to alleviate the associated pressure to engage whilst empowering awareness.

The two minor contributions of this work are: the approach toward generating synthetic push-notification engagement data and the approach of extracting persuasive principles from push-notification engagements to facilitate scrutability through SCRUPLE.

In addition to advancing the state-of-the-art, the generative method for creating push-notification engagements was also deployed within the research community to facilitate comparative evaluation of NMSs and User Modeling. An open-source OpenAI Gym environment *gym-push* was developed and seeded with synthetic push-notification engagement data to enable Reinforcement Learning agents, tasked with managing push-notifications for subscribers, to be trained and evaluated through standardised means. The methods proposed within ETHOS-Push are also currently being evaluated within industry via a new start-up borne of this research, named EmPushy, which hopes to provide improved push-notification experiences to larger audiences.

## 1.6 Research Methodology

A mixed methods research methodology was adopted for the work described in this thesis. A rigorous literature review, described in Chapter 2, of the push-notification and broader HCI domain, identified a number of limitations and gaps within



the space which could be further investigated to advance the field. The review directed the focus of the work toward an observational study of push-notification behaviour in-the-wild in which qualitative and quantitative data was gathered from participants through ESM questionnaires and mobile sensors to enable further investigation of notification behaviour and support the design and evaluation of generative models applied to synthetic generation of push-notifications.

The in-the-wild method was selected as it facilitated the collection of data in its natural environment whilst also supporting querying participants. Thus mixing both unobtrusive participant observation and just-in-time questionnaire methods which have shown to be effective in the domain of push-notification research [Meh+17b].

Quantitative methods were subsequently used on the collected data to extract insights regarding the persuasiveness of the pushed content, its impact on subscribers and the role of contextual features on notification actions. The data was also used to design and evaluate a novel NMS and generative models in lab-based experiments. The lab-based simulation experiments facilitated evaluation in a low-risk environment and provided empirical evidence for the proposed methods contributing to an empathetic NMS. This in turn informed the requirements of an intervention experiment which could ultimately validate the efficacy of the proposed empathetic NMS in a real-world environment.

Once the NMS was validated offline, both qualitative and quantitative methods were then once again deployed in an in-the-wild intervention with a large number of participants to ascertain whether the proposed components of the empathetic NMS had a statistically significant impact on the behaviour of push-notification subscribers. Qualitative questionnaires were deployed to participants and push-notification engagement metrics were tracked unobtrusively in real-time. This method, found to be lacking in the domain of push-notification research to date, was explicitly called for in a recent survey of push-notification technology [MM17] and enabled the lab-based results to be validated in-the-wild as well as providing additional data to be shared within the research community. Quantitative analyses were subsequently carried out to identify the statistical significance of the results as well as extract insights on the behavior of participants with respect to push-notifications and validate the efficacy of the proposed approach toward empathetic orchestration of scrutable push-notifications.

## 1.7 Thesis Structure

**Chapter 2** explores the state of the art literature with respect to push-notifications. Existing implementations of push-notifications are reviewed and a variety of challenges associated with the technology are identified. Following this, recent solutions applied within the domain of notification management are evaluated. Subsequent obstacles prominent in the literature, such as data availability and ethical constraints, are further discussed.

**Chapter 3** introduces two studies exploring push-notification engagement behaviour in-the-wild. In the first, an Android app was developed and deployed on participant devices to unobtrusively monitor notification engagement patterns as participants went about their daily lives. A method of abstraction was introduced and applied to sensitive notification features for the preservation of privacy. An Experience Sampling Method was implemented to extract participant perspectives of engagement behaviour etiquette. In the second study, intelligent systems and agents were developed to orchestrate the creation, presentation and delivery of web push-notifications. Through analyses of the data collected, a number of significant insights were identified between participants, notifications and the transient contexts in which notifications are received.

**Chapter 4** shifts the focus towards the evaluation of intelligent solutions for improved notification engagement. A variety of algorithms for predicting notification engagement actions are explored and evaluated offline using the in-the-wild data presented in Chapter 3. Following this, a number of generative models are introduced for generating synthetic notification engagement data which can be freely shared, scaled and conditioned for a wide variety of scenarios. The synthetic data generated is evaluated with respect to its quality and preservation of privacy as well as its utility when leveraged by NMS algorithms predicting opportune moments. The value of the generative approach is demonstrated through implementation during in-the-wild study of web push notifications.

**Chapter 5** describes an approach for extracting persuasive features from notification engagements to further inform and improve the performance of intelligent NMSs whilst additionally empowering subscriber awareness through a tool for scrutable reflection of past notification engagements. The impact of the persuasive principles and scrutable tool are assessed through an in-the-wild study of web push-notifications.

**Chapter 6** explores the text content of push-notifications, the influence of enticing text with respect to engagement and the autonomous generation of notification text content. A proposed method of autonomous moderation of overtly influential text is then evaluated in-the-wild and the impact on engagement discussed.

**Chapter 7** reports the conclusions of this thesis by describing the research objectives and achievements, the contributions made to the state of the art and identifies future work and directions for this research.

## 1.8 Publications

The research presented in this thesis has contributed to the following published workshop and conference papers to date:

### 1.8.1 Conference Papers

- K. Fraser, B. Yousuf, O. Conlan, "A Context-Aware, Info-Bead and Fuzzy Inference Approach to Notification Management", IEEE 7<sup>th</sup> Annual Ubiquitous Computing, Electronics & Mobile Communication Conference (UEMCON 2016) [**Best Paper Award**] [FYC16]
- K. Fraser, B. Yousuf, O. Conlan, "An in-the-wild and synthetic mobile notification dataset evaluation", Intelligent Systems Conference (IntelliSys 2017) [FYC17a]
- K. Fraser, B. Yousuf, O. Conlan, "Scrutable & Persuasive Push-Notifications", 14<sup>th</sup> International Conference on Persuasive Technology (PERSUASIVE 2019) [FYC19b]
- O. Conlan, K. Fraser, L. Kelly, B. Yousuf, "A User Modeling Shared Challenge Proposal", 10<sup>th</sup> Conference and Labs of the Evaluation Forum (CLEF 2019) [Con+19]
- R. Sutton, K. Fraser, O. Conlan, "A Reinforcement Learning and Synthetic Data Approach to Mobile Notification Management", 17<sup>th</sup> International Conference on Advances in Mobile Computing & Multimedia (MoMM 2019) [SFC19]

### 1.8.2 Workshop Papers

- K.Fraser, B. Yousuf, O. Conlan, "Synthesis & Evaluation of a Mobile Notification Dataset", EvalUMAP Workshop, 25<sup>th</sup> Conference on User Modeling, Adaptation and Personalization (UMAP 2017) [FYC17b]
- K. Fraser, B. Yousuf, O. Conlan, "Generation and Evaluation of Personalised Push-Notifications", EvalUMAP Workshop, 27<sup>th</sup> Conference on User Modeling, Adaptation and Personalization (UMAP 2019) [FYC19a]
- K. Fraser & O. Conlan, "Enticing Notification Text & the Impact on Engagement", UbiTtention Workshop, UbiComp 2020 [FC20]
- A. Madhusoodanan, A. Kumar M, K. Fraser, B. Yousuf, "Machine Learning Approach to Manage Adaptive Push Notifications for Improving User Experience", Workshop on Artificial Intelligence for Mobile and Ubiquitous Communication System, MobiQuitous 2020 [accepted]

# Chapter 2

## Background

### 2.1 Overview

This chapter describes a review of literature relating to push-notification technology and related fields. There were a number of objectives related to this review. The first objective was to identify where push-notification technology sits in the landscape of computer science literature. The second objective was to provide a systematic review of push-notification technology to capture past, current and potential future trends expressed in the state-of-the-art. The final objective was to identify gaps within push-notification literature which deserved further exploration, thus lending support and credence to the work subsequently undertaken and described in this thesis.

### 2.2 Methodology

The PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) scheme [Moh+09] was used for carrying out a systematic review of push-notification technology presented in this chapter. The ACM Guide to Computing Literature was selected as the source for the review as it is one of the primary and most comprehensive literature databases in the field of Computer Science, containing just over 2.9 million publications [ACM].

The publications included in this review were sourced at three time points. The first was January 2017, which included a filter criteria, *Years*, ranging from 2014-2016. The second time point was January 2020, which included a filter criteria, *Years*, ranging from 2017-2019. The final time point was January 2021 in which a single year criteria, 2020, was selected. As the three reviews were identical in all but year, the results have been combined and are presented as one all-inclusive review, ranging in *Year* from 2014-2020. This time frame was selected as:

- Other researchers in the domain of push-notifications and mobile HCI have carried out rigorous surveys on the evolution of push technology and context computing. Kunzler et al. [KKK17] completed an extensive survey ranging from years 1994 to 2017 in which they traced the origins and growth of *Context-aware Notification Management Systems* (CNMSs). Mehrotra et al. published, an extensive survey of push-notification technology in their book *Intelligent Notification Systems: A Survey of the State of the Art* [MM17] - dates were not explicitly stated for the review, but the work was published in 2017. Furthermore, Church et al. completed a survey of smartphone usage studies between years ranging from 2007 to 2014 [Chu+15]. Therefore, it was decided for this review to concentrate on years 2014-2020, as it would facilitate a small overlap with previous reviews (2014-2017) whilst also facilitating inclusion of new state-of-the art studies (2017-2020).
- Whilst push technology first became tangible in 2001 when Blackberry's Research in Motion (RIM) team implemented the first push service for alerting users to new email [MC06], mobile push technology was only adopted in 2009 (Apple) and 2010 (Google). *Rich notifications*, those which permit multimodal content, were just created in 2013/2014. Therefore, this technology is still in its infancy and, as such, studies examining the behaviour of users interacting with push-notifications fall to more recent years.

It should also be noted that, whilst the years 2014-2020 were selected for review, the snowball method [Woh14] was used to also include papers of significance that fell outside of this range. Table 2.1 demonstrates the selection criteria used during the review.

The purpose of this literature review was to identify relevant work in the field of computer science in which push-notification technology was a significant factor i.e. the focal point of a study, a supportive technology (e.g. behavioural

Table 2.1 Selection criteria and method for systematic review of push-notification technology

|                          |                                                                                                                                                                                             |                                                                                                                                                       |             |                         |
|--------------------------|---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-------------------------------------------------------------------------------------------------------------------------------------------------------|-------------|-------------------------|
| <b>Review Question</b>   | <i>How is notification technology applied, assessed and improved across computer science literature?</i>                                                                                    |                                                                                                                                                       |             |                         |
| <b>Literature Search</b> | <b>Source</b>                                                                                                                                                                               | The ACM Guide to Computing Literature                                                                                                                 |             |                         |
|                          | <b>Query</b>                                                                                                                                                                                | <i>Title:("notification" OR "push-notification" OR "push notification") OR Keyword:("notification" OR "push-notification" OR "push notification")</i> |             |                         |
| <b>Filter Criteria</b>   | <b>Years</b>                                                                                                                                                                                | 2014-2020                                                                                                                                             | <b>Type</b> | <i>Research Article</i> |
| <b>Exclusions</b>        | Titles and abstracts were examined to determine the theme, scope and contributions of each work. Those found not to include notification applications, assessments or impact were excluded. |                                                                                                                                                       |             |                         |
| <b>Evaluation</b>        | The full-text assessment included only articles which used, assessed or improved notification or related push technology.                                                                   |                                                                                                                                                       |             |                         |

interventions) or a confounding factor (e.g. smartphone/app engagement). The review question aimed to capture a holistic view of push-notification technology spanning multiple technology embedded domains including, but not limited to: Human-Computer Interaction; Mobile Computing; User Modeling; Artificial Intelligence; Personalisation; Persuasive Technology; and Digital Health. As such, the question which this review provides evidence toward is: *How is notification technology applied, assessed and improved across computer science literature?*

Minimal search terms were chosen to target notification literature explicitly in order to maintain a broad scope. A *notification* is defined in Merriam-Webster as *a written or printed matter which gives notice* [Mera]. With regard to technology, a notification is an alert used by devices to draw user attention to, generally new, information. The term "push-notification" is mostly used synonymously with "notification", however it can also be considered a sub-category, as the "push" distinguishes it as a piece of information that has been sent by a remote server, as opposed to being spawned locally on a device. For this reason, both "notification" and "push-notification" variants were used to define the parameters of the search query. Additionally, it was required that either term be present in the *Title* or as an *Author Keyword* to be applicable for review to ensure relevant publications were selected. The publications were filtered by year and content type. The total

Table 2.2 Result of systematic review of push-notification literature

| Year         | # Publications | # Research Articles | # Post-screening |
|--------------|----------------|---------------------|------------------|
| 2020         | 65             | 31                  | 25               |
| 2019         | 62             | 44                  | 39               |
| 2018         | 325            | 133                 | 35               |
| 2017         | 410            | 127                 | 32               |
| 2016         | 539            | 178                 | 38               |
| 2015         | 610            | 151                 | 26               |
| 2014         | 589            | 375                 | 28               |
| <b>Total</b> | 2600           | 1039                | 223              |

number of publications each year matching the selection query and the subsequent portion which were identified as *Research Articles* are shown in Table 2.2.

A total of 2,600 publications were identified, of which, 1,039 were research articles. Screening of the title and abstract of each record yielded the further exclusion of 816 records due to irrelevance with respect to push-notifications or a determination that no significant push-notification contributions were present (e.g. notifications were not the focus of the work and subsequently were not measured or adequately described or reported). Where applicable, when a determination could not be sufficiently made from the title or abstract, the full-text content was used to provide further evidence. Examples of titles which were excluded are:

- "Protecting Data on Smartphones and Tablets from Memory Attacks" [Col+15]
- "Annulus: A Dual Congestion Control Loop for Datacenter and WAN Traffic Aggregates" [Sae+20]
- "Dualdb: An Efficient LSM-based Publish/Subscribe Storage System" [QH17]

The resultant corpus after screening comprised 223 research papers. Figure 2.1 visualises the co-authorship and keyword co-occurrence discovered in the screened corpus of publications. Of the 703 authors present in the corpus, 73 co-authored (i.e. with another author present in the corpus) at least 2 publications (Figure 2.1a). From the visualisation, a core group of authors can be identified as having made significant contributions to push-notification technology since 2014. The largest set of connected co-author items discovered was 36 (Figure 2.1b).



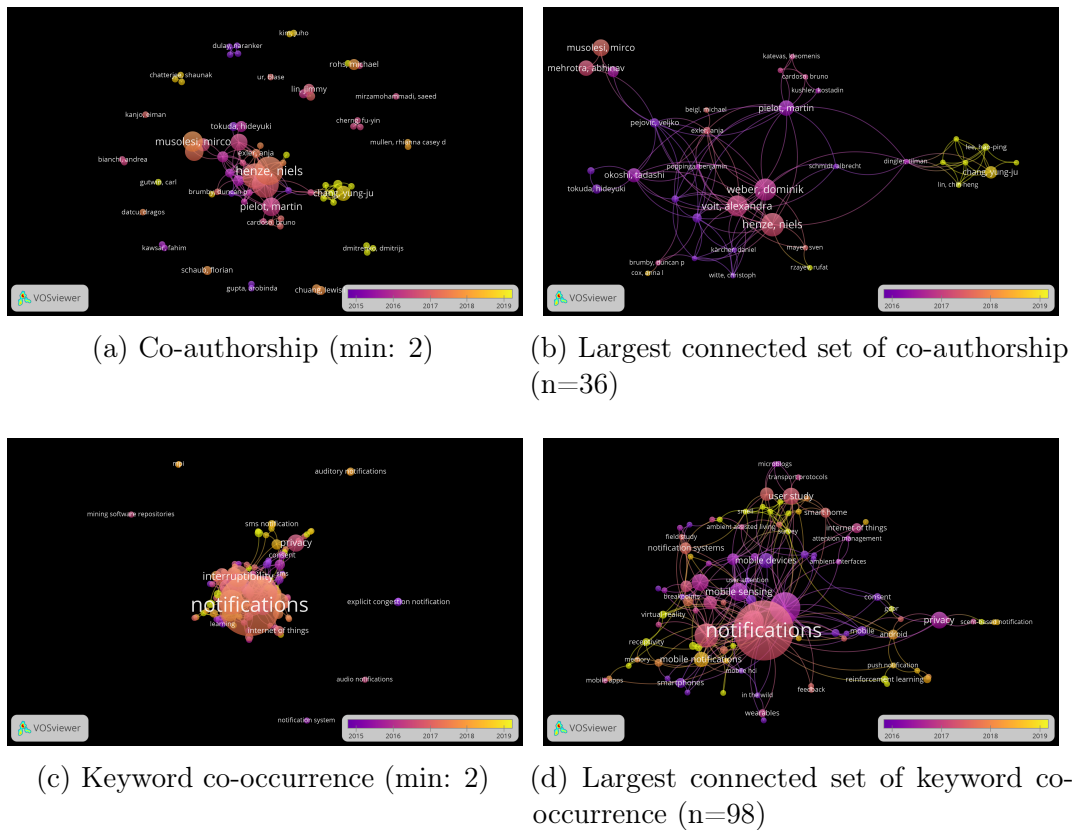


Fig. 2.1 Analysis of co-authorship and keyword co-occurrence within the post-screened corpus of literature referencing push-notification research 2014-2020.

Of the 702 keywords identified in the corpus, 109 met a co-occurrence threshold value of 2 (Figure 2.1c). A large core group of connected keywords was also identified in the literature, indicating a consistent narrative. The largest set of connected keyword items found was 98 (Figure 2.1d). Most prominent occurrences were references to "notifications" or "mobile devices" which was unsurprising given the domain of study, however the visualisations also highlight that the literature has focused on "virtual reality", "ambient notifications", "gdpr" and "reinforcement learning" in recent years. Other notable themes identified include "privacy", "consent", "receptivity", "mobile sensing" and "attention management".

The eligibility of each record was assessed for inclusion in the final review by evaluating the full-text of each article with respect to its significance toward *application*, *assessment* or *improvement* of push-notification technology:

- Application - notifications were applied to assist the research discussed in the article, but were not necessarily the main focal point e.g. used to deliver questionnaires in ecological momentary assessments, used as part

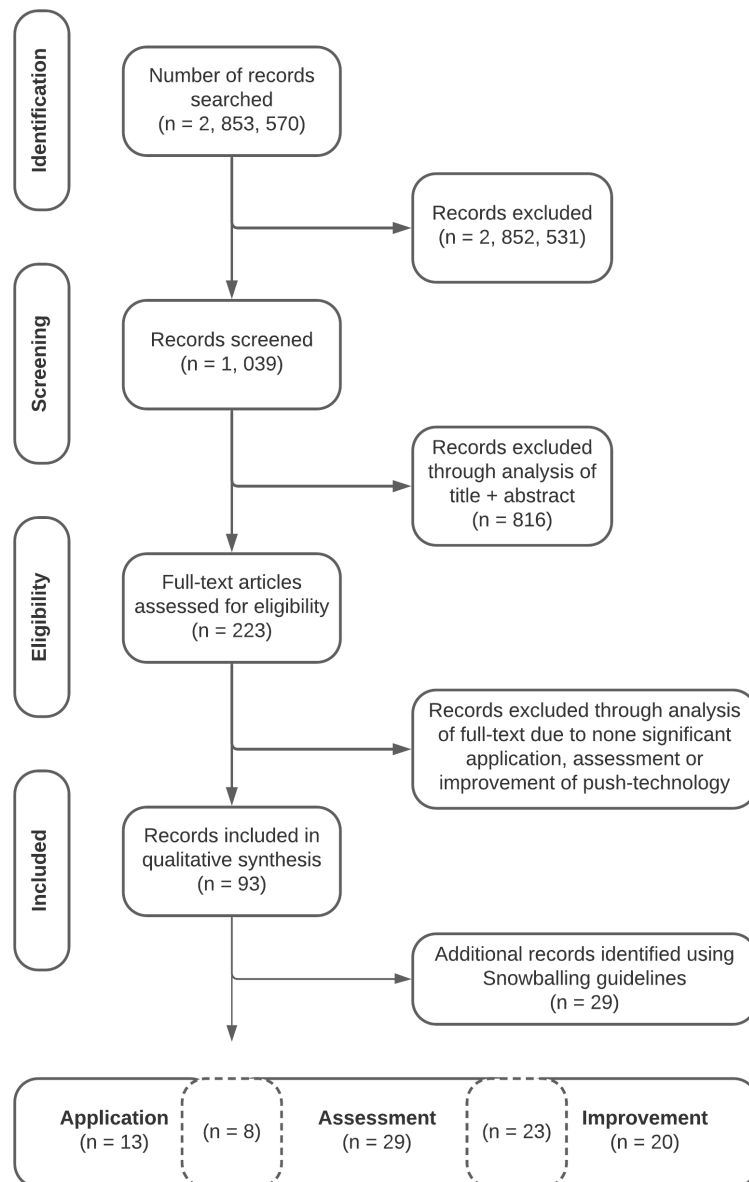


Fig. 2.2 Review method for identifying push-notification literature.

of behaviour interventions or applied as supportive technology in fields otherwise unrelated to technology.

- **Assessment** - notifications were a primary (or secondary) research focus of the article and their effect and/or behaviour was studied (no intervention) e.g. observational study of notifications in-the-wild.
- **Improvement** - notifications were a primary research focus of the article and improvements for the creation, delivery, presentation, interaction or measurement of notifications were presented.

Subsequently, 93 articles were selected for qualitative synthesis. The diagram in Figure 2.2 illustrates the method undertaken for identifying relevant literature for inclusion in the qualitative synthesis following PRISMA guidelines [Moh+09] and identifies the number of articles which fell into the respective category of application, assessment and improvement (Note: some records were found to belong to two categories e.g. notifications were applied as assistive technology and their effects were studied; notifications without interventions were assessed in addition to presentation of an improvement to the technology).

## 2.3 How Are Notifications Used?

Notifications are ubiquitous. They originated as a method of prompting users to check their email for new deliveries and have since been adopted by mobile, wearable and IoT technology for enabling apps, businesses and other users push information to other connected devices for a myriad of reasons.

Paulk et al. [PKA14] discuss a study in which they apply push-notifications to enable users be alerted when their favorite restaurant meals are being served in nearby locations. *TigerBites* autonomously scraped menus of the dining locations relevant to the Princeton University community and compared the results to a user's preferred meal options, subsequently sending email notifications to the students with location details of the restaurants serving their preferred breakfast, lunch and dinner dishes.

Consel et al. [CDS15] applied a *Unifying Notification System* to assistive services for the elderly as part of an assisted living platform. The notifications were used for such things as:

- Pill prompts - ensuring users remembered to consume medication
- Bill management - alerting users to upcoming bills
- Stove monitor - alerting user if the stove is unintentionally left unattended
- Social events - alerting user to senior social events

A six-month study of the platform deployed with 15 older adults interacting with the notifications showed high satisfaction scores (reported using the Quebec User Evaluation of Satisfaction with Assistive Technology questionnaire), particularly for those with poorer cognitive function.

Chen et al. [Che+16] proposed a *Content-oriented Notification Service* for use by different emergency service responders during disasters. Communication difficulties were identified when the services were communicating through their individual administrative methods. Therefore, a centralised notification system was proposed to simplify the communication by exposing a framework with which the responders could be assigned roles within the context of the disaster and push and receive notifications based on their dynamically assigned teams and roles.

Li et al. [Li+16] launched a large-scale study (760,935 incidents) regarding web server vulnerabilities to evaluate whether push-notifications could aid webmasters, and the public at large, by *reducing the duration a site remains compromised*. Direct communication, using push-notifications, significantly improved remediation when a compromise was detected and actually *decreases the time attackers retain access to a site*. The type of notification was significant in this context, with informative email notifications from an authority, WHOIS, resulting in a greater likelihood of remediation (54.6%) and direct alerts, via the Google Search console, improving the likelihood even further (+75%). As a result, authors called for greater coverage of highly concise and informative notifications to scale and improve remediation of hijacking incidents on the web.

Bashirieh et al. [Bas+17] applied notifications in the workplace to entice engagement with crowdsourcing campaigns advertising tasks such as collaborative design or knowledge creation. The notifications pushed from the *E-Crowd* application were designed to be contextually relevant with respect to employee break time and activity (e.g. walking) in order to maximise the likelihood of crowdsourced task completion. When deployed over a 2 month period with 83 employees it was found that timely notifications improved retention and participation in the tasks advertised. Significantly, the authors also acknowledged that 30% of task executions were completed outside office hours, which raised ethical concerns over the "right to disconnect". This in particular has become increasingly prominent over the past year as a large portion of the global workforce has been forced to work from home due to the COVID-19 pandemic, causing distinguishing lines between work and home life to become increasingly blurred.

Yaman et al. [Yam+17] discuss the ethical implications regarding the timing of notifications sent to users by software companies who wish to notify of intended use of data for experimentation purposes. 130 surveys were distributed to 4 different companies and between employees of differing roles. Results indicated there was a consensus that a notification should denote if and when Personally Identifiable Information (PII) was being used in experiments but other scenarios

not including PII caused division of opinion with regard to when a user would receive a notification, before or after the experiment had taken place. In particular, participants in management roles were cautious and tended toward always sending a notification to a user regarding inclusion in experimentation, whereas UX designers were found to favour sometimes misleading users during the experiment *if the validity of the results depend on it*. This work was particularly interesting as it highlights the ethical culture from wherein notifications are being pushed at scale. In this context, the notification was regarding inclusion in experimentation (e.g. A/B tests), but this begs the question whether the same ethical culture can be found in other such businesses pushing promotional notifications or notifications designed to entice use of an application and whether similar tactics of misdirection are considered for the sole purpose of achieving valid results, potentially at the expense of the end-user.

Samonte et al. [Sam+18] applied notification technology to ease administrative burdens in hospitals when disseminating information, such as test results, to patients. SMS notifications were used in conjunction with an online portal to notify patients when test results were available for online viewing - accessible via a patient ID and PIN number which were also pushed to them via SMS in order to reduce patient travel costs and registration times. Contextual rules were built into the *Online Hospital Document Management System* so that test-result notifications would only be pushed to outpatients, as the doctors of in-patients would already have access to the information and would share the information when necessary. The authors also evaluated the systems tangible benefits, such increased revenue and man-hours saved, as well as intangible benefits of an enhanced user experience and overall better healthcare provided.

Silvis et al. [SSL18] administered a notification system for alerting students-in-need when leftover food was available in order to prevent food waste. The proposed design would selectively prioritise students who may be facing food insecurity with outlets located at the University of Pittsburgh that had a food surplus destined for landfill sites. Of particular importance was the fairness of prioritizing notifications such that the students most in need were given the opportunity to attend events with surplus food first (i.e. they were first to receive a push-notification), as well as ensuring there was enough surplus food for those students who did receive a push (invite). The Fair (food) Insecure Probability Score value model was proposed to selectively determine which users received a push-notification and in what order. This use-case highlights the temporal relevancy of a push-notification. In this case, the "pusher" took responsibility to ensure that relevant notifications were

delivered in a fair and empathetic manner. However, this is not always the case in-the-wild. For instance, an airline pushing a seat sale promotion via notification to its massive audience of mobile subscribers is unlikely to apply similar empathy, but would instead either: send to all subscribers; or send to subscribers most likely to avail of the sale, whether or not purchasing the flight is in their best interest.

A common widespread application of push-notification technology is within mobile and web applications for the purpose of increasing user engagement with a platform and improving Key Performance Indicators (KPIs), such as page impressions (e.g. number of users viewing content) or conversion rates (e.g. number of users signing up to a newsletter or making a purchase). Airship, a Customer Engagement Platform, published recent statistics which showed that *average notification direct open rates increased 29% since the beginning of the COVID-19 pandemic* [Par21b]. Their analysis also illustrated that the context of delivery influenced the preferred communication channel (e.g. email vs SMS vs app notifications) of customers when they were being pushed information, with 32% deciding to unsubscribe if a poor push experience occurred (e.g. an out-of-context, disruptive or overly aggressive/enticing notification pushed).

Parker and Wang [PW16] explored what motivating factors influenced customers to engage with an app deployed within the domain of fashion retail. Their study, including qualitative interviews with 18 participants regarding their retail app engagements, found that *personalized services, convenience* and *efficiency* were the leading motivational factors for use. In particular, 9 participants indicated notifications regarding discounts were a motivating factor. However the channel in which notifications were received differed, with 6 choosing push-notification and 9 choosing email as their preferred channel of communication. Context and noisy/annoying notifications were factors which prevented participants from being willing to subscribe to notifications. The authors concluded that analysis tools such as *Google Analytics* should be used to identify preferred contexts in which to deliver notifications as *placing a targeted trigger in the path of a motivated person at the time of day they are most likely to be willing and able to engage* would improve the receptivity and use of retail apps and subsequently improve conversions.

As illustrated in these studies, notifications have been applied across many fields of interest: restaurant & dining; assisted living; emergency services; cybersecurity; workplaces; research; healthcare; food waste; and e-commerce, to name but a few. The benefits of this technology are clear, but they also draw further questions into light - such as those around privacy, consent and ethics. Many more

applications of push-notification technology exist in-the-wild resulting in a myriad of alerts ringing out, devices vibrating, lights flashing and other modalities set alight all to attract the fickle and fleeting attention of subscribers. The following section explores the wide-ranging impact these push-notifications have on wider society.

## 2.4 What Impact Do Notifications Have?

The examples of push-notification applications discussed in the previous section illustrate clearly that push-technology provides real value and is a benefit across multiple domains. But what are some of the tangible and intangible costs to using push-notifications? Do they outweigh the benefits? This section explores the impact of pushed notifications in greater detail and from multiple perspectives through the discussion of literature which observed and assessed the creation, delivery, presentation and measurement of notifications and associated behaviours.

### 2.4.1 Interruptions

Since the creation and evolution of the smartphone, and additional wearable technology soon thereafter, we now find ourselves physically, if not also emotionally, tethered to technology at almost every moment of the day. A consequence of this: technology has the opportunity and capability, through push-notifications, to interrupt us at any juncture. Most significantly, the "pusher" generally holds the discretion at which we are interrupted.

Cialdini, in his book *Influence: The Psychology of Persuasion* [CC07], ruminates over how a phone call ringing out whilst mid-conversation with a friend is not only a distraction, but also overwhelmingly enticing to answer, as with each ring, the opportunity to learn the new piece of information which it hides, becomes smaller and smaller.

Böhmer et al. [Böh+14] designed a mobile application, *CallHeads*, in 2013 which tracked interruptions made by incoming phone calls to evaluate the effects of using a non-obtrusive contact bubble UI in place of a full-screen notification for answering calls. The app was released in-the-wild and subsequent data was collected on 525 users for  $\approx 30$  days, during which time 88,516 incoming calls were identified. This translated to 3.3 calls, or interruptions, on average per user

per day. Considering that each interruption negatively impacts productivity and task completion [SB10; MN86; GM04], this highlights a growing concern and an argument for reducing our attachment to our smartphones, at least during hours in which we wish to be productive. The *CallHeads* app enabled the researchers to identify calls of users which, most significantly, interrupted them during use of their smartphone e.g. whilst typing a message or playing a game. It was found that 33.66% of incoming calls were interruptive in this context and 55.77% were answered, demonstrating the persuasive power discussed by Cialdini. The authors also identified another interesting behaviour - only 7.99% of the calls were explicitly dismissed (i.e. the user declined to answer before the call rang out), whilst 36.24% were allowed to ring out. This suggested a reluctance to explicitly snub the caller despite not wishing to take the call. The authors define this new behaviour as *passive decline*.

This attendance behaviour, or lack thereof, has also been identified with incoming notifications. Weber et al. [Web+18] undertook a large-scale, in-the-wild, observational study of the push-notification drawer of 3,953 mobile devices to explore how users interacted with alerts posted to the notification tray of their device. Their investigation showed that 3.4 notifications, on average, remained pending explicit attendance on user devices and that the type of notification category influenced the duration of nonattendance. Further work by Weber et al. uncovered reasons why notification attendance might be deferred. The *Snooze* app was deployed in-the-wild on 295 user devices enabling deferral of notifications for a user-defined set period of time. The study identified the main types of notifications deferred were SMS/Instant-message and Calendar/Reminders and that they were mainly deferred for attendance later that same day. A further smaller (n=16), but deeper, study with the same application yielded reasons for deferral as participants were interviewed after 7 days interacting with the app. Reasons for deferring a notification, as opposed attending it or explicitly dismissing it, included:

- not wanting to attend to a notification during work, but wishing to attend to it later (e.g. before and after work were found to be popular deferral periods)
- not wanting to switch current context to that of the notification, but wishing to attend later



- wanting to be reminded of a notification so an associated task would not be forgotten
- wanting to attend to the notification on a different device (e.g. arrived on smartphone, but wishing to attend to it on a desktop device)
- needing to complete other tasks before the notification could be attended

During interviews, 6 participants also stated they perceived being *never or rarely* interrupted by notifications, 7 felt they were *sometimes* interrupted and 3 felt they were *often* interrupted whilst all also admitting their responsiveness to notifications was generally immediate, particularly for *communication-related notifications*.

Similarly, Pielot et al. [PCD14] identified in their *In-Situ Study of Mobile Phone Notifications* involving 15 users over a one-week period, that users tended to receive 65.3 notifications on average per day and that instant messenger and email apps were reported to be the main source of perceived interruptions. The authors found that larger numbers of incoming notifications from messenger applications correlated with stronger feelings of *having to deal with a lot of messages and being overwhelmed*. In addition, qualitative reports from users indicated that social pressure was an influencing factor of notification attendance:

*WhatsApp was distracting a bit in the evening when I was out at the opera and then for birthday drinks with friends. It was not important to answer, but still a matter of politeness I had to answer in a reasonable time span.* []

Pielot interestingly, in an additional study [PR15], also explored the inverse effect of push-notification interruptions i.e. what happens when they are subsequently removed. Their study, which included 30 users who were asked to disable notifications for a period of 24 hours, showed that although levels of distraction fell, productivity increased and participants reported feeling less stressed, other less desirable behaviours and emotions emerged, such as increased anxiety over missing important information and increased self-interruption due to more frequently checking devices for missed notifications.

Exler et al. [EGB19] investigated, through a survey of 68 participants, differing physical contexts in which notifications would interrupt and disturb smartphone users. Understandably, locations such as Movie Theatres, Libraries, Restaurants and Gyms were amongst the physical contexts in which notifications would be perceived as interruptive, whilst notifications received in Bus/Subway stations,

Parking lots, Gas Stations and Meal Takeaways were perceived to be least interruptive.

Mehrotra et al. [Meh+15b] examine context with respect to interruption and notification receptivity in even greater detail. The authors designed an Android app, NotifyMe, to better understand *the influence of context and content on user response* to incoming notifications and deployed it on the device of 35 users over a period of 3 weeks which generated a dataset of  $\approx 70,000$  notifications and  $\approx 4,000$  questionnaire responses. The study showed that participants received, on average, 100 notifications (interruptions) per day, arriving *without any knowledge about recipients' willingness to get interrupted*. The user's activity at the time of notification delivery was identified as an influential factor for the subsequent response time (i.e. willingness to accept an interruption) of notifications. 80% of notifications were attended in under 10 minutes when traveling in a vehicle; 60% in under 10 minutes when still, walking or running; and least attended, understandable, whilst riding a bike. Additionally the context of the content was indicative of the receptivity to the notification interruption, with family and work related notifications having highest acceptance rates and notifications with categories of system, tools or media having lowest acceptance rates.

Anderson et al. [And+19] examined interruptibility between differing contexts of work and private use of smartphones by exploring the social roles assumed by users in both contexts and the accompanying associated behaviours. Sequences of app use were used to differentiate work and private smartphone activity and subsequent results showed *a significant statistical dependency between social roles and interruptibility preferences*. This suggests *that users tend to follow specific interruptibility strategies while enacting private and work roles*. For example, one participant was identified as a "combinator", as they were happy to be interrupted for either private or work related reasons whilst assuming their work role, but they implemented segmentation in their private role, being only receptive to private-related interruptions and not work-related interruptions during this time period.

It is clear from the literature that notification interruptions are a cause for concern. Both by the receiver, whose cognitive, emotional and productive state can be negatively influenced, but also by the pusher, whose inopportune attempts at communication can result in frustration, annoyance and subsequent notification un-subscribes or app un-installs.

### 2.4.2 Attention

Michael Goldhaber described attention as *an intrinsically scarce resource* when introducing his theory of an *Attention Economy* [Gol97]. In 1997 he predicted: *a continuing rapid rise in the number of people attached to the Web and trying to get attention through it; and that individual attention getters of all sorts will find it ever easier to get attention directly through the Web; and finally, that attention transactions, which already are far more numerous than monetary transactions, will come to dominate even further.*

Push-notifications have indeed made it easier for all "attention getters" to directly communicate with smartphone users and, with the number of apps and devices capable of nudging users growing rapidly, attention has become an ever more rare and valuable commodity. "Attention transactions" are increasingly indistinguishable from "monetary transactions" as the revenue streams which prop up a vast proportion of digital businesses depend on the active engagement of a global audience. As a consequence, the influx of notifications and information pushed at users results in reduced available attention and increased cognitive overload.

Research has shown that incoming notifications distract the user severely from their current tasks, regardless of whether the notification is given immediate attention [Tho+14; Lee+14b; Mar+15; War+17]. A significantly high influx in notifications has also been shown to resonate negatively with those receiving the notifications [Sah+14]. Push notifications from various social media sites, email accounts and other mobile apps, compete continuously for precedence. The problem with this scenario is that it occurs in real-time, with no awareness or empathy for the consumer of the information [Oul+12; Smi+14]

The UbiTtention Workshop, inaugurated in 2016 [Voi+16], focuses on exploring how *the flood of notifications on different computing devices and in smart environments* can be better managed to reduce cognitive overload due to pushed notifications. In their 2016 paper accepted at the workshop, Czerwinski et al. [Cze+16] discussed the competition for attention and the increasing concern that IoT devices may exacerbate the growing problem. The authors suggest improved timing and presentation mechanisms to *balance the costs and benefits of disruption vs the potential health improvements* when notifications are implemented as part of behavioural health interventions. In their 2018 workshop paper [GLP18], Gjoreski et al. discussed methods of utilising wearable technologies, such as a smartwatch, to infer the context of cognitive load for the purpose of ensuring attention would

remain unbroken during intense tasks. In their 2017 workshop paper, Mark et al. [MIC17] investigated attention in the workplace and the impact of disruptions caused by social media and other non-work related interruptions. During the study, 32 workers were observed over two weeks under two conditions: the first week was the control condition in which they worked as normal and the second week was the treatment condition in which they used an app restricting non-work-related website access e.g. Facebook and YouTube. The workers were measured for cognitive absorption, susceptibility to distraction and stress at the end of both weeks. Results indicated that with fewer distractions, participants were more focused and had increased productivity, but also felt more stressed due to their inability to indulge the impulse of checking in on social media. This demonstrates the difficulty of the attention management task. The challenge is to not only remove or reduce distractions, but in such a way that when they are removed or reduced, that the users are not negatively impacted by the sudden absence.

Exploring this point further, Kushlev et al. studied what happened when users, who had been exposed to push-notifications for extended periods, suddenly had those interruptions removed [KPD16]. For their study, 221 student participants were instructed to keep all notifications enabled and their smartphone nearby for a week. In the subsequent week, participants were instructed the opposite: to disable notifications and put their devices out of sight. Understandably, participants were found to be less productive and more distracted during the week of enabled notifications. Participants *felt higher levels of inattention and hyperactivity* during this period and results also showed that interruptions due to notifications were associated with self-reported symptoms of ADHD (Attention Deficit Hyperactive Disorder). The authors concluded that, whilst their study did not suggest that push-notifications were in any way a cause for ADHD, they do drive subscribers to feel more inattentive and hyperactive.

A scenario in which attention is vital and must be maintained is while driving. Giang et al. [Gia+15] studied how drivers engaged with smart devices (smartwatch and smartphone) whilst operating a vehicle and discovered that the smartwatch incited more glances per notification, on average, than smartphones. The authors hypothesised that, as the smartwatch was closer in proximity, it could be associated with greater opportunity for glancing at the screen to check for notifications but, as the screen is also smaller than a smartphone, it would also require more interaction to extract the same amount of information (e.g. to read the full notification). The study also discovered that brake response times were longer when a notification was just received via smartwatch, thus increasing the risk of an accident. Rajan

et al. [RSL16] also explored the effect of receiving notifications whilst driving by presenting notifications to participants in visual and aural form during the task. Results indicated that both auditory and visual notifications impeded driving performance and that notifications delivered during *low workloads* were attended best.

Attention is fiercely sought by devices, apps and the entities behind them pushing notifications. Although the information pushed provides value, it also bears costs on the subscriber who may not wish their finite resource of attention wasted on irrelevant distractions. Particularly during moments in which their cognitive load is already taxed or their psychological well-being is deteriorating.

## 2.5 How Is Notification Technology Evolving?

Since the conception and widespread adoption of the smartphone, there have been numerous studies examining its use in-the-wild and the relationship between the subsequent interruptions and distractions it incites as well as the ensuing behaviours and attitudes of users. In 2015, Church et al. [Chu+15] summarised state-of-the-art methods used within a number of smartphone usage studies in order to understand how experiments, both observational and interventions, in this domain varied depending on the parameters of the study. Mehrotra [MM17; MM20a] and Kunzler [KKK17] also undertook systematic reviews of push-notification technology and examined how different studies provided different perspectives on the features, contexts and factors of users relevant and most important to push-notification engagement outcomes.

Building upon this work, this section extends the review of Kunzler (which examined years 1994-2017), by summarising a selection of interruption and push-notification engagement studies executed in-the-wild and reported between the years 2017 and 2020 (inclusive). To also extend the work of Church, Table 2.3 illustrates the parameters of the in-the-wild studies, including details of participation, duration, recruitment, deployment, incentives offered, methods of experimentation used and the type of data collected. The aim of this was to provide an updated snapshot of push-notification in-the-wild studies present in literature over the past 4 years and compare both the study parameters and features pivotal for the continuous improvement of push-notification technology.

Table 2.3 Literature describing *in-the-wild* studies of mobile notifications, context-aware computing and interruptibility 2017-2020.

| Research goals                                                          | Reference | Year | Participants | Duration    | Deployment |   | Recruitment |   | Incentives |   | Method |   | Data Collect |    |
|-------------------------------------------------------------------------|-----------|------|--------------|-------------|------------|---|-------------|---|------------|---|--------|---|--------------|----|
|                                                                         |           |      |              |             | P          | S | L           | D | Y          | N | E      | I | Qn           | Ql |
| Device adapts to context to subtly maximise private interaction         | [Pea+17]  | '17  | 32;<br>58    | 1hr;<br>5wk | •          | • | •           | • | •          |   | •      | • | •            | •  |
| Investigate influencing factors for notifications in multi-device envs  | [MHM19]   | '19  | 24           | 21d         | •          |   |             | • |            | • | •      |   | •            |    |
| Evaluate context-aware notif. strategy for crowdsourcing tasks          | [Kan+20]  | '20  | 488;<br>432  | 1m;<br>6wk  | •          |   | •           |   |            | • | •      | • | •            | •  |
| Predict interruptibility intensity                                      | [YGL17]   | '17  | 22           | 4wk         |            | • | •           |   | •          |   | •      |   | •            | •  |
| Develop notification rule to reduce crowded IM notifications            | [SSL19]   | '19  | 17           | 2wk         |            | • | •           |   |            | • | •      | • | •            | •  |
| Evaluate novel interruption estimate and adapt notifications            | [OTT19]   | '19  | 382,<br>518  | 28d         | •          |   |             | • |            | • |        | • | •            |    |
| Context-aware notifications for drivers                                 | [SBV18]   | '18  | 3            | 2hr         |            | • | •           |   |            | • | •      |   | •            |    |
| Exploration of notification deferral                                    | [Web+18]  | '18  | 295;<br>16   | 1yr;<br>1wk | •          | • | •           | • | •          | • | •      |   | •            | •  |
| Infer moments for notification engagement                               | [Pie+17]  | '17  | 337          | 4wk         |            | • |             | • | •          |   | •      |   | •            | •  |
| Effect of contact, relationship and context on notification receptivity | [Lee+19]  | '19  | 34           | 4wk         | •          |   | •           |   | •          |   | •      |   | •            | •  |
| Influence of notification speculation on attendance                     | [CCS19]   | '19  | 33           | 2wk         | •          |   | •           |   | •          |   | •      |   | •            | •  |
| How notifications are handled on mobile devices                         | [PVP18]   | '18  | 278          | 4wk         |            | • |             | • |            | • | •      |   | •            |    |
| Investigate private vs work-related mobile use and interruptibility     | [And+19]  | '19  | 4            | 11wk        |            | • | •           |   |            | • | •      |   | •            | •  |
| Understand notification behaviour and evaluate intelligent prediction   | [Vis+19]  | '19  | 40           | 73d         | •          |   | •           | • |            | • | •      | • | •            | •  |
| Studying and improving interruptibility models                          | [TAW17]   | '17  | 224          | 178d        | •          |   |             | • |            | • | •      | • | •            |    |
| Exploring notification frequency and timing on app engagement           | [Mor+17]  | '17  | 77           | 6m          |            | • |             | • |            | • | •      | • | •            | •  |
| Influence of notification visuals on receptivity                        | [APT19]   | '19  | 6866         | 15d         | •          |   |             | • |            | • |        | • | •            | •  |
| Notification interaction behaviour across loc. and activity contexts    | [Meh+17b] | '17  | 26           | 2wk         | •          |   | •           |   | •          |   | •      |   | •            | •  |
| Identify casual links between emotion and mobile interaction            | [Meh+17a] | '17  | 28           | 20d         | •          |   |             | • | •          |   | •      |   | •            | •  |
| Understanding and mediating negative smartphone use                     | [RD20]    | '20  | 35;<br>20    | 3wk         | •          | • | •           |   |            | • | •      | • | •            | •  |

**Deployment:** P = Public; S = Selected; **Recruitment:** L = Local; D = Distributed; **Incentives:** Y = Yes; N = No;

**Method:** E = Experiment; I = Intervention; **Data Collected:** Qn = Quantitative; Ql = Qualitative

Pearson et al. [Pea+17] investigated a chameleon approach toward mobile phone interaction which was hypothesised to improve the privacy and safety of engaging with a mobile device in public while also reducing the risk of being perceived as rude or inattentive when checking in on a device amongst company. The long term objective of the study was that technology would eventually "vanish into the background". The authors describe the applicability of *subtle notifications* which, in contrast with current methods in which a device rings out or vibrates loudly to grab attention, would provide the user with a subtle cue indicating a notification arrived to minimise social disruption. One such cue applied, was the use of abstract colours to signify different incoming notifications, thus alluding to the content when seen by the user but yet kept private in case of prying eyes.

Qualitative results extracted during studies of the prototype deployed with participants illustrated keen acceptance of subtle notifications: *"it is really cool - much more private"*; *"it's very discreet, only you know what it's doing - your friends don't"*; and *"if the phone flashes I can smile about it because only I know I have a message"*. Users also wanted control over how the notifications would be visually presented to them: *"I would mostly use the yellow colour for incoming SMS messages. I could set different colours for different close friends"*. The authors found participants were attracted to the discreet notification method because of the improved privacy and control it provided over social disruption: *"its selling point is that I can decide when I get the notifications easily with my hand so there's no social interruption or being rude"*. Further investigation uncovered users need for more context and information when receiving the subtle notifications e.g. the flash of colour on screen indicated which app the notification was from, but did not allude to the sender or subject of the message - the Chameleon app addressed this by providing a further summary of the apps most recent messages when the colour was tapped.

Similarly, Chang et al. [CCS19] investigated how users speculate about notifications when initially alerted to their presence and the impact it has on subsequent engagement. The authors hypothesised that users, on sensing the arrival of a new notification via one of the alert modalities - vibration, ring tone or flashing LED lights, would begin to speculate about the app or sender which caused the notification to be delivered and why. Through a study implementing an Experience Sampling Method in-the-wild, the authors found that in 71% of cases, users speculated about the app and sender when alerted and that the sender of the notification was much more challenging to correctly identify. Phone events associated with notification speculation were also examined, with recent app use

found to be highly associated with speculating about a notification. In addition, speculation about a notification was also linked to greater attendance of notifications, in particular it was found the app which posted the notification played a significant role, as well as being linked to higher importance being placed on the notification if speculated about. The qualitative data gathered during the study further informed the basis of why and how users speculated about notifications. The authors found that the context, alert modality and temporal factors of notifications helped users form their speculations about the source app and sender.

Lee et al. [Lee+19] further explored the impact of sender relationships on the receptivity to notifications, as well as the interruption context. The authors used an Experience Sampling Method during an in-the-wild study with 34 participants in which measures of relationship, context and receptivity were calculated. For relationship measures, existing instruments were used such as the Unidimensional Relationship Closeness Inventory (to measure contact closeness) and the Mobile Maintenance Expectation (MME) scale (to measure user expectations with regard Instant Messaging (IM) responses). The results of the study found that context was considered more important than sender relationship for predicting receptivity measures such as *Interruptibility* (a measure of how interruptive notifications were) and *Opportuneness* (a measure of how good the notification delivery timing was). The authors also found that higher MME scores were associated with lower receptivity to notifications, illustrating that participants were less likely to engage with notifications when higher expectations of continued engagement were placed upon them and participants perceived notifications associated with higher response expectations to be "more interruptive and ill-timed".

In their in-situ study of 26 students, Mehrotra et al. [Meh+17b] investigated the contextual effects of location and activity on notification engagement behaviour. The authors found multiple notable findings. The first being that receptivity increased when users were engaged in activities such as chores or exercise, although there was some variability between users indicating individual preferences were a factor with regard to notification engagement during these times. Attentiveness to notifications was also found to be highest at "productive places compared to lazy places". The authors theorised that this was due to participants being more cognitively aware of incoming notifications related to their tasks at these places, such as work notifications from colleagues. The study also found that receptivity was lower when users were out in nature compared with urban environments and overall mobile device usage dropped when in places associated with nature,



distress or lazy activities. The results of the study illustrated that daily activity routines and place characteristics were indicative of subsequent attentiveness and receptivity to push-notification attendance.

Visuri et al. [Vis+19] also conducted an in-the-wild study of push-notification engagement on smartphone devices. The authors explored how the content of notifications influenced the perceived importance for users and proposed a method of intelligently mediating the presentation of delivered notifications accordingly. The study indicated, when observing user engagement without intervention, that "interaction fatigue" surfaced within users when overrun with incoming notifications and that, amidst the overload of information, the content considered important would get lost. This would also happen if "users neglect to interact" with incoming notifications i.e. if users let them build up in the notification drawer without tending to them on a regular basis. The authors also highlight the complexity of distinguishing what constitutes an important or valuable notification, noting that *the dismissal of a notification is not substantive evidence for deeming it unimportant*. During the study, once users had labeled a sufficient amount of notification data, an intervention was enabled, with consent, which mediated the delivery of notifications to participants i.e. did not present notifications to the user if they were predicted as "unwanted" based on context and notification content. A C4.5 classifier was trained, using the clicked and dismissed annotated notifications of participants, to predict classes "show" or "hide". The mean accuracy achieved by the deployed classifiers of 33 users during the study was 91.1%. The authors subsequently identified increases in CTR when the intervention was active, although they concluded further investigations of such interventions, and their effects on notification interactions, should be studied.

Okoshi et al. [OTT19] implemented an "adaptive notification scheduler" to identify opportune moments to push personalised and generic notifications. In their study, which deployed the system in-the-wild to 382,518 subscribers of the Yahoo! Japan Android app, they noted a 60.7% gain in click rate over a previous baseline system [Oko+17]. Their proposed method composed training a linear regression model on 474 context features, extracted from smartphone sensors, for interruption estimation to identify optimal breakpoints for delivering notifications to subscribers. Results of their study indicated that the approach was not only an improvement over their previous baseline, but also that it performed best when notifications were personalised to subscribers (e.g. the content aligned with their interests) or when notifications were classified as "breaking news".

Whilst Visuri et al. investigated the deployment of a Notification Management System on end-user devices for mediating notifications received, Okoshi et al. evaluated an adaptive notification scheduler for improving the *push* (and subsequent delivery) of notifications. The distinction between both methods is important. Visuri et al. solely optimise for the *subscriber* of notifications intending to reduce disruption, cognitive overload and improve the notification experience for smartphone users. Okoshi et al. in their study, optimise for the *pusher* (i.e. the app, business or marketing agency), to also improve the notification experience for subscribers, but as a by-product of improved KPIs, such as Click-through Rate. It is important to note this distinction as the goals of both NMSs may conflict in the future e.g. if the importance of KPIs, such as CTR, outweighs the importance of protecting subscriber attention and long term well-being. If such a scenario were to occur, an adversarial environment would be created in which an agent optimised to push notifications at a subscriber would be constantly trying to outmaneuver an agent optimised for mediating notifications on behalf of the end-user. One such example of this already occurring today is clickbait: notifications created with the promise of revealing certain information if opened, only shown to be misleading when the content studied within was not as promised. This problem was recently acknowledged in the domain of web push-notifications as Google issued an update [Goo20a] to the way in which the Chrome web browser handles notifications identified as manipulative and websites which users report (implicitly through behaviour, or explicitly) as sources of poor push-notification experiences.

The work by Okoshi et al. highlights a significant common factor prevalent within the push-notification literature analysed: the definition, extraction, implementation and scale of user, context and push-notification features harvested for data-driven notification management approaches. The majority of studies discussed capture features of context or content derived from smartphone users and their notification engagements. To supplement the analysis of in-the-wild push-notification engagement studies, Tables 2.4, 2.5 and 2.6 illustrate the most frequent features used within notification, attention and interruption management literature to date. It should be noted that whilst many studies use identical features, the methods of procurement are not always consistent. For example, the representation of time, such as when a notifications arrived to a device, were found to vary:

- timestamp recorded explicitly e.g. [CDM15]

- converted to bins (with differing times selected to represent morning, afternoon etc.) e.g. [OTT19] [TAW17]
- recorded as time elapsed since the start of the day e.g. [Kan+20]

Similar nuances were identified in other features, such as location (semantic categories, explicit latitudes and longitudes, or proximity to network towers), notification category (researcher defined or use of external source, such as the Google Play Store) and activity (self-reported, inferred using sensors or inferred using Android APIs). This makes it difficult to compare the performance of differing NMS approaches as they are trained upon inconsistent feature sets.

Despite this, these in-the-wild studies are crucial for understanding how users interact and engage with technology. However, the time and effort spent developing mobile software, recruiting participants and collecting data is a substantial part of all the studies outlined above. In addition, the subsequent data collected is rarely publicly released and shared with the research community, thus a major potential contribution of the work, and highly useful asset for the community, lies unutilized.

Reflecting on the above analysis, a pattern was identified within the literature relevant to improved push-notification delivery, interruptibility and receptivity:

1. The authors first developed an application with which to collect user data in-the-wild, both quantitative and qualitative.
2. On collection of this data, an analyses of the features indicated a select number with which target behaviours could be inferred (such as using the notification context to infer opportune delivery time).
3. Once identified, these features were then used to create an intelligent system which improved the data collected (e.g. the delivery time or method of notification delivery).
4. Finally, the intelligent system was validated offline, on a sample of the collected data saved for testing, or online, by introducing the system to a set of real users.

Depending on the type of study and procedures in place, additional factors which had an impact on the process above included:

- Seeking ethical approval to carry out the data collection process

Table 2.4 Summary of Notification features referenced in literature.

| Notification Features |                                                       |                                                                                                                                                                                      |                                                                                                                                                                                                                             |
|-----------------------|-------------------------------------------------------|--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| Name                  | Description                                           | Associated                                                                                                                                                                           | Reference                                                                                                                                                                                                                   |
| time                  | times of interest with respect to notification events | send time, arrival time, seen time, removal time, time of day, day of week, reaction time, starred time, snoozed time, attended time, checked time, consumed time, weekend indicator | [Meh+15b] [Ho+18] [MHM16a] [MHM19] [Kan+20] [YGL17] [SSL19] [OTT19] [YS20] [Web+18] [CCS19] [PVP18] [Sah+14] [Meh+16] [And+19] [Pie+15] [TAW17] [Mor+17] [CDM15] [APT19] [MHM16b] [Meh+17b] [Meh+17a] [PM14] [RD20] [CDM19] |
| contact               | the person/contact sending/receiving the notification | contact, sender relationship, sender importance, recipient, recipient relationship                                                                                                   | [SSL19] [CDM15] [CDM18]                                                                                                                                                                                                     |
| app                   | app which spawned the notification and its features   | name, category, significance                                                                                                                                                         | [MHM19] [Meh+15b] [Web+18] [CCS19] [Sah+14] [Meh+16] [And+19] [Vis+19] [Meh+17b] [RD20]                                                                                                                                     |
| content               | notification content                                  | title, subtitle, subtext, extra lines, extra text, message                                                                                                                           | [Meh+15b] [Meh+16] [Vis+19] [CDM15] [CDM19]                                                                                                                                                                                 |
| category              | category of the notification                          | type, ongoing, clearable                                                                                                                                                             | [CDM18] [CDM15] [Pie+15] [And+19] [CCS19] [Meh+15b] [MHM16a]                                                                                                                                                                |
| priority              | the priority level of the notification                |                                                                                                                                                                                      | [CDM18]                                                                                                                                                                                                                     |
| modality              | the alert modality of the notification                | sound, vibration, LED colour, device                                                                                                                                                 | [Meh+17a] [Meh+16] [CCS19]                                                                                                                                                                                                  |
| template              | visual format of the notification                     | text format, emojis, icons, images                                                                                                                                                   | [YS20] [APT19] [CDM19]                                                                                                                                                                                                      |
| action                | action taken on the notification by the subscriber    | response, outcome, clicked, engagement acceptance                                                                                                                                    | [Ho+18] [MHM16a] [Meh+15b] [Zha+18] [Lee+19] [CCS19] [Sah+14] [Meh+16] [Pie+15] [Vis+19] [TAW17] [Mor+17] [CDM15] [APT19] [MHM16b] [Meh+17a]                                                                                |

Table 2.5 Summary of Context features referenced in literature.

| Context Features |                                                                                     |                                                                                                                                                                                                                |                                                                                                                                                                          |
|------------------|-------------------------------------------------------------------------------------|----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|--------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| Name             | Description                                                                         | Associated                                                                                                                                                                                                     | Reference                                                                                                                                                                |
| location         | location of the device and, by extension, user                                      | latitude, longitude, cell towers, Places API, location popularity                                                                                                                                              | [Ho+18] [MHM16a] [Pea+17] [Kan+20] [SG16] [YGL17] [Meh+15b] [Cho+16] [Lee+19] [CCS19] [Meh+16] [And+19] [Pie+15] [Vis+19] [CDM15] [Meh+17b] [PM14] [UP16] [RD20] [CDM18] |
| activity         | activity of the user holding device as inferred, sensed or self-reported            | Activity Recognition API, activity accessibility, physical activity, task                                                                                                                                      | [CDM18] [RD20] [PM14] [Meh+17b] [CDM15] [And+19] [Meh+16] [CCS19] [Lee+19] [Pie+17] [SBV18] [Meh+15b] [OTT19] [YGL17] [SG16] [MHM19] [MHM16a] [Ho+18] [Vis+19]           |
| device state     | the contextual state of the device at various time points                           | screen on/off, unlocked, charging, volume, ringer mode, vibration pattern, network connectivity, battery level, airplane mode, audio jack, music playing, orientation, bluetooth, device type, device features | [Pea+17] [MHM19] [OTT19] [Meh+15b] [Pie+17] [Cho+16] [Lee+19] [CCS19] [PVP18] [And+19] [Pie+15] [Vis+19] [TAW17] [MHM16b] [UP16] [CDM18] [CDM19]                         |
| device sensors   | sensor readings recorded with the device at various time points                     | ambient noise, ambient light, accelerometer, gyroscope, distance traveled, humidity, microphone, proximity                                                                                                     | [SG16] [OTT19] [Meh+15b] [SBV18] [Pie+17] [Cho+16] [Meh+16] [Pie+15] [TAW17] [PM14] [UP16] [CDM19]                                                                       |
| online media     | information regarding physical or cognitive state of the user using online profiles | calendar events, chat room                                                                                                                                                                                     | [UP16] [SSL19] [SG16]                                                                                                                                                    |
| other            | other context features, tending to be self-reported                                 | company, emotion, stress, mood, social priority, obtrusiveness                                                                                                                                                 | [PM14] [CDM18]                                                                                                                                                           |

Table 2.6 Summary of Usage features referenced in literature.

| Usage Features     |                                                    |                                                                                                                                                                                                                       |                                                                                                                         |
|--------------------|----------------------------------------------------|-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-------------------------------------------------------------------------------------------------------------------------|
| Name               | Description                                        | Associated                                                                                                                                                                                                            | Reference                                                                                                               |
| notification usage | continuous engagement behaviour with notifications | ctr, notification drawer, notification center events, time since last engagement                                                                                                                                      | [MHM19] [Pie+17]<br>[Pie+15] [PVP18]<br>[WVH19]                                                                         |
| application usage  | continuous usage behaviour with apps               | duration foreground, duration background, app launch count, blacklisted apps, app category, app events, app use sequence, screen change event, significant app launch, non-significant app launch, time app last used | [Cho+16] [Zha+18]<br>[And+19] [Pie+15]<br>[MHM16b] [Meh+17b]<br>[Meh+17a] [RD20]<br>[Mor+17] [PVP18]<br>[Vis+19] [MD19] |
| device usage       | continuous device usage behaviour                  | num unlocks, time last used, num taps, long clicks, scroll, num apps installed, duration phone interactions                                                                                                           | [MD19] [RD20]<br>[Meh+17b] [MHM16b]<br>[Pie+17] [MHM19]<br>[Pie+15] [Meh+17a]                                           |

- Recruitment of participants, including obtaining informed consent and ensuring GDPR or other regulatory compliance
- A manual analysis and removal of potentially sensitive or confidential data

As the data sets collected were rarely publicly available to the research community, it is difficult to compare the different types of intelligent systems, reproduce results of the proposed algorithms and/or replicate the experiments introduced. As such, it cannot be assumed that the results obtained within the studies hold outside of the strict experimental constraints expressed (such as sample size, user demographics, location etc.). Mehrotra et al. [MM17] also identified a need for further large scale studies in their review of push-notification literature, however to address the issue of freely sharing benchmark data sets, with which intelligent systems can be trained and evaluated, the domain of generative modeling must be considered.

## 2.6 Generative Methods

From the previous section it can be seen that gathering information through mobile devices is a trend that continues to grow in the research community. Due to a myriad of sensors available to researchers through mobile devices, and the

rare rate of separation from their owners, they are an invaluable resource of rich user data [Rac+10]. One such research application for data sets harvested in this manner is the training of intelligent systems for management of mobile notifications [Meh+15b; KS; Oko+15; Qin+14; HI05]. However, obtaining rich data sets, such as those containing mobile notifications and contextual user data, from mobile devices does pose a problem as there are a number of factors which impede the ease of acquisition. Some such factors include development, ethics, security, transparency and scale. These issues also extend to other domains in which sensitive information is gathered, such as in healthcare, educational institutions, government agencies and private enterprises.

A potential alternative to using the sensitive data collected, through such mediums as ad-hoc mobile applications, is to generate synthetic data sets which imitate real world activity. Raab et al. [RND16] define synthetic data as the reproduction of *the essential features of the actual data with some or all of the values replaced by simulations from probability distributions*. Advantages include being able to share and reuse synthetic data freely without the same ethical implications associated with using real user data. Depending on the algorithm chosen, synthetic data can also be modified and scaled according to the needs of the user. While the flexibility and freedom of use are advantageous, some disadvantages also materialize with synthetic data, such as the possibility of weak representation of the true population from which real data is sampled and thus reduced utility of the data. Sections 2.6.1, 2.6.2 and 2.6.3 detail different examples of synthetic data being utilised and identify some of the current techniques for generation and evaluation.

It was chosen to focus on the generation of synthetic push-notifications for a number of reasons:

- Risk - deploying a Notification Management System (NMS), a system which decides to withhold or show a notification at a given moment, in-the-wild could be considered a risk. From the perspective of the app pushing notifications, a poor performing NMS which frequently alerts users could result in a lost customer or user audience. A NMS which withholds important notifications may also irritate the subscriber to the point of uninstalling the application. In addition to a lost audience, which may result in lost revenue for an app deployed in industry, the risk associated with a live deployment of a NMS could also be extreme - for instance a poor NMS delivering a notification to a user while they are driving, resulting in a crash. Synthetic

data facilitates simulations which can be run "offline", and as such, poses less risks when training and evaluating NMSs.

- **Speed** - creating, deploying and gathering a user base for a NMS can be a time-consuming task. There are many examples of mobile usage studies within HCI literature [Chu+15] whereby it is difficult to maintain high levels of participant engagement, within both short (initial onboarding) and longitudinal studies. Creation and use of synthetic data through developing generative models enables researchers to avail of data easily and without having to dedicate time and resources to developing and deploying their own mechanisms/apps for gathering data. In particular, conditional generative models also empower the user to select the type of data they wish to generate.
- **Comparative Evaluation** - comparing results within the HCI and User Modeling domains is a well known challenge [Con+19; Chu+15]. Differing participants, sample methods, data features and units of measurement between studies make it difficult for researchers to objectively compare results obtained from separate studies. Synthetic data, and the development of generative models powering simulation toolsets, such as OpenAI Gym, can allow benchmarks to be easily drawn between algorithms applied to the same data set and under similar conditions.

### 2.6.1 Synthea

Synthea [Wal+17] is an open source software packaged created by Walonoski et al. for generation and sharing of synthetic electronic health records. Particularly in the health care domain it is difficult to find open and readily available data with which to use for such purposes as developing clinical decision support systems. Ethical issues arise over the disclosure of potentially sensitive patient data. Therefore, researchers and developers wishing to extract useful insights from data such as electronic health records, must first address the issue of procuring such sensitive information. Less than satisfactory solutions include leveraging public health records, which are anonymized, but still pose a potential risk of breaching consent agreements pertaining to the use of data on collection. The granularity of this data is also limiting a limiting factor. The generation of realistic synthetic Electronic Health Records (RS-EHRs) is therefore favorable as it permits the data to be shared for educational, research and industrial purposes without



the aforementioned associated risks and constraints due to consent and privacy concerns.

With Synthea, Walonoski et al. introduce a method of generating EHR using only readily available public data and health statistics as a seed, in contrast with other such systems which are dependent on private EHR data. Additionally, the synthetic data is generated based on transparent models which can be continuously improved upon by facilitating experts, all of diverse clinical backgrounds, who collaborate on the system. Through the implementation of a temporal model, the generated data can represent a lifetime of EHR's for a patient as opposed to focusing on a single problem/disease.

The synthetic data generated by the Synthea system was validated by comparing the aggregated statistics from the generated patient records with publicly available health statistics. As an example, the progression and resultant care given to patients with type 2 diabetes was analysed resulting in the discovery of small discrepancies between the generated and real data (such as no diabetes onset in patients over 52 in the synthetic data set). However, as the system supports tuning via expert input, the model transition probabilities could be manipulated to reflect increasingly realistic results, though the authors acknowledge it as challenging due to the inherent variability of simulation runs. This highlights an additional challenge when generating synthetic data via this method: how can the results be reproduced and/or replicated?

### 2.6.2 Synthetic Data Vault

The Synthetic Data Vault (SDV) [PWV16] is a system proposed by Patki et al. to automatically create synthetic data using generative models of relational databases. The motivation of the SDV is to enable enterprises to freely share their internal data and be able to seek external aid with regard data analysis without needing to be concerned of privacy issues. For example, if an enterprise wished to release data on *Kaggle* to allow data scientists compete to solve a problem for them they may encounter issues with sharing the relevant data in a public domain, hence the only option would be to consult internal aid. The goal of the system, therefore, was to generate synthetic data which preserves the statistical distribution of the real data and also structurally adheres to the format of the original data such that software built on top would not break if the synthetic data was swapped in. Uniquely, the SDV applies generative modeling techniques to

the database itself, before any transformation or aggregations take place to fully empower the data scientists to utilise the raw features for their own purposes. This method was chosen to ensure the synthetic data generated would not be dependent on the problem/domain it was produced for and thus limiting the data scientists' exploratory impact.

To create a generative model, the SDV calculates both the probability distribution for values over each column of the relational database and also the covariance in order to describe the distribution shape and joint variability of every feature. A Gaussian distribution is assumed in most cases, but the Kolmogorov-Smirnov test is also implemented to adapt if necessary (e.g. switch to a uniform distribution). The multivariate Gaussian Copula is used to convert column distributions to standard normal before calculating covariances. The SDV is also extensible to multiple tables in a database using a method called Conditional Parameter Aggregation (CPA) which identifies tables which are dependent on others in the database and conditions their generation with respect to the parent table. This is useful if prior knowledge is known and subsequent values wish to be generated. For example, a user wishes to generate synthetic Airbnb transactions for a specific user segment classified by gender and location. The SDV also incorporates a method of synthetically generating categorical and datetime data types in addition to purely numeric. To achieve this the categorical values are converted to numerical values using the cumulative probability of the category value as an interval and a truncated Gaussian distribution, while the timestamp epoch of the datetime is used as the numeric value.

In order to evaluate that the synthetic data being generated remained realistic and of similar structure to that of the original data, Patki et al. used a number of publicly available relational data sets (e.g. From Kaggle: Airbnb and Rossman Store Sales) with which to use as seed for generating synthetic data. They then hired a number of data scientists to engage with both the original and synthetic data, evaluating the performance of features they created and predictive models they built using both data sets. On completion, statistical tests were then implemented to search for any statistical significance between those working on original (control) data sets and those working on synthetic. This method of validation, in contrast with Synthea, places emphasis on the utility of the data after synthesis as opposed to simply identifying whether statistical correlations remain between original and generated data. Patki et al. collected qualitative data from the data scientists partaking in their experiments in order to ascertain if any problems or adaptations occurred when working with the synthetic data as

opposed to the real data. Quantitative data, such as the number of new features created, was also collected for the purposes of determining the utility of the synthetic data sets.

Interestingly, Patki et al. also propose to address, as future work, whether there is a statistically significant drop in the accuracy of predictive models created from synthetic data with the insertion of noise. They suggest that, through the insertion of noise the underlying distributions of original data could be protected, thus increasing the privacy of the original data. Identifying the threshold at which the utility of the data is comprised due to noise would therefore be useful and could aid in generating synthetic data with improved privacy.

While SDV does provide an outlet with which synthetic data can be generated, there are limitations with regard to the type and structure of the generated data. For example, data types such as images, video and audio have not yet been explored. Additionally, there is no mechanism for the generation of text in the form of sentences or paragraphs. Including these types of data poses additional privacy concerns as they are not easily synthesized and would require additional steps to ensure that which is synthesized does not breach the privacy of the user. Key features of notifications are expressed in text, image and audio form, hence the SDV may not be expressive enough to generate synthetic notification entities with the necessary granularity for coaching intelligent systems.

### 2.6.3 Synthpop

Nowok developed an R package, Synthpop [Now15], to address the confidentiality constraints restricting free access and use of microdata. In particular, the authors indicate that national statistical agencies go to considerable lengths to collect valuable information on individuals and organisations, but can then only share the data internally due to the terms under which the data was collected. They state that their underlying aim for synthetic generation is to provide publicly available data sets enabling accurate inferences to be made without dependence on real data. The authors acknowledge that making inferences from synthetic data is challenging, hence, they focus on generating only synthetic test data derived from confidential data sets. The goal of synthpop is therefore to enable analyses and inference models to be created and tested on synthetic data, before then finally being applied to the real data for actual reporting. The authors highlight that

the method of validating results acquired from synthetic test data on real data is already being used in statistical agencies and is therefore a valid approach.

Synthpop creates a generative model by calculating conditional distributions over all features in a data set. Classification and regression trees (CART) are implemented to create models which are fitted feature by feature, with the next feature being conditioned on all previous features. Missing data is also not imputed before the synthesis process is executed so that the pattern of missing values is also represented in the synthetic data generated. Synthpop evaluates the synthetic data generated by comparing it with characteristics from the real data, such as calculating the frequency distributions of each variable in the data set. The package also uses models fitted to the synthetic data to evaluate quality. Comparisons are made between estimates made by models fitted on synthetic data and the original data.

## 2.7 Limitations within Literature

The previous sections have analysed, within push-notification and interruption management literature: how notifications are currently being applied across a diverse range of technologies and domains; the subsequent impact they have on notification subscribers; and the state-of-the-art methods currently proposed for managing the influx of notifications to offset adverse effects, such as distraction and anxiety. This section discusses insights from this analysis and identifies potential areas of limitation used to formulate the Research Questions, which the work in this thesis aims to address. The four main threads within literature in which limitations were identified were:

- Despite the value of timely and useful notifications, studies have shown that, when pushed without intelligence, notifications negatively impact productivity and task completion [SB10; MN86; GM04; Mar+15; Tho+14; Lee+14b; War+17], with contextual irrelevancy being a significant factor for notification dismissal or deferred attendance [Web+18]. Large influxes of notifications also tend to overwhelm, whilst their complete removal cause anxiety and an increased tendency to self-interrupt [PCD14; PR15; KPD16]. Studies also show that the relationship between the subscriber and those responsible for pushing the notification also fuels the perceived social pressure to respond [Lee+19; Meh+15b; CCS19]. Contexts, such as location [EGB19],

activity [And+19] and notification content [Meh+15b], play a significant role in subscriber receptivity to push-notifications. Chapter 3 of this work validates the significance of static context features identified in literature with respect to their impact on notification engagement but also further evaluates the contribution of transient contexts, the intricacies of which may have not been fully explored in literature to-date. In addition, Chapter 3 also examines the etiquette associated with attending to push-notifications and mobile devices by querying subscribers about their perception of good-behaviour in varying contexts. The work also adds to the body of research aimed at understanding notification subscriber behaviours by identifying the relationships between hope (goal-oriented thinking/planning), notification attendance and smartphone usage.

- Notification text content has been used in a number of key studies [Meh+15b; Meh+16; CDM15], however few examine the explicit text content within notifications to better understand the impact on engagement [Vis+19]. Equally, few have also investigated the explicit notification content with respect to its level of influence, intention for enticement and its potential to steer both subscriber and NMS actions. Indeed, this gap within push-notification literature was also recently identified by Liu et al. [Liu+19] as they noted aggressively influential push-notification content was found to be prevalent throughout all market verticals in which push-notifications are utilised. Concerns of technological implementations of dark patterns [Bri] within the field of interaction design, applied to manipulate, exploit and persuade users, have arisen in recent years [Rog+20]. A recent comprehensive review of HCI literature by Caraban et al. [Car+19] in particular highlighted the many overt and covert methods used to nudge users toward particular goals and evaluated the mechanisms used with respect to their level of transparency, autonomy and ethical use. The influence of dark patterns have yet to be explored from the perspective of push-notification text and, additionally, the subsequent impact persuasion has on NMSs deployed to protect subscribers. To address this, Chapters 5 and 6 discuss work which explored the influence of persuasion and enticement of push-notifications and describes an approach for mediating their impact.
- In addition, few push-notification studies have facilitated transparency for the subscriber with respect to the subsequent notifications which they receive, why they receive them and the influence entailed. Mehrotra et al.,

in one of the few studies which did address this gap in literature, proposed a mechanism in which notification engagements could be extracted, via association rule mining, and later studied and modified by subscribers [MHM17]. To add to this approach of transparent, human-in-the-loop, management of notifications, an approach for scrutinising notifications through an end-user tool was also discussed in Chapter 5.

- A primary reason few studies examine push-notification text content explicitly is due to the sensitive information it holds about subscribers. Even without including text content, push-notification engagement data have the potential to describe a lot about a subscriber - such as the apps they have installed on their device, contacts they received messages from and the times/durations in which they engaged with digital media. For this reason, few push-notification data sets are freely available within either academic or commercial domains. The NMSs discussed in this chapter have implemented data-driven approaches toward managing incoming notifications, which can be difficult to sustain as consent becomes difficult to extract from end-users, and provisions prevent, for good reason, the free transfer of data. To address these limitations and facilitate sustainable data-driven approaches, Chapter 4 explores the use a generative approach for creating synthetic push-notification engagement data which can be shared or used in place of real data to provide an additional layer of privacy for the subscriber or facilitate minimal use of real data.

## 2.8 Summary

This chapter presented a review, satisfying **RO1**, of the intersecting fields of push-notification, interruption management and context-computing literature and identified limitations which were deemed to provide an opportunity for further research and exploration which could positively contribute to the body of HCI, and in particular, push-notification, literature. This work also subsequently informed the consideration of the four Research Questions outlined in Chapter 1, **RQ1-RQ4**, aimed at advancing the state-of-the-art and provided evidence of a need for requirements, **R1-R4**, within the proposed ETHOS-Push approach.

# Chapter 3

## Understanding Notification Engagement In-the-wild

### 3.1 Overview

The original intention of a pushed notification was to provide updated information, initially regarding email [MC06], to a subscribed user. However, as push technology became increasingly ubiquitous, subscribers found themselves overloaded with notifications pushed from many devices and applications. Now, With user attention divided and spread so thin, performance and the ability to focus on simple tasks has declined dramatically [War+17; Mar+15; Lee+14b; Tho+14]. Research has shown that large numbers of incoming notifications resonate negatively with push recipients, but also, that switching off notification functionality increases levels of anxiety and invokes a Fear of Missing Out (FOMO) [PR16].

Push notifications, while a nuisance when delivered badly, are of great value to subscribers when used correctly. When push notifications are delivered at the right time, through the right device and with the right information, they can be a valuable resource for consuming time-sensitive information. Examples include changes to public transport schedules when commuting, sporting event highlights or even, within the present climate, public health advice due to potential exposure to a reported Covid-19 case [App20]. In addition, notifications can also remove the necessity of subscribers manually trawling through irrelevant information themselves to find pertinent or new information. In the medical

domain, notifications have been used to prompt reflection, engagement, reinforce better habits and quash feelings of anxiety [Bid+18; Doh+18].

Subscribers receive push-notifications from many sources, on many topics and in many contexts, but not all of them are useful or healthy to engage with at moments of delivery. This chapter discusses two in-situ studies which explored push-notification engagement behaviour *in-the-wild* with the purpose of understanding the different roles notifications, contexts and emotions play on notification engagement behaviour.

The key contributions of this chapter are:

1. An analysis of participant demographics and associations with smartphone usage and notification engagements
2. A method of representing the transient significance of a relationship between a notification subscriber and a known contact associated with a pushed notification
3. The identification of significant associations between contact/app significance and notification/context features
4. An analysis exploring the influence of environments and changing contexts on notification engagement behaviour
5. An exploration of perceived smartphone engagement etiquette in different contexts
6. The identification of significant associations between notification/smartphone engagement and goal-directed thinking

## 3.2 Native Mobile Push

This section discusses push-notifications in their native environment: the smartphone.

### 3.2.1 Data Collection

An Android application was developed as part of this in-situ study to analyse incoming push-notifications and the general smartphone usage of push-notification



subscribers. The WeAreUs app monitored push-notifications received by participants as they went about their daily lives and recorded their engagement behaviour, such as opening or dismissing a notification. The app also observed contextual states and general smartphone usage of participants at moments of interest and at regular intervals. The Ecological Momentary Assessment (EMA) methodology [SSH08] was used in this instance to sample participant notification behaviours and record their smartphone experiences for the purpose of identifying contextual associations and temporal sequences. The app generated questionnaires as part of an Experience Sampling Method (ESM) [CL14; VFK17], at moments when participants unlocked their device and received or acted on a notification.

Other methods of sampling participant perceptions of usage and engagement were considered, such as *Stimulated Recall* [Blo53]. With this method, notifications and engagements could have been observed unobtrusively in the background as participants went about their day. Participants could then have been later shown the notifications they received, actions taken on them and the context in which the actions occurred (e.g. using a map or other visualisations for aiding recall). Use of this method however, would have restricted the collection of certain contextual data recorded by the WeAreUs app in moments of interest, such as a participant's level of Hope as measured by the State Hope Scale. Responses using this method would also have been limited by a participant's autobiographical memory and could be vulnerable to recall bias. As such, the Experience Sampling Method was chosen for this study.

To entice use, the WeAreUs app enabled users to identify the contacts from which they received and accepted most notifications, encouraging them to reflect on the people they engaged with most via notifications throughout their day. Images reflecting the topics of notifications were also presented to the user to act as a brief snapshot of the trending *stories* happening in their lives at a given moment. Screenshots of the application are illustrated in Figure B.11 in Appendix B. The app collected data using two methods:

1. Background sensing - utilising Android's *Notification Listener Service* [And20g] and polling device sensors periodically using unobtrusive background services.
2. Experience Sampling Method - presenting questionnaires to the participant via notification for collecting both qualitative and quantitative feedback.

### 3.2.1.1 Background-Sensing

Two long running Android background services were created to unobtrusively monitor participant notification and smartphone usage engagement respectively:

1. Notification Bar - the *Notification Listener Service* provided by the Android SDK was used to identify moments when notifications were posted, updated or removed.
2. Usage Logging - at hourly intervals (once an internet connection was established), the state of the smartphone and its sensors, as well as other relevant aggregations of smartphone usage, were logged.

Table B.8 contains descriptions of each notification feature logged by the *Notification Listener Service* at moments of notification delivery and action. In order to calculate the action feature, a number of assumptions were necessary as the Android SDK did not, at the time of this study, provide methods for identifying whether the event triggered when a notification was removed from the device notification bar was subsequent to a participant opening or dismissing a notification. Therefore, to identify if a notification was opened, a background service polled the participant’s app usage history to determine if the app associated with the notification was opened after the notification was removed. If so, the notification was classified as opened, otherwise, dismissed. Other studies have used identical techniques for detecting notification actions [Vis+19; Meh+16]. It was also necessary to cache notifications that were posted to the notification bar and not yet acted upon by the user. This was due to the habit of apps updating the information of active notifications in the notification bar. To prevent duplicate notifications being logged, these updates were added to the existing notification recorded at the time it was first posted to the notification bar, thus preserving the original feature values, such as delivery time, which disrupted the participant.

The WeAreUs app measured the elapsed time between different notification engagement points of interest, defined by Mehrotra et al. [Meh+16]: seen time, decision time and response time. The seen time was measured as the time it took for a participant to notice a notification after it had arrived on their (locked) device. The *Notification Listener Service* logged the time the notification arrived to the locked device and measured the time elapsed until the device was next unlocked. If a notification arrived when the device was unlocked, the notification was assumed to be noticed immediately and the seen time was recorded as 0. In reality, this may not always hold true, as a participant could leave their device

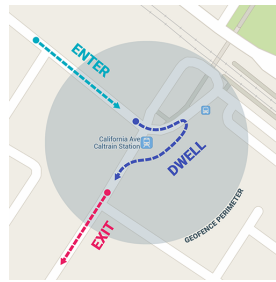


Fig. 3.1 Example of a geofence implemented using the Android SDK (image from Android’s Documentation [And20c])

unlocked and unattended, thus not noticing a notification for a period of time. Similarly a participant might unlock their device and not noticed a newly delivered notification. However, without further querying participants or using additional invasive sensors (e.g. eye-tracking), it was not possible to improve the accuracy of this feature. The decision time was measured as the time it took for a participant to act upon (open/dismiss) a notification once they had noticed it. The response time was measured as the total time taken to act on a notification after it arrived on the participant’s device (the sum of seen and decision times).

The *isClearable* and *isOngoing* features were used to identify notifications appearing in the notification bar which were not created for engagement, but used by apps for functional purposes such as running a background task or downloading updates. For example, the *Spotify* app uses a notification to show audio controls and playback stats. This notification is set to "ongoing" and is updated multiple times as tracks and playback position change. Similarly, the Google Play Store and other apps make use of a notification to display download progress. The Android system updates are also displayed as a notification which can’t be swiped away (dismissed). These notifications, while of some interest, are not particularly useful for identifying notification engagement behaviours as they appear silently and do not engage users.

Google’s Natural Language API [Clo20b] was used to extract keywords from the explicit notification text content and to infer the topic of the notification. Additional pre and post-processing steps were performed on the text to ensure phone numbers, email addresses, person names and other Personally Identifiable Information were not present. The full list of possible topic categories extracted from text content is available in the Natural Language Documentation [Clo20a]. Google’s Places API [Clo20d] was used to convert participant geolocation data to place type categories, such as *cafe* or *university*. Geofences (see Figure 3.1) of radius 50 meters were also constructed around places in which participants

received or acted on notifications to enable engagements, spanning different regularly visited locations, to be consolidated and allowing further inferences regarding context to be made. The *latitude* and *longitude* features were used only for registering geofences and identifying the *placeType* feature. They were stored only for the duration of the study on the device of the participant. The full list of possible place types inferred using geolocation coordinates is available in the Places Documentation [Clo20c]. The *activity* feature was created using Android’s Activity Recognition API [And20a; And20f] which would infer one of seven possible activities associated with the participant using the device sensors, such as running, on a bicycle or still. The *numUpdates*, *isContactSig* and *isAppSig* features are discussed fully in Sections 3.2.5.4, 3.2.5.1 and 3.2.5.2 respectively.

Table B.9 illustrates the usage statistics collected at hourly intervals by a background service. In recent updates to the Android SDK, apps were restricted from running background services for long periods [And20b]. Additionally, changes were implemented to optimise battery usage such that app services which received infrequent engagement from users were killed. Therefore an app was deployed in the cloud which would ping devices with the WeAreUs app installed at hourly intervals and launch a background service to query the *UsageStatsManager* [And20h] for aggregated usage metrics from the past hour. The reference to *significant* apps in features such as *totalSignificantAppsLaunchCount* refers to the number of launches of apps inferred by the WeAreUs app as significant based on past usage. A full description of the inference process is described in Section 3.2.5.2.

### 3.2.1.2 Experience Sampling Method

Two different ESM questionnaires were used to collect both qualitative and quantitative insights from participants regarding their actions and cognitive state in particular moments of interest:

1. Notification Action - at moments when a notification was acted upon (opened or dismissed), participants were prompted to answer questions pertaining to the notification and their current context.
2. Screen Unlock - at moments when they unlocked their smartphones, participants were prompted to answer questions regarding their current state-of-mind and their immediate context.

Table B.10 lists the descriptions of features collected using the questionnaire presented to participants immediately after acting on a notification. The goal of

this questionnaire was to gather participant thoughts regarding their perception of behaviour toward responding to a notification recently removed. Not all features listed in the table were presented in each generated questionnaire. The WeAreUs app randomly included either the *behaviour*, *relevance* or *mode* features so that the questionnaire was brief and focused on one particular facet of notification engagement at a time. A limit was also placed on the number of questionnaires that could be presented to participants on a single day. A maximum of 8 questionnaires, spaced at least 2 hours apart, was chosen as a limit to reduce the potential annoyance they might cause participants and to ensure they remained engaged with the study. A 5-point Likert scale was used for gathering quantitative feedback from participants and additional free-text fields were provided for explaining the responses if a participant felt it was necessary.

Table B.11 lists the features extracted from the second ESM questionnaire type presented to participants at moments they unlocked their smartphone. The *Hope* feature was calculated using the State Hope Scale questionnaire [Sny+96]. Section 3.2.6.3 describes the feature in full detail. On initial sign-up to the app, participants were also asked a number of demographic questions and given the option to sync their device contact lists to further inform the research. Table B.12 illustrates the relevant information gathered.

### 3.2.2 Participant Recruitment

Participants were recruited for the study by advertising the Android application throughout our research center and university. The app was made freely available on the Google Play Store, therefore it was accessible for download by the general public also - the benefit being participants would get a visual representation of their notifications through the application. No limit was placed on the number of participants wishing to join, although participants under the age of 18 were prohibited from opting-in to the study due to the difficulty involved in obtaining additional parental consent.

### 3.2.3 Privacy & Ethics

Before release, the project underwent rigorous analysis by the School's Research Ethics Committee due to the sensitive nature of the data being logged and use of participant questionnaires. The study was granted approval. Within the released

app, participants were given the option to opt-in for research on sign-up and also from the settings menu. On electing to opt-in participants were shown Participant Information and Consent documents (which outlined the project in detail, including processes regarding data gathering, security and life-time of use) and were asked to grant consent before proceeding. Only those who granted consent and were at least 18 years of age were given the option to opt-in and be included in the study. The participants could also choose to leave the study at any point with no penalty.

A number of considerations were given toward preserving participant privacy. Sensitive notification features such as contact names, notification text and locations were abstracted. For example, Google’s Natural Language API was used to classify the topic of the notification text and store in place of the full notification text. It was also used to extract people’s names. If a person’s name was found, it was compared with the participant’s phone contact list and a binary value indicating a matched contact was stored in place of the actual contact name. GPS coordinates were sent to Google’s Places API to identify the place a participant received, opened or dismissed a notification. The category of place was stored in place of the location coordinates.

Participants were also given the option to blacklist applications from which notifications would not be logged, in order to preserve privacy. 4 participants availed of this functionality, blocking a total of 6 apps between them. 5 participants also chose not to sync their phone or social contact lists, the remaining did, with their contact lists ranging from 58 to 1544 contacts.

### **3.2.4 Dataset**

The app was available to download on the Google Play Store for a period of 3 months. 18 participants (2 female, 13 male, 3 unknown) ranging in ages from 21 to 64 downloaded the WeAreUs app and granted consent to be included in the research study. However, notification data was missing or incomplete for 5 participants (1 female, 3 male and 1 unknown) due possibly to the battery optimisation schemes of certain devices preventing long running background tasks, such as the Notification Listener Service, to continue running indefinitely on a device. These 5 participants were excluded from analysis. The final dataset comprised of the remaining 13 participants includes 35,314 in-the-wild notifications and 7,123 smartphone usage logs. 309 ESM questionnaire responses were recorded

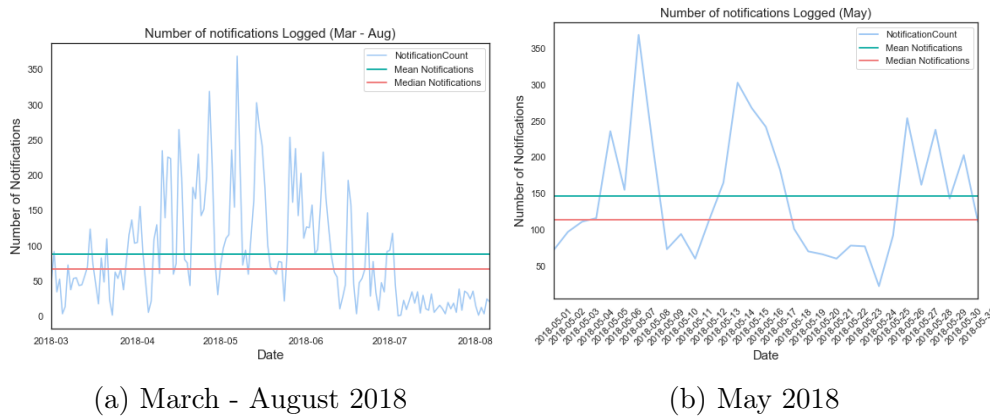


Fig. 3.2 Notifications recorded per day across all participants during the WeAreUs study period

of 2,242 sent, representing a response rate of  $\approx 14\%$ , between the months of March and August 2018.

Figure 3.2 illustrates notifications recorded per day for all participants over the study period. The number of notifications per day peaks mid-study, with over 350 notifications recorded. The number of notifications fall toward August as participants leave the study and the level of engagement with the WeAreUs application drop. The number of notifications tend to peak during weekends and fall back toward median values during the week. The month of May is illustrated as an example in Figure 3.2b. Notification numbers peak at over 350 on May 6th, possibly due to May 7th being a Bank Holiday in Ireland. Similarly, notification numbers peak again the following weekend (12th/13th) and again during the last weekend of the month (26th/27th). The weekend of the 19th/20th doesn't exhibit a spike in the number of notifications. This could be for any number of reasons, but lends evidence toward the need to move away from fixed rules governing notifications and credence for adaptive notification systems.

### 3.2.5 Context and Notification Engagement

An analysis of participant characteristics, smartphone usage and their propensity toward push-notification engagement can be found in section B.1. In this section, the relationship between participant context and notification engagement behaviour is explored. Notifications collected in-the-wild and corresponding ESM questionnaire responses were further assessed to provide insight into actions taken

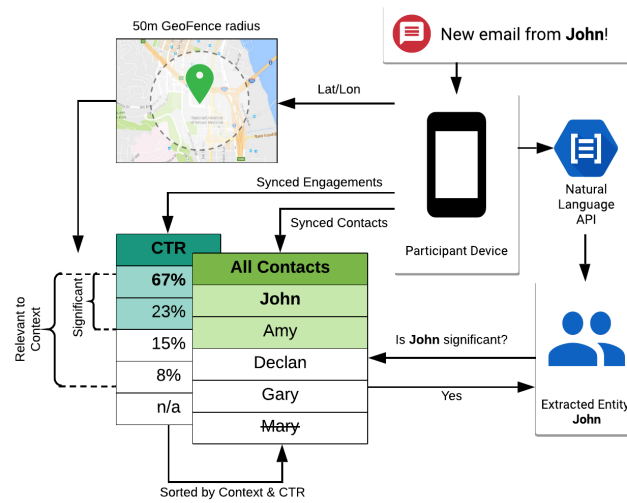


Fig. 3.3 Method used to identify a significant contact at moments of notification delivery

toward notifications by participants in different cognitive and physical states. The contributions of this section are summarised in Section 3.2.5.6.

### 3.2.5.1 The Role of Sender Relationship on Notification Engagement

Eight of the top ten apps generating the majority of notifications during the study were communication apps, such as *WhatsApp*, *Gmail* and *Snapchat*. Notifications from these apps are generally not pushed by brands or marketing teams attempting to re-engage customers, but by colleagues, friends and family members. This section explores the relationships between the personal contacts associated with notifications and the subsequent engagements that followed across different contexts and scenarios.

#### Identifying a Significant Contact

Other studies [PCD14; Meh+16] have highlighted the subtle changes in notification engagement when pushed by different sources with varying types of relationship to the receiver, such as friend, machine or colleague. In contrast, this study went a step further and assumed that a participant's relationship with a contact is not fixed to a particular category, but instead varies depending on their current context, such as physical place. Therefore, the WeAreUs app used the following method, illustrated in Figure 3.3, to determine whether or not a delivered notification contained a significant contact:



1. Retrieve all contacts of the participant (Phone, Facebook and Gmail contact lists synced on sign-up)
2. Identify the current place and all previous notification engagements (open/dismiss) which occurred in that place, (achieved using the geofence data stored with archived participant engagements)
3. Filter the list of contacts by keeping only those associated with a notification previously delivered to the participant in the context of the current place
4. Sort the remaining list of contacts based on the number of associated notifications the participant previously opened from them in the context of the current place (contacts from which the participant opened most notifications in the current place now appear first in the sorted list)
5. Apply Named Entity Recognition to the text content of notifications, when delivered, to extract names of potential contacts
6. Compare the extracted names with contact names of the sorted list using the Jaro-Winkler Similarity algorithm implemented using the Apache Commons Library [Com20]
7. If a match occurs in the top 15% of the sorted list, the contact is considered significant to the participant in that context

A limitation of this method included the difficulty in distinguishing contacts stored under similar names and participant contacts using different names across different platforms (e.g. a *Facebook* username differing from a *Snapchat* handle for the same contact). Although, it could be argued that this adds a layer of context, as the same contact might be considered significant to the participant on one platform but not as much on another (e.g. a notification from a colleague on *Snapchat* may not be as significant during work hours as a *Gmail* notification from that same colleague). Moreover, the context used to filter relevant contacts was restricted only to physical place. Additional contextual data points could have been used to further determine if a contact was relevant at moments of notification delivery, thus improving the accuracy of inferring a significant contact.

## Findings

The WeAreUs app was able to calculate contact significance at moments of notification delivery for 12,494 push-notifications. There were 2,076 cases where

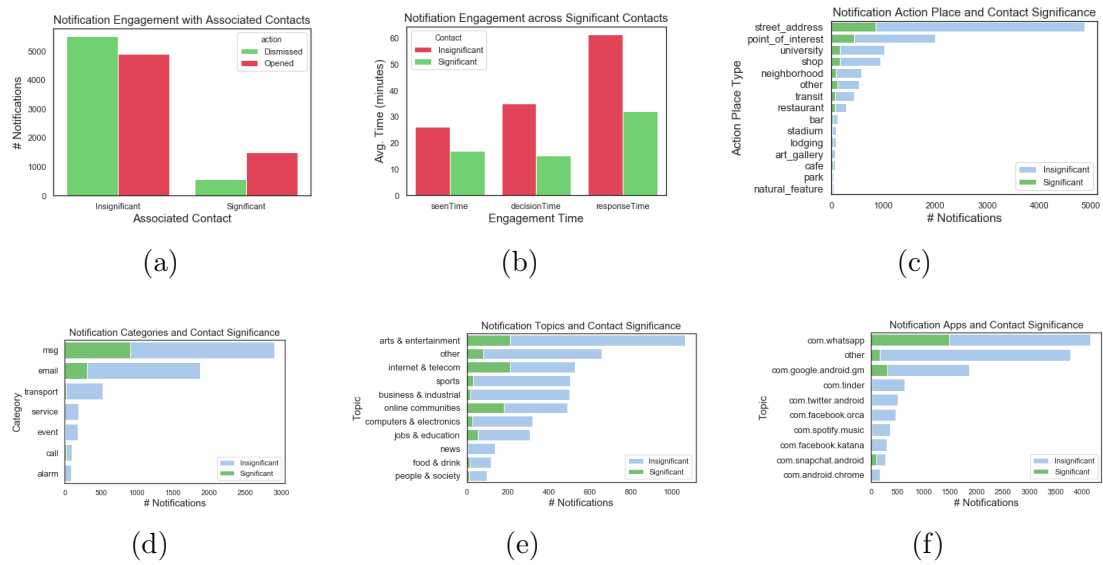


Fig. 3.4 Engagement behaviour of notifications associated with significant and insignificant contacts inferred at moments of notification delivery

the contact was significant and 10,418 cases where the contact was not significant with respect to participant context. 72.4% of notifications were opened when a significant contact was found compared to 47.1% when the contact was found not significant to the participant in their current context, a statistically significant difference in proportions of .25,  $p < .001$ . This illustrates the importance of both contact and context with respect to subsequent actions taken on notifications by participants in the study (Figure 3.4a).

A Mann-Whitney U test was run to determine if there were differences in the time taken by participants to notice notifications (seen time) from significant contacts. The median seen time was statistically significantly lower (11.95 minutes) for notifications with significant contacts associated than for those without a significant contact associated (17.28 minutes), as detailed in Table 3.1. Figure 3.4b illustrates that participants noticed notifications which included significant contacts quicker. An explanation for such a result might include participants anticipating, due to habits or communicative norms, an incoming notification from a significant contact and thus they check (unlock) their device regularly for updates. Alternatively, a significant contact may have been associated with a particular alert modality such as vibration pattern, ringtone, or flashing LED color, enabling a participant to distinguish between insignificant and significant contacts without directly interacting with their device.

Table 3.1 Engagement times for notifications with differing associated contact significance

| Feature       | isContactSignificant |                    | Mann-Whitney<br>U | $z$    | $p$   |
|---------------|----------------------|--------------------|-------------------|--------|-------|
|               | False<br>(Mdn mins)  | True<br>(Mdn mins) |                   |        |       |
| seen time     | 17.28                | 11.95              | 1252952           | -4.820 | <.001 |
| decision time | 2.02                 | 2.32               | 11103695          | 1.931  | 0.053 |
| response time | 8.93                 | 6.20               | 9882629           | -6.206 | <.001 |

To further analyse engagement times, a Mann-Whitney U test was run to determine if there were differences in the overall time taken by participants to respond to notifications (response time) between those with and without a significant contact associated. Distributions of response time for notifications with and without significant contacts associated were similar. The median response time was statistically significantly lower for notifications with significant contacts associated (6.20 minutes) than for those without a significant contact associated (8.93 minutes).

An important feature used when calculating the significance of a contact present in a notification was a participant's context at times of delivery. But participants often received notifications in one context and acted upon them in another. For example, an email notification could be delivered when the participant is at home but acted on (opened/dismissed) when at university. The context in which the user received and acted upon the notification were quite different. As such, analysing the relationship between places of notification action and whether the notification included a significant contact, calculated at the moment of delivery, sheds light upon places in which participants respond to significant and insignificant contacts with respect to their original contexts.

Within 11,215 notifications, the WeAreUs app succeeded in identifying both the place category and the significance of a contact at the moment of action (open/dismissal). A chi-square test of homogeneity was run, with an adequate sample size established according to Cochran [Coc54]. The two multinomial probability distributions were not equal in the population,  $\chi^2(14) = 78.218p < .001$ .

Observed frequencies and percentages of the types of places a participant acted upon a notification, and whether the notification was from a significant contact, are presented in Table B.13. Post-hoc analysis involved pairwise comparisons using multiple z-tests of two proportions with a Bonferroni correction. Statistical significance was accepted at  $p < .003$ . There were statistically significant differences in the proportion of notifications acted upon in cafes with and without a significant contact associated ( $n = 26$ , 1.3% versus  $n = 40$ , 0.4%). Similarly, the proportion of notifications acted upon at points of interest with a significant contact associated had statistically significant differences to those acted on without ( $n = 439$ , 21.1% versus  $n = 1563$ , 17.1%). There were no other statistically significant differences in the proportions of significant contacts present in notifications across place types where engagement actions occurred.

Analysing the time of day and day of week in which participants received notifications from significant contacts also yielded insight into notification engagement behaviour. From 12,494 notifications, the WeAreUs app identified the time of day, day of week and the significance of a contact. A chi-square test of homogeneity was run which indicated the two multinomial probability distributions were not equal in the population,  $\chi^2(4) = 57.134$   $p < .001$ . Observed frequencies and percentages of the time of day participants acted upon a notification, and whether the notification was from a significant contact, are presented in Table B.1. Post-hoc analysis involved pairwise comparisons using multiple z-tests of two proportions with a Bonferroni correction. Statistical significance was accepted at  $p < .01$ . There were statistically significant differences in the proportion of notifications received from significant and insignificant contacts during *early-morning*, *morning* and *evening* categories. In particular, participants received the highest proportion of notifications from significant contacts in the evening, 32% and from insignificant contacts during the mornings and afternoon, 31.6%.

A second chi-square test of homogeneity was run for days of the week. Again, the two multinomial probability distributions were not equal in the population,  $\chi^2(6) = 75.180$   $p < .001$ . Observed frequencies and percentages of days of the week participants acted upon a notification, and whether the notification was from a significant contact, are presented in Table 3.2. Post-hoc analysis involved pairwise comparisons using multiple z-tests of two proportions with a Bonferroni correction. Statistical significance was accepted at  $p < .007$ . There were statistically significant differences in the proportion of notifications received from significant and insignificant contacts on *Mondays*, *Tuesdays* and *Saturdays*. In particular, participants received the highest proportion of notifications from

Table 3.2 Observed frequencies and percentages of notifications responded to on different days with respect to contact significance

| Day         | isContactSignificant |                | <i>p</i> |
|-------------|----------------------|----------------|----------|
|             | False                | True           |          |
| <b>Mon</b>  | 1965<br>(18.9%)      | 290<br>(14.0%) | <.001*   |
| <b>Tues</b> | 1520<br>(14.6%)      | 250<br>(12.0%) | 0.002*   |
| Wed         | 1305<br>(12.5%)      | 266<br>(12.8%) | 0.719    |
| Thur        | 1301<br>(12.5%)      | 296<br>(14.3%) | 0.027    |
| Fri         | 1374<br>(13.2%)      | 261<br>(12.6%) | 0.447    |
| <b>Sat</b>  | 1568<br>(15.1%)      | 438<br>(21.1%) | <.001*   |
| Sun         | 1385<br>(13.3%)      | 275<br>(13.2%) | 0.953    |

significant contacts on Saturdays, 32%, and from insignificant contacts on Mondays, 31.6%. These results are interesting as they lend credence to preconceptions of social Saturday evenings spent communicating with significant contacts and Monday mornings spent dealing with communications from less significant contacts.

### 3.2.5.2 The Importance of Pushing with the Right App

Following the findings of the previous section in which many key associations were found between participants, their contacts and notification engagements, this section shifts attention toward the relationship between participants and the apps interrupting their lives with push-notifications. A large part of modern mobile marketing involves developing push-campaigns. Teams create these campaigns using reports and data to direct them toward an optimal set of notification feature values, such as the category, vibration pattern or time for delivery, that will result in maximising the CTR of notifications sent to their subscribers. Choosing the correct platform or app to send a push is pivotal. The app largely dictates the type of audience the notification is received by and under what terms. For example, consider a brand that has followers on Facebook, newsletter subscriptions on email and the ability to send direct messages via WhatsApp's Business API. All three are very different channels of communication and subscriber expectations

of notifications might be very different for each (e.g. WhatsApp used only for personal updates, emails for general updates). Each app will generate a push-notification for any post/message sent on their platform. But which app reflects the persona the marketing team wishes to target with their push? Which app is most relevant to the tone of the message? Which app will subscribers be most forgiving to if the notification annoys them? Questions such as these ravage marketing teams to a point where fear of getting it wrong, prevents any push from being sent.

### Identifying a Significant App

A method for identifying a significant app was put forth by Pielot et al. in their in-situ study of push-notifications [Pie+15]. The WeAreUs app uses this same method for calculating the importance of an app to a participant at moments of notification delivery.

1. Log the package names of all apps installed on a participant's device at first sign-up
2. During the background task run at hourly intervals, calculate the average app launch count i.e. the average number of times an app moved to the foreground of the participant's screen
3. During the same background task, categorise an app as significant if it has been launched more times than the average launch count
4. When a notification is delivered, extract the app package name and check the log indicating whether it is currently a significant app of the participant

As the background service which calculated app significance was run at hourly intervals, the set of significant apps at any one time reflected the most recent smartphone usage behaviours of participants. A limitation however, was the absence of other contextual features in determining app significance. In particular, the method of using places and Geo-Fences for context and using the CTR of notifications, delivered by apps identified as relevant, in those contexts. The method implemented in the WeAreUs app is illustrated in Figure 3.5.

### Findings

The WeAreUs app was able to calculate app significance at moments of notification delivery for 13,933 push-notifications. There were 7,464 cases where an app was significant and 6,469 cases in which an app was inferred as not significant. 69.3%

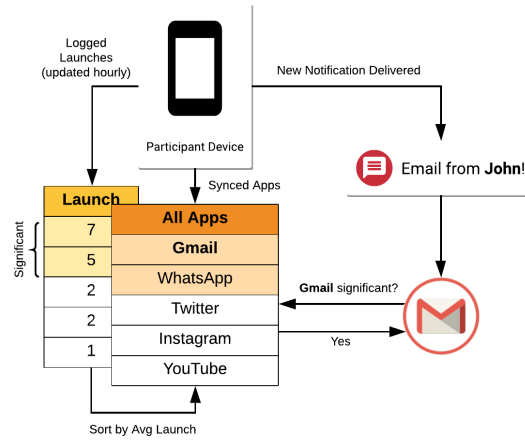


Fig. 3.5 Method of identifying app significance

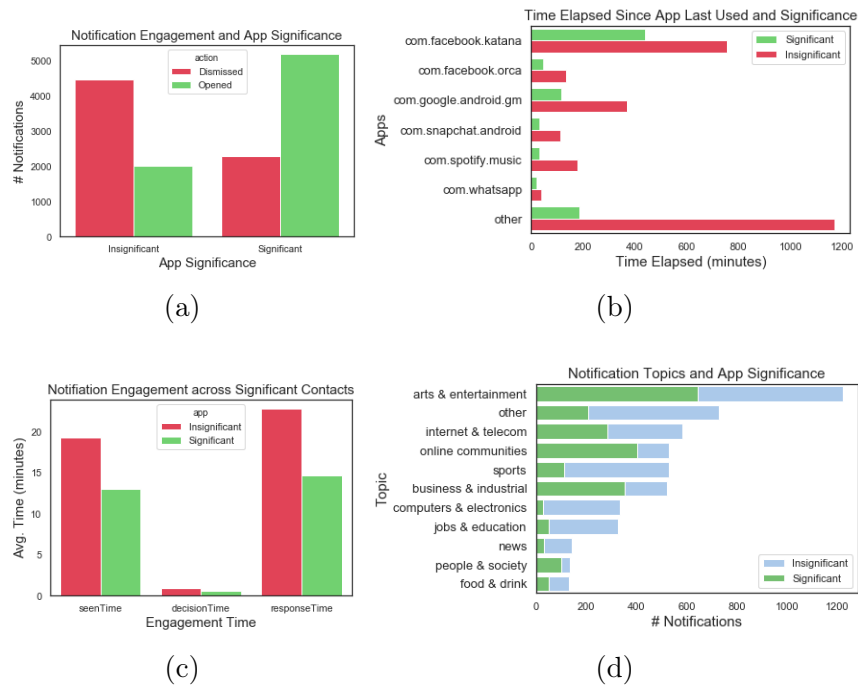


Fig. 3.6 Notification engagement behaviours across significant and insignificant apps

of notifications inferred as being pushed from a significant app were opened compared to 31.1% of apps pushed by insignificant apps. A chi-square test for association was conducted between app significance and actions (open/dismiss) taken toward notifications. All expected cell frequencies were greater than five. There was a statistically significant association between app significance and action,  $\chi^2(1) = 2029.506, p < .001$ . The association was moderately strong, Cramer's  $V = .328$ . Figure 3.6a illustrates many more notifications were opened by participants when an app was inferred as significant as opposed to insignificant at moments of delivery. Following findings from the previous section, a chi-square test for association was also conducted between app significance and contact significance. A small to moderate, statistically significant, association was identified between both features,  $\chi^2(1) = 588.876, p < .001$ , Cramer's  $V = .217$ . Of the 10,418 notifications found to not have a significant contact associated, 47% were inferred to be from significant apps. Of the 2,076 notifications found to have a significant contact associated, 76.2% were inferred to be from a significant app. These results illustrate that app significance could be used as another factor of determining whether a contact is significant to the context.

A Mann-Whitney U test was run to determine if there were differences in elapsed time between a notification delivery and the time an app was last used, between apps inferred as significant and insignificant. Distributions of the elapsed time were similar for both groups. The median elapsed time was statistically significantly lower for significant apps than insignificant apps, see results in Table 3.3. Additionally, seen and response times were also statistically significantly lower for apps inferred as significant, illustrated in Figure 3.6c. Again, the alert modality with which notifications were delivered could be a contributing factor to reduced seen time. For example, Figure B.9 in Appendix B illustrates that *Snapchat* uses yellow predominantly for the LED colour flashing when notifications are delivered, thus participants would be able to distinguish the app, and its relative significance, quicker. In contrast, decision times for notifications pushed by significant apps were statistically significantly longer than insignificant apps, indicating that participants took more time deliberating over notifications sent by significant apps before acting on them. The number of times a notification was updated to reflect new information was also statistically significantly larger for apps inferred as significant perhaps suggesting these apps contained information that changed or was supplemented to regularly, e.g. notifications of a WhatsApp group chat or updates on football match scores.



Table 3.3 Results of Mann-Whitney U tests run to determine if there were differences between significant and insignificant apps on a number of continuous notification features.

| Feature                     | isAppSignificant |            | $U$       | $z$        | $p$      |
|-----------------------------|------------------|------------|-----------|------------|----------|
|                             | False (Mdn)      | True (Mdn) |           |            |          |
| app last used (mins)        | 122.11           | 24.42      | 3358236.5 | -21.838462 | <.001*   |
| <b>number of updates</b>    | 2                | 3          | 3291867.5 | 7.788007   | <.001*   |
| relevance (likert scale)    | 2                | 3          | 7         | 0.655186   | 0.727273 |
| seen time (mins)            | 19.23            | 13         | 2955624.5 | -8.393946  | <.001*   |
| <b>decision time (mins)</b> | 1.68             | 2.51       | 26853136  | 11.448545  | <.001*   |
| response time (mins)        | 11.70            | 6.74       | 20745365  | -14.346190 | <.001*   |

Table 3.4 Results of chi-square tests of homogeneity run to determine if there were differences in probabilities between apps inferred as significant and insignificant.

| Feature           | $\chi^2$ | $df$ | $p$   |
|-------------------|----------|------|-------|
| topic             | 938.145  | 10   | <.001 |
| category          | 2120.56  | 3    | <.001 |
| LED colour        | 1024.473 | 5    | <.001 |
| vibration pattern | 256.834  | 5    | <.001 |
| priority level    | 934.523  | 4    | <.001 |

A number of chi-square tests of homogeneity were run to determine if there were statistically significant differences in probabilities between apps inferred as significant and insignificant on multiple multinomial notification features: topic, category, LED colour, vibration pattern and priority level. The test results can be found in Table 3.4.

Observed frequencies and percentages of notification topic, and whether the notification was pushed by a significant app, are presented in Table B.14. Post-hoc analysis involved pairwise comparisons using multiple z-tests of two proportions with a Bonferroni correction. Statistical significance was accepted at  $p < .005$ . There were statistically significant differences in the proportions of many topics of notification received from significant and insignificant apps. Of particular interest were the difference of proportions found in notifications with *business & industry*, *online communities* and *people & society* topics. The proportion of notifications with these topics was higher across apps inferred as significant as opposed to proportions found across insignificant apps.

Observed frequencies and percentages of notification categories, and whether the notification was pushed by a significant app, are presented in Table 3.5.

Table 3.5 Crosstabulation of app significance and notification category (significance accepted at  $p < .001$ ).

| Category   | isAppSignificant |                 | $p$       |
|------------|------------------|-----------------|-----------|
|            | False            | True            |           |
| email      | 1895<br>(52.8%)  | 161<br>(5%)     | $<.001^*$ |
| <b>msg</b> | 1019<br>(28.4%)  | 2546<br>(78.3%) | $<.001^*$ |
| other      | 656<br>(18.3%)   | 537<br>(16.5%)  | $<.001^*$ |
| recommend  | 16<br>(0.4%)     | 7<br>(0.2%)     | 0.026     |

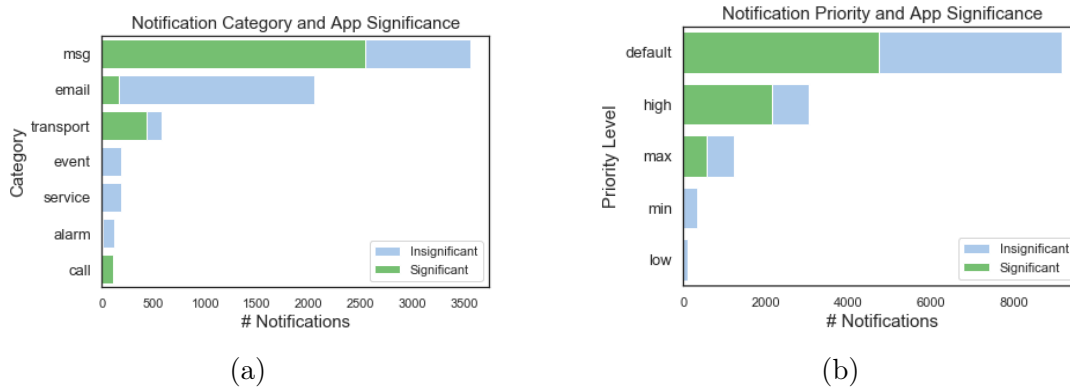


Fig. 3.7 Distribution of notification category and priority level across significant and insignificant apps

Post-hoc analysis involved pairwise comparisons using multiple z-tests of two proportions with a Bonferroni correction. Statistical significance was accepted at  $p < .001$ . There were statistically significant differences in the proportions of all but one categories of notification received from significant and insignificant apps: *recommendations*. Of particular interest were the difference of proportions found in notifications with the *msg* category as the proportion of notifications with this category were higher across apps inferred as significant as opposed to proportions found across insignificant apps. The result illustrates that a large share of notifications received from significant apps were categorised as messages. This is illustrated also in Figure 3.7a as notifications with a *msg* category far outnumbered all others. Also of interest is that notifications with email set as category were more frequently attributed to insignificant apps.

### 3.2.5.3 Environmental factors

The WeAreUs app recorded a number of features using device sensors which further described the environments in which participants resided at moments of notification delivery and action. The features include: place, ambient noise and ambient light.

#### Places

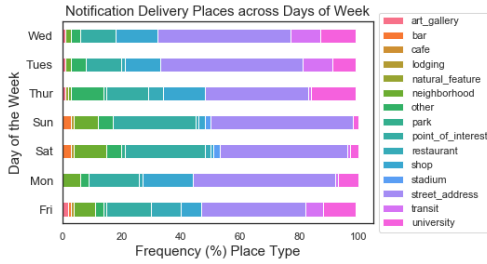
The place type was recorded using the device GPS coordinates at moments of notification delivery and action. The coordinates were sent to Google's Places API, which returned a place type for the participant's physical location. GPS coordinates for places were stored securely on a participant's device for the purpose of constructing a geofence around the places they frequented such that notification engagement metrics for particular places could be stored.

The WeAreUs app successfully identified a place type at the moment of delivery for 4,491 notifications. 2,790 (62.1%) of these notifications were subsequently dismissed and 1,701 (37.9%) were later opened. It is important to note that notification actions may have taken places other than the delivery location (see Figure B.4 and Section 3.2.5.5 concerning changes of context between moments of notification delivery and action). The purpose of this analysis is to identify the place types in which participants received notifications and the subsequent action taken on them. Table 3.6 illustrates which action was later taken on notifications delivered at differing place types. In particular, proportions of notifications accepted in the lodging and natural feature place types were statistically significantly different to other place types,  $p < .001$ . Post-hoc analysis involved pairwise comparisons using the z-test of two proportions with a Bonferroni correction. The results showed that proportions of notification actions made in the university, street address and lodging place types were all statistically significantly different. Notifications delivered at the lodging place type had the highest subsequent CTR, 76%, and notifications delivered at the park place type had the lowest, 20%.

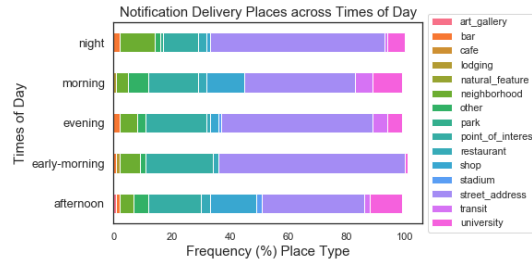
The WeAreUs app also successfully identified a place type at the moment of action for 11,485 notifications. 6,009 (52.3%) of these notifications were dismissed and 5,476 (47.7%) were opened. Table 3.7 illustrates which action was taken on notifications at differing place types. In particular, proportions of notifications acted on in bar, cafe, lodging and university place types were statistically significantly different,  $p < .001$ , as identified by a post-hoc analysis involving pairwise comparisons using the z-test of two proportions with a Bonferroni correction. The cafe place type had the highest notification CTR, 78.8%, closely followed by

Table 3.6 Place types in which notifications were received and the subsequent action taken

| Place Delivered       | Action    |    |        |    |
|-----------------------|-----------|----|--------|----|
|                       | Dismissed |    | Opened |    |
|                       | N         | %  | N      | %  |
| art gallery           | 15        | 63 | 9      | 38 |
| bar                   | 29        | 60 | 19     | 40 |
| cafe                  | 2         | 50 | 2      | 50 |
| <b>lodging</b>        | 4         | 24 | 13     | 76 |
| natural feature       | 4         | 36 | 7      | 64 |
| neighborhood          | 158       | 63 | 93     | 37 |
| other                 | 469       | 57 | 356    | 43 |
| park                  | 16        | 80 | 4      | 20 |
| point of interest     | 459       | 56 | 368    | 45 |
| restaurant            | 75        | 68 | 35     | 32 |
| stadium               | 20        | 65 | 11     | 36 |
| <b>street address</b> | 1264      | 64 | 704    | 36 |
| <b>university</b>     | 275       | 78 | 80     | 23 |



(a)



(b)

Fig. 3.8 Place types at moments of notification delivery and action

lodging, bar and natural feature. The lowest notification CTRs were in university, 28.8%, park and street address place types.

A chi-square test of independence was conducted between place types, the time of day and the day of the week a notification was acted upon. There was a statistically significant association between both the time of day and day of week for different place types,  $\chi^2(56) = 1311.689, p < .001$  and  $\chi^2(84) = 3007.076, p < .001$ , respectively. The association was small to moderately strong between the day of the week and place type, Cramer's  $V = .209$ , and small between the time of day and place type, Cramer's  $V = .169$ .

Table 3.7 Place types in which notifications were acted on (opened/dismissed)

| Place | Action            | Action    |      |        |      |
|-------|-------------------|-----------|------|--------|------|
|       |                   | Dismissed |      | Opened |      |
|       |                   | N         | %    | N      | %    |
|       | art gallery       | 35        | 46.1 | 41     | 53.9 |
|       | bar               | 39        | 30.5 | 89     | 69.5 |
|       | cafe              | 14        | 21.2 | 52     | 78.8 |
|       | lodging           | 22        | 24.4 | 68     | 75.6 |
|       | natural feature   | 14        | 32.6 | 29     | 67.4 |
|       | neighborhood      | 330       | 51.5 | 311    | 48.5 |
|       | other             | 293       | 52.8 | 262    | 47.2 |
|       | park              | 36        | 69.2 | 16     | 30.8 |
|       | point of interest | 826       | 41.2 | 1178   | 58.8 |
|       | restaurant        | 137       | 46.6 | 157    | 53.4 |
|       | shop              | 363       | 38.1 | 589    | 61.9 |
|       | stadium           | 39        | 40.2 | 58     | 59.8 |
|       | street address    | 2874      | 58.2 | 2064   | 41.8 |
|       | transit           | 198       | 44.9 | 243    | 55.1 |
|       | university        | 789       | 71.2 | 319    | 28.8 |

### Ambient Noise

The ambient noise feature was recorded using open source sensor libraries [Lat+13b] of the *EmotionSense* (ES) platform [Rac+10]. The microphone of a participant's device was used to record 500 millisecond clips of audio at delivery and action moments. The average amplitude was recorded, converted to decibel and the audio file was then discarded. The decibel values were further assigned to categories for greater understanding of ambient noise levels, similar to Mehta et al. [MZC12].

Goodman and Kruskal's  $\lambda$  was run to determine whether the action taken toward a notification by the participant could be better predicted by knowledge of the ambient noise present in the participant's environment at the moment of action. Goodman and Kruskal's  $\lambda$  was .056. This was a statistically significant reduction in the proportion of errors due to using ambient noise as a predictor of participant actions toward a notification,  $p < .001$ . Figure 3.9a illustrates actions taken on notifications in differing noise environments. The majority of notifications were dismissed in environments with noise levels similar to quiet rooms and washing machines, while the majority were opened in environments with noise levels similar to normal talking, city traffic and lawnmowers.

A Kruskal-Wallis H test was also run to determine if there were differences in the ambient noise levels between the 15 (most frequent) place types at moments

when notifications were acted upon by participants. Distributions of ambient noise levels were not similar for all groups. The distributions of ambient noise was statistically significantly different between place types,  $\chi^2(14) = 3250.437, p < .001$ . Figures 3.9d and 3.9c depict ambient noise levels at places of notification delivery and action. The bar and stadium place types understandably had the highest ambient noise levels,  $\approx 85\text{dB}$ , and the natural feature place type had the lowest ambient noise level,  $\approx 41\text{dB}$ . The majority of place types had ambient noise levels similar to talking or city traffic.

Additionally, Kendall's tau-b correlation was run to determine the relationship between ambient noise levels and notification engagement times (seen, decision and response). There were no statistically significant associations between ambient noise and response time, however there were weak, negative associations between ambient noise, seen and decision times which were statistically significant. The strongest of these associations was between ambient noise levels at moments notifications were delivered and the amount of time elapsed before the notification was seen,  $\tau_b = -.111, p < .001$ , weakly illustrating that in higher ambient noise environments, notifications were noticed quicker. This could suggest that participants were cognitively aware they may not hear notification alerts on arrival and subsequently check for notifications more frequently than in lower ambient noise environments. Figure 3.9b shows average seen, decision and response times recorded for each ambient noise category. Ambient noise levels similar to a quiet room resulted in longer seen, decision and response times while levels similar to city traffic resulted in shortest times.

### Ambient Light

Similarly, the ES libraries were also used to record ambient light intensity levels at moments of notification delivery and action. Light intensity (luminance) values were also further labeled into categories for better understanding of device environments at moments of interest. Elazhary et al. [Ela+17] identify two sources, defined by Microsoft and Android [Bri20; And20e], for discretising continuous luminance values collected through a mobile device. Figure 3.10a illustrates notification engagements in differing ambient light environments.

Goodman and Kruskal's  $\lambda$  was run to determine whether the action taken toward a notification by the participant could be better predicted by knowledge of the ambient light intensity of the participant's environment at the moment of action. Goodman and Kruskal's  $\lambda$  was .057. This was a statistically significant

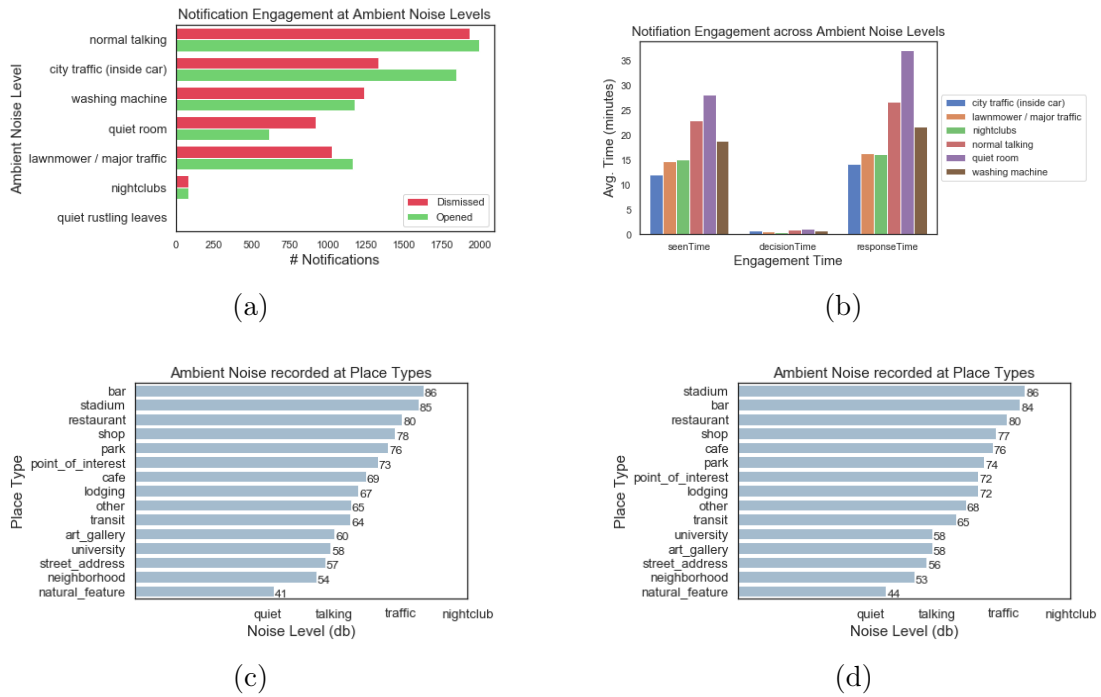


Fig. 3.9 Ambient noise at moments of notification delivery and action

Table 3.8 Results of Kendall's tau-b correlation for determining relationships between ambient light and notification engagement features

| Feature       | moment of delivery |        | moment of action |        |
|---------------|--------------------|--------|------------------|--------|
|               | $\tau_b$           | $p$    | $\tau_b$         | $p$    |
| seen time     | <b>-.145</b>       | <.001* | -0.033           | .001*  |
| decision time | 0.02               | 0.03*  | -0.006           | 0.334  |
| response time | -0.087             | <.001* | -0.01            | 0.081  |
| relevance     | -                  | -      | <b>-0.503</b>    | 0.042* |

reduction in the proportion of errors due to using ambient light as a predictor of participant actions toward a notification,  $p < .001$ .

A Kruskal-Wallis H test was run to determine if there were differences in the ambient light intensity levels between the 15 (most frequent) place types at moments when notifications were acted upon by participants. Distributions of ambient light intensity levels were not similar for all groups. The distributions of ambient light intensity was statistically significantly different between place types,  $\chi^2(14) = 1417.013, p < .001$ .

A Kendall's tau-b correlation was run to determine the relationship between ambient light intensity levels, notification engagement times (seen, decision and response) and the relevance of a notification as noted by participants at moments

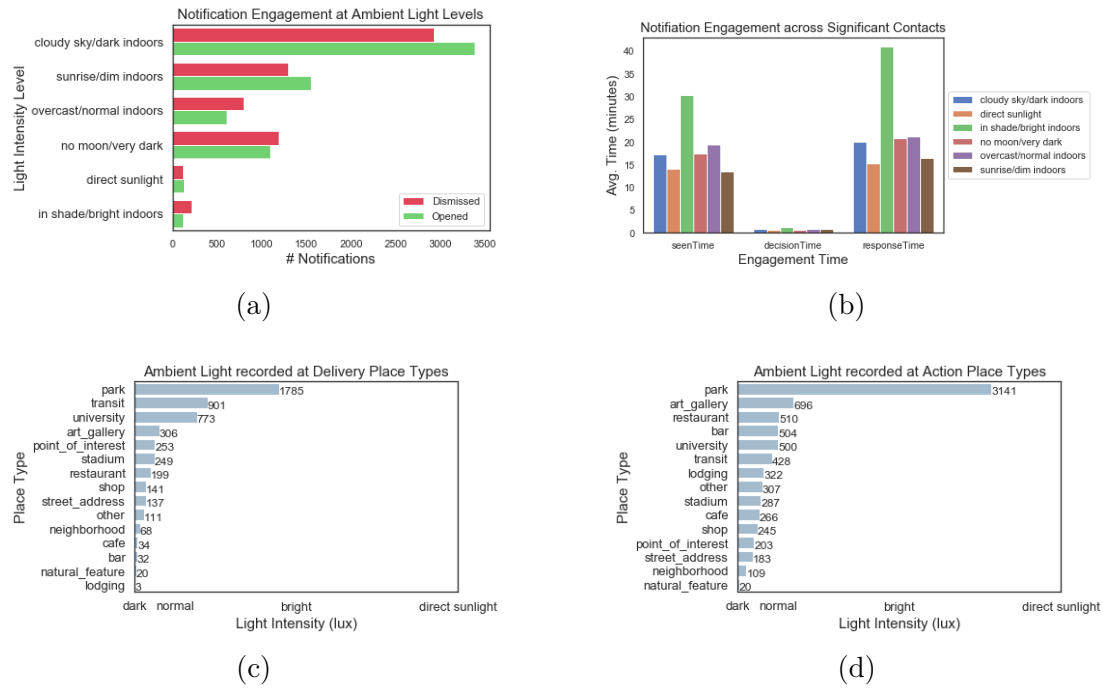


Fig. 3.10 Ambient light at moments of notification delivery and action

of action. Decision and response times had negligible statistically significant associations with ambient light intensity at moments of notification delivery and action. The time elapsed before a notification was noticed (seen time) had a statistically significant weak negative association with ambient light intensity. Interestingly, there was a statistically significant, moderate negative association of  $-0.503$  between ambient light intensity and the relevance of a notification. This illustrates that participants, during this study and in lower light environments received more relevant notifications, or in higher light environments received less relevant notifications. Figure 3.10b highlights that notifications engaged with in environments with ambient light levels similar to *shade* or *bright indoors*, had longer engagement times compared to other ambient light environments. Environments with ambient light levels similar to *dim indoor* lighting had the shortest engagement times. Additionally, Figures 3.10c and 3.10d illustrate the ambient light intensity levels at moments of notification delivery and action in different places.

### 3.2.5.4 Updated Notifications

When notifications were delivered to participant device's, the WeAreUs app stored it in an active notification cache, until such time as the participant made an action



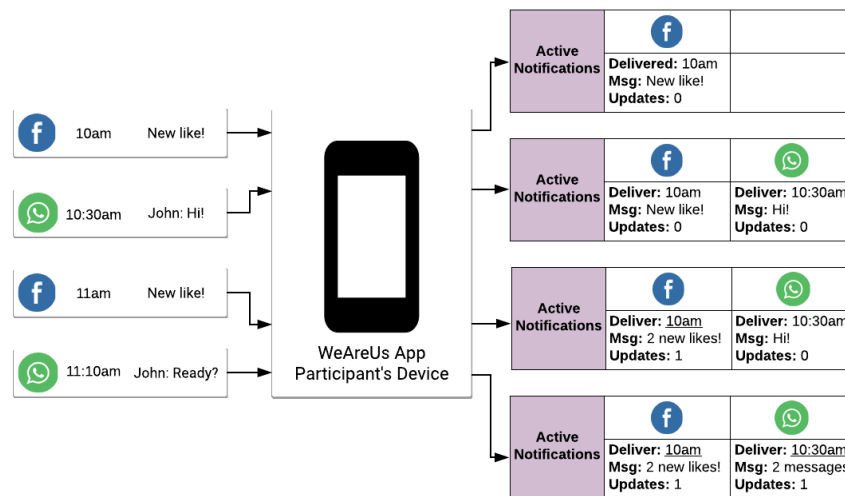


Fig. 3.11 Method of tracking updated notification content implemented in WeAreUs app

(open/dismiss) on it. Between the moments of delivery and action, apps had the opportunity to update the existing notification. Such updates have the ability to alert users multiple times to the same notification but with updated content. An example of this are instant messaging apps updating the notification to reflect the most recent message posted (which is sometimes accompanied by a counter which identifies the number of unread messages). When updates occurred during the WeAreUs study, a search of the active cache was triggered for existing notifications using the app package name and unique notification identifiers, such as the id tag. The app name could not be used alone, as some apps deliver different types of notification for different events. For example, *Twitter* uses separate notification types for tweets and direct messages sent via their platform. If a notification is found in the active cache, the *numberOfUpdates* feature was incremented and the additional information contained within the new notification was added to the existing stored content. For example, any text contained within the new notification that was not in the original, was appended. The original time and date the notification was delivered was retained and used as the notification delivery time. In this manner, the number of times an app updated (alerted/distracted) participants, before an action occurred, was recorded. Figure 3.11 illustrates this method using a simple example.

A Mann-Whitney U test was run to determine if there were differences in the number of times a notification was updated and the action (open/dismiss). Distributions of the number of updates for open and dismiss actions were not

similar. The number of updates for opened (mean rank = 7017.49) and dismissed (mean rank = 6919.63) notifications were not statistically significantly different,  $U = 23901004.5$ ,  $z = 1.684$ ,  $p = 0.092$ . Regardless of the subsequent action taken on the notification, constant updates of notifications are a source of distraction for notification subscribers. Analysing the apps, topics and behaviours toward notifications which produce the most number of updates, and therefore distractions, can provide further insight of the types of notifications creating the majority of smartphone interruptions. Figure 3.13a illustrates the actions taken on notifications which had no, one or multiple updates during the period of time they were in the active notification cache. Note that the majority of notifications were opened when the number of updates were equal to or less than one and that the largest proportion of notifications delivered were not updated.

A Kruskal-Wallis H test was conducted to determine if there were differences in the number of updates between different apps. Distributions of notification updates were not similar for all apps. Higher update values were found in *Chrome*, *Messenger* and *WhatsApp* apps, as reflected in their mean ranks of 12201, 8407 and 7302 respectively. The mean ranks of the number of updates were found to be statistically significantly different between groups,  $\chi^2(9) = 2405.413$ ,  $p < .001$ . A Kruskal-Wallis H test was also conducted to determine if there were differences in the number of updates between different notification topics. Distributions of notification updates were not similar for all topics. Higher update values were found in the *Internet & Telecoms*, *Online Communities* and *People & Society* topics, which was reflected in the largest 3 mean ranks of 3332, 3344 and 2919 respectively. The mean ranks of the number of updates were found to be statistically significantly different between groups,  $\chi^2(10) = 587.342$ ,  $p < .001$ . Figures 3.13c and 3.13b illustrate the average number of updates occurring in notifications generated by apps and associated with distinct topics, respectively. The *Online Communities* topic is, understandably, associated with the most notification updates. Surprisingly, the *news* topic generates the least number of updates. An explanation could be that the content of news articles changes infrequently compared to instant messaging conversations. Once published and a notification sent, a news app may not send an additional notification to notify subscribers of the event of modifying an article for instance. Of course, multiple exceptions might occur, such as breaking news, or sports updates. *Gmail* was found to be the app updating the majority of notifications delivered, followed by other social network and communication apps. *Twitter* and *Sky Sports* were two of the apps generating the least number of notification updates on average.

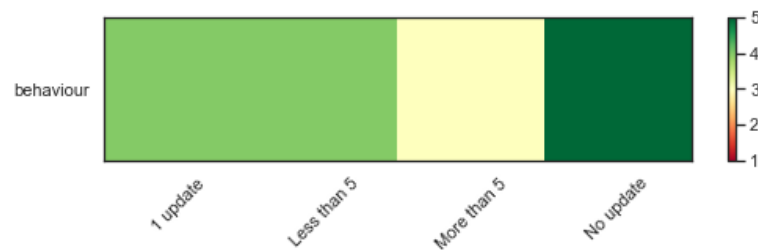


Fig. 3.12 Notification behaviour perception (1:VeryPoor; 2:Poor; 3:Neutral; 4:Good; 5:VeryGood) over notifications with differing levels of updates

The WeAreUs app sent participants questionnaires at moments of notification engagement, such as when opening or dismissing a notification. As part of the questionnaire, participants were asked to indicate on a 5-point Likert scale ranging from 1 (very poor behaviour) to 5 (very good behaviour), whether they perceived their action toward the notification, just opened or dismissed in their current context, to be good or bad behaviour. The number of responses to the behaviour question was low (21 responses), thus no test was run to determine if there were statistically significant differences in the perceived behaviour between notifications which were updated and those which were not. However, Figure 3.12 indicates that the perceived behaviour for updated notifications was lower than that of notifications without updates. Thus illustrating that participants, on reflection, seemed to think their behaviour was better when confronted with notifications that had not been updated (sounding additional visual or audible alerts) by the app which delivered them, and slightly worse by comparison when updates occurred. This might be an indication that constant updates/interruptions of a notification could lead subscribers to yield when abstaining from checking a notification, e.g. while occupied with another task.

### 3.2.5.5 Changing Contexts

The WeAreUs app recorded features relating to participant context at moments in which notifications were delivered to their device. The app logged the same set of context features at moments participants acted on a delivered notification. By comparing the context features at both time points, changes in participant context between moments of delivery and action were identified.

The set of context features recorded at both time points which were analysed are:

$\{ activity, airplaneMode, noise, batteryLevel, bluetoothHeadphones, bluetoothOn,$

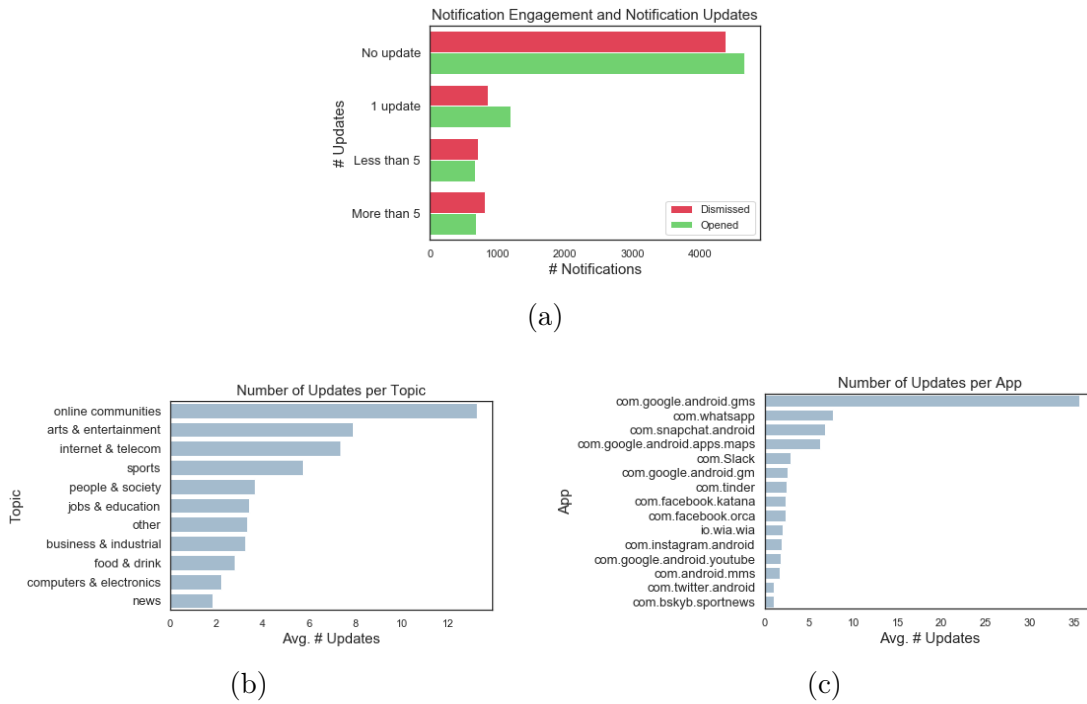


Fig. 3.13 The relationship between multiple notification updates and notification behaviours

*charging, connectionType, headphonesIn, luminance, microphoneMute, musicActive, numberBluetoothDevices, placeCategory, proximity, ringerMode, roamingStatus, stepCount, volumeAlarm, volumeMusic, volumeNotification, volumeRinger, timeOfDay, dayOfWeek* }

A dichotomous variable was created for each feature in the set above indicating whether or not the value had changed between delivery and a participant acting on the particular notification. An additional *anyContextChanged* variable was also created to indicate if any feature in the set had changed between the two time points. A test of two proportions was carried out between each individual context feature and the action (open/dismiss) taken on the notification to identify potential associations between differing contextual states and notification engagements. The results are recorded in Table 3.9.

The test of two proportions used was the chi-square test of homogeneity. Fisher's exact test was used in cases where there was not a sufficient sample size, denoted by (*F*) in the Table. There was a statistically significant difference between the proportions of notifications opened and dismissed for 13 notification features, including the *anyContextChanged* feature, highlighted by an asterisk in the table. *N* indicates the number of notifications which were used in the test for each feature. Not all notifications could be used for each test as the WeAreUs

Table 3.9 Identifying differences between opened and dismissed notification context states using test of two proportions (\* denotes significant results; (F) indicates Fisher’s Exact Test was used due to insufficient sample size)

| Feature               | Notifications<br>Opened (%) |               | Majority<br>Direction | <i>N</i> | <i>p</i>   |
|-----------------------|-----------------------------|---------------|-----------------------|----------|------------|
|                       | Did Not<br>Change           | Did<br>Change |                       |          |            |
| activity              | 49.472                      | No change     | -                     | 6539     | NA         |
| airplaneMode          | 49.838                      | 0             | -                     | 11442    | 1 (F)      |
| noise                 | 56.25                       | 38.120        | -                     | 4848     | 0.136      |
| <b>batteryLevel</b>   | 40                          | 36.559        | decreases             | 4848     | 0.014*     |
| <b>bluHeadphones</b>  | 52.018                      | 38.950        | not->active           | 13933    | <.001*     |
| bluetoothOn           | 51.608                      | 51.525        | -                     | 13933    | 0.935      |
| <b>charging</b>       | 52.675                      | 42.225        | not->charging         | 13933    | <.001*     |
| connectionType        | 38.344                      | 35.271        | -                     | 4848     | 0.323      |
| <b>headphonesIn</b>   | 52.832                      | 31.164        | out->in               | 13933    | <.001*     |
| <b>lightIntensity</b> | 30.281                      | 40.507        | decreases             | 4848     | <.001*     |
| micMute               | 51.59                       | No change     | -                     | 13933    | NA         |
| <b>musicActive</b>    | 53.377                      | 39.130        | not->active           | 13933    | <.001*     |
| <b>numBluDevices</b>  | 36.635                      | 30.017        | decreases             | 4389     | 0.002*     |
| placeCategory         | 37.944                      | 38.009        | -                     | 4469     | 0.979      |
| <b>proximity</b>      | 37.409                      | 40.705        | decreases             | 4848     | 0.045*     |
| ringerMode            | 38.161                      | 40            | -                     | 4848     | 0.790      |
| roamingStatus         | 38.507                      | 62.5          | -                     | 4776     | 0.274 (F)  |
| stepCount             | 37.753                      | 35.035        | -                     | 4702     | 0.169      |
| <b>voluAlarm</b>      | 38.104                      | 100           | decreases             | 4848     | 0.003* (F) |
| voluMusic             | 38.542                      | 35.305        | -                     | 4848     | 0.144      |
| <b>voluNoti</b>       | 37.997                      | 50            | decreases             | 4848     | 0.035*     |
| <b>voluRinger</b>     | 37.984                      | 50.667        | decreases             | 4848     | 0.025*     |
| timeOfDay             | 51.311                      | 53.783        | -                     | 13933    | 0.065      |
| <b>dayOfWeek</b>      | 51.376                      | 59.162        | Sun->Mon              | 13933    | 0.003*     |
| <i>anyChange</i>      | 68.706                      | 44.540        | -                     | 13957    | <.001*     |

app didn't successfully capture values for all features at both delivery and action time points. In particular, on delivery of notifications, if the device was locked, background services were limited and identifying some features, such as place type, consistently was found to be difficult. The *majority direction* column attempts to indicate the average direction in which the value changed, when a change was detected. In the case of discrete features, such as *dayOfWeek*, the frequency at which a value changed to another was used to indicate majority direction. For example, the most frequent change identified in the *dayOfWeek* feature was a notification delivered during Sunday and acted on (opened/dismissed) during Monday.

### **Change in context detected**

The *anyContextChanged* feature identifies notifications which were delivered in one context and acted on in another. Of 4,103 notifications which did not have any change in context, 68.7% were opened in contrast with 44.5% of the 9,854 notifications which were acted on after a change. This demonstrates that changing contexts could be a contributing factor of notification dismissal. Apps have the opportunity to remove notifications that have been delivered, if they feel they are no longer relevant. An improved notification experience could be to remove a notification when a change in context occurs and re-deliver once a stable context has been inferred. Additionally, the direction of the context change could also be used to improve the relevancy of notifications such that if a change in context occurred and the change direction was favorable to the notification, it would still be delivered. For example, headphones switched from in to out or music switched from on to off.

### **Battery Level**

CTR of 2,285 notifications with no change of battery level was 40%. When a change in battery level was detected, in 2,563 notifications, the CTR was 37% and the direction was found to be, on average, decreasing. Notifications with no changes in battery level between delivery and action times suggest they were either dealt with quickly on arrival or there was little smartphone activity between both time points. A drain of battery might be indicative of a participant engaged with other activities on their device, thus the interruption may not have been an opportune moment. Figures 3.14a and 3.14b illustrate small changes of battery level ( $\leq 5\%$ ) which occurred between notification delivery and action. The colour and strength of edges indicate the number of notifications in which the change occurred and whether the resulting notifications were opened or dismissed. A

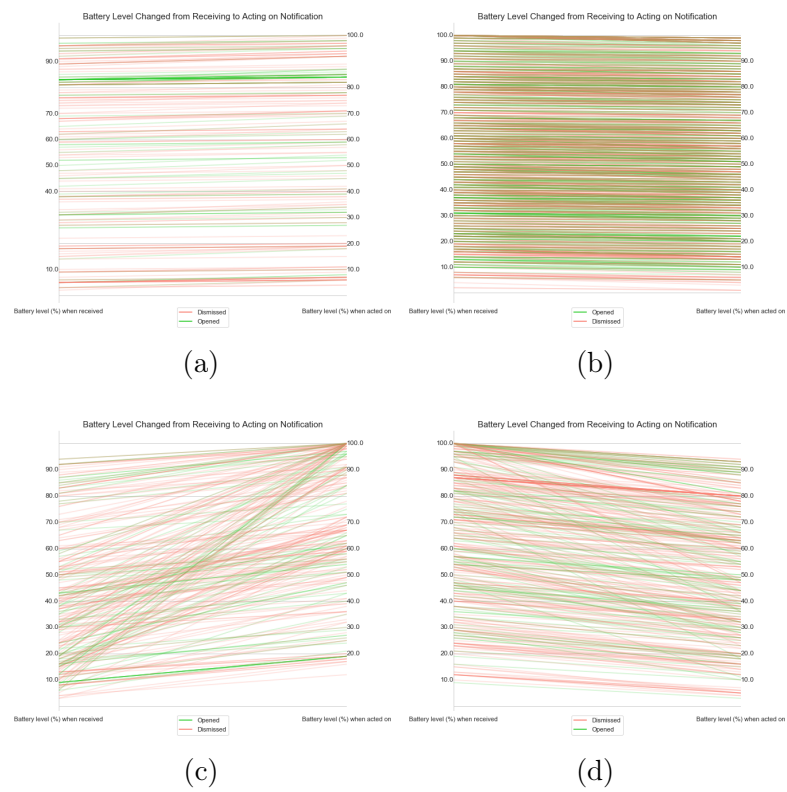


Fig. 3.14 Changes in battery level between moments of notification delivery and action

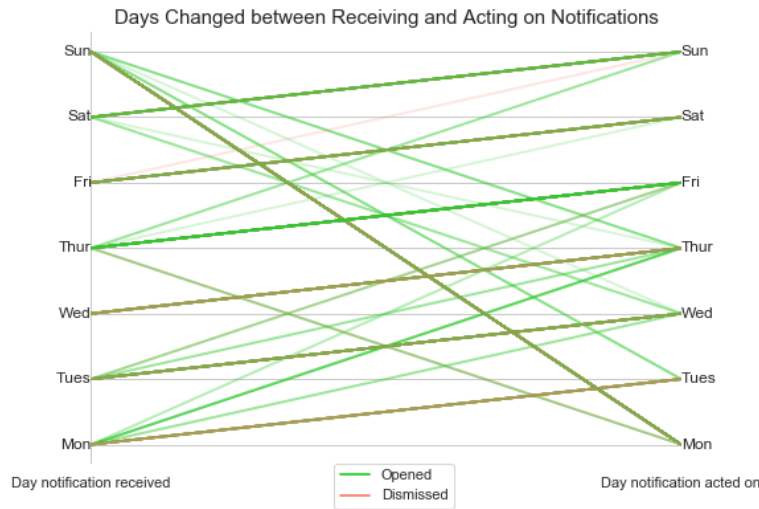


Fig. 3.15 Changes in day between moments of notification delivery and action

large portion of notifications were opened when devices were charging (increasing) at  $\approx 80\%$ . Figures 3.14c and 3.14d illustrate larger changes of battery level ( $>5\%$ ) which occurred between notification delivery and action. Notifications which arrived at lower battery levels and were acted on after the device was fully charged can easily be deciphered as the edges increase to 100%, although no significant action stands out. When decreasing, battery levels moving from  $\approx 90\%$  to  $\approx 80\%$  tended towards dismissal. Figure 3.14d also demonstrates that few notifications actions coincide with very large decreases in battery level.

### Day of Week

There was no change of day between delivery and action in 13,551 notifications and the CTR was 51%. A change was detected in the remaining 382 notifications, of which, 59.2% were opened. The most frequent direction of change was a notification delivered during Sunday and acted on during Monday. The slight increase in the open rate could suggest notifications left unattended during leisurely Sunday routines, irrelevant to weekend activities, but nonetheless important. The build up of notifications would then be attended to on Monday mornings once a new week began. Figure 3.15 illustrates changes of day between notification delivery and action. The strongest edges suggest the majority of notifications are left unattended for a max of one day change, however some notifications were not acted upon for longer periods, such as Sunday to Thursday. Notifications delivered during Thursday and acted on during Friday are shown to be the most favourable change with Monday to Tuesday been the least favourable, perhaps due to the approaching weekend.



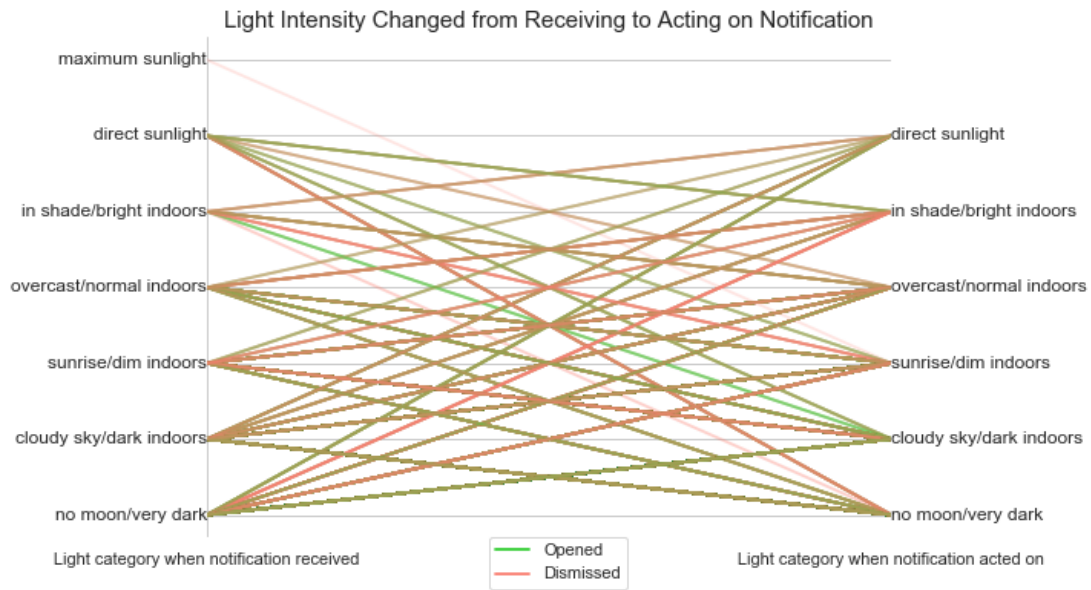


Fig. 3.16 Changes in ambient light category between moments of notification delivery and action

### Light Intensity

No change in ambient light intensity level was detected in 1,103 notifications, of which 30% were opened. In contrast, 3,745 notifications were delivered and acted on in differing ambient light intensity environments, 41% of which were opened. Figure 3.16 illustrates the changes which occurred in ambient light between delivery and action (discrete categories used for ease of understanding). The majority direction of changes were decreasing. The strong green edge between bright indoors and dark indoors indicates a large portion of notifications were opened when this change of ambient light was detected. Similarly, notifications delivered in very dark environments were mostly dismissed when subsequently acted on in bright indoor environments. Apps could utilise the detected change of ambient light to further personalise the delivery of notifications such that participant preferences are met.

#### 3.2.5.6 Key Findings

- 72.4% of notifications with a significant associated contact were opened compared to 47.2% without.
- Notifications with a significant contact associated were seen and responded to quicker, on average, than those without a significant contact associated

- 69.3% of notifications pushed from a significant app were opened compared with 31.3% pushed from insignificant apps
- 76.2% of notifications with a significant contact were found to be pushed from a significant app
- Although seen and response times were shorter for notifications pushed from significant apps, participants took 50 seconds longer, on average, to decide how to act on a notification sent from a significant app
- Participants opened most notifications in cafes, lodgings and bars and dismissed most notifications in universities, parks and street addresses
- The CTR of notifications was lowest in environments with noise levels similar to quiet rooms and washing machines, while environments with noise levels similar to normal talking and city traffic had higher CTRs
- Environments with dark indoor or dim indoor ambient light levels had higher CTRs than normal indoor, very dark and bright indoor levels
- Chrome, Messenger and WhatsApp apps updated notifications more frequently than any other app
- Of the notifications in which a change in context was detected between delivery and action time points, 44.5% were opened compared with 68.7% opened when no change in context was detected

### 3.2.6 Smartphone Engagement Etiquette

In this section we investigate participant perceptions toward good notification engagement behaviour. We also analyse participant levels of hope at moments of smartphone interaction and study the relationship between hope, context and engagement. The contributions of this section are summarised in Section 3.2.6.4.

#### 3.2.6.1 Perceived Behaviour

In section 3.2.5.4, the relationship between participant behaviour in moments of acting on a notification and the number of times a notification was updated (re-alerting the participant to its presence) was discussed. The behaviour feature collected was an indicator of how well a participant thought they acted in a

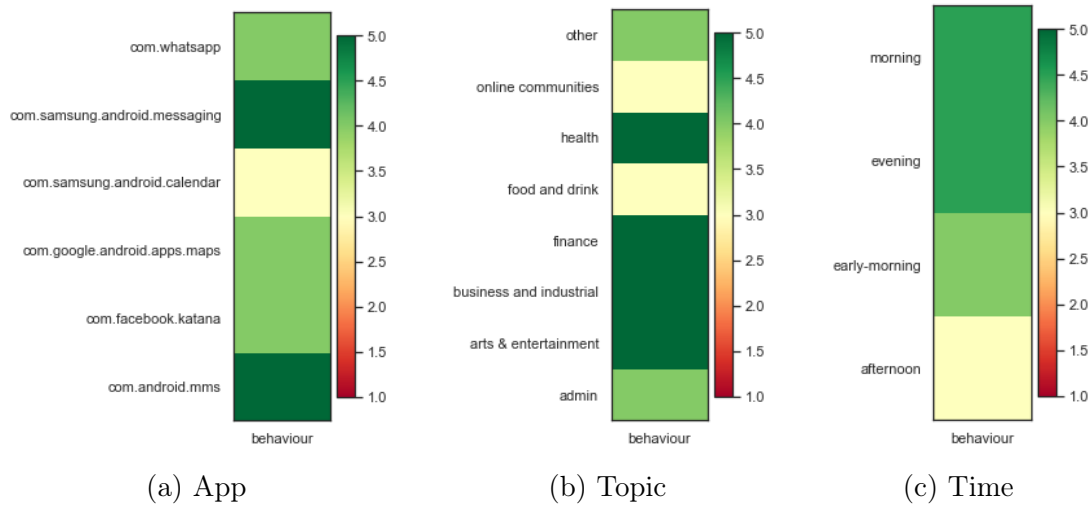


Fig. 3.17 Scores of participant behaviour when acting on notifications of differing types (1:VeryPoor; 2:Poor; 3:Neutral; 4:Good; 5:VeryGood)

moment of opening or dismissing a notification. The purpose was to ascertain what behaviours toward different notification types and in different contexts, were deemed good or poor behaviour from the perspective of participants. Behaviour was measured using a 5-point Likert scale ranging from 1, very poor behaviour, to 5, very good behaviour. Unfortunately, it was noted that the number of responses collected via the ESM questionnaire on perceived behaviour was low (21 responses). Therefore, this section will only briefly discuss the responses received with respect to notification engagements during the study, with the aim of informing future work in this area.

Of the 21 responses recorded, 20 were associated with a notification being opened and 1 with a notification being dismissed. The maximum behaviour reported, which was also found to be the most frequent (11), was 5 (very good) and the minimum reported was 3 (neutral). Figures 3.17a, 3.17b and 3.17c illustrate average behaviour responses for notifications delivered with differing apps, topics and times of day. The *calendar* app received the lowest average behaviour score and participants identified their actions toward messaging apps to be strongest. Actions toward notifications with *Food & Drink* and *online communities* topics were perceived as neither good or bad behaviour and topics such as *health* or *finance* scored strongest. Behaviour in the *morning*s and *evening* were highest on average over *afternoon*.

Behaviour with respect to a selection of contexts in which notifications were received are visualised in Figures 3.18a, 3.18b and 3.18c. Notifications with an associated *machine* sender (i.e. notification which was not sent directly by a

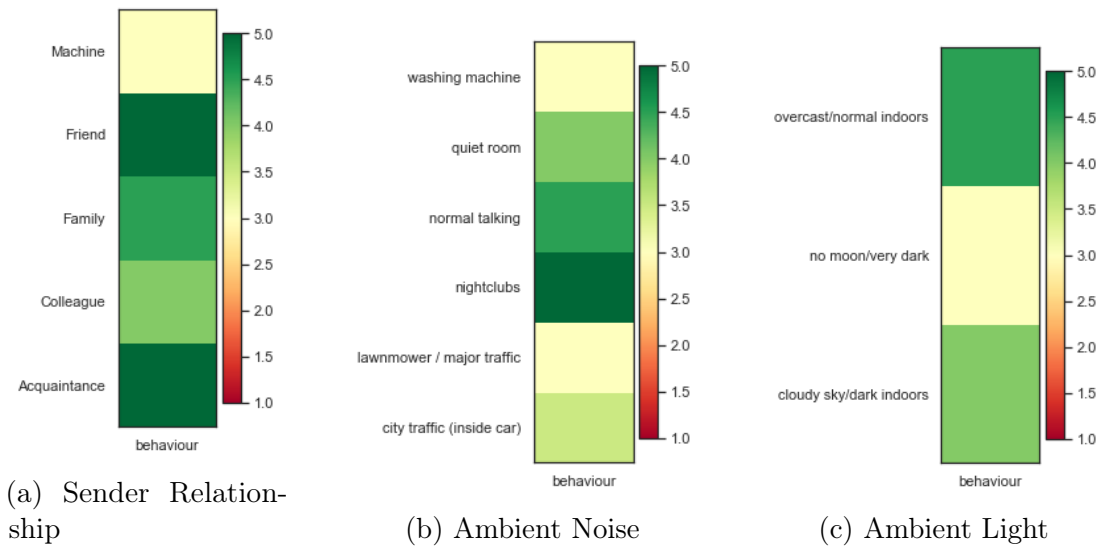


Fig. 3.18 Scores of participant behaviour when acting on notifications in differing contexts (1:VeryPoor; 2:Poor; 3:Neutral; 4:Good; 5:VeryGood)

human) had the lowest behaviour score of *neutral*, while notifications associated with a *friend* or *acquaintance* sender were acted on with best behaviour. Average behaviour in environments with ambient noise levels similar to nightclubs was perceived better compared to environments with noise levels similar to washing machine or lawnmowers. Finally, notifications delivered in environments with very dark ambient light levels were associated with poor notification actions compared with normal indoor lighting which had notification engagements associated with better perceived behaviours.

### 3.2.6.2 Smartphone Use in Company

In the previous section, we discussed participant perceptions of behaviour at moments of push-notification engagement in order to identify contextual differences between positive and negative behaviours with regard push-notifications. The WeAreUs app also spawned questionnaires in moments participants unlocked their device and included questions regarding the company present in their immediate environment as well as recording levels of necessity for using their device in their given context. The goal of collecting this information was to identify trends in etiquette of smartphone use. Five options for company were presented to participants: strangers, friends, family, colleagues and alone. The WeAreUs app collected 122 responses indicating the type of company present at moments of unlocking a device. *Alone* was the most frequent response (48) followed by family

(47), colleagues (19), friends (6) and strangers (2). Necessity for using a device in the given context was measured using a 5-point Likert scale ranging from 1, very unnecessary, to 5, very necessary. The WeAreUs app recorded 120 responses for necessity. *Necessary* was the most frequent response (35) followed by *very unnecessary* (30), *very necessary* (21), *neutral* (18) and *unnecessary* (16).

A Kruskal-Wallis H test was conducted to determine if there were differences in necessity levels between different types of company present when devices were unlocked. Necessity levels were statistically significantly different between the different categories of company present,  $\chi^2(3) = 3.326, p = .017$ . Figure 3.19b illustrates the average levels recorded for necessity of device use when in different company. Participants deemed their smartphone use, when in the company of *friends*, as least necessary. Smartphone use in the presence of colleagues was also less than neutrally necessary. Interestingly, smartphone use in the company of family was, on average, slightly more necessary than neutral. Figure 3.19a illustrates the average necessity of smartphone use at different times of the day. *Night* use was the only time period which had an average necessity level less than neutral, perhaps indicating that participants were aware that smartphone use at night could be an unhealthy habit. Figure 3.19c demonstrates the different places in which necessity of smartphone use was recorded. *Stadium*, *night club* and *cafe* place types had the minimum average necessity scores, perhaps indicating that smartphone use in these social settings was unnecessary.

The average necessity was also analysed over other aggregated notification engagement statistics. A Spearman's rank-order correlation found the relationship between average necessity levels and average ambient noise levels at moments of notification delivery, to have a, statistically significant, strong negative correlation,  $r_s = -.812, p = .05$ . This illustrates that participants of the study who, on average, used their device only when necessary also tended to receive notifications in environments with lower levels of ambient noise. Conversely, participants of the study who, on average, used their device unnecessarily, also tended to receive notifications in environments with higher levels of ambient noise. A statistically significant strong positive correlation,  $r_s = .899, p = .015$ , was also identified between average necessity level and average notification response time, such that participants who reported higher average necessity scores for smartphone use also had higher notification response times. This could indicate behaviour of intermittent smartphone use resulting in times to notice, and subsequently respond to notifications going up.

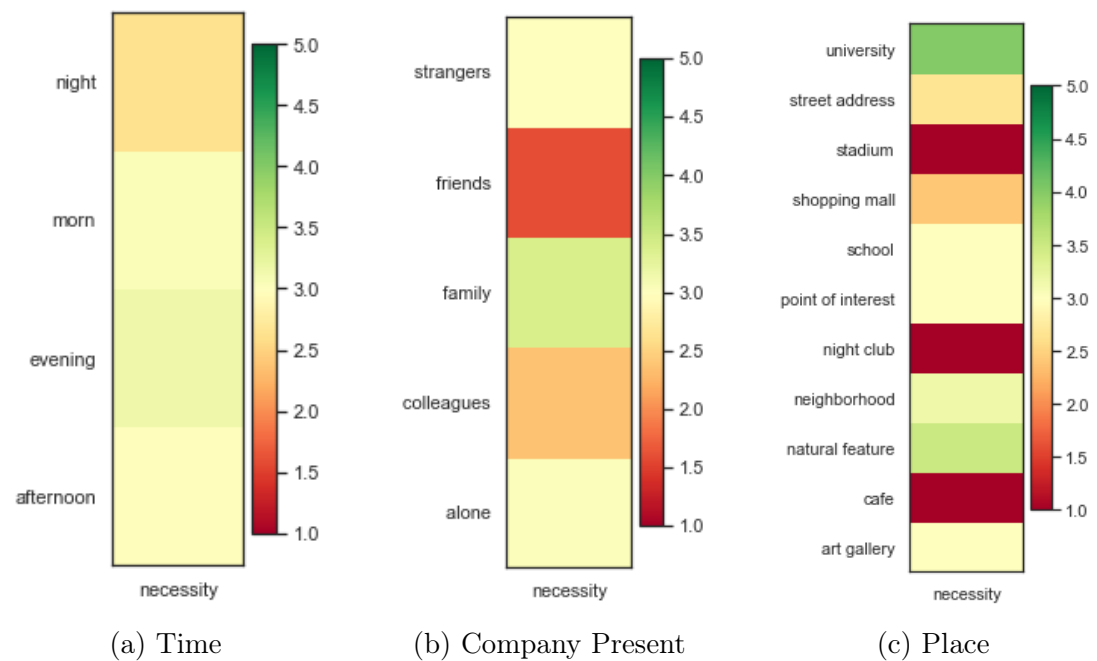


Fig. 3.19 Scores for necessity of smartphone use in different contexts (1:VeryUn-necessary; 2:Unnecessary; 3:Neutral; 4:Necessary; 5:VeryNecessary)

### 3.2.6.3 Hope, Smartphone Usage & Notifications

Snyder et al. developed the State Hope Scale as a "brief, internally consistent, and valid self-report measure of ongoing goal-directed thinking" [Sny+96]. In their work, hope is defined "as a cognitive set comprising agency (belief in one's capacity to initiate and sustain actions) and pathways (belief in one's capacity to generate routes) to reach goals". The WeAreUs app presented the 6-item State Hope Scale to participants at moments in which they unlocked their device in order to identify their levels of goal-directed thinking in moments of direct interaction with their smartphone. An 8-point Likert Scale was used for each item ranging from 1, *Definitely False*, to 8, *Definitely True*. Table 3.10 records the instructions and questions proffered to participants at these moments. The score for Pathways could range from 3 to 24 and was calculated by summing items 1, 3 and 5. The score for Agency could also range from 3 to 24 and was calculated by summing items 2, 4 and 6. Total Hope was the summation of Pathways and Agency scores, ranging from 6 to 48. Higher scores represented higher levels of Pathways, Agency and goal-directed thinking, respectively. This section analyses the responses of the State Hope Scale collected throughout the study with respect to smartphone usage and notification behaviour metrics.

Table 3.10 The State Hope Scale presented to participants at moments they unlocked their smartphone

---

|                                                                                                                                                                            |                 |                   |                   |                  |                  |                |                  |
|----------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-----------------|-------------------|-------------------|------------------|------------------|----------------|------------------|
| Using the scale provided, please select the number which best describes<br>how you think about yourself right now<br>(in the moment of unlocking your smartphone for use). |                 |                   |                   |                  |                  |                |                  |
| Please take a few moments to focus on yourself and what is going on in<br>your life at this moment.                                                                        |                 |                   |                   |                  |                  |                |                  |
| Once you have this "here and now" set, go ahead and answer the 6 questions                                                                                                 |                 |                   |                   |                  |                  |                |                  |
| 1                                                                                                                                                                          | 2               | 3                 | 4                 | 5                | 6                | 7              | 8                |
| Definite<br>False                                                                                                                                                          | Mostly<br>False | Somewhat<br>False | Slightly<br>False | Slightly<br>True | Somewhat<br>True | Mostly<br>True | Definite<br>True |
| <hr/>                                                                                                                                                                      |                 |                   |                   |                  |                  |                |                  |
| 1 If I should find myself in a jam, I could think of many ways to get out of it                                                                                            |                 |                   |                   |                  |                  |                |                  |
| 2 At the present time, I am energetically pursuing my goals                                                                                                                |                 |                   |                   |                  |                  |                |                  |
| 3 There are lots of ways around any problem that I am facing now                                                                                                           |                 |                   |                   |                  |                  |                |                  |
| 4 Right now, I see myself as being pretty successful                                                                                                                       |                 |                   |                   |                  |                  |                |                  |
| 5 I can think of many ways to reach my current goals                                                                                                                       |                 |                   |                   |                  |                  |                |                  |
| 6 At this time, I am meeting the goals that I have set for myself                                                                                                          |                 |                   |                   |                  |                  |                |                  |

---

## Findings

Of the 122 ESM questionnaires responses collected at moments participants unlocked their device, the State Hope Scale was sufficiently completed in 110 instances (4 responses were incomplete and dropped from the analysis) by 7 different participants. Figure 3.20 illustrates the average Pathways, Agency and Hope scores collected over a number of days of the study. The majority of scores are high, although a number of days in which participants reported less goal-directed thinking at moments of device interaction could be easily identified. Agency, "which taps the individual's perceived capacity for initiating and maintaining the actions necessary to teach a goal" [Sny+96], was found to score, on average, 21 (min=3, max=24, Mdn=21) throughout the study indicating that participants were mostly confident their actions were goal-oriented during moments of smartphone use. Pathways, "which taps the perceived ability to generate routes to one's goals", was also found to be 21 on average (min=3, max=24, Mdn=21) indicating that participants were confident of navigating toward their goals at moments of smartphone use. Responses spanned the full range possible for both subscales, with some participants not confident their actions or plans contributed toward their goals during some moments of unlocking their device, indicated by the lowest

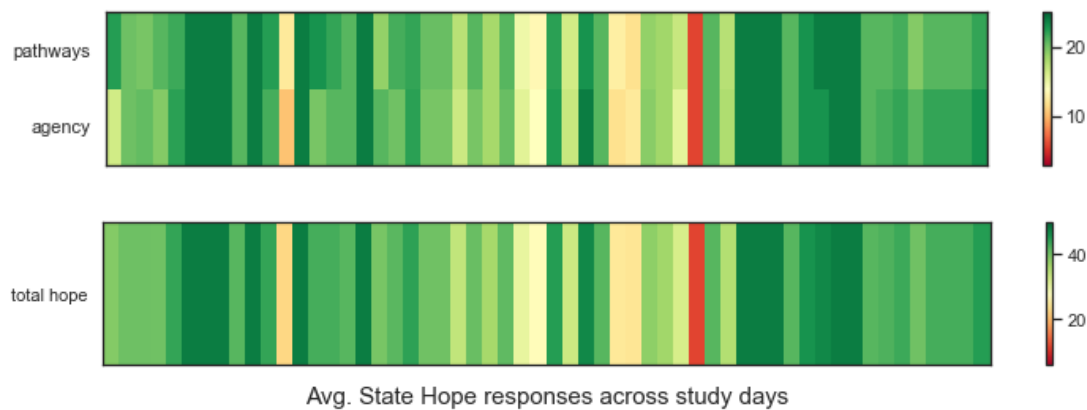


Fig. 3.20 Average State Hope responses over days of the WeAreUs Study

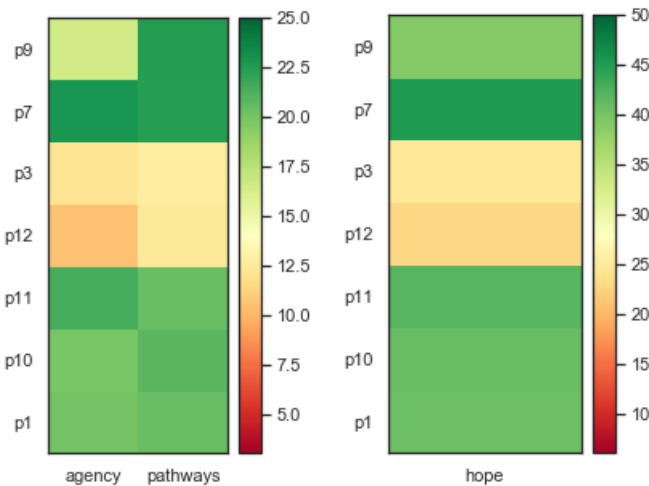


Fig. 3.21 Average State Hope scores per participant during the WeAreUs Study

possible Agency and Pathways scores collected. Figure 3.21 illustrates the average state Agency, Pathways and Hope scores for each participant. Two of the seven participants responding to the State Hope Scale questionnaires had relatively low scores compared to other participants, ranging between 20 and 25 out of a possible score of 50 on the Hope scale. Agency and Pathways scores were mostly similar in all participants but for p9, in which their Agency scored much lower than their Pathways, perhaps indicating that their actions did not quite reflect their goal-directed plans.

Using aggregated statistics of participant responses over the course of the study, relationships could be analysed between average scores for Hope and other notification engagement and smartphone usage metrics. Table 3.11 illustrates the significant relationships between state Agency, Pathways, Hope and other aggregated metrics. There was a statistically significant, strong negative associa-



tion between average battery level, sampled at moments of notification delivery, and Agency. Therefore participants who had lower levels of battery also scored themselves higher in Agency thinking, or "goal directed determination" [Sny+96]. One contributing factor could be attributed to participants frequently using their smartphone for executing actions toward goals, thus their smartphone battery remains consistently drained. However, alternative explanations could include participants infrequently charging their device due to initiating and maintaining goal-directed actions without using their smartphone. Conversely, participants with higher levels of battery on average, had lower reported levels of goal directed determination, which, depending on the perspective, could indicate infrequent smartphone use and therefore less actions taken toward goals. Alternatively, frequent charging due to excessive smartphone use could have distracted participants from taking actions toward their goals. In order to accurately identify the true meaning, future work would need to explore the causation for the significant associations identified. Average battery levels sampled at moments of notification action also had a statistically significant strong negative association with Agency. In addition, a statistically significant strong negative association was also identified with state Hope indicating that participants with lower average battery levels, at moments of notification action, also had higher levels of goal-directed thinking. Or alternatively, participants with higher average battery levels, at moments of notification action, also had lower reported levels of goal-directed thinking.

There was a statistically significant, strong positive association between the number of apps installed on participant devices and average state Pathways reported throughout the study. Participants with a higher number of apps installed also reported higher scores for Pathways thinking (planning to meet goals). The presence of more apps would give a smartphone user more functionality and, potentially, more opportunity to complete goal-directed tasks. The action of having installed many apps would therefore indicate a plan toward achieving goals using the apps installed and may therefore be a factor contributing to this association.

There were statistically significant, strong positive associations found between all state Hope measures reported by participants and the average elapsed time between an app being used and it sending a notification. Participants interrupted more often by a notification from an app shortly after using it reported lower average scores for Agency, Pathways and Hope. Conversely, participants with longer average periods of elapsed time between use and a notification had higher reported scores. The notification interruptions so soon after spending time within

Table 3.11 Results of Spearman’s rank-order correlation for identifying relationships between average State Hope and notification engagement behaviours. ( del. = moment of delivery; act. = moment of action)

| Feature            | Agency |        | Pathways |        | Total Hope |        |
|--------------------|--------|--------|----------|--------|------------|--------|
|                    | $r_s$  | $p$    | $r_s$    | $p$    | $r_s$      | $p$    |
| avg. battery del.  | -0.757 | 0.049* | 0.072    | 0.878  | -0.571     | 0.180  |
| avg. battery act.  | -0.847 | 0.016* | -0.018   | 0.969  | -0.786     | 0.036* |
| # apps installed   | 0.518  | 0.233  | 0.827    | 0.022* | 0.450      | 0.310  |
| avg. app last used | 0.775  | 0.041* | 0.775    | 0.041* | 0.821      | 0.023* |
| avg. seen time     | 0.937  | 0.002* | 0.577    | 0.175  | 1          | <.01   |
| avg. notif. volume | 0.844  | 0.017* | 0.302    | 0.511  | 0.657      | 0.109  |

an app may have negatively impacted a participants ability to reflect, act and plan toward their goals and could be a contributing factor. Alternatively, the notifications may have prompted participants to return to an app and re-motivate and remind them of their goals after a prolonged absence or period of procrastination. Once again, the causation should be explored in future work to identify the true factors contributing to this correlation.

A statistically significant, strong positive correlation was found between the seen time (time elapsed between the moment of notification delivery and the moment a participant unlocked their device) and average state Agency, indicating that participants in the study with higher reported average levels of goal-directed determination also took less time to unlock their device, on average, once a notification was delivered. This could suggest participants who were more active in initiating and maintaining actions toward their goals were also enabled by incoming notifications via their smartphone, although a number of confounding factors could also exist. Finally, a statistically significant, strong positive association was also identified between the average notification volume of participant devices and average state Agency. Participants who had a higher average notification volume set also reported higher average levels of Agency, perhaps also indicative of the previous suggestion that notifications were enabling goal-directed determination as participants wished to be sure of being alerted to incoming notifications. The causation study involving these features is left as future work.

Following the previous section in which smartphone necessity was analysed with different present company, Figure 3.22 illustrates how the State Hope Scale features varied when company presence differed. A Kruskal-Wallis H test was conducted to determine if there were significant differences in state Agency, Pathways and Hope scores between differing types of company. All scores were statistically

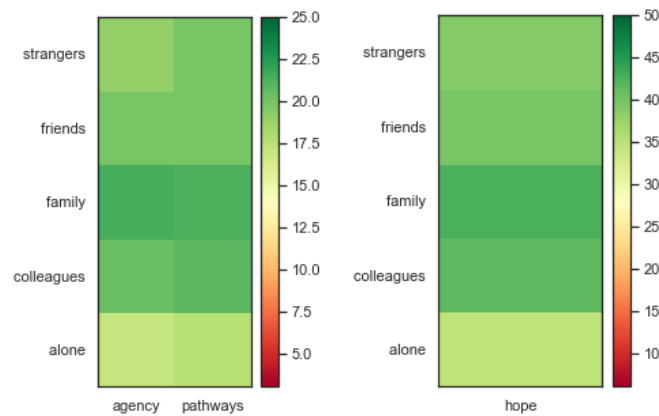


Fig. 3.22 Average State Hope scores recorded with different company present

significantly different between the different company categories,  $p = 0.019$ ,  $0.011$  and  $0.022$  respectively. As illustrated, levels of reported state Hope were lowest when alone and highest when in the company of family. Agency thinking was slightly stronger than Pathways thinking when with family members and slightly lower than Pathways thinking when in the company of strangers.

#### 3.2.6.4 Key Findings

- On average, participants perceived smartphone use while in the company of family as necessary and in the company of friends as unnecessary
- Smartphone use at night was, on average, least necessary compared to other times of day
- Smartphone use in cafes, night clubs and stadiums was perceived as least necessary
- On average, at moments of unlocking their device, participants scored highest for ongoing goal-directed thinking while in the company of family and lowest when alone
- A strong positive correlation was found between the number of apps installed and Pathways thinking (planning to meet goals)
- A strong positive correlation was found between seen time and Agency thinking (goal-directed determination)

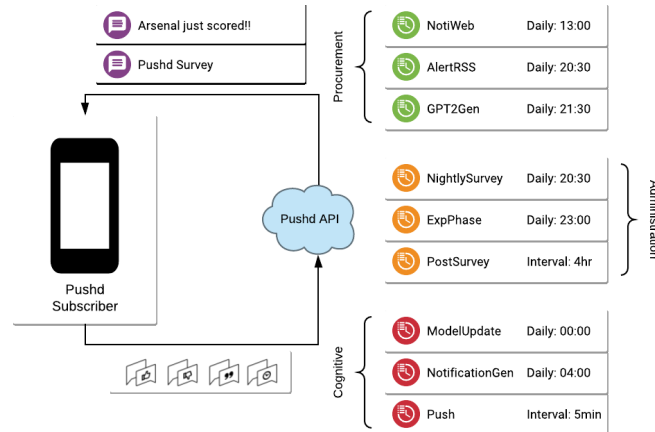


Fig. 3.23 Pushd study design

### 3.3 Web Push

Web push-notifications are an example of the expansion and ubiquity of *push technology*. These are push-notifications subscribed and delivered through internet browsers. They are increasingly device agnostic compared to their native mobile relatives as browsers are found in many other technologies in addition to the mobile device, such as desktop computers, televisions, game consoles, smartwatches and many others. In some ways, this creates a much more direct route to the subscriber, as web push notifications do not require a specific app, just simply a browser, for push capability to be possible - although there are some caveats (such as iOS incompatibility). The following sections introduce a study of web push-notifications in-the-wild which was conducted to evaluate individual components of the ETHOS-Push framework, but also provides insight on sentiment toward web push, its effectiveness and the behaviours which it instills in its subscribers.

#### 3.3.1 Data Collection

The *Pushd* study was designed to study web push-notification engagement in-the-wild, observe and assess the shortcomings of push technology in-the-wild and inform the design of components which satisfy requirements of the ETHOS-Push approach outlined in Chapter 1. A high-level diagram of the push-notification service created for this study is illustrated in Figure 3.23.

Participants were asked to subscribe to notifications on the Pushd website. The website acted as an interface between the participants and the the Pushd API which controlled a number of tasks:

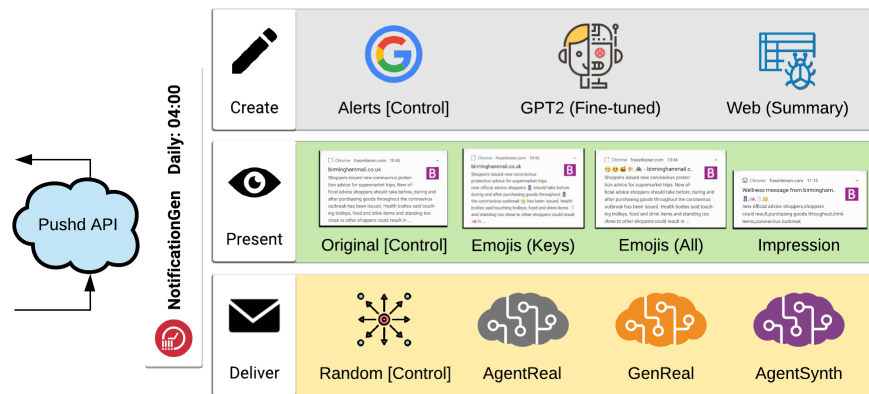


Fig. 3.24 The 3 distinct phases of the Pushd study evaluating different components of the ETHOS-Push framework

1. Procurement - this component scheduled tasks to curate a number of push-notification data sets on a daily basis. Each type of notification data set was procured via a different method, which will be discussed in the following sections.
2. Administration - this component managed the distribution of questionnaires to participants at particular moments of the study, tracking their progress through different phases of the study and ensuring participants were subscribed / unsubscribed at their discretion.
3. Cognition - this component managed the continuous update of user models, scheduled tasks for push-notification inference and distributed the study's push-notifications in real-time.

The Pushd web service worker, running in the background of a participant's device and managed by their browser (e.g. Google Chrome in the case of this study), enabled push-notifications to be presented to participants once delivered to their device from the Pushd API. The service worker also identified engagement actions such as open (explicit click on the notification) or dismiss (an explicit dismiss or swipe away of the notification).

The Pushd study was divided into 3 phases in order to evaluate individual components of the ETHOS-Push framework: creation (curation), presentation and delivery. Each participant was randomly assigned to a distinct group for each phase of the study and would receive push-notifications relevant to their grouping. Figure 3.24 illustrates the available groups in each phase. The first phase of the study was evaluating push-notification creation or source. During this phase, participants would receive push-notifications generated from three distinct

source data sets. The first source, the control, was the Alerts data set. This was comprised of push-notifications from Google Alerts, a content change detection and notification service. The second source data set was created by a generative model, GPT-2, which was fine-tuned on the Google Alert data set and conditioned to generate enticing or unenticing notifications relevant to a given topic. The third source was a custom web crawler which identified text on a predefined set of websites and adapted it for consumption through push-notification. The methodology and implementation of curation for each group will be explored in greater depth in Chapter 6. Figure 3.25a illustrates the groups of phase 1. Note that the presentation and delivery variables remain fixed throughout this phase of the study.

The second phase of the study was designed to evaluate push-notification presentation. During this phase of the study, participants would receive push-notifications in 1 of 4 distinct formats. The source of the push-notifications for this phase was the control of phase 1, the Alerts data set. The first format group, the control, had no modifications to the push-notification presentation. The second group, emojis-key, had keywords replaced with contextually relevant emoji's. The third group, emojis-all, used a line of emojis representing the entire push-notification message as the first line of content. The final group, impression, used a combination of keywords, emojis, topics and icons to create an impression of the push-notification content such that it could be quickly consumed by participants. The methodology and implementation of inference for each presentation format will be discussed in detail in Chapter 5. Figure 3.25b illustrates the groups of phase 2. Note that creation and delivery variables remain fixed throughout this phase of the study.

The third phase of the study was designed to evaluate push-notification delivery. During this phase of the study, participants would receive push-notifications at different times and in different presentation formats depending on inferences made by 1 of 4 distinct agents. The notifications were once again drawn from the same source, the Alerts data set, for all groups. The first agent, the control, implemented a random delivery method. This was consistent with phase 1 and 2, where push-notifications were also delivered at random intervals. The second agent, AgentReal, implemented a Decision Tree algorithm and was trained using real notification engagements of each participant. The third agent, GenReal, implemented a generative model, was trained using real notification engagements of participants and was conditioned to generate delivery times and presentation formats of notifications thought to be positively received for a participant on a

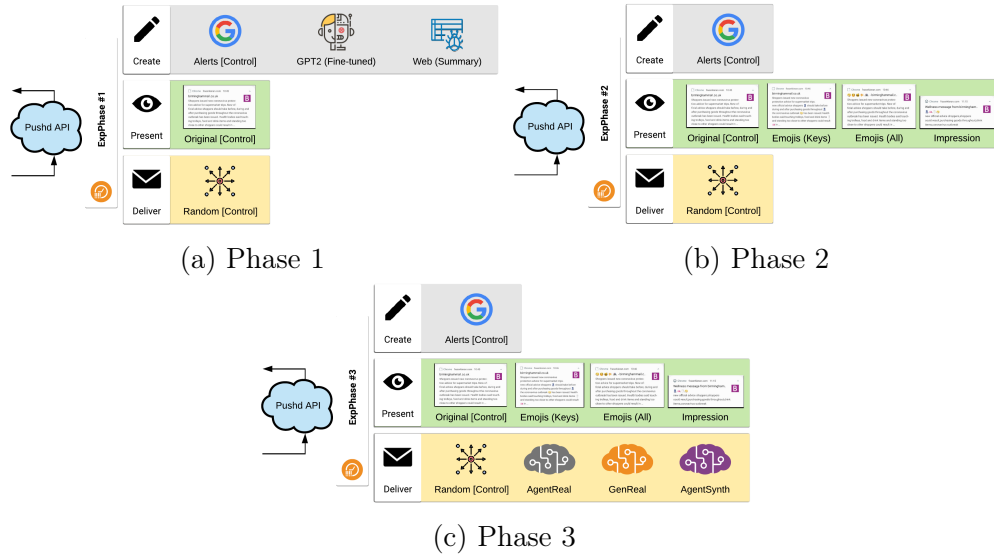


Fig. 3.25 The creation, presentation and delivery combinations for each phase of the Pushd study.

given day. The fourth and final agent, AgentSynth, also implemented a Decision Tree algorithm but, in contrast to AgentReal, it was trained using synthetic notification engagements created by a generative model. The methodology and implementation of agents for each group will be discussed at greater depth in Chapter 4. Figure 3.25c illustrates the groups of phase 3. Note that the creation variable remains fixed throughout this phase of the study.

Once allocated to their groups, participants received push-notifications through their *Google Chrome* web browser on their mobile device as they went about their normal activities and their engagement actions were recorded. In addition to the engagement actions, 3 questionnaires were also presented to participants at significant stages of the study, as managed by the *Administration* component of the Pushd API scheduler:

1. Pre-study questionnaire
2. Mid-study questionnaire
3. Post-study questionnaire

### 3.3.1.1 Pre-study questionnaire

This questionnaire was presented to participants prior to receiving any web push-notifications. The purpose of this questionnaire was to record participant demographic data and understand their current preferences toward push-notifications,

such as their preferred topics or sources for information. It also contained competency questions regarding identification of emoji’s and enticing *clickbait-like* text. Two additional questionnaire items were also included:

1. The Mindful Attention Awareness Scale (MAAS) - this questionnaire was used to measure the dispositional mindfulness and receptive awareness of participants and their ability to be attentive to the present moment. This measure was added to assess whether dispositional mindfulness could be an influencing factor for receptivity to push-notifications and subsequent engagement behaviours.
2. The Curiosity and Exploration Inventory (CEI-II) - this questionnaire was used to measure trait curiosity, or, *the motivation to seek out knowledge and new experiences and willingness to embrace the novel, uncertain and unpredictable nature of everyday life*. This measure was added to assess the relationship between curiosity and receptivity to notifications, and in particular, those notifications enticing engagement through exploitation of the curiosity gap (e.g. use of clickbait-like notification titles).

The full list of questions included in the pre-study, MAAS and CEI-II questionnaires are found in Tables B.15, B.16 and B.17 respectively.

### 3.3.1.2 Mid-study Questionnaire

The mid-study questionnaire was presented to participants during all 3 phases of the study. This questionnaire was used to assess what influence various push-notification features had on participants and record their perceptions toward certain aspects of the notification content, presentation and delivery methods. In particular, the *credibility* feature was used to assess the ability of a generative model to create realistic push-notification text content. The *credibility* measure included was first used by Solaiman et al. during their investigation of the social impact of large language models [Sol+19]. Also included were questions regarding participant engagement which could be used to assess whether participant’s prior convictions toward web push-notification would hold true and whether their perceptions of engagement differed greatly from reality. The full list of questions included in the mid-study questionnaire and their intended purpose are listed in Table B.19.



### 3.3.1.3 Post-study Questionnaire

The post-study questionnaire was presented to participants on completion of phase 3 of the Pushd study. This questionnaire was used to extract concluding impressions of the web push-notifications received throughout the study. The ability of the generative models to create synthetic notification engagement data was also qualitatively assessed in this questionnaire by asking participants to identify synthetic engagements amidst real engagements which occurred during the study. In addition, further qualitative perceptions of the notification presentation formats was also requested as was participants own sense of time spent interacting with the notifications delivered. The full list of questions included in the post-study questionnaire and their intended purpose are listed in Table B.20.

### 3.3.2 Participant Recruitment

Participants for the Pushd study were crowdsourced through the Prolific Academic platform and received £3 in remuneration for their efforts. The sample from which participants were drawn was restricted based on the following requirements for the study:

- English speakers - this was a requirement as the push-notification content curated was in English only. Future work could explore translation and support for multiple languages.
- Android phone operating system - this was a requirement as the push-notification presentation formats were likely to differ based on the operating system running the browser. Thus, for consistency of assessing the effects of notification format, a single operating system was chosen to ensure each participant of a group would see the same presentation format.
- Mobile device - this was a requirement as, while web push-notifications are ubiquitous across mobile and desktop, this study wished to address mobile push-notification engagement alone.
- Over 18 years of age

In addition to these limitations placed on the participant sample, a pre-screen test was also created to ensure participants could receive push-notifications after subscribing to the service. The Web Push APIs are relatively new and issues

with delivery are common, even among industry leaders in the space. A host of reasons can prevent a push-notification from being successfully delivered after initial subscription to the push-notification service, such as: too little memory left on the subscribed device; accidental deletion of the cached memory for the subscribed website; native device notification settings; or even past notification engagement behaviour. Google Chrome in particular has focused on removing malicious push-notifications by studying user engagement with the subscription interface and deciding whether a user is likely to respond positively to a request for subscription. Therefore, the pre-screen test asked participants to subscribe to the service and, once subscribed, a test push-notification was sent. Each participant was required to explicitly open the push-notification to signify they had received it and were able to engage with it. Only after successful completion of the pre-screen test, were participants able to proceed to the pre-study questionnaire.

### 3.3.3 Privacy & Ethics

Similar to the WeAreUs study, this study also underwent a rigorous analysis by the School's Research Ethics Committee prior to launch due to the sensitive nature of the demographic data requested of participants and observation of their push-notification engagement habits. The study was granted approval.

On sign-up to the study, participants were required to read Participant Information and Informed Consent documents which outlined the study in detail, including processes regarding collection, security and the intended life-span of data collected. Only those who granted consent were given the option to be included in the study.

Participants were given the option to opt-out of the study at any point by clicking a button on the Pushd website. On clicking the button the push-notification token, used to deliver push-notifications to the participants browser, would be destroyed. Participants were not asked for any identifier (e.g. email, username etc.) on initial sign-up to the study. Instead, a randomly generated key was used to identify and link a participant to their push-notification engagements. This random key was stored as a cookie in the participants web browser. On opting-out of the study, the key would be used to delete any and all previously collected data relating to the participant during the study, before being destroyed itself.

### 3.3.4 Dataset

The Pushd study was launched in November 2020 and had a duration of two weeks. 226 participants signed up to the pre-screen test through the Prolific Academic platform. Of the 226 participants, 166 successfully completed the task of engaging with a test push-notification signifying they could receive and action push-notifications of the study. These 166 participants were then invited to complete the pre-study questionnaire, of which 131 participants submitted successfully. Subsequently 105 of these participants proceeded to the engage with the study and 24,682 notifications were then sent over the following two weeks. 797 mid-study questionnaires were submitted during this period. On conclusion of the study, all participants received the post-study questionnaire, which all 105 participants proceeded to answer and submit.

Due to incomplete data associated with a small number of participants, the final data set was restricted to participants who had:

- answered a pre-study, post-study and at least 1 mid-study questionnaire
- engaged with at least 5 delivered push-notifications

Therefore, the final data set of the Pushd study was made up of 91 participants, 19,852 pushed notifications, 725 mid-study questionnaire responses and 91 questionnaires submitted pre & post-study.

The final allocation of participants to phase groups (illustrated in Figure 3.24) are stated in Table 3.12. There was a minimum participation target of 69 participants engaging with the study for the full period. This was calculated using Cochran's Formula [Coc07], assuming a sample size with 90% confidence level and 10% precision. Participant numbers between 50 and 100 are in line with similar studies in the field of push-notifications which have proven to be statistically effective [MHM16a; Meh+17a]. The final sample size of 91 participants has an associated confidence level of 95% and 10% Margin of Error.

### 3.3.5 Understanding Participant Characteristics & Attitudes toward Web Push

This section explores the pre-study questionnaire completed by participants by analysing their demographics, personal characteristics, push-notification preferences and competency at identifying certain features of push-notifications.

Table 3.12 Final participant group allocations during each phase of Pushd study.

| Phase | Group           | # Participants |
|-------|-----------------|----------------|
| 1     | Alerts (ctrl)   | 32             |
|       | GPT-2           | 32             |
|       | Web-Sum         | 27             |
| 2     | Original (ctrl) | 23             |
|       | EmojiKey        | 20             |
|       | EmojiAll        | 26             |
|       | Impression      | 22             |
| 3     | Random (ctrl)   | 26             |
|       | AgentReal       | 23             |
|       | GenReal         | 22             |
|       | AgentSynth      | 20             |

The opening two questions of the questionnaire recorded participant gender and age-group: 57% of respondents were Female and 43% were Male; 39% were aged between 18-29, 37% between 30-39, 17% between 40-49 and 7% between 50-59. Two additional questionnaires were included as part of the pre-study questionnaire. These were the Mindful and Attention Awareness Scale (MAAS) and the Curiosity & Exploration Inventory (CEI-II). The results obtained from both are illustrated in Figures 3.26a and 3.26b respectively. These two measures, mindfulness and curiosity, along with gender and age group, were used to provide further insight into the notification engagement preferences and behaviours of participants.

Of the 91 participants, when asked if they had altered their notification settings on their device, 56% had not, 10% did not know and 34% confessed they had. This was an important feature to record, as the resulting engagements may have been effected due to interventions by on-device notification settings (e.g. Android introduced *Android Adaptive Notifications* in their release of Android 10).

When asked to identify an emoji between two texts, one of which contained a smiley faced emoji, the other without any emoji, all respondents correctly identified the text containing an emoji. This competency was important to track as certain notification formats presented in phase 2 and questions asked in the post-study questionnaire, depended on the understanding of an emoji. This is further discussed in Chapter 5.

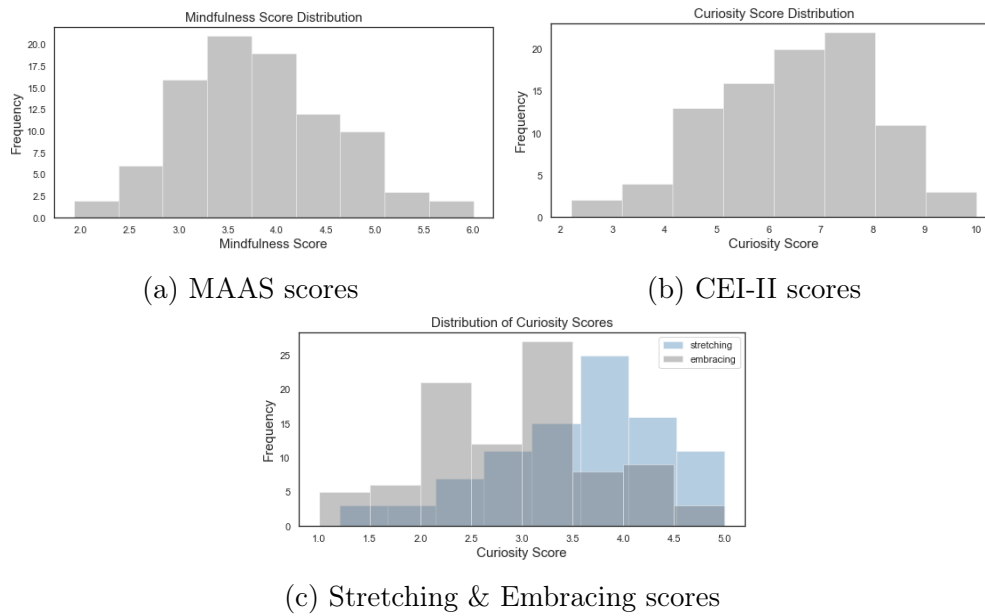


Fig. 3.26 Results from Mindful and Attention Awareness Scale (MAAS) and the Curiosity & Exploration Inventory (CEI-II) responses.

Figures 3.27a and 3.27b illustrate the 5-point likert scale responses to the questions *How likely are you to subscribe to push-notifications on the web?* and *To what degree are you comfortable with your data being used for automated decision making?* respectively. Participants were less likely to subscribe to web push-notifications and they were marginally less comfortable with their data being used for automated decision making.

Participants were shown two titles and asked to identify which one was deliberately enticing. The titles presented were extracted from a corpus of annotated clickbait and non-clickbait titles. The results, illustrated in Figure 3.27c, show that not all participants were able to identify clickbait-like enticing text when it was presented. This is by no means a comprehensive test of participant competency in identifying enticing text, but is broadly indicative. Further exploration with respect to the age of participants and their response showed that, of the participants that correctly identified enticing text, the majority were in the youngest age-bracket.

Participants were asked to select topics and websites from which they would and would not like to receive push-notifications. The average number of preferred topics chosen per participant was 3, with some participants subscribing to as many as 20 and as few as 2. 27 unique topics were selected in total. The most popular topics chosen were *comedy*, *entertainment* and *food & drink*. The

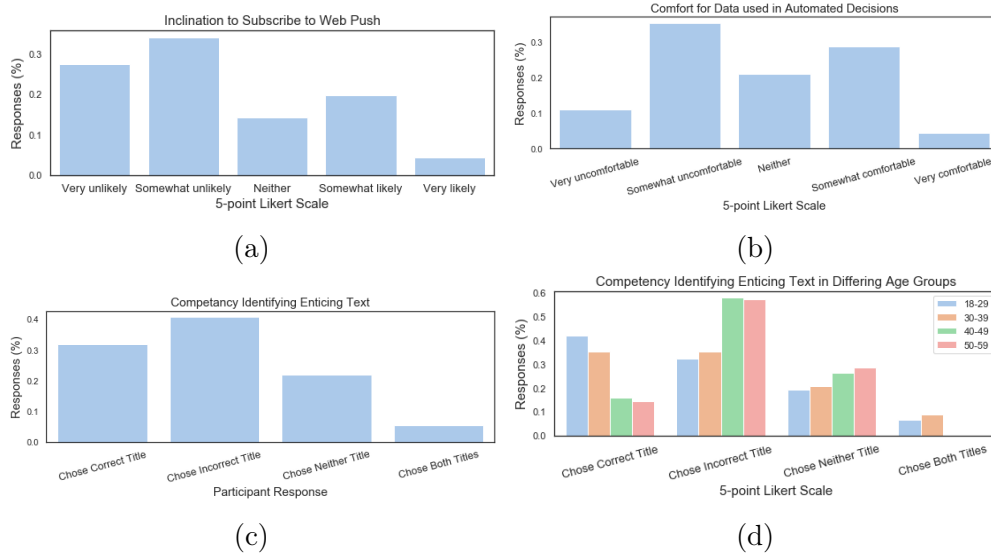


Fig. 3.27 Participant sentiment toward a) subscribing to web push-notifications b) sharing data for automated decision making; Participant identification of enticing text c) aggregated responses d) responses by participant age.

least popular were *impact*, *weddings* and *college*. Participants subscribed to 4 websites on average, with some subscribing to as many as 15 and as few as 2. The most popular websites chosen to receive push-notifications from were *bbc.co.uk*, *theguardian.com* and *amazon.com*. Some of the least popular websites were *weightwatchers.com*, *ticketmaster.com* and *runnersworld.com*. As well as subscribing to websites, participants were also given the option to select websites which they would not like to receive push-notifications from. Participants chose to "unsubscribe" from 12 websites on average, with some participants selecting as many as 15 and as few as 2. This, in contrast to the subscribed websites, shows that participants were much more concerned with ensuring they did not receive unwanted push-notifications. Figure 3.28 illustrates the most popular subscriptions.

Finally, the pre-study questionnaire also tasked participants with engaging positively (opening) or negatively (dismissing) with a small set (10) of web push-notifications to record an initial Click-through Rate (CTR) benchmark before the study began. The average CTR was 30%, with some participants opening as many as 70% and others indicating they would not have opened any.

During the two weeks of the study, 19,852 push-notifications were sent to participants. Not all notifications pushed were subsequently delivered. There are a myriad of reasons for this, including: device turned off at the time of the push; poor internet connectivity; device power saving settings; do-not-disturbed turned on;

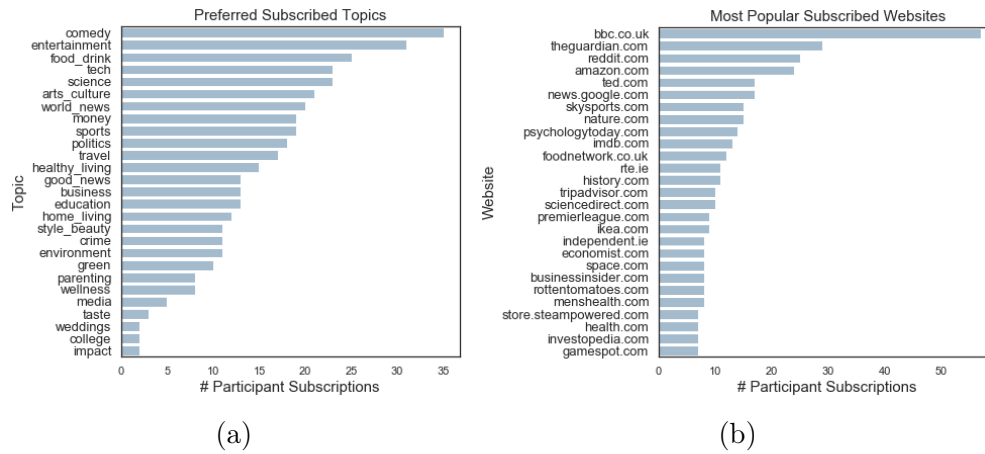


Fig. 3.28 Most popular a) topic and b) website subscriptions

and many others. 14,368 (72.4%) were successfully delivered to participant devices. In addition, not all push-notifications delivered were engaged with by participants. There are also a number of reasons for this, such as: notifications left unattended in the notification bar; device turned off before an engagement occurred; browser application terminated (process killed) before an engagement occurred; and many others. 12,082 (60.9%) were explicitly engaged with (opened or dismissed) by participants. Therefore, in the subsequent analyses in the following Chapters, where not otherwise stated, only those notifications which were explicitly acted on by the participants were considered.

## 3.4 Limitations

This chapter discussed an in-situ study of notification engagement behaviour through a mobile app designed to track notification and smartphone usage while participants went about their daily lives. The study also implemented an Experience Sampling Method to collect direct responses from participants. On reflection of this work, a number of limitations were identified and categorized as: technological; experimental design; or experimental execution.

### 3.4.1 Technological Limitations

Perhaps the most significant technological limitation of this WeAreUs study was that Android was the only platform supported by the WeAreUs app and the *Google*

*Play Store* was the sole distribution point. It was chosen not to create an identical iOS app as the push-notification functionality of the platform didn't support the same omniscient capabilities, required for holistic logging of notification and app behaviours, that Android did, which was pivotal to the study. Therefore, the profile of potential participants was first limited to those who owned and used an Android device. Another limitation, briefly mentioned previously in the chapter, were the protocols implemented in the *Oreo* version of the Android OS tasked with optimising battery usage by regulating app background services. This resulted in restricted functionality of the WeAreUs app's background tasks at key moments, such as notification delivery or the device being unlocked. Partial data was collected at these time points if the OS was currently restricting the app from querying sensors, such as GPS, or sending API requests, such as to Google's Places API. A limitation also arose due to the relative infancy of notification technology and the minimal functionality offered to developers for effective implementation. The Android SDK did not give developers a method of identifying the action taken on notifications delivered to devices, such as opened or dismissed. Therefore, a custom method was implemented (discussed in Section 3.2.1) which may have resulted in false positive or negative actions being recorded.

### 3.4.2 Experimental Design Limitations

The number of questionnaires delivered to participants on a single day was limited to 8, regardless of whether they were answered or not. This design choice was made to limit the number of interruptions and annoyances participants experienced a day due to the WeAreUs app, such that engagement over longer periods would be tolerated. However, the limited number of responses to questionnaires in the study could indicate that participants should have been queried more aggressively to ensure a greater amount of data was extracted. Conversely, a large amount of detailed usage data and context data was collected, using sensors such as the microphone to record ambient noise. This increased the number of permission dialogues presented to potential participants on sign-up to the app which may have contributed to the relatively low engagement numbers. Another limitation arising from experimental design was the use of notifications for delivering the ESM questionnaires. Delivering the questionnaires via notification may have unintentionally caused or invoked irregular smartphone or notification engagement behaviour in participants. For example, a participant may have unlocked their



device or attended to all types of notifications more frequently simply because they wished to complete questionnaires for the study quickly. Another limitation, related closely to privacy and ethics, was that the full notification text content was not saved for analysis. This decision was made to protect the privacy of participants, however it does reduce the granularity of the resulting data set collected and its usefulness, particularly for NLP research and applications. For the same reason, participant contacts, and individual social relationships with those contacts, were not saved. On reflection, pre and post-study questionnaires could have also provided additional demographic and qualitative data regarding notification behaviour and participant perceptions toward the WeAreUs app’s functionality, such as the visual representations of *stories* created using their notification behaviours and the inferred significant contacts in differing places.

### 3.4.3 Experimental Execution Limitations

The number of participants that engaged with the WeAreUs study is a limiting factor of this work. However, the number is similar to other state-of-the-art in-situ studies in the domain of notification engagement [MHM16a; Meh+15b; Meh+16], in which final analyses included 16, 17 and 20 participants, respectively. A mitigating factor is the granularity of data collected, which, as mentioned, may have been a reason for the low participation in the study. Following this, gender diversity is also a limitation of this work. While the study was advertised across many channels within and also outside the university, the number of female participants who signed up and engaged with notifications was low. On reflection, this study could also have been extended to run for much longer than the original 6 months in which it was active. A longitudinal study design might have been better suited and could have reduced or eradicated the low participation and gender diversity limitations.



# Chapter 4

## Generation & Simulation of Push-Notifications

We have the algorithms to save millions of lives, but only if we dismantle data silos.

---

*Jeremy Howard*

### 4.1 Overview

The previous chapter identified, through studying the engagement action taken by notification subscribers, inefficiencies in the delivery of push-notifications. The high dismissal rates indicate valuable time and energy was invested by subscribers attempting to manage the influx of notifications which they received. In a digital age, when our attention is divided and spread thin, performance and focus on simple tasks has declined dramatically [Tho+14; Lee+14b; Mar+15; War+17]. Addictive tendencies due to interactions with technology have become widespread, spawning new terms such as NoMoPhobia [Kin+13], FOMO [Prz+13] and Phantom Vibration Syndrome [Lin+14]. Particularly in younger generations, the social pressure to respond to messages and alerts immediately is immense [Elh+16; Lee+14a]. The designers of technology are partly to blame as subtle features are embedded in applications with the sole purpose to hook users, increase

engagement and form new, potentially unhealthy, habits [Oul+12]. Smartphone notifications are a prime example of a design hook used to attract users back on to, and engage with, a platform. When used in an ethical and empathetic manner, notifications can be a powerful tool for improving end-user experiences. However, when implemented in an exploitative or unintelligent manner, intended or not, notifications tend to do more harm than good [Smi+14].

This chapter explores and evaluates solutions for improving the push-notification experience for both subscribers and creators. This was achieved by first assessing the ability of algorithms to predict engagement (open/dismiss) actions toward notifications. A NMS with the ability to identify whether a notification would be opened or dismissed by a subscriber, in a given context, could throttle deliveries to favourable moments. Training such a NMS would require large amounts of historical notification engagement data, which as discussed in previous chapters, is difficult to obtain, particularly at scale and in balance (i.e. both positive and negative engagement action scenarios represented equally in the data set). Therefore, an approach toward generating synthetic notification-context pairs was explored through implementation of a number of generative models. The models were evaluated with respect to their utility for training NMSs as well as their ability to maintain real-world associations, albeit with enough deviation to offset privacy concerns and facilitate exploration of previously unforeseen notification engagement scenarios.

The contributions of this chapter are as follows:

1. An analysis of algorithmic performance when predicting future subscriber engagement actions toward notifications
2. An approach for generating synthetic notification and context data
3. An evaluation of synthetic notification data with respect to statistical consistency and agent utility

## 4.2 Predicting Notification Engagement Actions

Measuring the value of a notification to a subscriber is not a straightforward task. The most logical, and perhaps simplistic, approach is to use the binary feature *action*, which indicates whether a subscriber opened or dismissed a notification. From this feature, the Click-Through-Rate (CTR) of notifications can

be determined and used as a measure of success or failure for notifications which have been pushed to a subscriber. The CTR is the total number of notifications opened divided by the total number of notifications delivered. However, there are a number of scenarios, both positive and negative, not captured by this metric. For example, a notification may provide adequate information in the title and ticker text that subsequently does not require further subscriber engagement. Therefore, a notification may have been positively received by a subscriber yet also dismissed. Conversely, a notification opened by a subscriber may also not result in a positive user experience. For example, in scenarios where a notification could be classified as clickbait (which is discussed further in the following chapters). For these reasons, notification marketing campaigns tend to use other metrics for measuring the success or failure of notifications pushed at subscribers, such as subsequent time spent on a platform or purchases made after a push. In contrast, it could be argued that the metrics used by NMSs working on behalf of those receiving notifications should be derived from physical, cognitive and emotional health features of the subscribers to ensure notifications presented are positively contributing toward their well-being.

This section evaluates the simpler approach of predicting notification engagement actions, open and dismiss, which could be leveraged by a NMS for identifying favourable moments for notification delivery. Predicting the engagement action of a notification is a common task repeated and reported in push-notification literature, albeit without a standard set of open data sets or features on which replicable benchmarks can be derived, thus perhaps weakening comparative evaluations. The following sections will address this through the creation of a synthetic notification data set as well as report the performance achieved by a number of algorithms, to be used as a benchmark for future work.

### 4.2.1 Notification Data & Features

The WeAreUs notification data, analysed in the previous chapter, was used to evaluate the performance of a selection of machine learning algorithms at predicting notification engagement actions. The data set was comprised of 13,933 notifications spanning a wide range of notification categories, topics and apps. A subset of features ( $n = 46$ ), informed by those referenced in push-notification literature ( Tables 2.4 2.5 and 2.6), and comprised of subscriber context (at

moments of delivery) and notification content/settings (selected by notification creators) were chosen as input to the models. The features were:

**Context:** { activity, airplane mode, bluetooth headphones, bluetooth on, charging, connection type, headphone jack, music active, place, proximity, ringer mode, roaming status, app significance, contact significance, screen state, average noise, battery level, light intensity, bluetooth devices nearby, step count, device volume levels (alarm, music, notification, ringer), time app last used, number of updates, time from first update to last update, day of week, hour of day, minute of hour, holiday }

**Notification:** { app, category, topic, LED colour, priority, show when, sound, vibrate pattern, visibility, number of alerts }

The notification features set by notification creators included the *app*, *category*, *topic*, *LED colour*, *priority* and *visibility*. The app developer or marketing team pushing a notification decides how to populate these features of a notification, albeit with some limitations as, for example, there are a finite number of priority levels available to choose from. Additional features which notification creators also have control over are: *time of day* and *day of week*. However, the time and day a pushed notification is delivered can sometimes be outside of their control. For example, if a subscriber's device is turned off, without internet connectivity or their device model is restricting background tasks (e.g. in power saver mode), the delivery of a notification may be delayed. Thus these two features can also be categorised as contextual. The remaining features in the subset are features relating to the context of the subscriber at moments of notification delivery. Using both context and notification features enables a machine learning model to learn scenarios in which subscribers are receptive to certain types of notification and predict whether a notification will be opened in a given context.

## 4.2.2 Predicting Notification Engagement Action

A number of standard machine learning models, identified within literature and industry, were selected for predicting the action of opening or dismissing a delivered notification. The act alone of opening or dismissing a notification does not necessarily imply a positive or negative engagement. However, if opened, it does demonstrate a subscriber's willingness to divert their attention to and engage with the app delivering the notification in a given moment - which is useful for a

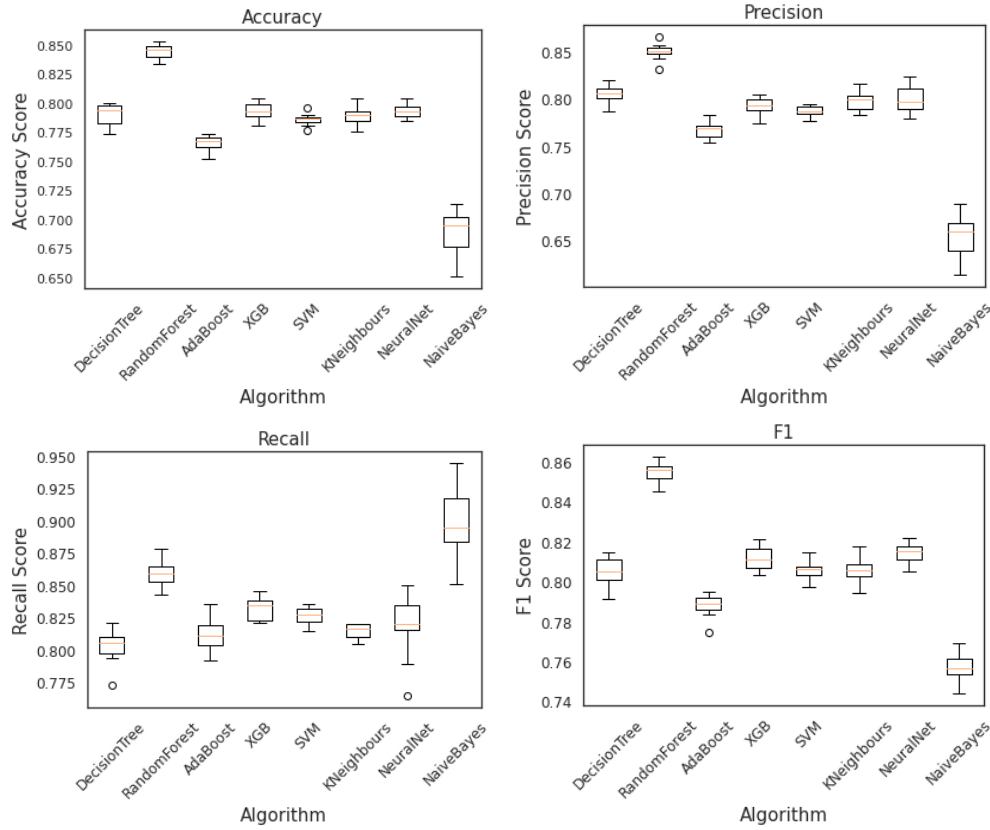


Fig. 4.1 Algorithm performance predicting notification engagement actions

NMS to model as it could help identify similar moments in which subscribers are most receptive to investing time on a platform if nudged by a notification. The algorithms evaluated for learning notification engagements were: Decision Tree [Bre+84], Random Forest [Bre01], AdaBoost [FS97], XGBboost [CG16], Support Vector Machine [Sch+00], K-Nearest Neighbors [Gol+04], Neural Net [MP43] and Naive Bayes [Zha04]. These algorithms in particular were identified within recent research as potential, or part, solutions applied to various push-notification inference tasks, including interruption and action prediction [Meh+15b; Pie+14; KFG18; OTT19; Oko+17; TAW17]. Therefore, the evaluation of this work further builds upon the state-of-the-art within the push-notification management space by offering a comparative evaluation (albeit with a known caveat that the sample of data is quite different). However, as will be discussed, the results of the algorithms applied to synthetic notification data can and will facilitate standard evaluation benchmarks to be drawn on a data set that can be freely shared with the research community for future work. The performance metrics chosen for evaluating each algorithm were *accuracy*, *precision*, *recall* and *f1 score*. Stratified 10-fold cross

validation was used for evaluating the performance of each algorithm on the WeAreUs notification engagement data. The results are illustrated in Figure 4.1.

The Random Forest algorithm performed best across accuracy (84%), precision (85%) and F1 score (85%) metrics, followed closely by the Decision Tree and XGBoost. However, the Naive Bayes algorithm was superior with respect to recall, achieving 90% compared to Random Forest at 86%. Recall is a particularly important metric with respect to real-world deployment of a NMS as the cost of a *False Negative* is quite high. For example, a notification which is important to the subscriber, perhaps containing emergency or time-sensitive information, but subsequently withheld by the NMS, due to incorrectly classifying the notification as irrelevant to the subscriber in a given moment, could breed mistrust in the relationship between machine and end-user as the subscriber might feel important information is being withheld from them. A high recall value indicates fewer False Negative occurrences. The precision metric is important with regard to the utility of a NMS as higher values indicate lower False Positives occurring. A False Positive, in this case, being a NMS delivering an irrelevant or badly timed notification to a subscriber who subsequently dismisses it. Thus not improving the notification engagement experience. Whilst lower precision values are not (as much) a cause for concern as lower recall values, a NMS which is not effectively delivering notifications at improved moments cannot expect to be used by a subscriber. The Naive Bayes algorithm achieved worst results for precision (66%), thus indicating it is a relatively weak NMS compared to other algorithms. As such, f1-score, which represents a balance between recall and precision, was used as an indicative measure of a well balanced NMS. Random Forest, Decision Tree and XGBoost were identified as the top performing algorithms predicting notification engagement action. Whilst far from perfect, the results illustrate that push-notification actions can be predicted to a useful degree of accuracy by observing a combination of key context and content features, which when deployed in-the-wild, can improve the receptivity and acceptance of notifications which are pushed only at moments of predicted positive action.

Further examination of the features used for predicting notification engagement action and their relative importance within the task was carried out using *scikit-learn's* implementation of *Permutation Importance* [Ped+11; Bre01]: *permutation feature importance is defined to be the decrease in a model score when a single feature value is randomly shuffled*. A Random Forest classifier was fit on 80% of the notification engagement data and permutation importance values of each feature were calculated using the remaining 20% held-out set of engagements.



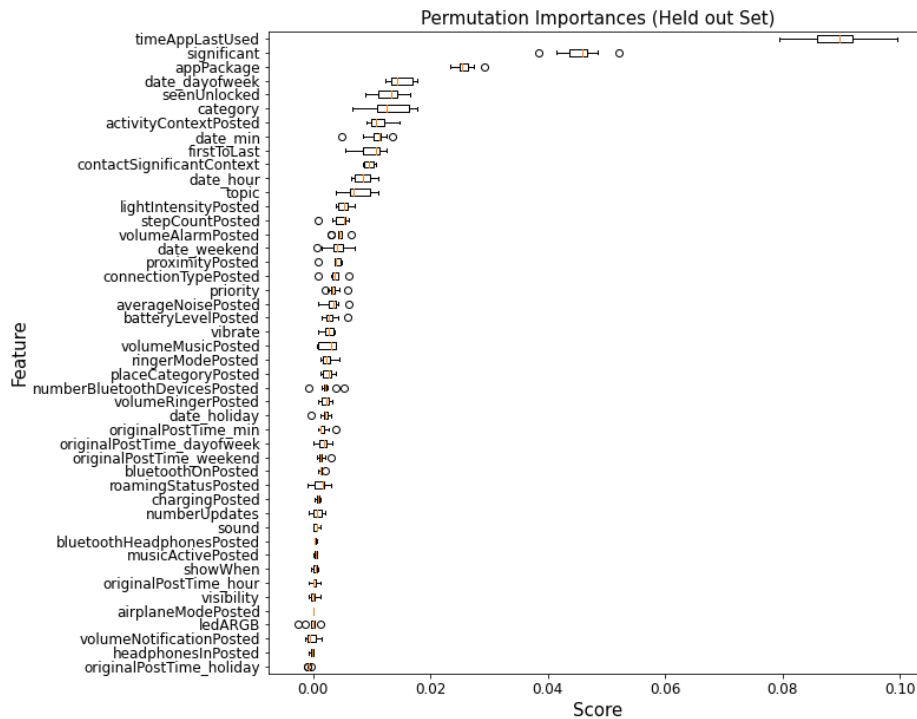


Fig. 4.2 Permutation feature importance when predicting notification engagement actions

Figure 4.2 illustrates the importance ranking of features for the task. The most important features identified were related to the app sending the notification. The time between using an app and receiving a subsequent notification was ranked most important followed by app significance (a boolean indicating if an app was frequently over time) and the app package name. Other contextual features also were also ranked highest, such as the day of the week, contact significance and whether or not the notification arrived when the subscriber’s device was unlocked. Other notification attributes set by the notification pusher were also found to impact the model performance, such as the category of notification, the topic and priority value set.

Due to a large number of features having a relatively similar low ranking score, indicating their contribution to the performance of the model was low, a reduced subset of features ( $n = 17$ ) was selected for exploring generative modeling. This subsequently reduced the dimension of the encoded notification vector from 314 (full feature) to 196 (feature subset), optimising both compute resources and time. The reduced subset of features was evaluated with the same models outlined above and results indicated a negligible drop in performance between the full feature set

and the reduced subset (MAE of accuracy, precision, recall and f1 scores between feature sets were 0.007, 0.01, 0.01 and 0.006 respectively).

**Feature subset:** { activity, app, app significance, category, contact significance, LED colour, priority, topic, holiday, screen state, day of week, time app last used, weekend, hour of day, minute of hour, time from first update to last update, number of updates }

### 4.3 Generating Synthetic Notification-Context Pairs

Procuring notification engagement data in-the-wild is difficult and time consuming, as discussed in previous chapters. Few open-source push-notification data sets could be found referenced in literature due to the sensitivity of the content present in the notification and the context in which it is delivered. This section presents an approach toward applying generative models to push-notifications for the purpose of creating synthetic notification engagement data imitating real-world relationships and distributions found between notifications, contexts and subscriber engagements. A number of model variations were chosen for implementation (WGAN, WGAN-GP, CWGAN, GPT-2, SynthPop) and were evaluated with respect to their ability to generate realistic data, preserve associations and maximise utility for training NMSs, such as those presented in the previous section.

#### 4.3.1 Generative Adversarial Networks

Generative Adversarial Networks (GAN) [Goo+14] are a type of unsupervised-learning algorithm comprised of two models, e.g. two multilayer perceptrons (MLPs), competing against each other (Figure 4.3). The first MLP, the generator (G), takes random noise (prior distribution) as input and attempts to generate realistic data samples. The second MLP, the discriminator (D), takes a data sample as input and attempts to identify whether it is real or fake, the output being a scalar value of probability of the sample being real data,  $[0, 1]$ . During training, D is given alternating real and generated values. Ideally, the loss associated with both networks converges as G learns to generate better data samples and D learns to better discriminate between real and synthetic samples. At the optimum point, D can no longer discriminate between the real and generated data, hence G can

be used to generate data which imitates the distribution of real values. It can be formulated as a *two-player minimax game with value function*  $V(G,D)$  [Goo+14]:

$$\min_G \max_D V(D, G) = \mathbb{E}_{x \sim p_{data}(x)} [\log D(x)] + \mathbb{E}_{z \sim p_z(z)} [\log(1 - D(G(z)))] \quad (4.1)$$

In which:

$x$  = real data variable

$z$  = latent variable randomly sampled from prior  $p_z$

$G(z)$  = maps latent variable  $z$  to the data space

$D(z)$  = outputs a probability of input being sampled from real distribution  $p_{data}$

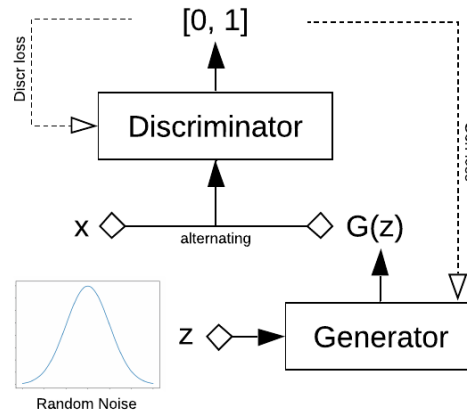


Fig. 4.3 Training a Generative Adversarial Network

Since its conception and subsequent publication by Goodfellow et al. in 2014, many variations of GANs have been developed [Hin18; Luc+17]. One such variation is the Wasserstein GAN (WGAN) proposed by Arjovsky et al. [ACB17]. This variation comprised a number of differences to the original GAN which improved the stability of training, key of which was the application of Wasserstein Loss which facilitated greater distinction between real and synthetic samples. When minimized, this reduced the distance between the distributions of real data used for training and synthetic data created by the generator. The discriminator (D) was renamed a *critic* (C) as it no longer produced a probability in the range  $[0, 1]$  which classified samples as real or synthetic, but instead it produced a scalar score prediction in the range  $[-\infty, \infty]$  representing how real a sample was. This helped approximate the Wasserstein distance between real and generated samples, which when minimised, indicated the generator was close to the real data distribution.

In order to ensure stability when training, Arjovsky et al. also introduced weight-clipping to force the critic to operate as a *1-Lipschitz continuous function*, thus limiting *the rate at which the predictions of the critic can change between two samples* [Fos19]. A further GAN variation proposed by Gulrajani et al. [Gul+17] in 2017, the WGAN-GP, introduced a gradient penalty as an alternative to weight clipping to ensure the critic conformed to a 1-Lipschitz continuity function.

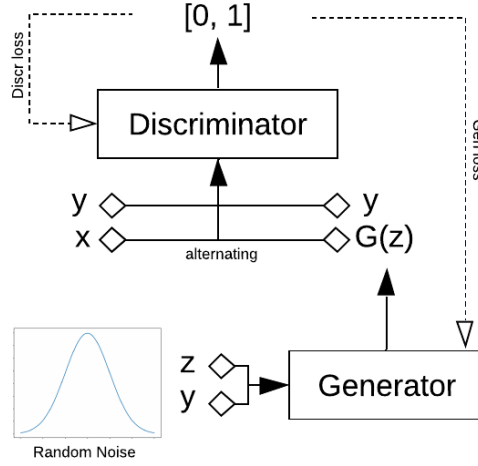


Fig. 4.4 Training a Conditional Generative Adversarial Network

Goodfellow et al. also briefly introduced the concept of a conditional GAN as future work, positing that  $p(x|c)$  could be learned by adding conditional input to the generator and discriminator networks (Figure 4.4). Conditional Generative Adversarial Nets (CGAN) were later defined by Mirza et al. [MO14] to facilitate and *direct the data generation process*. This enabled synthetic samples to be generated conditionally and thus suit a situational need. Applied to push-notifications, this could mean generating a notification conditioned on the action being opened, or generating notifications conditioned to suit a certain context, such as place or time. This was formalised as:

$$\min_G \max_D V(D, G) = \mathbb{E}_{x \sim p_{data}(x)} [\log D(x|y)] + \mathbb{E}_{z \sim p_z(z)} [\log(1 - D(G(z|y)))] \quad (4.2)$$

where:

$y$  = additional information input to condition networks  $G$  and  $D$

For the purpose of generating synthetic notifications, comprised of both continuous and discrete notification-context features, implementations of the WGAN and

WGAN-GP were applied to learn the distributions of push-notification engagement features. These models were state-of-the-art at the time of experimentation, showed improved training stability to generate realistic data samples, and supported synthesis of both continuous and discrete variables which was essential for this work as notifications comprise both types of feature (e.g. topic of notification = discrete; number of notification updates = continuous). A Conditional WGAN-GP was also tasked with generating synthetic push-notification data sets to evaluate the potential for controlling the notifications generated. This type of model could be used to automate the task of marketing teams creating push-notifications personalised to particular subscribers whilst conditioning on contexts or outcomes, such as positive engagements. In addition to these three, two alternate models were also implemented to facilitate comparative evaluation of the generative models performance:

- SynthPop - a well-established R package created by Nowok et al. [NRD+16], SynthPop creates a generative model by calculating conditional distributions over all features in the data set. Classification and regression trees (CART) [CGM98] are implemented to create models which are fitted feature by feature, with the next feature being conditioned on all previous features.
- Random Sampling - this generative method represents the baseline. For every feature of the data set, a random value was sampled from real data to create multiple instances of randomly assembled notification-context-engagement pairings. To generate continuous variables, values were sampled from within the min/max range identified in the real data. Discrete variables were randomly sampled from the unique set of categories present within the real data.

### 4.3.2 Evaluation Metrics

The WeAreUs data described in the previous chapter was used for training the generative models selected and it was additionally used for evaluating the utility of the subsequent synthetic data generated. The subset of features, identified in subsection 4.2.2, were selected for generation due to their significance when predicting subsequent notification engagements.

The main focus of this work was to ensure that realistic synthetic notification data was created and that it would provide utility to intelligent models trained for

Table 4.1 Evaluation metrics, identified in literature and implemented in industry, which quantify Quality, Utility and Privacy of synthetic data.

| Measure | Features   | Metric                        | Origin and/or Use                 |
|---------|------------|-------------------------------|-----------------------------------|
| Quality | continuous | Pearson Correlation           | [Bea+19; How+17; Kal19]           |
|         | all        | Propensity Score              | [Woo+09; Sno+16]                  |
|         | discrete   | Mutual Information            | [Cov99; GT09; Vie21; PSH17; MN18] |
| Utility | all        | Synthetic Ranking Agreement   | [JYV18; JYS18; Ros+20]            |
|         | all        | Feature Importance Similarity | [Bea+19; Vie21; WMZ10; JYS18]     |
|         | all        | Train on Real, Test on Synth  | [FGG20; JYS18; MN18]              |
| Privacy | all        | Hitting Rate                  | [LWY19]                           |
|         | all        | Distance to Closest Record    | [Par+18; LWY19]                   |

intelligent notification management tasks. In addition to the utility and quality, and to facilitate comparative evaluation and sharing, this work also evaluated the synthetic data with respect to its potential for exposing the privacy of individuals from which the real data was gathered. Table 4.1 illustrates a consolidated list of applicable metrics related to measuring the quality, utility and privacy of synthetic data. The metrics are being actively used in both research and industry domains. The following sections detail the construction and interpretation of the metrics selected with respect to evaluating synthetic push-notifications.

### Pearson Correlation

Pairwise Pearson correlations were used to evaluate whether or not the generative models were capable of preserving the statistical properties of continuous notification features [Bea+19; How+17; Kal19]. This was achieved by extracting correlations between all continuous features,  $n$ , within the real and synthetic data sets respectively and subsequently calculating the Mean Absolute Error (MAE) between the pairwise correlations of both data sets to consolidate results into a single score for utility. Smaller values indicate continuous synthetic data generated closely matches real distributions.

$$r_{MAE} = \frac{1}{n} \sum_{i=1}^n |r_{synth} - r_{real}| \quad (4.3)$$

where:

- $r_{real}$  = pairwise pearson correlations of real data
- $r_{synth}$  = pairwise pearson correlations of synthetic data
- $n$  = number of pairwise comparisons

### Propensity Score

Propensity score Mean-Squared Error ( $p_{MSE}$ ) [Woo+09; Sno+16] was used as a metric for assessing synthetic data quality as it measures the probability of a treatment being assigned given observed covariates [Aus11]. In this instance, the covariates are the features of notification-context engagements and the treatment is an assignment to one of two classes, real or synthetic. The score obtained, as discussed by Snoke et al., then represents *an estimate of distinguishability*. Smaller propensity scores indicate that the synthetic data can not be easily distinguished from real data, thus suggesting greater similarity between both sets and higher quality synthetic data.

$$p_{MSE} = \frac{1}{N} \sum_{i=1}^N (\hat{p}_i - c)^2 \quad (4.4)$$

where:

$\hat{p}$  = propensity score i.e. probability of sample  $i$  belonging to real or synth class

$N$  = total samples in combined data set i.e.  $N_{real} + N_{synth}$

$c$  = the proportion of synth samples in combined data set i.e.  $\frac{N_{synth}}{N}$

### Mutual Information

Mutual Information, the *reduction in uncertainty about one random variable given knowledge of another* [LR09], was used to evaluate the pairwise co-dependence of the discrete notification features in both the real and synthetic data sets respectively. The MAE between the real and synthetic data sets was calculated for a singular score representing discrete data quality i.e. the capability of the generative model to preserve relationships between discrete elements. Smaller values indicate that the discrete synthetic data generated closely matches the distributions and relationships found in the real data [Cov99; GT09; Vie21; PSH17].

$$MI(X; Y) = \sum_{x \in X} \sum_{y \in Y} p(x, y) \log \frac{p(x|y)}{p(x)p(y)} \quad (4.5)$$

$$MI_{MAE} = \frac{1}{n} \sum_{i=1}^n |MI_{synth_i} - MI_{real_i}| \quad (4.6)$$

where:

|              |                                          |
|--------------|------------------------------------------|
| $X, Y$       | = a pair of discrete variables           |
| $p(x), p(y)$ | = probability of observing $x, y$        |
| $p(x y)$     | = probability of observing $x$ given $y$ |
| $n$          | = number of pairwise comparisons         |

### Synthetic Ranking Agreement

Synthetic Ranking Agreement (SRA), proposed by [JYV18; JYS18], measures the degree to which the generated data preserves the relative performance of algorithms trained upon it compared with performance achieved using real data. The hypothesis is that differing algorithm performance should be consistent between the real and synthetic data sets if utility is to be maximised i.e. different algorithms trained to predict push-notification engagement should have identical performance rankings when both synthetic or real data is used. For the implementation of SRA in this work, formalised in Equation 4.7, five algorithms were chosen for predicting notification engagement action (open/dismissed): Random Forest; AdaBoost; XGBoost; Naive Bayes; and K-Nearest Neighbors. The indicator function resolves to 1 when the ordering of algorithms is identical between real and synthetic data and resolves to 0 when the ordering differs. Higher values indicate that algorithm rankings are preserved between both data sets and that use of the synthetic data set provides a level of utility for intelligent agents consistent with that of real data.

$$SRA(\{A_i\}_{i=1}^L, \{C_i\}_{i=1}^L) = \frac{1}{L(L-1)} \sum_{j=1}^L \sum_{k \neq j} \mathbb{I}((A_j - A_k) \times (C_j - C_k) > 0) \quad (4.7)$$

where:

|              |                                                                      |
|--------------|----------------------------------------------------------------------|
| $L$          | = number of predictive models (five used in this work)               |
| $A_i$        | = performance of model $i$ when trained and tested on real data      |
| $C_i$        | = performance of model $i$ when trained and tested on synthetic data |
| $\mathbb{I}$ | = indicator function                                                 |

### Feature Importance Similarity

Similar to the previous metric, the Feature Importance Similarity (FIS) utility score represents the degree to which the generative model is capable of preserving rankings, in this case: the rankings of important features extracted when training for a prediction task [Bea+19; Vie21]. The task chosen was the prediction of



push-notification actions (open/dismiss) given notification and context features. The model chosen for prediction was Random Forest, as it both demonstrated the best performance at predicting engagement action and supports extraction of feature importances. Two methods of ranked list comparison were considered for evaluating the ranked feature importance lists created when the classifier was trained and tested upon real and synthetic data respectively: Kendall's Tau [Sen68] and Rank-Biased Overlap (RBO) [WMZ10]. RBO was selected in this instance as it facilitates placing greater weight for certain ranking positions whilst Kendall's Tau does not. As illustrated in Figure 4.2, there is a clear distinction between the importance of the top 3 features and the remaining features when training to predict notification engagement action. Whilst the correct positioning of all features in the ranked list is the goal to ensure consistency across real and synthetic data, using RBO and providing higher weight to the top 3 importance positions ensures greater penalty is given to the generative model's feature importance similarity score if it fails to explicitly correctly rank the most important features. It can be argued that minor shifts in ranking between the less important features is not as relevant, particularly as there is little distinguishing distance between those feature importance scores. Higher RBO values indicate greater similarity between the ranked importance of features extracted using synthetic and real data and subsequently suggests greater consistency and utility of the synthetic data.

$$FIS = RBO_{EXT}(S, T, p, k) = \frac{X_k}{k} \cdot p^k + \frac{1-p}{p} \cdot \sum_{d=1}^k \frac{X_d}{d} \cdot p^d \quad (4.8)$$

where:

- $S, T$  = ranked lists of feature importances extracted from real and synth sets
- $d$  = depth of top positions to be weighted
- $p$  = tunable parameter determines the contribution of top d ranks, range (0, 1)
- $k$  = evaluation depth i.e. length of lists
- $X$  = size of overlap between S and T up to depth

**Train on Synthetic, Test on Real** Esteban et al. [EHR17] proposed the Train on Synthetic, Test on Real (TSTR) approach to measure the utility of synthetic data applied to a prediction task. The approach requires training an agent on real and synthetic data respectively and testing both trained agents on a held-back set of real data. A comparison of the performance of both agents then describes the utility of using synthetic data for the task. The approach has

also been implemented by Hazy [Vie21; FGG20] who have additionally proposed the *Predictive Capability* (PC) metric which consolidates the result into a single utility score. F1 score was chosen as the metric for performance comparison in this work. Values at or above 1 indicate identical or improved utility when using synthetic data.

$$PC = \frac{F1 - score_{synth}}{F1 - score_{real}} \quad (4.9)$$

### Hitting Rate

The hitting rate in this instance, is a metric which can allude to potential exposure of real data within synthetic data created by the generative model. Lu et al. [LWY19] define it in their work as *the ratio of the number of identical records between the original data and synthetic data and the total number of records*. As the generative model attempts to maximise the similarity between the real data it is trained upon and the synthetic data it generates, there exists a possibility that the generative model simply learns to regurgitate explicit samples from the real data set. The hitting rate therefore identifies the percentage of exact real sample matches found in the synthetic data. This metric is relevant in this work as push-notification data describes end-user attributes (such as apps installed), preferences (receptivity to certain topics, apps and contacts) and contexts (times of day or places using a smartphone), which could be leveraged by attackers to re-identify them if combined with other sources of information. Higher hitting rates indicate greater similarity between the real and synthetic data generated but also a greater exposure risk. Therefore, a balance must be obtained between quality and minimising exposure risk, which is an active research topic in the domain of generative modeling [EN14; MS04; NS08].

### Distance to Closest Record

The Distance to Closest Record (DCR) privacy metric describes how close the synthetic data resembles the real data from which it was generated. Once again, there is a trade off between maximising the quality and utility of the synthetic data generated and maximising the privacy of individuals from which the data came. Park et al [Par+18] and Lu et al. [LWY19] use Euclidean Distance to measure the distance between records of synthetic and real samples. In this work, as both continuous and discrete features are present in the notification engagement data sets, Gower distance [Gow71] was used to represent the similarity between samples. Pairwise comparisons were made between each sample of the real and

Table 4.2 Differing model implementation hyperparameters used when generating synthetic push-notifications

| Model  | Hyperparameter    | WGAN                                      | WGAN-GP                                    | CWGAN-GP                                   |
|--------|-------------------|-------------------------------------------|--------------------------------------------|--------------------------------------------|
| Gen    | Neural Net        | Feed-forward                              | Feed-forward                               | Feed-forward                               |
|        | Hidden Layer      | 1 fully connected                         | 1 fully connected                          | 1 fully connected                          |
|        | Activation Hidden | ReLU                                      | ReLU                                       | ReLU                                       |
|        | Activation Output | Softmax, Sigmoid                          | Softmax, Sigmoid                           | Softmax, Sigmoid                           |
|        | Loss function     | Wasserstein Loss                          | Wasserstein Loss                           | Wasserstein Loss                           |
|        | Optimizer         | RMSProp                                   | Adam                                       | Adam                                       |
|        | Learning rate     | 0.0001                                    | 0.0002                                     | 0.0002                                     |
| Critic | Neural Net        | Feed-forward                              | Feed-forward                               | Feed-forward                               |
|        | Hidden Layer      | 1 fully connected                         | 1 fully connected                          | 1 fully connected                          |
|        | Activation Hidden | ReLU                                      | ReLU                                       | ReLU                                       |
|        | Activation Output | Linear                                    | Linear                                     | Linear                                     |
|        | Loss function     | Wasserstein Distance<br>+ weight clipping | Wasserstein Distance<br>+ Gradient Penalty | Wasserstein Distance<br>+ Gradient Penalty |
|        | Optimizer         | RMSProp                                   | Adam                                       | Adam                                       |
|        | Learning rate     | 0.0001                                    | 0.0002                                     | 0.0002                                     |

synthetic data sets and the mean and standard deviation of the minimum distance was recorded. A mean value of 0 would indicate that a real record was copied directly to the synthetic data set, thus potentially indicating a potential privacy risk. In the case of mean values larger than 0, if standard deviation values are also high it would also indicate samples exist which could closely resemble real data, thus also potentially reducing the privacy held by the synthetic data.

### 4.3.3 Implementation, Synthesis Results & Discussion

The settings and hyperparameters used in the implementations of the WGAN, WGAN-GP and CWGAN-GP developed for this work are illustrated in Table 4.2. As a preprocessing step, the discrete notification and context features were transformed to one-hot encoded vectors and combined with continuous notification and context features scaled within the range of  $[0, 1]$ . The dimension of the encoded vector was 198. The CWGAN-GP had a slight variation for the preprocessing step as the model required an additional conditional input parameter,  $y$ . For this work, the notifications generated were conditioned on the app name. Therefore the dimensions of the condition vector,  $y$ , and notification-context vector,  $x$ , were 105 and 93 respectively. The WGAN-GP, highlighted in literature as the best model suited to the task of generating tabular data such as the notification-context pairs, was additionally examined with respect to varying:

Table 4.3 WGAN-GP model utility, quality and privacy scores for varying batch size (b) and latent dimension ( $\ell$ ) (Top 10 at maximising utility of synth data). Best values achieved emboldened (max or min depending on metric).

| $\ell$     | b     | FIS           | PC            | SRA        | $p_{MSE}$     | $r_{MAE}$     | $MI_{MAE}$    | HR | $DCR_{AVG}$   | $DCR_{STD}$   |
|------------|-------|---------------|---------------|------------|---------------|---------------|---------------|----|---------------|---------------|
| (64 - 128) | (128) | 0.7079        | 0.8711        | <b>1.0</b> | <b>0.0096</b> | <b>0.0363</b> | 0.0323        | 0  | 0.0861        | <b>0.0138</b> |
| 128        | 128   | 0.8727        | 0.8832        | 0.8        | 0.0120        | 0.0575        | 0.0330        | 0  | 0.0826        | 0.0168        |
| 16         | 128   | <b>0.8730</b> | 0.8469        | 0.8        | 0.0108        | 0.0815        | 0.0303        | 0  | 0.0920        | 0.0163        |
| 128        | 16    | 0.7165        | 0.8505        | 0.9        | 0.0496        | 0.1025        | 0.0278        | 0  | 0.0855        | <b>0.0138</b> |
| 32         | 16    | 0.8251        | <b>0.8991</b> | 0.7        | 0.0967        | 0.0976        | <b>0.0252</b> | 0  | <b>0.0974</b> | 0.0186        |
| 128        | 64    | 0.8516        | 0.8687        | 0.7        | 0.0832        | 0.0426        | 0.0259        | 0  | 0.0892        | 0.0178        |
| 16         | 64    | 0.8530        | 0.8464        | 0.7        | 0.0474        | 0.0390        | 0.0286        | 0  | 0.0910        | 0.0162        |
| 64         | 64    | 0.7690        | 0.8577        | 0.7        | 0.0173        | 0.0521        | 0.0299        | 0  | 0.0948        | 0.0183        |
| 16         | 32    | 0.8322        | 0.8772        | 0.6        | 0.0961        | 0.0518        | 0.0274        | 0  | 0.0938        | 0.0179        |
| 32         | 128   | 0.8288        | 0.8705        | 0.6        | 0.0231        | 0.0474        | 0.0321        | 0  | 0.0895        | 0.0183        |

- latent dimension - the size of the latent space i.e. the input noise vector
- batch-size - the number of training samples used per iteration (weight update)
- training epochs - the number of passes of the entire training data set
- seed data - the amount of real data used to train the model

Subsequently, the most suitable latent dimension, batch-size and training epoch were selected based on the performance of the WGAN-GP and applied to the WGAN and CWGAN-GP models for comparison of quality, utility and preservation of privacy. As discussed in the previous section, models that perform well are those that maximise utility: Predictive Capability ( $PC$ ); Synthetic Ranking Agreement ( $SRA$ ); and Feature Importance Similarity ( $FIS$ ), whilst minimizing the difference in quality between real and synthetic data sets: Propensity Mean Squared Error ( $p_{MSE}$ ); Pearson Mean Absolute Error ( $\rho_{MAE}$ ); and Mutual Information Mean Absolute Error ( $MI_{MAE}$ ). The models should also minimize Hitting Rate ( $HR$ ) and strike an acceptable balance between quality and Distance to Closest Record ( $DCR$ ).

Tables 4.3 and 4.4 contain the quality, utility and privacy scores obtained for the WGAN-GP using different combinations for latent dimension and batch size, sorted by maximum utility (Table 4.3) and minimum loss of quality (Table 4.4) respectively (top 10 combinations selected). The number of training samples used for all variations was 10,000 and comprised engagements from all users. The WGAN-GP model implementation with a latent dimension of 64 and batch size of

Table 4.4 WGAN-GP model utility, quality and privacy scores for varying batch size (b) and latent dimension ( $\ell$ ) (Top 10 at minimising quality loss between real and synth). Best values achieved emboldened (max or min depending on metric).

| $\ell$    | b   | FIS           | PC            | SRA        | $p_{MSE}$     | $r_{MAE}$     | $MI_{MAE}$    | HR | $DCR_{AVG}$   | $DCR_{STD}$   |
|-----------|-----|---------------|---------------|------------|---------------|---------------|---------------|----|---------------|---------------|
| (64, 128) |     | 0.7079        | 0.8711        | <b>1.0</b> | <b>0.0096</b> | 0.0363        | 0.0323        | 0  | 0.0861        | <b>0.0138</b> |
| 64        | 64  | 0.7690        | 0.8577        | 0.7        | 0.0173        | 0.0521        | 0.0299        | 0  | 0.0948        | 0.0183        |
| 128       | 128 | 0.8727        | <b>0.8832</b> | 0.8        | 0.0120        | 0.0575        | 0.0330        | 0  | 0.0826        | 0.0168        |
| 32        | 128 | 0.8288        | 0.8705        | 0.6        | 0.0231        | 0.0474        | 0.0321        | 0  | 0.0895        | 0.0183        |
| 64        | 32  | 0.7301        | 0.8466        | 0.7        | 0.0434        | <b>0.0336</b> | 0.0331        | 0  | 0.0818        | 0.0174        |
| 16        | 64  | 0.8530        | 0.8464        | 0.7        | 0.0474        | 0.0390        | 0.0286        | 0  | 0.0910        | 0.0162        |
| 16        | 128 | <b>0.8730</b> | 0.8469        | 0.8        | 0.0108        | 0.0815        | 0.0303        | 0  | 0.0920        | 0.0163        |
| 32        | 64  | 0.6343        | 0.8782        | 0.7        | 0.0404        | 0.0512        | 0.0313        | 0  | 0.0841        | 0.0167        |
| 32        | 32  | 0.6645        | 0.7433        | 0.5        | 0.0442        | 0.0540        | 0.0302        | 0  | <b>0.0952</b> | 0.0175        |
| 128       | 32  | 0.8103        | 0.7499        | 0.5        | 0.0753        | 0.0487        | <b>0.0274</b> | 0  | 0.0915        | 0.0175        |

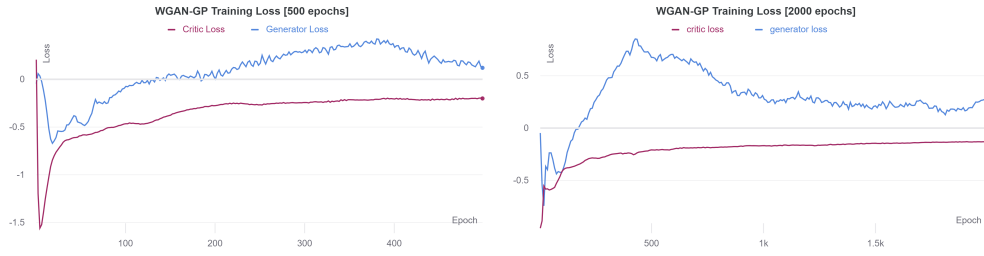


Fig. 4.5 WGAN-GP generator and critic loss convergence whilst training.

Table 4.5 WGAN-GP model utility, quality and privacy scores for varying number of training epochs. Best values achieved emboldened (max or min depending on metric).

| epochs | FIS           | PC            | SRA        | $p_{MSE}$     | $r_{MAE}$     | $MI_{MAE}$    | HR | $DCR_{AVG}$   | $DCR_{STD}$   |
|--------|---------------|---------------|------------|---------------|---------------|---------------|----|---------------|---------------|
| 100    | 0.4953        | 0.7587        | 0.4        | 0.0812        | 0.1748        | 0.1084        | 0  | <b>0.1344</b> | 0.0289        |
| 500    | <b>0.9092</b> | <b>0.8948</b> | 0.5        | 0.0204        | 0.0578        | 0.0506        | 0  | 0.1054        | 0.0205        |
| (1000) | 0.5381        | 0.8722        | <b>0.9</b> | 0.0142        | <b>0.0368</b> | 0.0455        | 0  | 0.1018        | 0.0176        |
| 2000   | 0.7465        | 0.8321        | 0.4        | <b>0.0104</b> | 0.0634        | 0.0368        | 0  | 0.0972        | 0.0179        |
| 5000   | 0.8175        | 0.8766        | 0.8        | 0.0537        | 0.0462        | 0.0328        | 0  | 0.1002        | <b>0.0163</b> |
| 10000  | 0.6672        | 0.8801        | 0.4        | 0.0224        | 0.0533        | <b>0.0299</b> | 0  | 0.0968        | 0.0180        |
| 20000  | 0.6787        | 0.8853        | 0.6        | 0.1079        | 0.0700        | <b>0.0299</b> | 0  | 0.0846        | 0.0166        |

128 showed overall best performance when minimising quality loss and maximising utility. Although it did not achieve the highest DCR, this was to be expected as it achieves better quality (i.e. closer imitation of real data). Nonetheless, the mean distance and HR suggest it did not replicate any notification-context pairings in the data set and therefore did, to an extent, preserve privacy (this is also shown in the following sections when compared to the random baseline).

Table 4.6 Utility, quality and privacy scores of WGAN-GP model trained using varying amount of real samples as seed. Best values achieved emboldened (max or min depending on metric).

| $N$   | FIS           | PC            | SRA        | $p_{MSE}$     | $r_{MAE}$     | $MI_{MAE}$    | HR | $DCR_{AVG}$   | $DCR_{STD}$   |
|-------|---------------|---------------|------------|---------------|---------------|---------------|----|---------------|---------------|
| 100   | 0.5452        | 0.7556        | 0.6        | 1.0000        | 0.1213        | 0.1456        | 0  | <b>0.1320</b> | 0.0232        |
| 1000  | 0.7069        | 0.8073        | 0.6        | <b>0.0247</b> | 0.0837        | 0.0542        | 0  | 0.0929        | 0.0186        |
| 5000  | 0.7073        | 0.8735        | <b>0.9</b> | 0.0274        | 0.0739        | 0.0382        | 0  | 0.0887        | 0.0172        |
| 10000 | <b>0.7394</b> | <b>0.8839</b> | 0.8        | 0.0250        | <b>0.0375</b> | <b>0.0325</b> | 0  | 0.0860        | <b>0.0136</b> |

Figure 4.5 depicts the loss of the generator and critic of the WGAN-GP as they are trained adversarially over 500 and 2000 epochs respectively. The loss of both networks converge indicating the generator creates increasingly better samples as the critic learns to improve approximating the Wasserstein distance between real and generated samples. After convergence, the loss of both networks remains relatively stable suggesting both networks have reached an impasse in their minimax game. Table 4.5 illustrates the scores achieved by the synthetic data generated by the WGAN-GP model for varying numbers of training epochs. The FIS, PC and  $DCR_{AVG}$  scores peak after just 500 epochs, followed by SRA and  $r_{MAE}$  after 1000.  $MI_{MAE}$  is the only metric that constantly decreases as the number of training epochs increase, suggesting that it benefits most from the extended training time, whilst the other metrics do not improve consistently. SRA for instance, fluctuates after 1000 epochs and  $p_{MSE}$  is significantly worse when trained using 20,000 epochs compared to 1000. Taking this analysis into consideration, the synthetic data generated by the WGAN-GP was subsequently compared with notifications and contexts generated by the WGAN and CWGAN-GP models with a fixed number of training epochs, latent dimension and batch size of 1000, 64, 128 respectively.

The available sample size was also an important factor for consideration in this work. Identifying a minimal threshold of real seed data with which synthetic data can be generated to an adequate standard of quality, utility and privacy for the task of managing notification engagement is important for ensuring minimal intrusion and exploitation of subscriber data. It is particularly relevant to Article 5, 1c of the General Data Protection Regulation [Reg16] which states: *Personal data shall be adequate, relevant and limited to what is necessary in relation to the purposes for which they are processed ('data minimisation')*. Subsequently, the WGAN-GP was trained using varying amounts of real notification engagement data and evaluation metrics were extracted from the ensuing synthetic data generated.

Table 4.6 illustrates the scores attributed to synthetic data generated by the model when trained with 100 (1%), 1000 (10%), 5000 (50%) and 10,000 (100%) samples respectively and compared with 10,000 samples of real notification engagement data. The effect of sample size also informs the deployment of generative models in-the-wild in scenarios where archived engagement data is not available (i.e. a cold-start problem) as it describes how quickly the model learns to produce adequate samples. Assuming a mobile subscriber receives  $\approx 60$  notifications per day, 100 samples could be collected to seed the model within 2 days and 1000 samples in  $\approx 2$  weeks. The PC metric recorded in Table 4.6 illustrates that the WGAN-GP was able to achieve  $\approx 0.8$  Predictive Capability when trained on just 2 weeks of notification engagement data. The FIS, PC,  $r_{MAE}$  and  $MI_{MAE}$  scores were naturally best when maximal data was provided as seed, but some of the largest improvements were also identified between cases of 100 and 1000 seed samples used suggesting that generative models could be called to action after an initial 2 day - 2 week period of notification engagement observation.

The previous analyses assessed the performance of the WGAN-GP model using different latent dimension, batch size and training epoch settings as well as under varying states of data availability. The synthetic data discussed in the following analysis was generated using fixed latent dimension, batch size, training epoch and data availability values: 64; 128; 2000; and 100%; respectively, in order to

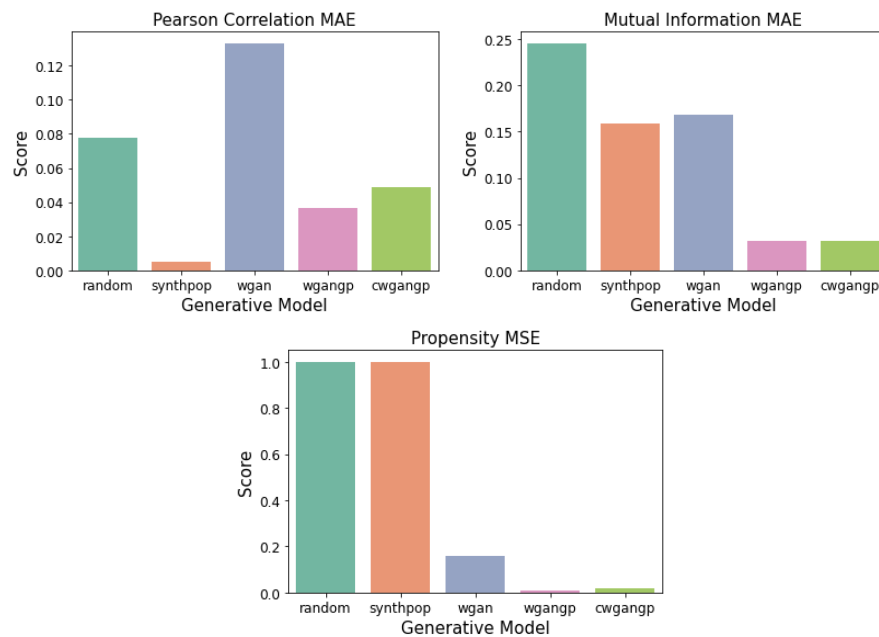


Fig. 4.6 Comparison of synthetic data quality scores achieved by a selection of generative models (lowest scores indicate better synthetic data quality).

facilitate comparative evaluation between the selection of models introduced in subsection 4.3.1.

Figure 4.6 illustrates the scores achieved by the selected models for the three metrics measuring synthetic data quality. As the synthetic notification data gets closer to resembling its real counterpart, the quality scores approach 0. The synthpop model was best at generating continuous data (i.e. it had the lowest pearson correlation MAE score) which closely resembled real distributions. The WGAN-GP and CWGAN-GP models were slightly less effective but still achieved an improvement over the random baseline. The WGAN model achieved worst results, indicating the continuous values it generated were inconsistent with the real notification-context data e.g. it could not accurately model the number of times a notification was updated or the amount of time elapsed between app use and a notification delivery. The WGAN-GP and CWGAN-GP models were best at preserving the associations present within discrete notification-context features as suggested by achieving lowest  $MI_{MAE}$  scores. Additionally, the synthetic data generated by both models was very closely indistinguishable from real data by a classifier as indicated by  $p_{MSE}$  scores approaching 0.

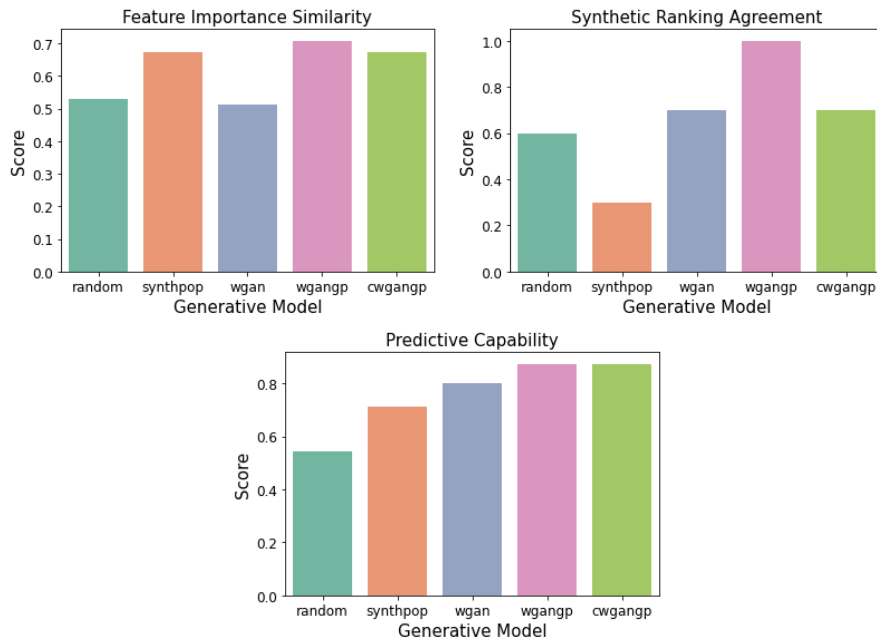


Fig. 4.7 Comparison of synthetic data utility scores achieved by a selection of generative models (higher scores indicate better synthetic data quality).

Figure 4.7 illustrates the scores achieved by the selected models for the three metrics measuring synthetic data utility. Higher utility scores indicate intelligent agents behave consistently when trained upon synthetic or real notification data



respectively. The WGAN-GP and CWGAN-GP models score highest in Feature Importance Similarity, followed closely by synthpop. The highest scores approximated 0.7 indicating that the agents trained on synthetic data identified features to be of similar importance to those found by agents trained on real notification data, albeit with room for improvement. Implications of using synthetic data with low FIS scores could be that features actually important to notification engagements are mistakenly discarded due to the inability of the generative model to replicate their importance in the synthetic data. The WGAN-GP achieved the highest score for Synthetic Ranking Agreement which indicates the battery of algorithms tasked to predict notification engagement actions were ranked identically, by performance, when trained on synthetic and real data respectively. The data generated by the synthpop model fell below the random baseline for this metric indicating that good replication of the continuous features alone, which the model excelled at, was not sufficient for the task of predicting notification engagement action which places equally high importance on discrete features, such as the app producing the notification. The WGAN-GP and CWGAN-GP models were also strongest for Predictive Capability, with scores of over 0.8 indicating that, when trained on synthetic data alone and tasked with predicting real engagement actions, intelligent agents could achieve performance up to 80% achieved when trained on real data. Tables 4.3 and 4.5 illustrate that this value can approach 0.9 with different batch size, latent dimension and training epoch parameters.

Figure 4.8 illustrates the scores achieved by the selected models for privacy metrics measuring the degree to which real samples were exposed within the synthetic data. The synthpop model had a slightly elevated Hit Rate compared to the GAN models and random baseline, however the number of explicitly exposed samples was relatively small (14 of 10,000 notifications exposed). The Distance to Closest Record metric illustrates a much richer state of potential exposure as it

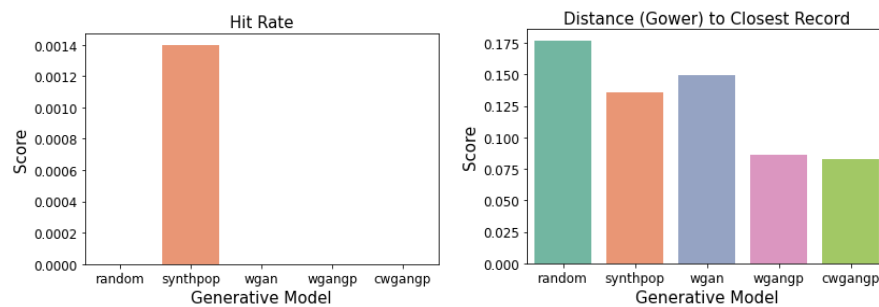


Fig. 4.8 Comparison of synthetic data privacy scores achieved by a selection of generative models.

describes how similar the synthetic notifications are to the real notifications, even if not explicitly duplicated - this is significant as notifications which may have a slightly different continuous feature value ( $\pm 1$  update for instance) would not be represented in the HR metric but still pose a risk of exposure). The random baseline, in this scenario, provides a benchmark for data which does not expose any (or very few) real notification engagement associations and relationships, thus the synthetic data generated by this model can be considered least at risk of exposure. Despite wishing to reduce the risk of exposure, as synthetic samples with a high DCR are less similar to real samples, they have a negative impact on quality and subsequent utility. Therefore optimal DCR values are suggested to be between the random baseline and 0. The WGAN-GP and CWGAN-GP strike the best balance of reducing exposure risk whilst maintaining quality and utility as their DCR scores lie midway between the random baseline and 0. The synthpop and WGAN models lie closer to the random baseline, suggesting less risk of exposure, but also decreased quality and utility which could be seen in Figures 4.6 and 4.7 respectively. Models implementing Differential Privacy facilitate throttling privacy and utility to achieve desired levels of both [DR+14].

Based on the analysis of the quality, utility and privacy metric results discussed in this section, the WGAN-GP stands out as the strongest model for generating synthetic notification-context engagements compared with a state-of-the-art alternative (synthpop) and a random baseline. The CWGAN-GP, whilst not generating synthetic data as consistently as the WGAN-GP (e.g. lower SRA and higher  $r_{MAE}$  scores), is highlighted as the next best alternative with the added incentive of facilitating conditional generation.

Whilst this section has discussed the performance of the selected generative models when creating synthetic push-notifications, the proposed metrics, method of evaluation and implementations of the selected generative models could also be applied within other domains to create other synthetic data sets. Therefore the proposed process could be defined as *domain agnostic*. Indeed, GANs have been applied in many domains for generating structured, time-series, image and video data *ganzoo*. The metrics proposed in this work facilitate evaluating the synthetic data generated under three key components equally: quality, utility and privacy. However, depending on the domain, additional metrics or methods of weighting could be used to evaluate whether the data is fit-for-purpose (e.g. domains which are not concerned with privacy might weight the utility and quality metrics higher than privacy). In addition, the use of the CWGAN could be of particular use in domains which require specific types of data points to be

generated on request. With respect to the notification domain, an example would be generating a notification fit for a given context. An example within the domain of finance could be generating a sequence of banking events that are classified as fraud, thus facilitating the balancing of a highly skewed data set (assuming fraudulent events occur less frequently than legitimate banking events).

### 4.3.4 Gym-push

An example use case for the generative approach discussed in this chapter was leveraging it for the creation of synthetic data sets to facilitate the sharing of push-notification data with the wider research community. There are a number of benefits associated with this, such as enabling comparative evaluation and a standard set of features and benchmarks with which NMSs can be trained. But it also facilitates transfer learning in which agents applied to the problems associated with push-notifications may aid other domains associated with interruptibility management and context computing. An open-source, custom, OpenAI Gym [Bro+16] environment, *gym-push*<sup>1</sup>, was created and seeded with synthetic notification data to facilitate comparative evaluations of User Models during the EvalUMAP Workshop series [Con+19]. The generative models and data sets were also released<sup>2</sup> to support further research in the area of NM.

## 4.4 Evaluating Synthetic Notifications In-the-wild

The previous sections introduced and evaluated an approach for generating synthetic push-notification data which could facilitate the training of intelligent agents without exposing subscribers to the same risks associated with fully disclosing their real data. This section further evaluates the utility of the generative approach by describing the deployment of the WGAN-GP and CWGAN-GP models, introduced and evaluated in the previous section, in-the-wild to enhance the delivery of web push-notifications. During this experiment, requirements, **R3** and **R4**, of the ETHOS-Push approach were implemented and evaluated providing empirical evidence to inform **RQ2** and satisfy objectives, **RO3** and **RO4**.

<sup>1</sup><https://github.com/kieranfraser/gym-push>

<sup>2</sup><https://ethos-push.github.io/>

### 4.4.1 Experiment Approach

Chapter 3 briefly introduced the Pushd study in which subscribers were pushed a variety of web push-notifications over a period of two weeks. During the first 10 days of the study, notifications were pushed from a variety of sources, presented in different formats and delivered at random times of the day. This facilitated an initial explorative search of subscriber preferences toward web push-notifications in differing contexts and provided an accumulation of seed data with which to train intelligent agents to optimise notification management tasks. During the final 4 days of the study, 4 agents were tasked with deciding, for a given day, the time (hour) and presentation template with which to deliver and present web-push notifications to subscribers. The presentation templates, in this context, are the differing formats in which notifications were displayed to participants (e.g. plain text, with contextual emojis or summarised content). The presentation templates will be discussed further in Chapter 5. The four agents, illustrated in Figure 4.9, were defined as follows:

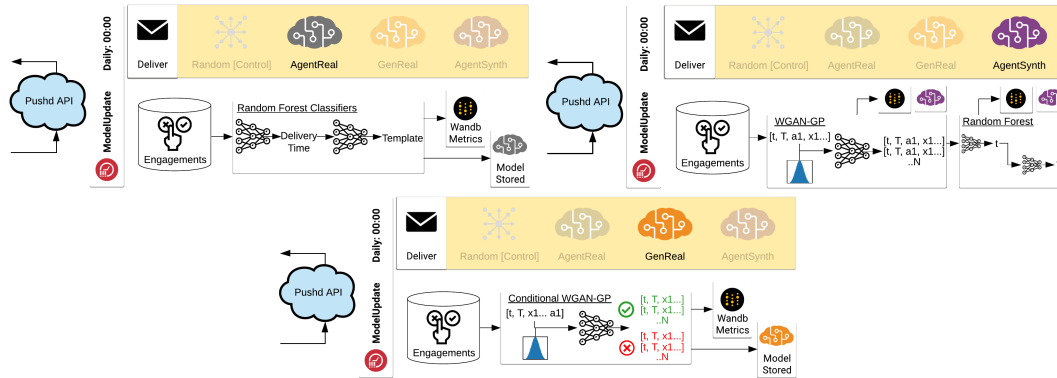


Fig. 4.9 Intelligent agents tasked with managing notification delivery.

- *Random* - this agent randomly selected an hour of the day and presentation format to deliver the notification.
- *AgentReal* - this agent was created using two Random Forest multi-class classifiers (one each for time and presentation template), trained on the previous 10 days notification engagements and tasked with predicting an optimal time and template to deliver and present notifications such that they would be opened. The inputs ( $x$ ) to the first model, predicting the time of day, were the features:

{ category, website domain, enticement, sentiment, topic, day of week, action  
}

The inferred preferred delivery time was appended to  $x$  and used as input to the second classifier for predicting the template with which notifications would be presented.

- *AgentSynth* - this agent was identical to *AgentReal* but for the data used to train the classifiers which was synthetic and generated using a WGAN-GP trained upon the previous 10 days of notification engagements. The WGAN-GP took a vector comprised of notification time ( $t$ ), template ( $T$ ), action ( $a$ ) and notification features ( $x$ ) and generated 2000 synthetic notifications to be used for training the classifiers instead of using real engagement data.
- *GenReal* - this agent was created using a conditional generative model, CWGAN-GP, and was trained upon the previous 10 days of notification engagements. At inference, it was fed a conditional value for action ( $a$ ), *opened*, thus ensuring the data generated would be associated with a positive engagement. The output of the generative model included times and presentation templates inferred to be associated with positive engagements. These times and templates were randomly sampled to provide an inferred preferred delivery time and template for presentation.

#### 4.4.2 Results

13,718 notifications and their associated engagements in varying contexts were accumulated during the 10 days preceding the deployment of the intelligent agents. During the four days in which the agents were deployed, they were responsible for delivering 5,939 notifications to 83 participants. 4,013 notifications were successfully delivered to the participant devices and 3,276 were explicitly engaged with (opened/dismissed). The number of participants assigned to the *Random*, *AgentReal*, *AgentSynth* and *GenReal* agents respectively were: 25, 22, 15, 21. Although 91 participants were engaged with the study (as referenced in Table 3.12), it was found that 5 did not engage during the final 4 days when the delivery agents were deployed and thus were removed from the analysis. Additionally, it was found that 3 further participants had response times under 10 seconds for all notifications, which suggested potential autonomous interaction, thus were

also removed. The following sections discuss quantitative and qualitative results regarding the quality of the synthetic data generated in-the-wild, the impact of differing delivery approaches and the perceptions of subscribers toward both the synthetic data and delivery approach.

#### 4.4.2.1 Synthetic Data In-the-wild

Table 4.7 displays the quality, utility and privacy metrics of synthetic data generated for each participant of the *AgentSynth* group. As each feature generated was discrete, Pearson’s correlation ( $r_{MAE}$ ) metric evaluating the quality of continuous feature generation, was excluded. Additionally, a number of scores could not be calculated for some participants with low engagements due to class imbalance, which highlights a limitation of the metrics and their application in environments with rare occurrences of target features, in this case: positive notification engagements.

Over the course of the initial 10 days of the study, each participant had engaged with, on average, 106 notifications. The Predictive Capability average score indicates that the synthetic data being generated at that point, was 87% as useful for the classification agents as the real data. The Feature Importance Similarity and Synthetic Ranking Agreement average scores, approximating 0.6, suggest that some features relevant to the agents trained on the real data were not as well represented in the synthetic version, thus potentially reducing its utility. Additionally, the Propensity Mean Squared Error score indicates that the synthetic data was distinguishable when a classifier was applied. The following section will expand on this point further. The Mutual Information score suggests that, although lower scores were achieved when trained on larger amounts of data (see Figure 4.6), the model was nevertheless capable of replicating associations from the smaller web push-notification engagement data sets, as demonstrated in Figure 4.10. The Hitting Rate flagged that 22 engagement samples extracted from participant 2 also appeared in their 2000 sample synthetic data set, which was a relatively small exposure and could be attributed to the low number of subscribed topics and websites which would increase the potential for repeated permutations of feature values. The Distance from Closest Record average and standard deviation values also indicate that the synthetic data generated is not close enough to that of the real data to suggest high risk to the privacy of participant data.

In addition to calculating the quality, utility and privacy metrics for evaluating the synthetic data generated by the WGAN-GP in-the-wild, participants were

Table 4.7 Quality, Utility and Privacy metrics of synthetic data generated during experiment.

| user | FIS   | PC    | SRA  | $p_{MSE}$ | $MI_{MAE}$ | HR    | $DCR_{AVG}$ | $DCR_{STD}$ |
|------|-------|-------|------|-----------|------------|-------|-------------|-------------|
| 1    | n/a   | n/a   | n/a  | 1         | 0.175      | 0.000 | 0.127       | 0.031       |
| 2    | n/a   | n/a   | n/a  | 1         | 0.223      | 0.011 | 0.109       | 0.034       |
| 3    | n/a   | n/a   | n/a  | 1         | 0.234      | 0.000 | 0.113       | 0.034       |
| 4    | 0.553 | 0.726 | 0.4  | 1         | 0.224      | 0.000 | 0.175       | 0.034       |
| 5    | 0.598 | 0.885 | 0.8  | 1         | 0.220      | 0.000 | 0.128       | 0.034       |
| 6    | 0.652 | 0.732 | 0.8  | 1         | 0.203      | 0.000 | 0.118       | 0.033       |
| 7    | 0.742 | 0.726 | 0.9  | 1         | 0.186      | 0.000 | 0.149       | 0.039       |
| 8    | 0.584 | 0.821 | 0.7  | 1         | 0.340      | 0.000 | 0.127       | 0.037       |
| 9    | n/a   | n/a   | n/a  | 1         | 0.197      | 0.000 | 0.133       | 0.040       |
| 10   | 0.711 | 0.812 | 0.9  | 1         | 0.205      | 0.000 | 0.120       | 0.030       |
| 11   | 0.652 | 0.634 | 0.1  | 1         | 0.269      | 0.000 | 0.122       | 0.038       |
| 12   | 0.602 | 1.065 | 0.4  | 1         | 0.229      | 0.000 | 0.120       | 0.033       |
| 13   | n/a   | n/a   | n/a  | 1         | 0.232      | 0.000 | 0.093       | 0.034       |
| 14   | 0.621 | 1.489 | 0.4  | 1         | 0.267      | 0.000 | 0.134       | 0.045       |
| 15   | 0.648 | 0.819 | 0.7  | 1         | 0.228      | 0.000 | 0.108       | 0.034       |
| AVG  | 0.63  | 0.87  | 0.61 | 1         | 0.229      | 0     | 0.125       | 0.035       |
| STD  | 0.05  | 0.25  | 0.27 | 0         | 0.04       | 0.002 | 0.02        | 0.004       |

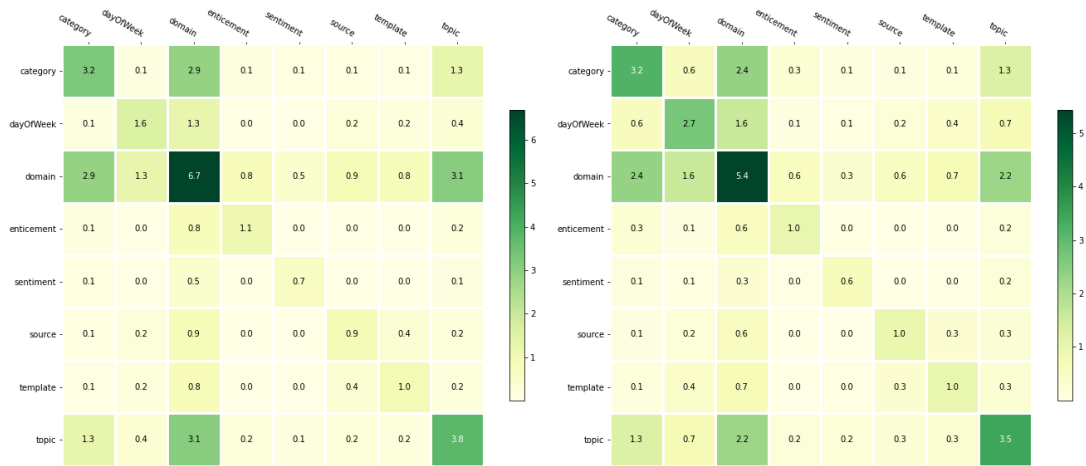


Fig. 4.10 Mutual information between real notification engagement features (left) and synthetic notification engagement features (right)

also tasked with distinguishing between their real engagement data and synthetic engagement data generated by the model. Visualisations of real data and synthetic data were placed side by side and participants were asked to select which one they thought most resembled their own notification engagement actions. Eight visualisations were shown in total, two each (A and B) for four notification features: topic, website, day and time. Each visualisation was a sorted bar graph showing the top topics and websites associated with notifications they opened or the top

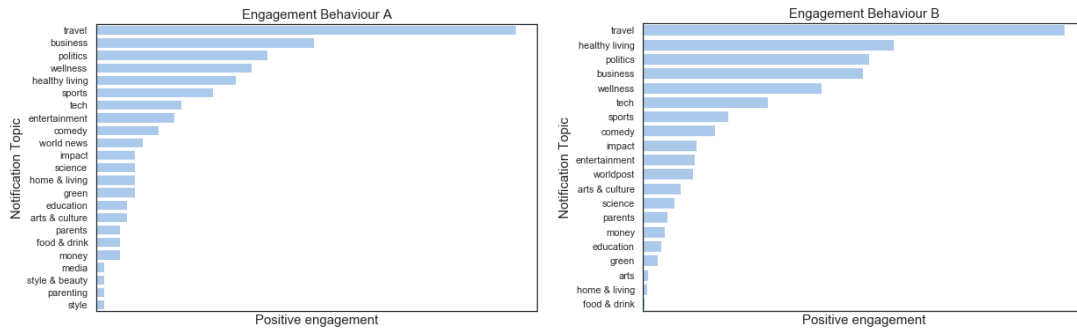


Fig. 4.11 Visualisations shown to participants when discerning real from synthetic data

days and times they opened them (Figure 4.11 illustrates visualisations presented for discerning real from synthetic data using the notification topic feature). The hypothesis of this experiment was that if participants themselves could not discern their own engagement behaviour from the synthetic data generated, then the generative model was successfully creating realistic samples. It would also indicate that, if a participant could not discern a difference, any discrepancies between the real and synthetic data were actually behaviours the participant perceived as feasible, thus suggesting the synthetic data provided additional useful information not necessarily captured within the real data. Synthetic data was only generated for participants of the *AgentSynth* group, therefore, participants from the remaining three groups were shown two visualisations which were both created using their real engagement data, with the expected result being participants would select visualisation A over visualisation B, approximately 50% of the time. Table 4.9 displays the results from a chi squared goodness-of-fit test evaluating the proportion of correct guesses for each feature over the differing scenarios. The first two tests assumed the proportion of participants correctly distinguishing real data would be 50% i.e. the synthetic data was indistinguishable from real data. The final test assumed an unequal proportion, which instead of assuming a proportion of 50% for both correct and incorrect guesses, assumed a proportion equal to that found when participants of the Real vs Real group were guessing between their two samples of real data. This test was included to evaluate the synthetic data proportions in cases where responses from the Real vs Real group deviated from equal proportions. Significant values (taken as  $p < 0.05$ ) indicate that the number of correct responses deviated from the assumed proportions, which in turn suggests that participants tended to perceive one visualisation as more representative of their notification engagement data than the other, or



the inverse. The topic and time features did not inspire statistically significant deviations from the assumed proportions, suggesting that participants in both groups could not distinguish which visualisation was more representative of their notification engagements. Participant answers did deviate from the assumed equal proportions when distinguishing between website engagements for both the Real vs Real and Synth vs Real groups. Table 4.8 confirms that participants were better able to distinguish the correct answer more frequently in both cases. As the  $p$  value for the Synth vs Real (unequal proportions) does not suggest a statistically significant difference between the responses in the Synth vs Real group and the Real vs Real group, this suggests that it may not have been the quality of the synthetic data that caused this result, but rather some confounding factor within the website feature.

The absence of significant differences in proportions does not necessarily imply that the synthetic data was indistinguishable from the real data. Therefore, the odds ratio was also used to measure whether the presentation of Synth vs Real engagement data increased the probability of correctly identifying real engagements as opposed to the scenario in which Real vs Real engagement data presented. If the synthetic data is indistinguishable from real data, the odds ratio should approximate 1. Values greater than 1 suggest that participants were more likely to distinguish real from synthetic data than the participants of the Real vs Real group suggesting that quality of the synthetic data for that feature was not as representative of the real data on which the WGAN-GP model was trained. Values less than 1 indicate that participants were less likely to distinguish between synthetic and real data than the Real vs Real group. The odds ratio for both the day and time features approximated 1, suggesting quality synthetic data generated. The odds ratio for the topic feature was less than one, also suggesting quality synthetic data. The odds ratio for the website feature was 3.86 however, suggesting that for this feature, participants of the Synth vs Real group were more likely to correctly distinguish between real and synthetic data, which correlates with the chi squared goodness-of-fit analysis.

#### 4.4.2.2 Delivery Agents In-the-wild

Table 4.10 illustrates how participants of the study engaged with notifications when delivery times and presentation templates were selected by the intelligent agents deployed in-the-wild. Notifications delivered by the *GenReal* agent, which comprised the WGAN-GP trained on real notification engagement data of the par-

Table 4.8 The odds ratio of correctly distinguishing real notification engagement data.

| Feature | Response<br>(* = correct) | Data shown    |              | Odds Ratio | 95% CI         |
|---------|---------------------------|---------------|--------------|------------|----------------|
|         |                           | Synth vs Real | Real vs Real |            |                |
| topic   | A                         | 13            | 31           | 0.417      | 0.149 - 1.171  |
|         | B*                        | 7             | 40           |            |                |
| website | A*                        | 19            | 59           | 3.864      | 0.471 - 31.699 |
|         | B                         | 1             | 12           |            |                |
| day     | A                         | 12            | 44           | 1.086      | 0.394 - 2.997  |
|         | B*                        | 8             | 27           |            |                |
| time    | A                         | 11            | 39           | 0.997      | 0.368 - 2.703  |
|         | B*                        | 9             | 32           |            |                |

Table 4.9 Chi squared goodness-of-fit test evaluating proportion of participants correctly distinguishing real from synthetic data.

| Feature | Real vs Real<br>(equal prop.) |       | Synth vs Real<br>(equal prop.) |       | Synth vs Real<br>(real prop.) |       |
|---------|-------------------------------|-------|--------------------------------|-------|-------------------------------|-------|
|         | $\chi^2(1)$                   | $p$   | $\chi^2(1)$                    | $p$   | $\chi^2(1)$                   | $p$   |
| topic   | 1.14                          | 0.285 | 1.8                            | 0.18  | 3.58                          | 0.058 |
| website | 31.11                         | <.001 | 16.2                           | <.001 | 2.04                          | 0.153 |
| day     | 4.070                         | 0.044 | 0.8                            | 0.371 | 4.109                         | 0.043 |
| time    | 0.69                          | 0.406 | 0.2                            | 0.655 | 0.81                          | 0.369 |

ticipants, had the highest average (10%) and maximum (87%) engagement values across all agents. This was followed closely by the *Random* and *AgentSynth*, which showed little differentiation in CTR (7-8%). The agent trained on real notification engagement data was the least effective agent (4%). The engagement results could be further dissected by evaluating engagements based on whether participants had modified any notification settings on their device (a question asked during the pre-study questionnaire). This is an important aspect to study, as intelligent delivery agents or NMSs will generally have to contend with modifications and custom device-side rules set by users or external interventions by other agents with differing agendas i.e. a multi-agent system. Therefore, delivery agents tasked with finding optimal and empathetic delivery moments will need to consider both the subscriber's preferences and the potential impact of external agents. The *GenReal* agent incited the highest CTR (15%) of notifications pushed at participants with default settings enabled. This was followed by *AgentSynth* (11%), a result which contributes positively toward the case for using synthetic data to train NMS

agents deployed in-the-wild. Little differentiation was found between the *Random* and *AgentReal* agents. In contrast, engagement levels were much reduced when custom notification settings were set by participants. *AgentSynth* delivered the highest number of notifications which were subsequently opened (3%) followed by the *Random* and *GenReal* agents, however, there was little to distinguish between agents.

The relatively high standard deviation across all scenarios also indicates a high level of CTR variability between participants of the 4 agent groups, suggesting that the delivery agents do not perform equally well for all participants thus indicating that other influencing factors could be at play. This demonstrates the challenge of evaluating notifications in-the-wild as highlighted by [MM20a; Chu+15]. Additionally, whilst CTR can be a good indicator of a successful push, it does not necessarily indicate an agent was successful at delivering notifications at optimal or empathetic moments. Successfully extracting a metric to track this is difficult without constantly querying participants after each push. The WeAreUs study, described in Chapter 3, implemented an ESM approach to understand user perceptions toward a notification at the moment of delivery - however this approach can be taxing on participants [Meh+15a] and result in few engagements. In this study, a Technology Assisted Recall (TAR) [Kar20] approach was taken

Table 4.10 Web push-notification engagement during deployment of intelligent delivery agents.

| Group      | All<br>(n=3751) |         |     |     |     |
|------------|-----------------|---------|-----|-----|-----|
|            | Num.<br>Users   | CTR (%) |     |     |     |
|            |                 | mean    | min | max | std |
| Random     | 25              | 8       | 0   | 65  | 15  |
| AgentReal  | 21              | 4       | 0   | 38  | 9   |
| AgentSynth | 15              | 7       | 0   | 55  | 15  |
| GenReal    | 21              | 10      | 0   | 87  | 21  |

| Group      | Default Settings<br>(n=2607) |         |     |     |     | Custom Settings<br>(n=1144) |         |     |      |     |
|------------|------------------------------|---------|-----|-----|-----|-----------------------------|---------|-----|------|-----|
|            | Num.<br>Users                | CTR (%) |     |     |     | Num.<br>Users               | CTR (%) |     |      |     |
|            |                              | mean    | min | max | std |                             | mean    | min | max  | std |
| Random     | 21                           | 9       | 0   | 65  | 17  | 4                           | 2       | 0   | 9.1  | 7   |
| AgentReal  | 11                           | 8       | 0   | 38  | 12  | 10                          | 0       | 0   | 15.4 | 1   |
| AgentSynth | 8                            | 11      | 0   | 55  | 20  | 7                           | 3       | 0   | 14.3 | 13  |
| GenReal    | 13                           | 15      | 0   | 87  | 25  | 8                           | 2       | 0   | 20.1 | 9   |

Table 4.11 Change in Net Promoter Score when intelligent agents deployed.

| Group      | $\Delta$ NPS |                  |                 |
|------------|--------------|------------------|-----------------|
|            | All Settings | Default Settings | Custom Settings |
| Random     | +0.29        | +0.25            | +0.29           |
| AgentReal  | +0.25        | -0.19            | +0.69           |
| AgentSynth | <b>+1.46</b> | <b>+1.43</b>     | <b>+1.5</b>     |
| GenReal    | -0.28        | -1.27            | -0.28           |

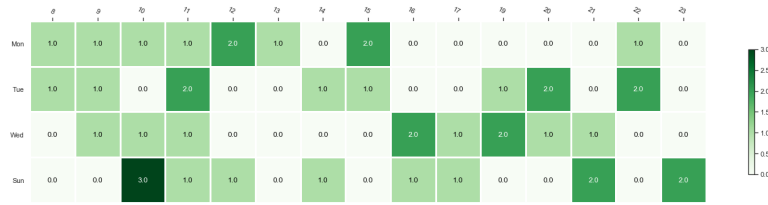
Table 4.12 Evaluating agents trained on synthetic vs real notification engagement data.

| Group      | Model        | users | Avg. F1-score | Mean Rank | $U$ | $z$    | $p$   |
|------------|--------------|-------|---------------|-----------|-----|--------|-------|
| AgentReal  | presentation | 21    | 0.974         | 25.48     | 11  | -4.741 | <.001 |
| AgentSynth | presentation | 15    | 0.873         | 8.73      |     |        |       |
| AgentReal  | deliveryTime | 21    | 0.194         | 13.9      | 269 | 3.578  | <.001 |
| AgentSynth | deliveryTime | 15    | 0.362         | 25.93     |     |        |       |

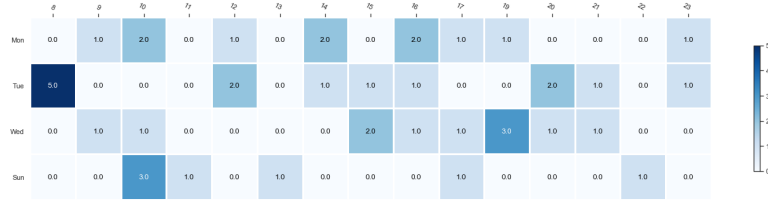
aimed to minimise the number of times a participant was interrupted and queried for feedback. A nightly questionnaire was delivered to participants in which they could submit a Net Promoter Score (NPS) representing how likely they were to recommend the notification subscription service to a friend based on their experience of notifications received that particular day. To offset recall bias, the notifications a participant received during the day were displayed along with the action they chose to take (open/dismiss) and the time taken. Table 4.11 illustrates the change in NPS from the initial 10 day period with random delivery time selection, to the 4 day period in which the intelligent delivery agents were deployed. The slight increase of NPS elicited by the *Random* agent can be considered the baseline, as participants within the group did not benefit from any change in delivery method. Thus *AgentReal*, despite the small increase in NPS, is considered to have had little impact on improving the notification experience for participants. In contrast, *AgentSynth* significantly increased the NPS, thus demonstrating that, paired with increased CTR, the agent was successful in delivering an improved notification experience to participants. Surprisingly, the *GenReal* agent had the opposite effect, despite achieving the highest CTR engagement - perhaps once again demonstrating that CTR alone is not completely sufficient for capturing positive notification experiences.

With respect to evaluating the approach of training a NMS with synthetic notification engagement data as opposed to real data, Table 4.12 demonstrates

that *AgentSynth* was, on average,  $\approx 10\%$  less effective than *AgentReal* in the task of inferring the optimal presentation template, as demonstrated by the F1-score achieved by both models, but was significantly more effective when tasked with identifying the optimal delivery time. Figures 4.12a and 4.12b show a comparison between moments a participant decided to open notifications (during the 10 day period preceding agent deployment) and moments which *AgentSynth* decided to push notifications to the participant, respectively. Wednesday in particular, demonstrates how the agent learned the participants delivery preferences, albeit with room for improvement if given more days to learn.



(a) Moments participant opened notifications (preceding agent deployment)



(b) Moments *AgentSynth* pushed notifications (during agent deployment)

## 4.5 Limitations

Limitations of the work described in this chapter include a deeper evaluation of notification features, and in particular text features, which could be utilised for improved management of notifications as well as a broader evaluation of applicable agents for NM tasks. The following two chapters attempt to address this by exploring and deriving additional features for improved inference of notification actions. A limitation also exists in the approach toward generating and evaluating synthetic notification data sets, as explicit text content was excluded from the analysis. Chapter 6 attempts to address this by implementing and evaluating a generative model capable of creating synthetic notification text content. Finally,

the relatively short period of time in which notification delivery agents were deployed and evaluated is also a limitation of this work, however, engaging participants over longer periods of time, within which they are regularly interrupted (as with push-notification studies), is recognised as an ongoing challenge in this research domain [Meh+15a].

## 4.6 Summary

This chapter introduced a generative approach which could create synthetic notification data of adequate quality, utility and privacy for subsequent use in training intelligent agents tasked with finding empathetic moments for notification delivery and optimal methods of notification presentation. The identification of notification features significant for predicting future action contributed to fulfilling **RO2**, and provided evidence for **RQ1**, as it facilitated further understanding of notification engagement behaviour. The proposed generative approach contributed to **RO3** by promoting sustainable methods for empathetic, data-driven notification management. The ensuing evaluation of the generative approach also provided empirical evidence for **RQ2** as did the deployment of the generative models in-the-wild. The deployment of intelligent agents orchestrating the delivery of web push-notifications in-the-wild also contributed to **RO4**. The methods outlined in this chapter heeded **R2** and **R3** of the ETHOS-Push approach, demonstrating in part, the value of an empathetic approach toward push-notification management.

# Chapter 5

## Scrutable Push-notifications

There is no expedient to which a man will not resort to avoid the real labor of thinking.

---

*Sir Joshua Reynolds*

### 5.1 Overview

In today's attention economy, an abundance of information is pushed from every direction and device, seeking to constantly engage regardless of the emotional state or health of those targeted. Push-notifications are an example of a powerful design tool used to persuade engagement [Oul+12]. The intent of a push-notification is to add timely value, however due to their inherent design and unintelligent management, notifications contribute toward adverse smartphone use and behaviours resulting in poor digital health (e.g. NoMoPhobia [Kin+13] and FOMO [Elh+16]). There is a natural conflict between the apps pushing and subscribers receiving notifications. Those behind the notification push are usually seeking to increase and improve engagement on their platforms by nudging subscribers. Subscribers have their own goals and aspirations in life which may not align with spending additional time, money or effort on a particular platform or service. When notification creation, presentation and delivery is done right, both the *pusher* and *receiver* of notifications benefit in some way e.g. the subscriber receives a push

they find serendipitous and spends time engaging with the pushing app, perhaps resulting in a purchase, thus both the app and subscriber extract value from the push event. However, more frequently when the push goes wrong, it leads to annoyance, unsubscribes or poorer relationships with technology e.g. the push is dismissed by the subscriber leading the *pusher* to resort to increasingly persuasive methods to entice engagement, particularly when competing within a crowded space i.e. with other app notifications (enter the *attention economy*).

This chapter explores an approach of extracting well-defined persuasive principles from push-notifications to facilitate the enlightenment of subscribers, promote awareness of notification engagement behaviour and contribute toward improving the prediction capability of NMSs. This work was motivated by **RO2** and **RO4** as it facilitates greater understanding of the influence of push-notifications and evaluates empathetic mechanisms/tools for the subscriber, thus also informing **RQ3** and **RQ4**. The methods proposed and evaluated in this chapter also help satisfy **R3** and **R4** of the ETHOS-Push approach. As such, the key contributions of this chapter are:

1. a method of feature engineering scrutable, persuasive push-notification features, based on well defined principles of persuasion
2. an evaluation of the scrutable features in a simulated environment demonstrating predictive power and potential (albeit with caveats) for autonomous notification creation
3. an evaluation of scrutable features in-the-wild and their utility for aiding subscriber understanding of push-notification engagement and potential exploitation
4. a method and evaluation of autonomous formation of push-notification content

## 5.2 Engineering Persuasive & Scrutable Features of Push-notifications

In a recent survey [OM18] of Persuasive Technology (PT), *health and wellness* was identified as a domain in which emerging trends of PT were particularly apparent. The study found smartphone and push-notification technology were



used in many of the documented studies for behavioural change intervention, 92% of which reported positive outcomes. One such applicable study, Morrison et al. [Mor+17], explored the impact of timing and frequency of notifications on user responses and subsequent use of a health-intervention app. They concluded that frequent notifications encouraged greater exposure to the content beneficial for subscribers to consume within the app. Smith et al. [SDM16] studied the impact of personality on choosing a persuasion type for personalised reminders in melanoma patients. Cialdini's 6 principle's of behaviour [Cia87] were used for identifying the suitability of a reminder in a given situation. Comparably, Thomas et al. [JMO17], also used Cialdini's principles to craft personalised messages for encouraging healthy eating. This section discusses the use of Cialdini's principle's in a similar manner. Persuasive features were extracted from push-notifications and used to autonomously generate personalised *persuasive* notifications which, when used for empathetic purposes such as in behavioural health interventions, could improve the livelihood of subscribers e.g. personalised and persuasive notifications encouraging exercise. In addition the principles and subsequent features were also leveraged to create a method of scrutinising the potentially influential aspects of notification engagement, thus enabling transparency and a means of explaining persuasive push-notifications. The following sections discuss the data used for this study, the process of extracting scrutable persuasive features and an offline evaluation in which the application of the features for intelligent notification management was assessed.

### 5.2.1 Dataset

Chapter 3 presented an in-the-wild study in which a smartphone application was created to capture push-notifications for the purposes of understanding positive and negative push-notification engagement patterns and aiding informed development of intelligent systems which could improve the engagement experience of push-notifications. 11 participants with notifications totalling 8,586 engagement-context pairs were selected from the WeAreUs data set for this study. The selection process was as follows:

- Reduce the WeAreUs data set to notifications with no missing values for the subset of features chosen to inform the engineering of persuasive notification features

- Select participants who had engaged with *at least* 100 notifications from the reduced pool of notifications

## 5.2.2 Feature Engineering

Chapter 3 analysed and identified notification features which had statistically significant contributions toward subsequent notification engagement by participants. The features included:

$\{ app, category, priority, subject, time, day, numUpdates, contactSignificant \}$ .

On further inspection, these features did not readily reveal the persuasiveness of a notification i.e. why and how a subscriber might be influenced or enticed to open it. For example, the time and day of a notification delivery might seem inconsequential to a subscriber not consciously keeping track of their engagement, but in actual fact, it may have been intentionally pushed at that time due to previous positive engagement associated with the time or due to recently browsing the source application. An objective of this study was to create a mechanism which would enable subscribers to identify persuasive facets of the notifications they receive in order to promote self-awareness of notification-engagement behaviour and prevent destructive habits (e.g. engaging with a notification just because it was pushed and not because it was relevant or useful) forming when designers seek to exploit users through persuasive hooks.

Consequently, 6 features of persuasiveness (P1-P6) were derived with respect to push-notifications using Cialidini's 6 principle's [CC07] as a guide and the raw notification features as seed. Cialidini's principles and the interpretations for push-notifications are defined as follows:

### 5.2.2.1 Authority

*People follow and respect requests made by an authority.* With this principle, Cialidini was referencing our inclination to accept the word of an authority figure, regardless of its merit, even if that authority is only symbolic. For example, an actor, who previously played the role of a doctor on TV (*Marcus Welby, M.D.*) for several years, boosted the sale of a coffee brand when he was hired to describe it as a healthier alternative in an advert. Cialdini states:

*Because we typically misperceive the profound impact of authority (and its symbols) on our actions, we are at the disadvantage of being*

*insufficiently cautious about its presence in compliance situations. A fundamental form of defense against this problem, therefore, is a heightened awareness of authority power. When this awareness is coupled with a recognition of how easily authority symbols can be faked, the benefit will be a properly guarded approach to situations involving authority-influence attempts. [CC07]*

**Notification Features:** A combined measure of notification *priority*, *number of updates* and *contact significance* were selected to express the *authority* of a notification.

**Assumptions:** The priority level of a notification is a variable set by the app which should indicate the importance of a notification. The app therefore in this case is taken to be the authority on whether the content contained within the notification is important. The Android OS is also an authority in this respect, as it is responsible for alerting the subscriber with a specific modality dependent on the priority level chosen. The number of times a notification is updated and used to re-alert the subscriber is at the joint discretion of the Android OS and source application, both of which are considered an authority. In addition, the *contact significance* feature, described in 3.2.5.1, determines if a contact is significant to the context of delivery. This is an example of an intelligent algorithm being in a position of authority, a scenario which is already becoming frequently ubiquitous.

**Procedure:** The authority feature,  $P_1$ , was calculated as an equal weighted sum of the chosen three authoritative notification components (Equation 5.4).  $P_1$  was calculated for each notification,  $i$ , of each user,  $u$ . The priority feature comprised a number of levels which differed across Android versions. Therefore, the priority of a notification ( $priority_i$ ), was expressed as a fraction of the maximum priority ( $priority_{max}$ ) level possible (Equation 5.1). Notifications with higher priority assigned would be associated with higher levels of authority. Notifications updated more than once would indicate multiple alerts by the Android/App authority, thus increasing urgency and perceived importance of the notification, similar to receiving second or third notices for an unpaid bill. If the notification was updated more than once, the influence of authority would be raised (Equation 5.2). This was subsequently true when a contact was identified as significant in the context of the delivered notification (Equation 5.3).

$$a_1 = \frac{priority_i}{priority_{max}} \quad (5.1) \quad a_2 = \begin{cases} 1, & \text{if } updates > 1 \\ 0, & \text{otherwise} \end{cases} \quad (5.2)$$

$$a_3 = \begin{cases} 1, & \text{if } contact = True \\ 0, & \text{otherwise} \end{cases} \quad (5.3) \quad P_1 = \frac{a_1}{3} + \frac{a_2}{3} + \frac{a_3}{3} \quad (5.4)$$

### 5.2.2.2 Scarcity

*People will place higher value on something that is rare.* With this principle, Cialdini references how something of limited availability is desirable. But he also references how loss, or potential loss, is an influential driver of decision making. For example, Cialdini describes the scenario of receiving a phone call whilst in a conversation with a friend. The person receiving the phone call has two options, continue speaking with their friend, or answering the call. Answering the call is compelling as its availability is limited (the call will ring out and the potential information gained is lost) whereas the conversation with the friend present is readily available.

**Notification Features:** A combined measure of notification *app*, *category*, *subject*, *delivery time* and *delivery day* were selected to express the *scarcity* of a notification.

**Assumptions:** Certain feature values of notifications occur more frequently than others and, as such, subscribers can develop habits for instantly dismissing or opening regularly occurring notifications. Notifications which occur rarely don't incite the same *Click*, *Whirr* [Cia87] response of trigger to action, but instead relate closer to the loss aversion cognitive bias in which the fear of missing out (FOMO) on potentially new information entices further exploration. The app, category and subject features were chosen as they are linked directly to the visual content of the notification, thus would be less frequently seen, quite literally. The day and time of delivery were also selected as notifications occurring out of normal delivery patterns suggest urgency or a unique event, which in itself is a persuasive prospect.

**Procedure:** The scarcity feature,  $P_2$ , was calculated as an equal weighted sum (Equation 5.6) of the chosen five ( $n = 5$ ) features ( $f$ ) which, depending on the number of occurrences of the feature value found in a notification ( $S_f$ ), could indicate the level of persuasiveness due to scarcity. For a sequence,  $s$ , of notifications delivered to a user,  $n$ , the sum of occurrences of each feature value,  $v$  (e.g. *WhatsApp* for app, or *Email* for category), was calculated and represented

as a fraction of the sequence length, which was subsequently subtracted from 1 to allocate larger scores to feature values with smallest occurrences, i.e. rarer notifications.

$$S_f = 1 - \frac{\sum_{i=1}^n [s_i = v]}{n} \quad (5.5)$$

$$P_2 = \sum_{f=1}^n \frac{S_f}{n} \quad (5.6)$$

### 5.2.2.3 Liking

*People prefer to say yes to the request of someone/something they like.* With this principle, Cialdini was referencing our inclination to react positively (*Click, Whirr*) toward something if it is packaged attractively, is familiar to us, similar to us or is associated with something we like. For instance, studies have shown that we are more willing to help another if they are dressed in the same fashion or vote for a politician simply because their name seems familiar. We are equally likely to dislike something if it's associated with a negative experience, misdirected or not - such as anger at a weatherman for predicting rain.

**Notification Features:** A combined measure of notification *app*, *category*, *subject*, *delivery time* and *delivery day* were selected to express the *likeness* of a notification.

**Assumptions:** Notifications presented in an attractive or eye-catching format can be persuasive. For example, use of emoji's in push-notifications has shown to improve CTRs [Pus19; TK16], simply because the attractive visual addition has greater influence on our emotions than a block of text - this will be explored further in following sections of this chapter. Notifications which are closely aligned to our interests are also naturally persuasive, partly due to similarity - if the notification subject is one we ourselves have a particular interest in, we are more likely to pursue it than if it were generic. As noted by Cialdini, a potential protection against untoward liking for the sake of exploitation, is a measure of the strength of liking in a given context. If the liking is significantly higher than normal it might then suggest unbeknownst persuasive influence. The app, category and subject notification features were chosen once again to measure likeness as they describe the source and content of the notification which subscribers had varying affinity towards depending on the sequence of values selected. In addition, the day and time were also selected to provide context to the measure.

**Procedure:** The likeness feature,  $P_3$ , was calculated using a similar method to *scarcity*, but for the set definition,  $s$ , in Equation 5.7 which now comprises only notification feature values ( $v$ ) which were previously acted on positively by the

user (notifications which were opened). The sum of identified features was then divided by the total number of notifications (opened and dismissed),  $n$ . Individual feature likeness values were also not subtracted from 1 as larger values should indicate larger likeness levels. Simply, it calculates the CTR for each feature value and the likeness measure,  $P_3$ , is then an equally weighted sum of each selected feature ( $n = 5$ ) (Equation 5.8).

$$l_f = \frac{\sum_{i=1}^{|s|} [s_i = v]}{n} \quad (5.7)$$

$$P_3 = \sum_{f=1}^n \frac{l_f}{n} \quad (5.8)$$

#### 5.2.2.4 Social Proof

*People tend to do what they see their peers doing.* With this principle, Cialdini was referencing our inclination to follow the crowd. If a crowd exhibits identical behaviour, we perceive that behaviour to be more likely correct and follow it on *autopilot*. For example, this principle is prominently used in sitcoms when *laugh tracks* are used to entice audiences to laugh longer and find content more humorous, regardless of its actual merit. This principle plays on our thirst for validation in moments following indecision or hesitancy - even if our decision or behaviour is incorrect, if it was identical to a crowd at large, it is validated. Implementations of the social proof principle can be commonly found on e-commerce websites in which users are sent nudges or notifications highlighting other similar buyers in their area who have "just recently" made a purchase or benefited from the product/service for sale<sup>123</sup>. Cialdini also describes how similarity plays a role in social proof. *We are more inclined to follow the lead of a similar individual than a dissimilar one* [Cia87]. This theory is prominent today with the rise of social media influencer marketing:

*Influencer marketing can be defined as a process of identifying and activating individuals who have an influence over a specific target audience or medium, in order to be part of a brand's campaign towards increased reach, sales, or engagement.* [SS17]

Influencers are targeted by brands and marketing agencies as optimal points of dissemination for their products or services because of their similarity and affinity

<sup>1</sup><https://wp-social-proof.com/>

<sup>2</sup><https://wpdeveloper.net/plugins/notificationx/>

<sup>3</sup><https://wpfomify.com/ref/161/>

with their followers, which is generally within a particular market vertical or niche, thus attracting audiences with common similarities. Cialdini identified two scenarios in which social proof can misdirect: the first, if the data from which we derive correct behaviour of a crowd is actually false (e.g. the laughing track); the second, is when pluralistic ignorance occurs i.e. each individual of the crowd identifies the crowd behaviour as wrong, but conforms to the crowd behaviour as they assume everyone else believes it is right. In both cases, transparency and self-reflection of facts are both required for social proof misdirection mediation:

*Because automatic pilots can be engaged and disengaged at will, we can cruise along trusting in the course steered by the principle of social proof until we recognize that a piece of inaccurate data is being used. Then we can take the controls, make the necessary correction for the misinformation, and reset the automatic pilot... With no more cost than a bit of vigilance for plainly counterfeit social evidence, then, we can protect ourselves nicely. [Cia87]*

**Notification Features:** A combined measure of notification *app*, *category*, *subject*, *delivery time* and *delivery day* were selected to express the *social proof* of a notification. **Assumptions:** Although actions toward notifications occur relatively privately, general notification engagement can still be viewed through the public eye. For instance, receiving a notification during a meeting or eating dinner with family and choosing to tend to it or not sets a standard of social norm in that context for those present, particularly if the person receiving the notification is an influential figure (e.g. manager or parent). Universal social standards also come into play, particularly for notifications generated due to direct social interaction e.g. tending to an instant message from *WhatsApp* immediately as it might be construed as rude if left unattended [Pie+14]. For the purpose of this work, the social proof measure assumed a general model of notification engagement behaviour. On receiving a particular notification, the method would determine how others, the crowd, acted having also received similar notifications. In such fashion, a social consensus on how to engage with particular notifications was used to represent the social proof.

**Procedure:** The social proof feature,  $P_4$ , was calculated using an identical method to *liking*, but for the set definition,  $s$ , in Equation 5.9 which now comprises only notification feature values ( $v$ ) which were previously acted on positively (opened) by *all other subscribers*. The sum of identified features was then divided by the total number of notifications (opened and dismissed) by *all other subscribers*,  $n$ .

Simply, it calculates the CTR of all other subscribers for each feature value and the social proof measure,  $P_3$ , is then an equally weighted sum of each selected feature ( $n = 5$ ) (Equation 5.10).

$$sp_f = \frac{\sum_{i=1}^{|s|} [s_i = v]}{n} \quad (5.9) \quad P_4 = \sum_{f=1}^n \frac{sp_f}{n} \quad (5.10)$$

#### 5.2.2.5 Commitment & Consistency

*People tend to follow through on their word and uphold behaviours associated with their own self-image.* With this principle, Cialdini was referencing our desire to keep our belief, words and actions consistent, regardless of whether it is logical or prudent. One such example offered is a parent committing to purchasing a particular toy for Christmas, only to find it has been sold out, so instead purchasing another toy for the occasion. Early in the New Year, the parent will follow through on their earlier commitment and purchase the additional toy. Toy stores implement a marketing strategy to incite this behaviour by ensuring select desirable toys are understocked and therefore sold out come the holiday period. Cialdini states:

*Although consistency is generally good, even vital, there is a foolish, rigid variety to be shunned. It is this tendency to be automatically and unthinkingly consistent that Emerson referred to. And it is this tendency that we must be wary of, for it lays us open to the maneuvers of those who want to exploit the mechanical commitment → consistency sequence for profit. [Cia87]*

**Notification Features:** A combined measure of notification *app*, *category*, *subject*, *delivery time* and *delivery day* were selected to express the *consistency* for a notification.

**Assumptions:** Understandably, in the context of notification management, there is limited data with which to fully gauge preconceived beliefs of subscribers. However, previous engagements with notifications could give an indication toward the subscriber's rigidity of action in certain contexts e.g. always dutifully tending to emails quickly or engaging positively with a newly installed fitness app. Additionally, the simple action of installing an app on a device is in itself a conscious commitment to consume the content of the app. Therefore, these two measures were used to represent consistency. The first measure identified the CTR of past



notification engagements identical to the incoming notification (e.g. identical app, topic, category etc.), giving an indication as to the influence of consistency for positive engagement with the incoming notification. The second measure contributing to the consistency feature was calculated by identifying whether an incoming notification was delivered through an app installed proactively by the subscriber or an app pre-installed on the device. Simply giving a subscriber an awareness that they are being completely consistent or never consistent with certain notification types is in itself, a form of protection against misdirected use of the consistency principle and, as Cialdini describes, the *Click, Whirr* of reacting on autopilot.

**Procedure:** The consistency feature,  $P_5$ , was calculated as an equal weighted sum, Equation 5.13, of two components,  $c_1$  and  $c_2$ . The first component,  $c_1$ , identifies the subset,  $s$ , of past notifications identical to the incoming notification and calculates the ratio of positive engagements over total engagements in that set (the CTR). The second component,  $c_2$  identifies if the incoming notification was delivered via an app installed by the subscriber or otherwise. Both components are weighted equally and summed to represent a measure of consistency.

$$c_1 = \frac{\sum_{i=1}^{|s|} [s_i = open]}{|s|} \quad (5.11) \quad c_2 = \begin{cases} 1, & \text{if } installed = True \\ 0, & \text{otherwise} \end{cases} \quad (5.12)$$

$$P_5 = \frac{c_1}{2} + \frac{c_2}{2} \quad (5.13)$$

#### 5.2.2.6 Reciprocity

*People feel obliged to return a favour.* With this principle, Cialdini was referencing our inclination to fulfill our debts and *repay, in kind, what another person has provided us*. This principle can be easily found in every human society, stemming perhaps from our tendency to share what resources we have with one another, for the purposes of survival in early times (e.g. hunter gatherers sharing food) to current economic growth (exchange of goods and services). Within society, the rule of reciprocity is considered good manners and those failing to oblige may be looked upon poorly, by both others and themselves, such is our inherent disposition toward fairness and obligation - which is a persuasive incentive to abide by the rule. Interestingly, not only are we inclined to repay favours, we are also more inclined to repay a concession if one is first offered to us. For example, if an

initial large request is made of us, which we refuse, and the requester concedes to offer a smaller request - we are generally persuaded to also concede and accept the smaller request simply to fulfill the principle of reciprocity, regardless of whether the smaller request is one which we wish to accept.

**Notification Features:** A context feature expressing the time elapsed since a subscriber last used an app was used to represent the reciprocity measure.

**Assumptions:** In this study, a reciprocity measure for a notification was associated with the recency of use of an app. An engagement with an app could be considered a favour to the app owner and therefore is repaid in kind or rewarded by offering additional valuable information shortly after use through a push-notification. Alternately, a push-notification may be sent on the assumption that the user gained value from an app interaction and would therefore be more receptive to a push-notification as an acknowledgement that the app recently provided value. A good example of this in-the-wild are in-app notifications nudging users to give them a rating in the relevant App store. This nudge is generally programmed to appear *just after* the app has completed some positive task for the user and, as such, leverages the principle of reciprocity to facilitate the user to repay the value imparted in kind. Other features may have also been included to represent the inferred influence of reciprocity in a notification, such as an analysis of text content to identify contexts in which reciprocity might be applicable (e.g. social media messages - following back a user who has followed you), or the frequency of notifications sent by an application (a subscriber refusing to engage with the first may concede and engage with a second), however this is left for future work.

**Procedure:** The reciprocity feature,  $P_6$ , was calculated using exponential decay. It was reasoned that the power of reciprocity would reduce as time elapsed between engaging with the app and any subsequent push-notification delivery. As such, Equation 5.14 was used to calculate the measure of reciprocity, with  $a$  representing maximum reciprocity (equal to 1 in this case, as each persuasive principle was weighted equally),  $b$  representing the decay rate and  $x$  representing the elapsed time (in minutes) between app engagement and a subsequent push-notification from the app. Based on research examining significant response times with respect to engagement action [MHM16a], a decay rate of 0.02 was chosen. Table 5.1 illustrates the relationship between elapsed time and reciprocity power. As can be seen, notifications sent immediately after app engagement are associated with higher reciprocity scores, which decays over time as value derived from the app loses relevance.

Table 5.1 Reciprocity scores for increasing elapsed time between app engagement and notification delivery using decay rate of 0.02.

| $a$ | $b$ | $x$ | $P_6$ |
|-----|-----|-----|-------|
| 1   | .02 | 1   | .98   |
| 1   | .02 | 5   | .90   |
| 1   | .02 | 15  | .7    |
| 1   | .02 | 30  | .5    |
| 1   | .02 | 60  | .3    |
| 1   | .02 | 120 | .1    |

$$P_6 = a(1 - b)^x \quad (5.14)$$

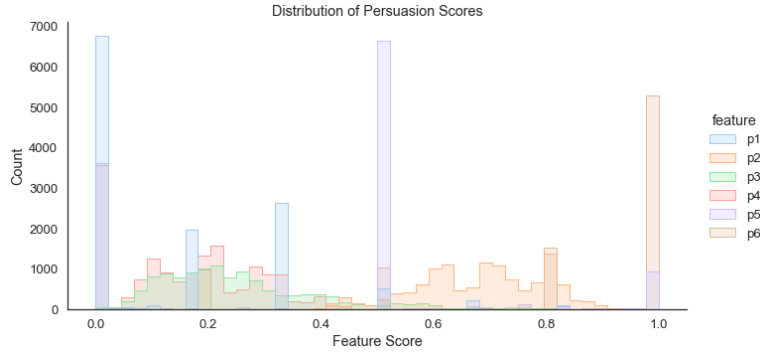
### 5.2.3 Total Persuasion

The total persuasion metric,  $P_T$ , represents the full persuasive power of a notification pushed at a subscriber and is the sum of each individual persuasive feature derived from Cialdini's persuasive principles:  $P_1 + P_2 + P_3 + P_4 + P_5 + P_6$ . The  $P_T$  metric ranges from 0 - 6, with higher values indicating stronger levels of persuasion inherent in a notification.

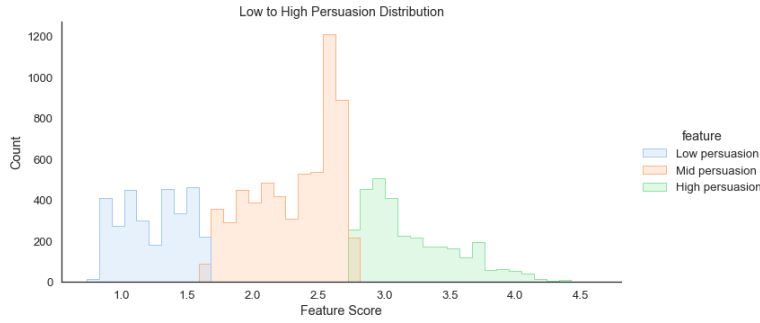
### 5.2.4 Offline Evaluation of Persuasive & Scrutable Features

Once extracted, the 6 persuasive principle's ( $P_1 - P_6$ ), derived with respect to push-notifications, were evaluated using a selection of machine learning algorithms to ascertain the effectiveness of the features toward predicting whether an incoming push-notification was opened or dismissed. Additionally the persuasive features were also evaluated using *Mean Decrease Impurity* (MDI) [Lou+13] for identifying the level of importance of each feature when predicting open or dismiss actions. The hypothesis was that the derived persuasive features should improve the performance of predicting notification CTRs, as highly persuasive notifications should indicate a higher likelihood of positive notification engagement. Subsequently, the persuasive features should also be identified as of higher importance when attempting to predict opens/dismissals for this same reason, highlighting their value toward performance.

The distributions of scores obtained for each persuasive feature,  $P_1 - P_6$ , extracted from the WeAreUs notification data set are illustrated in Figure 5.1a.



(a) Distribution of the 6 persuasive features extracted from WeAreUs notifications.



(b) Ranges of low, mid and high persuasion within notifications of WeAreUs study.

$P_2 - P_4$  vary continuously over the range 0 to 1, whilst  $P_1$ ,  $P_5$  and  $P_6$  experience less diversity, highlighting a potential limitation in the expressiveness of these features. The combined metric,  $P_T$ , was found to range from 0.73 to 4.6, suggesting that achieving a maximum  $P_T$  may be unrealistic. The majority of notifications had  $P_T$  scores between 1.7 and 2.7, indicating this range may be indicative of average persuasive scores found within push-notifications. Therefore, engagement actions taken on push-notifications with low, mid, and high persuasive scores (as defined in Figure 5.1b) by subscribers were examined to investigate the impact of persuasiveness, as defined in the previous section, on notification engagement actions. A Friedman test was run to determine if there were differences in notification CTR between the low, mid, and high persuasive notification groups. Pairwise comparisons were performed with a Bonferroni correction for multiple comparisons. Notification CTR was found to be statistically significantly different when notifications of differing persuasiveness were pushed at subscribers,  $\chi^2(2) = 16.55, p < .001$ . A Post-hoc analysis revealed a statistically significant difference in CTR between notifications with low ( $Mdn_{CTR} = 24\%$ ) and high ( $Mdn_{CTR} = 84\%$ ) persuasiveness ( $p < .001$ ). However, no significant difference

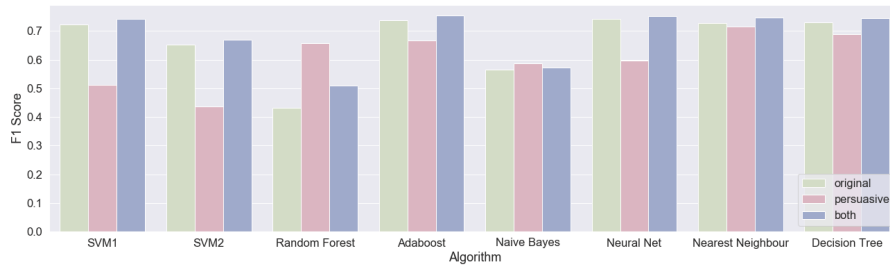


Fig. 5.2 F1 scores of selected algorithms when predicting notification action using differing feature data

in CTR was identified between low and mid ( $Mdn_{CTR} = 48\%$ ) or mid and high suggesting extreme persuasive values to be more indicative of engagement action.

As the analysis demonstrated that the persuasive features were a good indicators of notification engagement, it follows that synthetic notifications generated and conditioned to have a high  $P_T$ , would be opened more frequently by the receiver. A conditional *Wasserstein Generative Adversarial Network with Gradient-Penalty* (WGAN-GP), introduced in the previous chapter, was therefore used to synthetically generate push-notifications with high  $P_T$  values. Subsequently, the synthetic notifications were then classified as opened or dismissed by the selection of machine learning algorithms trained on the original notification data (simulating a real world scenario) and the results were compared against a random benchmark and a held-back set of original notification data.

Figure 5.2 illustrates the results obtained when a selection of algorithms were used to predict the open or dismissal of a notification. Three scenarios are shown for comparison. In the *original* scenario, each algorithm was trained and tested using only the raw features of the notification. In the *persuasive* scenario, only the extracted persuasive features ( $P_1 - P_6$ ) were used for training and testing. The last scenario, *both*, is the union of all features used in the first two scenarios. The F1 score was taken as the metric for measuring performance due to the necessity for balancing *Precision* and *Recall* as the cost of *false positives* and *false negatives* could dissuade use in real-world implementations (e.g. false positives suggest a notification which wouldn't be opened by the user is still pushed at them; false negatives suggest a notification which would have been opened is withheld). Stratified 10-fold cross validation was used for each scenario. As can be seen from the results, the scenario in which both the raw and persuasive features were used together yields best performance across all algorithms but for Naive Bayes and Random Forrest, in which the persuasive scenario performs best.

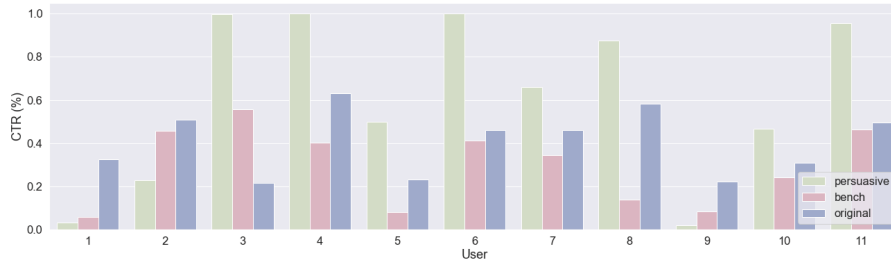


Fig. 5.3 Comparison of CTR of users using differing notification data sets.

Whilst only a marginal increase in most cases over use of raw features alone, this demonstrates that the persuasive features add value in performing predictions of open/dismissal of notifications.

Although the combination of raw and persuasive features performed best, by ascertaining feature importance via MDI implemented through the Scikit-Learn *ExtraTreesClassifier* [Ped+11], select persuasive features were identified as significant in predicting engagement actions.  $P_3$  and  $P_6$ , *Liking* and *Reciprocity* respectively, were ranked highest alongside the app and the day of the week in which notifications were delivered. An additional benefit of the persuasive features is the scrutable element they add to engagements, as they were derived based on well defined principle's of persuasion. Therefore, a system implementing features  $P_1 - P_6$  could illustrate their values to end-user's when explaining automated decisions or facilitating self-reflection and steering behavioural change - this will be discussed later in this chapter.

For each user, 1000 notifications with high  $P_T$  values were generated using the conditional WGAN-GP generator. The conditionally generated notifications were then tested using the selected algorithms which were trained on real notification data for each user. A data set of 1000 randomly generated notifications was used as a benchmark and the original notifications were also used for comparison. Figure 5.3 illustrates the CTR of each scenario for all users and illustrates that the synthetically generated persuasive notifications typically had a much higher CTR than those randomly generated, or those originally sent to the user. The implications of this result are that persuasive notifications may be generated on demand and in such a way, based on Cialidini's principle's, that they are primed for increasing positive subscriber engagement. Notifications generated using this approach could be used to motivate positive behavioural change at opportune moments whilst ensuring, through the scrutable persuasive features, that the subscriber is aware of the type of persuasion they are being subjected to.

## 5.3 Evaluating Persuasive & Scrutable Push In-the-wild

The previous sections have introduced an approach of extracting persuasive principles with respect to push-notifications with the aim of better understanding the influence of incoming interruptions via notification and enhancing the capability of intelligent agents predicting engagement actions. This section further validates the utility of the persuasive features defined in this chapter by discussing their application to an additional study of in-the-wild web push-notifications. Within this study, the persuasive features contributed to a tool for subscribers which facilitated the scrutiny of persuasive facets of incoming notifications. The SCRUPLE (SCRUtable Push LEdger) tool, which satisfies **R4** of the ETHOS-Push approach, was motivated by **RO4** and its evaluation provides empirical evidence for **RQ3**. Also discussed in this section, is an approach toward contextual adaptation of the visual presentation of notifications in-the-wild and its impact on engagement. The implementation and evaluation of the adaptive engine satisfied **R4** of the ETHOS-Push approach, which was again motivated by **RO4**, and contributed toward evidence informing **RQ4**.

### 5.3.1 Experiment Approach

Chapter 3 introduced a study of web push-notifications in-the-wild. Prior to starting the study, participants were randomly allocated to one of two groups. Participants of the first group, the control, engaged with notifications as they arrived throughout the study and answered nightly questionnaires regarding various aspects of the notifications which they received, such as their perceived level of enticement or diversity. The second group, the *scrutable* group, were given access to an additional feature within their nightly questionnaire that attempted to highlight and explain the persuasive facets of notifications with respect to their past engagements, in order to enlighten them of particularly influential notification elements and the subsequent impact on it had on notification engagement behaviour. The purpose of this experiment was to assess if and how notification engagement changed when subscribers were made aware of the persuasive influence of notifications pushed at them. In addition, during this period participants were also subjected to differing notification presentation formats, which varied in their ability to adapt to the context. The subsequent engagements

of users with push-notifications of differing templates were assessed to identify influencing factors and assess whether enticement or persuasiveness could be offset through the visual representation of notifications, which would support an empathetic experience.

### 5.3.1.1 Persuasive Feature Definition

The persuasive features,  $P_2 - P_6$ , defined in Section 5.2.2 of this chapter were extracted from each notification at moments of delivery to later inform participants of the *scrutable* group. The *Authority* persuasive feature,  $P_1$ , was replaced with an alternative principle defined by Cialdini named: *Unity*. *Unity*, as defined by Cialdini [Cia16], represents *shared identity*, such as that shared by members of the same family, members of the same club or people from the same district or country. The shared unity, sense of identity, belonging or alignment with a group was found by Cialdini to be a highly powerful persuasive factor. Applied to push-notifications, unity can be expressed by websites and topics actively sought and subscribed to, as it indicates a user has an identity with the pushing entity, in so far as they are willing to receive regular communication from them. *Unity* was used in place of *Authority* in this experiment as, unlike with mobile push-notifications which had some authoritative entities, such as notifications generated by the Android system, web push-notifications have no similar central authority. If the browser company, such as Google in the case of *Chrome*, were to push-notifications to users, *Authority* would be applicable, but this was not in scope for this experiment.

### 5.3.1.2 The SCRUPLE

The SCRutable Push LEgger was a tool proposed to enlighten subscribers of the persuasive elements within the push-notifications they received and to promote awareness of their engagements with respect to those notifications. In order to achieve this, the tool bridges two cognitive modes of thinking: automatic and reflective. Receiving a notification cues the "click, whirr" automatic response described by Cialdini [Cia87] which makes it easy for subscribers to be nudged toward engagement. Implementing a means to reflect on the engagement decisions however, promotes awareness and empowers subscribers to decide whether their current course of engagement actions are in their best interest. The Merriam-Webster dictionary defines *scruple* as: *an ethical consideration or principle that*



*inhibits action* [Merb]. The SCRUPLE was created with this definition in mind, with the purpose to empower subscribers to scrutinise the persuasive facets of incoming notifications, reflect on the behaviour they incite and be more aware of how they might steer engagement decisions one way or another, ethically or not. By coaxing subscribers to think reflectively as opposed autonomously and by facilitating transparency of nudges, the SCRUPLE tool also fits the framework proposed by Hansen and Jespersen for identifying and classifying manipulative nudges [HJ13].

The tool was constructed by simply listing each notification a participant engaged with throughout the day and displaying, when clicked, a visual and textual representation of the most persuasive elements of that notification with respect to the participant. Figure 5.4a illustrates the list of reconstructed notifications from which the participant could open to learn more about how it was specifically persuasive to them, based on their past engagements. Figure 5.4b illustrates the most persuasive feature found within a selected notification and its accompanying textual explanation seeking to enlighten subscribers as to how they may have been persuaded to engage with the notification. Figure 5.4c illustrates how all persuasive features inherent in a notification were displayed to participants with accompanying explanations for further enlightenment. The text displayed for each persuasive feature and accompanying examples are defined in section C.1 in the Appendix.

Participants in the scrutable group were asked to evaluate use of the SCRUPLE tool with respect to four features of interest:

1. **Understandable:** *How understandable were the explanations for each notification's influence?*

This question aimed to capture how well participants understood the explanations on persuasive features offered by the SCRUPLE tool. This was measured using a five-point Likert scale ranging from 1, very unclear to 5, very clear.

2. **Trust:** *After using the tool, how trusting are you of the apps/websites which sent you notifications?*

This question aimed to ascertain if use of the SCRUPLE tool had an impact on the trust participants had with the websites they subscribed to. This was measured using a five-point Likert scale ranging from 1, very untrusting to 5, very trusting.

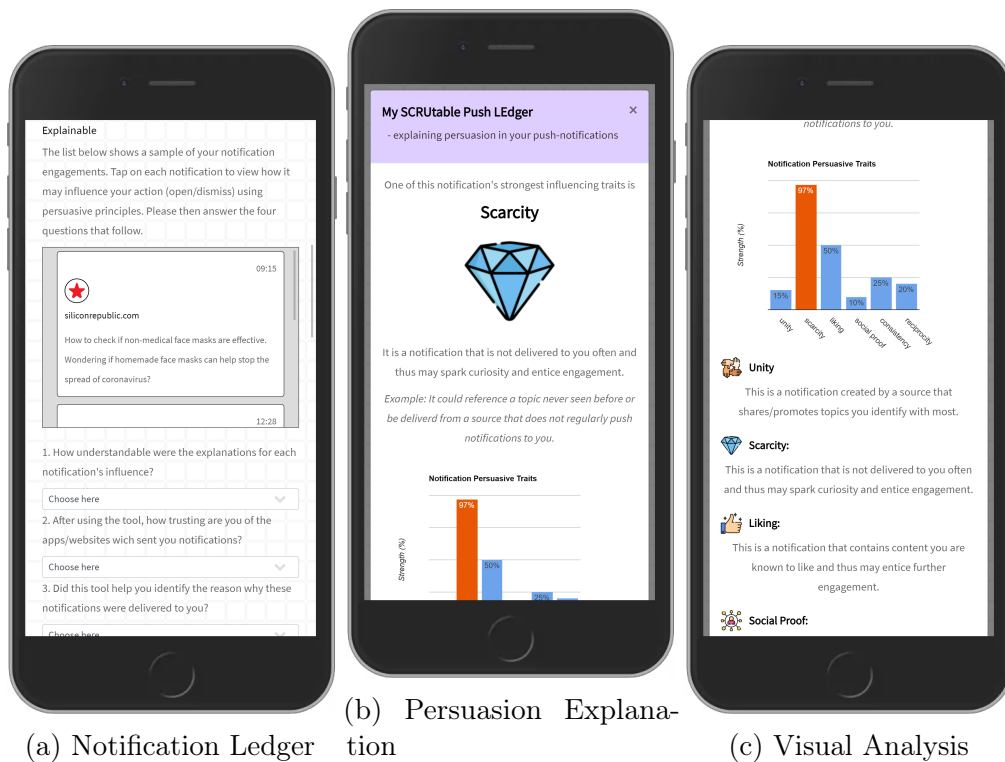


Fig. 5.4 Screenshots of the SCRUPLE.

3. **Helpful:** *Did this tool help you identify the reason why these notifications were delivered to you?*

This question aimed to identify if the SCRUPLE tool helped participants better understand notification behaviour. This was measured using a five-point Likert scale ranging from 1, very unhelpful to 5, very helpful.

4. **Useful:** *Is knowing the reason why these notifications were delivered to you, useful to know?*

This question aimed to identify the usefulness of the tool and extract indication of whether it would be used by subscribers in-the-wild. This was measured using a five-point Likert scale ranging from 1, very useless to 5, very useful.

### 5.3.1.3 Adaptive Presentation Templates

During the *Pushd* study introduced in Chapter 3, participants were, over a period of 4 days, pushed notifications using different adaptive visual representations. The default (control) representation of a notification is illustrated in Figures 5.5a and

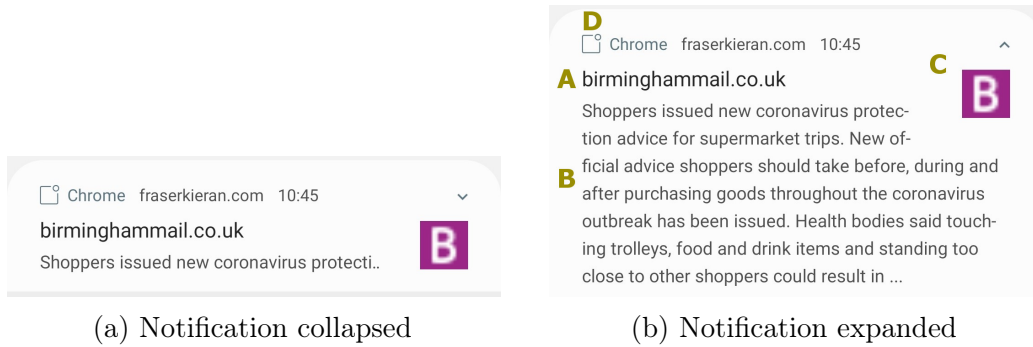


Fig. 5.5 Notification presentation examples.

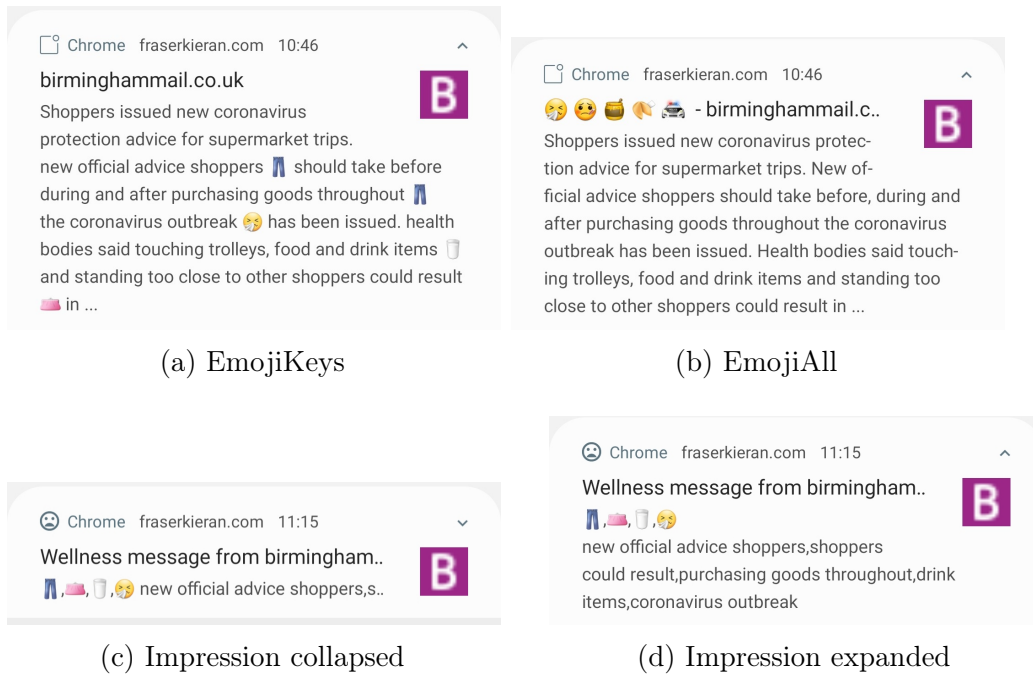


Fig. 5.6 Notification presentation examples.

5.5b, which display the notification in collapsed and expanded mode respectively. The collapsed mode is viewed when participants access the notification bar of their device. The expanded mode of a notification is viewed when participants swipe down on the content of a notification with an intention to view more information. The configurable notification title, content and icon and badge of web push-notifications are highlighted as A, B, C and D in Figure 5.5b for reference. During this study, 3 alternate notification representations were defined. The template definitions are as follows:

- *EmojiKey*

This notification template was identical to the control template except with the addition of contextual emojis inserted after keyword entities within the

notification content. This was achieved by first extracting keywords and converting them to vector representations [Dev+18]. Cosine similarity was then applied to find emoji keyword descriptors (also converted to vector form) [Wij+17] which were similar to the identified keywords. The template for the *EmojiKey* visual representation is displayed in Figure 5.6a.

- *EmojiAll*

This notification template was similar to *EmojiKey* but constructed and displayed slightly differently. Instead of identifying keywords and converting to vectors, the approach taken for this template was to convert the whole text within the notification to a vector representation, this was achieved using sentence embedding [RG19], and again apply cosine similarity to identify emojis which best visually represented the notification as a whole. The string of emojis was then displayed in the title of the notification to give subscribers a visual indication of content without needing to expand the notification and read further. The template for the *EmojiAll* visual representation is displayed in Figure 5.6b.

- *Impression*

The final notification template applied the *EmojiAll* approach to extract contextually relevant emojis, but placed them on the first line of the notification content instead of in the title. The title of the notification then adapted to the content of the notification by displaying the topic addressed. The badge of the notification was used to reflect the sentiment of the text: a smiling face emoticon represented positive sentiment, a frowning face emoticon represented negative sentiment and a neutral face emoticon representing neutral sentiment. The remainder of the text content was also replaced with keywords to facilitate an impression of a notifications content as opposed to a block of text. The template for the collapsed and expanded *Impression* visual representation are displayed in Figures 5.6c and 5.6d, respectively.

During the period of the experiment in which participants were pushed differing notification representation templates, the source of notification content and the method of delivery remained fixed.

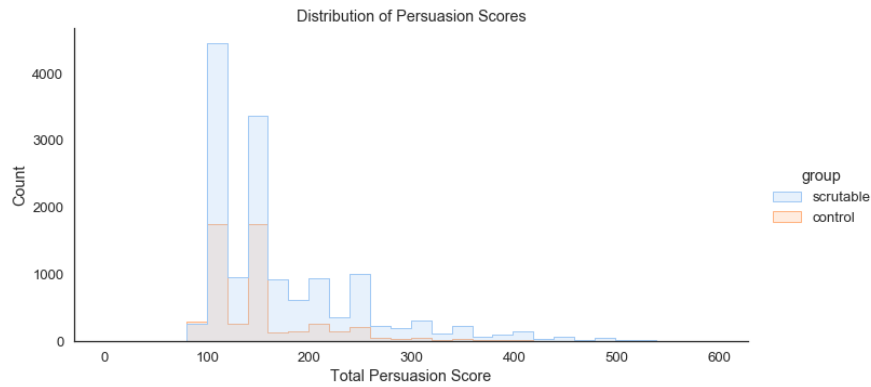
### 5.3.2 Results

The following sections discuss the results obtained during the *Pushd* study with respect to the deployment of the SCRUPLE tool, the use of adaptive presentation templates and the utility of persuasive features applied to push-notifications.

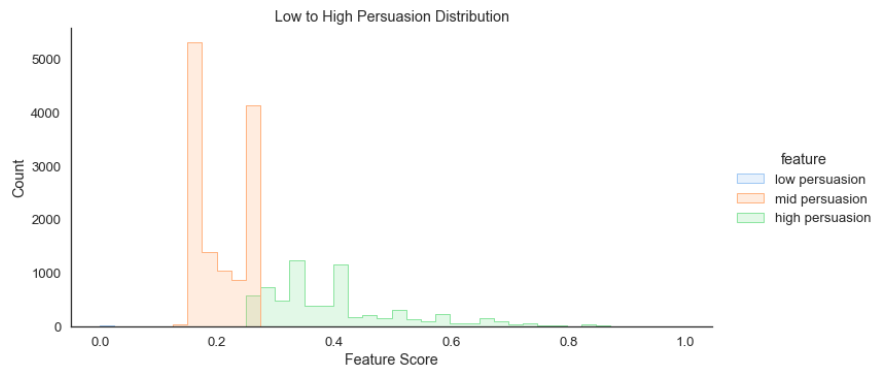
#### 5.3.2.1 Persuasive Notification Engagement In-the-wild

The total persuasive score,  $P_T$ , was calculated for all 19,852 notifications pushed at participants during the study. Figure 5.7a illustrates the distribution of scores for participants in the scrutable group (i.e. 24 participants given the SCRUPLE tool) and those in the control group (i.e. 67 participants not given the SCRUPLE tool). The plot suggests that  $P_T$  was similarly distributed for both groups, which enables comparative evaluation between those using the tool and those not using the tool, as both groups dealt with the same proportions of notifications with differing levels of persuasiveness. The distribution of  $P_T$  was positively skewed, with the majority of notifications having low  $P_T$  scores of 0.1-0.25. Assuming the majority of notifications exhibit normal levels of persuasive influence (i.e.  $IQR = mid\ persuasion$ ), ranges of  $[0, 0.1)$ ,  $(0.1, 0.25)$  and  $[0.25, 1)$  were selected to represent low, mid and high persuasion as illustrated in Figure 5.7b.

A Spearman's rank-order correlation was run to assess the relationship between notification CTR and the persuasive features. Table 5.2 shows the correlations identified across participants in both the control and scrutable groups. Statistically significant, moderate to strong, positive correlations were found between notification CTR and *unity*, *liking*, *commitment* & *consistency* and  $P_T$ , indicating that higher scores for the persuasive features were associated with higher positive notification engagement actions by participants. A statistically significant moderate association was also identified between CTR and scarcity in the control group, which interestingly was not replicated in the group using the SCRUPLE tool. Although the association in the scrutable group between CTR and scarcity is not statistically significant ( $<0.05$ ), the sign of Spearman's rank-order correlation coefficient indicates that, when presented with a scarce notification whilst using the SCRUPLE tool, participants may have been less likely to open notifications. The results illustrated in the table indicate the significance and utility of the persuasive features, as they both indicate likely engagement actions of notifications and whilst explaining the type of persuasion that may be of influence.



(a) Distribution of  $P_T$  extracted from web push-notifications during the *Pushd* study.



(b) Ranges of low, mid and high persuasion within notifications of the *Pushd* study.

Fig. 5.7 Distribution of persuasion within Web Push notifications.

Table 5.2 Results of Spearman's rank-order correlation between mean feature scores and mean participant CTR.

| Feature      | Control |       | Scrutable |       |
|--------------|---------|-------|-----------|-------|
|              | $r_s$   | $p$   | $r_s$     | $p$   |
| unity        | .973    | <.001 | .974      | <.001 |
| scarcity     | .436    | .033  | -.208     | .091  |
| liking       | .512    | .011  | .830      | <.001 |
| social proof | n/a     | n/a   | .026      | .837  |
| commitment   | .548    | .006  | .658      | <.001 |
| reciprocity  | .121    | .573  | .138      | .267  |
| $P_T$        | .783    | <.001 | .745      | <.001 |

A Friedman test was run for both the control and scrutable groups to determine if there were differences in CTR between notifications associated with low, mid

Table 5.3 Results of Friedman test to determine differences in participant CTRs for differing levels of persuasive notification.

| Group     | Median CTR (%) |           |            | $\chi^2$ | $p$   |
|-----------|----------------|-----------|------------|----------|-------|
|           | Low $P_T$      | Mid $P_T$ | High $P_T$ |          |       |
| Control   | 0              | 6.15      | 100        | 19.158   | <.001 |
| Scrutable | 0              | 1.57      | 42.727     | 20.857   | <.001 |

and high persuasiveness,  $P_T$ . The hypothesis was that, participants using the SCRUPLE tool would be less easily influenced to engage with persuasive notifications due to increased awareness and enlightenment ascertained through use of the visualisations and explanations within the tool. A statistically significant difference in notification CTR was identified between levels of persuasion for both groups. The results in Table 5.3 illustrate that participants using the SCRUPLE tool exhibited much lower median CTRs when notifications were in the mid to high  $P_T$  range, suggesting that increased awareness toward notification engagements, facilitated by the SCRUPLE tool, tempered the receptivity of participants to web push-notifications.

Participant responses to the questions regarding the *helpfulness*, *coherence*, *usefulness* and *the impact on trust toward the notification source*, of the SCRUPLE tool are illustrated in Figure 5.8. The results indicate that the tool, and by association the persuasive principles applied to push-notifications:

- helped participants identify why they received certain notifications
- was easy for participants to understand (i.e. with regard to the visualisations and explanations of the persuasive principles applied to push-notifications which were presented)
- was deemed to have imparted useful information with respect to their notification engagement behaviour
- did not cause the trust between the notification source (i.e. the *pusher* of the notification) to deteriorate or improve

A Spearman's rank-order correlation was also run to assess the relationship between participant characteristics (such as age, gender etc.) and responses given for the utility of the SCRUPLE tool. There was just one statistically significant correlation identified which was between participants' propensity to trust websites after using the SCRUPLE tool and participants initial indication as

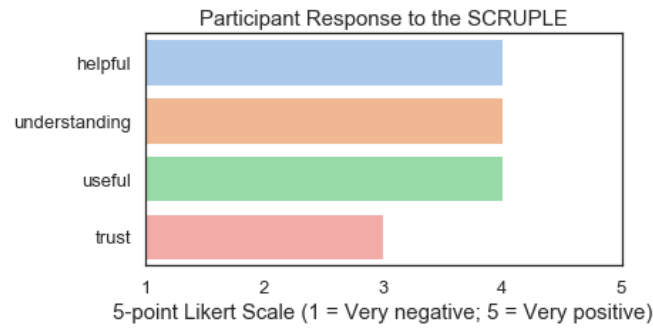


Fig. 5.8 Participant feedback on use of the SCRUPLE tool for enlightened push-notification engagement.

to the likelihood of subscribing for web push-notifications,  $r_s(67) = .260$ ,  $p < .034$ . Although the association was weak, it indicated that participants of the study who were more inclined to subscribe to web push-notifications from the outset, were also more likely to be trusting of the websites after using the SCRUPLE tool.

### 5.3.2.2 Impact of Adaptive Presentation of Push-notifications

A Wilcoxon signed-rank test was conducted to determine the effect of the differing adaptive notification templates on CTRs of participants. Table 5.4 illustrates the shift in participant CTRs when they were presented with one of the proposed adaptive templates after 4 days of being presented with the default notification representation (control). Each template incited statistically significant reduced median CTRs. Participants who were only shown the control template had a -2.68 median decrease in CTR, suggesting that participants may have become less engaged in the study as time went by.

This result is particularly interesting for the *EmojiKey* and *EmojiAll* groups as it conflicts with the sentiment that emoji use increases notification engagement [Pus19]. Other factors, such as the frequency of use or ratio of emojis to text, could be other confounding factors in this instance and future work should consider a wider array of features. The *EmojiKey* and *EmojiAll* representations concentrated on contextualising the emojis to the content of the notifications, however other such contextual adaptations could be made, such as adaption to the physical context of the user (their location, current activity).

The response times, defined as the time from notification delivery (arrival of notification to the participants' device) to participant action (open/dismissal), between the differing notification templates were assessed using a Kruskal-Wallis



test. Median response times were statistically significantly different between the different notification template groups,  $\chi^2(3) = 8.491, p = .037$ . The default template evoked the longest median response time (108s), followed by the *EmojiAll* (51s) and *Impression* (47s) templates. The *EmojiKey* had a significantly shorter response time (11s). As the default representation was made up purely of text content, it may have taken participants longer to read and decide on a subsequent action to explore further or dismiss. In contrast, the *EmojiAll* and *Impression* representations contained summary features, minimising the need for participants to dive further into the notification to get an impression of the content, which may explain the reduced response time.

Table 5.5 highlights the change in NPS awarded by participants between periods of receiving notifications with only the default template and receiving notifications with the adaptive template. The negative change for NPS (-1.6) awarded by participants who only saw the default representation of notifications throughout the experiment illustrates participants' frustration with the lack of intelligence in notification presentation, and potentially the experience of push-notifications as a whole - demonstrating, once again, the need for improved empathy within notification systems. The *Impression* template garnered a small increase in NPS, illustrating the participants may be more approving of the abstract, but informative, information it provided. The *EmojiKey* template, which was similar to the default representation, but for keyword emojis inserted, was received worse by participants and the *EmojiAll* template had incited no change, which given the decrease of the control group, could be considered positive. These results of NPS were found to not be statistically significant, and as such, further studies should explore the impact of the adaptive representations on subscriber receptivity and satisfaction.

## 5.4 Limitations

A limitation of the approach toward extracting persuasion from push-notifications, proposed in this chapter, was the use of a finite set of notification engagement features. In particular, the exclusion of the text feature is a limitation of the work presented in this chapter. Persuasive language present in notification text fields may be indicative of future engagement action and could be leveraged to inform and enlighten subscribers to promote awareness of how they are potentially steered.

Table 5.4 Results of Wilcoxon signed-rank test to determine differences in participant CTRs when differing notification representations presented.

| Group      | Median CTR (%) |          | Mdn diff | $z$    | $p$   |
|------------|----------------|----------|----------|--------|-------|
|            | Control        | Template |          |        |       |
| Control    | 6.67           | 2.2      | -2.68    | -2.464 | 0.014 |
| EmojiKey   | 8.13           | 1.26     | -3.88    | -2.77  | 0.005 |
| EmojiAll   | 10.65          | 4        | -4.56    | -2.92  | 0.004 |
| Impression | 11.02          | 4.63     | -4.8     | -2.311 | 0.021 |

Table 5.5 Results of Kruskal-Wallis H test to determine differences in CTRs or NPS scores between groups presented with differing notification representations.

| Group              | Control | EmojiKey | EmojiAll | Impression   | $\chi^2$ | $p$   |
|--------------------|---------|----------|----------|--------------|----------|-------|
| $\Delta$ NPS score | -1.6    | -0.5     | 0        | <b>+0.25</b> | 5.845    | 0.119 |

The following chapter will address notification with respect to *information-gap* theory, however, future work should consider Cialdini’s principles applied to the text content of notifications.

## 5.5 Summary

This chapter introduced an approach of extracting scrutable persuasive features from push-notifications for the purpose of improving predicted action outcomes and enabling users to reflect on the persuasive characteristics. The results illustrated that persuasive features could be extracted and visualised and that performance in predicting notification action outcomes could be improved using the persuasive features. Recommended persuasive notification could also be generated based on the persuasive features and were also shown to increase the Click Through Rate (CTR) of users when simulated.

This section further validates the utility of the persuasive features defined in this chapter by discussing their application to an additional study of in-the-wild web push-notifications. Within this study, the persuasive features contributed to a tool for subscribers which facilitated the scrutiny of persuasive facets of incoming notifications. The SCRUPLE (SCRUtable Push LEDger) tool, which satisfies **R4** of the ETHOS-Push approach, was motivated by **RO4** and its evaluation provides empirical evidence for **RQ3**. Also discussed in this section, is an approach toward

---

contextual adaptation of the visual presentation of notifications in-the-wild and its impact on engagement. The implementation and evaluation of the adaptive engine satisfied **R4** of the ETHOS-Push approach, which was again motivated by **RO4**, and contributed toward evidence informing **RQ4**.



# Chapter 6

## Exploring the Effects of Enticing Notification Text Content

Without knowing the force of  
words, it is impossible to know  
more

---

*Confucius*

### 6.1 Overview

Push-notifications were originally created to inform users they had received a new email[MC06]. However over the past decade, with the number of smartphone apps and ubiquitous devices capable of pushing notifications rising, the number of scenarios in which notifications are pushed has expanded. Notifications have progressed from being pushed via a single channel with a single intended purpose (updates on email), to spanning multiple channels, each with a different purpose (e.g. social network interactions, meeting reminders, sensor readings, public transport updates etc.). The motivation behind push-notifications has also advanced. Push-notifications, once intended as a simple update, are now crafted with an intention to incite particular behaviours, such as increased platform engagement or purchasing. As such, notifications are now created with the goal of maximising positive engagement and, to achieve this, marketing techniques are implemented

to purposefully influence subscribers to open notifications and improve marketing campaign metrics, such as Click-Through-Rates (CTRs). One such marketing technique leveraged for improved notification engagement is the *Curiosity Gap*. The Curiosity Gap exploits Loewenstein’s Information-Gap theory [Loe94] which suggests that curiosity peaks when a small amount of information is revealed, but key parts are withheld. An example of a notification title exploiting the curiosity gap might be: *Business Schools That Will Dominate In The 21st Century*. In this text title, key information, the names of the business schools, are withheld to entice subscribers to open the notification and find the answer. Similarly, a common Facebook notification: *You have a new friend suggestion*, also exploits the curiosity gap, as key information, the persons name, is omitted from the notification tempting users to open and engage with the Facebook platform to find the answer. Indeed, Kang et al. [Kan+09] illustrate, through experimentation with trivia questions, that curiosity due to information-gaps can elicit an inclination to pay for missing information, such is the power of its influence.

In this chapter, existing data sets comprised of both clickbait and non-clickbait news article headlines were leveraged to create a classifier for identifying enticing notification text content. Potthast et al. define *Clickbait* as:

*A certain kind of web content advertisement that is designed to entice its readers into clicking an accompanying link. Typically it is spread on social media in the form of short teaser messages.* [Pot+16]

Additionally, Chakraborty et al. [Cha+16] further define clickbait headlines as exploiting the curiosity gap, enabling attention distraction and contributing to readers cognitive overload. Clickbait implies the *teaser message* does not accurately represent the content within, whereas the curiosity gap can be applied while still accurately representing content, albeit with missing information in the headline. Therefore, *enticing notifications* are defined, in the context of this thesis, to be *notifications with text similar to article headlines identified as clickbait, indicating potential exploitation of the curiosity (info) gap*. Subsequently, for the purposes of identifying enticing notifications, a corpus of clickbait headlines was used for predicting a level of enticement for notifications with text similar to those found in clickbait.

The contributions of this chapter, motivated by **RO1/4** which provide support to **RQ1/4**, are as follows:

1. An analysis of enticing notification text found *in-the-wild* and the impact on engagement

2. A method of training and evaluating a NMS using notification text
3. An (offline) evaluation of the influence of enticing notifications on NMSs and a subsequent method of mediation
4. Procurement, analysis and (offline) evaluation of an additional notification data set containing both enticing and non-enticing text
5. Application and (offline) evaluation of a generative model for autonomous, conditional, creation of enticing push-notification text
6. An in-the-wild implementation of autonomous notification curation methods aimed at further exploring the influence of enticement and evaluating autonomous summarisation as a method of mediation

## 6.2 Enticing Notification Text In-the-wild

Chapter 3 introduced a study of smartphone push-notifications *in-the-wild*. Over 35,000 push-notifications were observed and logged over a period of 6 months. Through studying engagement actions (open/dismissal) of these notifications, it emerged that participants dismissed almost as many notifications as they received. The analysis of Chapter 3 uncovered a number of relationships between notification features and participant contexts which could have contributed to the positive and negative engagement actions taken toward notifications. The text content of the notifications delivered was not explored. This section aims to explore the notification text present in notifications delivered during the WeAreUs study and measure the influence enticing text may have on engagement actions.

### 6.2.1 WeAreUs Text Feature

For the purpose of preserving the privacy of participants in the WeAreUs study, the full title, ticker and content texts of the notification were not recorded, but were instead used to create a privacy-compliant text feature as follows:

- The title, ticker and content text elements of each notification were combined
- The topic of the notification was inferred and recorded

- [illegible]

The resultant notification text feature was made up of key phrases representing the title, ticker and content text elements of the notification. A number of participants in the WeAreUs study did not have an adequate number of text features collected via the app, thus restricting the analysis to 11 participants and 8,440 notifications. The sentiment of notification text was inferred using the TextBlob [Lor20] python package and Figure 6.2b illustrates the majority of text found in notifications was neutral, if oriented slightly positive. Figure 6.2a illustrates the distribution of notifications with differing character lengths. The vast majority are under 100 characters which is unsurprising as the purpose of notifications, by design, is to quickly update users with new information e.g. an event or a task. Character limits for title, ticker and content text elements are also imposed on notifications depending on the technology (mobile/browser) used. Using scikit-learn’s [Ped+11] feature extraction tools, n-grams were extracted from the notifications text feature to further analyse the types of words found in the notification data set. The word cloud in Figure 6.1 illustrates a selection of the most frequent words found in the texts. Terms such as: *message*, *photo* and *likes* are among the most frequent which alludes to social media messaging



being a prominent driver of notifications. Other frequent terms such as *email*, *event* and *order* hint at notifications related to work or administration tasks. The frequency of *traffic* demonstrates that notifications related to timely information, such as commute updates, also made up a large portion of notifications sent to participants. Figure 6.2c illustrates the most frequent bigrams (two consecutive words) found in notification text, with terms such as *message messages*, *light traffic*, *location cloudy*, *wi fi* and *just eat* among them, again demonstrating that a large portion of notification messages pertain to instant messaging, commute updates as well as other time-dependent information - in this case: food delivery. Figure 6.2d illustrates the most frequent trigrams (three consecutive words) found in notification text, again demonstrating commute and travel updates as most frequent with items such as: *price updates flights*, *hotel montana paris*, *traffic delays congestion*, *thu travel tips*.

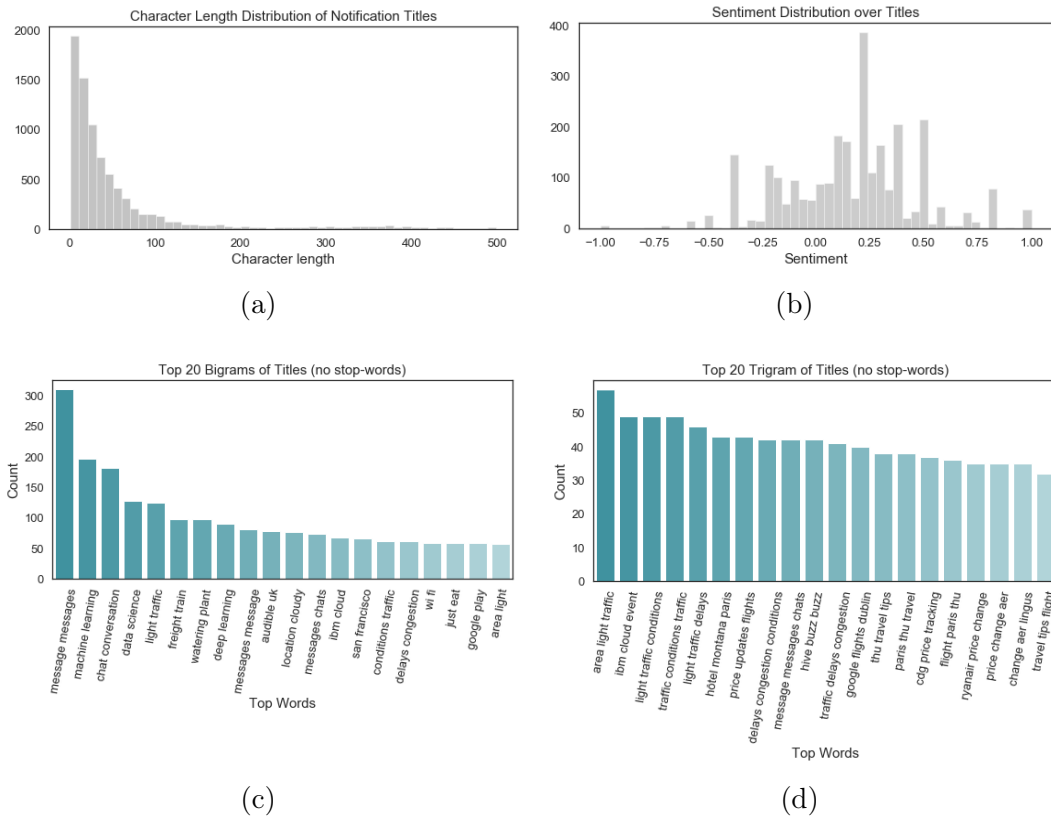


Fig. 6.2 Analysis of the privacy-preserving text feature from the WeAreUs notification data set

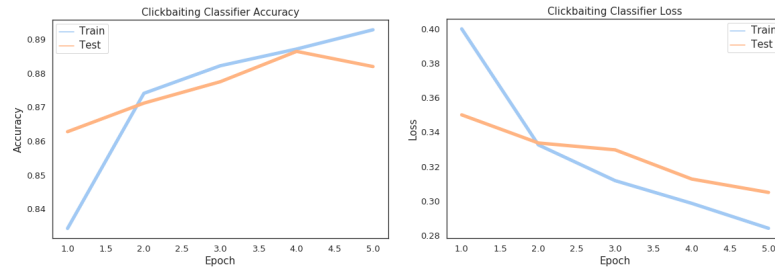


Fig. 6.3 Notification enticement classifier accuracy and loss

Table 6.1 Clickbait and Non-clickbait examples

| Clickbait | Headline                                                      |
|-----------|---------------------------------------------------------------|
| Yes       | I Was A Weatherman For A Day, And This Is What I Learned      |
| Yes       | Business Schools That Will Dominate In The 21st Century       |
| Yes       | 13 Reasons To Be In Love With Your Life                       |
| Yes       | The Struggle Of Adulthood Perfectly Summed Up In Two Images   |
| No        | Chris Hughton appointed permanent manager of Newcastle United |
| No        | James Harden is Our NBA MVP at the All-Star Break             |
| No        | Royal Wedding delayed as Prince plans for Pope's funeral      |
| No        | Intel launches Centrino 2 aiming for performance and mobility |

### 6.2.2 Extracting Notification Enticement

Potthast et al. [Pot+17] hosted the *Clickbait Challenge* in 2017 to enable the comparable evaluation of clickbait detectors and released a data set comprised of annotated clickbait examples as part of their challenge. Chakraborty et al. [Cha+16] gathered a data set comprised of both clickbait and non-clickbait samples from sources such as *Wikinews* and *BuzzFeed*. For this work, the annotated clickbait data sets of Potthast et al. and Chakraborty et al. were chosen to train a classifier for identifying enticing notifications. The combined data set was split into a training set of 47,538 clickbait/non-clickbait samples and a test set of 4,000 samples. An example of clickbait and non-clickbait samples are illustrated in Table 6.1. A Multilayer Bidirectional Recurrent Neural Network (RNN) with two stacked GRU layers was chosen as a model for classifying notification text as enticing. An embedding layer was also added and initialised with GloVe's [PSM14] 100 dimension pre-trained word vectors. The model was trained using a batch size of 128 for 5 epochs, reaching an accuracy of 89.2%. The training and validation accuracy and loss are illustrated in Figure 6.3.



Table 6.2 Observed frequencies of notifications, with and without enticing text, delivered through differing apps

| App                   | isEnticing      |                | <i>p</i> |
|-----------------------|-----------------|----------------|----------|
|                       | False           | True           |          |
| com.android.chrome    | 100<br>(1.7%)   | 24<br>(1%)     | 0.019    |
| com.facebook.katana   | 187<br>(3.1%)   | 30<br>(1.2%)   | <.001*   |
| com.facebook.orca     | 141<br>(2.3%)   | 212<br>(8.7%)  | <.001*   |
| com.google.android.gm | 1233<br>(20.5%) | 310<br>(12.7%) | <.001*   |
| com.snapchat.android  | 53<br>(0.9%)    | 14<br>(0.6%)   | 0.149    |
| com.tinder            | 227<br>(3.8%)   | 210<br>(8.6%)  | <.001*   |
| com.twitter.android   | 272<br>(4.5%)   | 91<br>(3.7%)   | 0.104    |
| com.whatsapp          | 2189<br>(36.5%) | 1290<br>(53%)  | <.001*   |
| other                 | 1603<br>(26.7%) | 254<br>(10.4%) | <.001*   |

tral), were analysed with respect to the app through which they were delivered to participants. A chi-square test of homogeneity was run, with an adequate sample size established according to Cochran [Coc54]. The two multinomial probability distributions were not equal in the population,  $\chi^2(8) = 660.599$ ,  $p < .001$ . Observed frequencies and percentages of notifications delivered per app for enticing and neutral notifications are presented in Table 6.2. Post-hoc analysis involved pairwise comparisons using multiple z-tests of two proportions with a Bonferroni correction. Statistical significance was accepted at  $p < .006$ . There was a number of statistically significant differences in the proportion of enticing notifications delivered compared to neutral notifications delivered for differing app types. In particular, *WhatsApp* delivered a much higher proportion of enticing notifications than neutral. Conversely, *Gmail* delivered a much lower proportion of enticing notifications than neutral.

3,658 notifications, classified as enticing or neutral, were analysed with respect to the topic extracted from the notification text. A chi-square test of homogeneity was again run, with an adequate sample size established. The two multinomial

probability distributions were not equal in the population,  $\chi^2(10) = 354.948$ ,  $p < .001$ . Observed frequencies and percentages of notifications delivered per topic for enticing and neutral notifications are presented in Table 6.3. Post-hoc analysis involved pairwise comparisons using multiple z-tests of two proportions with a Bonferroni correction. Statistical significance was accepted at  $p < .005$ . There was a number of statistically significant differences in the proportion of enticing notifications delivered compared to neutral notifications delivered for differing topics extracted. A much larger proportion of *Business & Industry* notifications delivered were enticing (19.4%) as opposed neutral (5.9%). *Food & Drink* and *Online Communities* topics also had higher proportions of enticing notifications. Perhaps understandably, the proportion of enticing notifications with an inferred topic of *News* was less than the proportion of neutral *News* notifications, indicating that news sources may not have been generating many notification titles resembling clickbait. A similar conclusion could be drawn for *Sports* notifications.

4,381 notifications, classified as enticing or neutral, were analysed with respect to the category of notification delivered. A chi-square test of homogeneity was run, with an adequate sample size. The two multinomial probability distributions were not equal in the population,  $\chi^2(3) = 244.394$ ,  $p < .001$ . Observed frequencies and percentages of notifications delivered per category for enticing and neutral notifications are presented in Table 6.4. Post-hoc analysis involved pairwise comparisons using multiple z-tests of two proportions with a Bonferroni correction. Statistical significance was accepted at  $p < .013$ . There were statistically significant differences in the proportion of enticing notifications delivered compared to neutral notifications delivered for *email*, *msg* and *other* categories. The proportion of neutral notifications delivered with a category of *email* was much higher than enticing notifications of the same category. However, the opposite was true of notifications with a category of *msg*. Thus indicating that not all notifications of the communication type were equally responsible for generating clickbait-like text.

8,440 notifications, classified as enticing or neutral, were analysed with respect to the LED colour emitted to alert participants to their presence. A chi-square test of homogeneity was run, with an adequate sample size established. The two multinomial probability distributions were not equal in the population,  $\chi^2(6) = 181.997$ ,  $p < .001$ . Observed frequencies and percentages of notifications delivered per LED colour for enticing and neutral notifications are presented in Table 6.5. Post-hoc analysis involved pairwise comparisons using multiple z-tests of two proportions with a Bonferroni correction. Statistical significance

Table 6.3 Observed frequencies of notifications delivered, with and without enticing text, and their inferred topic

| Topic                   | isEnticing     |                | <i>p</i> |
|-------------------------|----------------|----------------|----------|
|                         | False          | True           |          |
| arts & entertainment    | 592<br>(23.9%) | 268<br>(22.7%) | 0.114    |
| business & industry     | 145<br>(5.9%)  | 229<br>(19.4%) | <.001*   |
| computers & electronics | 126<br>(5.1%)  | 21<br>(1.8%)   | <.001*   |
| food & drink            | 48<br>(1.9%)   | 48<br>(4.1%)   | <.001*   |
| internet & telecoms     | 259<br>(10.5%) | 184<br>(15.6%) | <.001*   |
| jobs & education        | 183<br>(7.4%)  | 63<br>(5.3%)   | 0.255    |
| news                    | 80<br>(3.2%)   | 23<br>(1.9%)   | 0.142    |
| online communities      | 215<br>(8.7%)  | 154<br>(13.1%) | <.001*   |
| other                   | 391<br>(15.8%) | 109<br>(9.2%)  | <.001*   |
| people & society        | 70<br>(2.8%)   | 47<br>(4%)     | 0.006    |
| sports                  | 369<br>(14.9%) | 34<br>(2.9%)   | <.001*   |

Table 6.4 Observed frequencies of delivered notifications, with and without enticing text, and their assigned categories

| Category  | isEnticing      |                 | <i>p</i> |
|-----------|-----------------|-----------------|----------|
|           | False           | True            |          |
| email     | 1245<br>(43.6%) | 310<br>(20.3%)  | <.001*   |
| msg       | 1375<br>(48.2%) | 1016<br>(66.6%) | <.001*   |
| other     | 225<br>(7.9%)   | 199<br>(13%)    | <.001*   |
| recommend | 10<br>(0.4%)    | 1<br>(0.1%)     | 0.148    |

Table 6.5 Observed frequencies of delivered notifications, with and without enticing text, and their assigned LED colours

| LED colour  | isEnticing      |                 | <i>p</i> |
|-------------|-----------------|-----------------|----------|
|             | False           | True            |          |
| blue        | 222<br>(3.7%)   | 42<br>(1.7%)    | <.001*   |
| green       | 348<br>(5.8%)   | 169<br>(6.9%)   | 0.047    |
| magenta     | 8<br>(0.1%)     | 4<br>(0.2%)     | 0.732    |
| other       | 23<br>(0.4%)    | 4<br>(0.2%)     | 0.107    |
| transparent | 5009<br>(83.4%) | 1864<br>(76.6%) | <.001*   |
| white       | 346<br>(5.8%)   | 339<br>(13.9%)  | <.001*   |
| yellow      | 49<br>0.8%      | 13<br>(0.5%)    | 0.169    |

was accepted at  $p < .007$ . There were three statistically significant differences in the proportion of enticing notifications delivered compared to neutral notifications delivered using differing LED colours. The proportion of neutral notifications delivered using *blue* and *transparent* LED colours was greater than the proportion of notifications delivered which were enticing, for both colour categories. In contrast, the proportion of enticing notifications delivered using a *white* alert modality was greater than the proportion of neutral notifications delivered under the same LED colour. As discussed in Chapter 3, the LED colours of notifications are generally indicative of the app delivering the notification. In particular, notifications using a *white* LED colour tend to be generated from *WhatsApp*, which, as the analysis between enticement and app indicated, was the largest source of enticing notifications.

8,413 WeAreUs notifications, classified as enticing or neutral, were analysed with respect to the priority level assigned to notifications. The priority level is generally assigned by the notification creator or app owner, depending on the type of notification being pushed. There are 5 discrete possible values for priority level min, low, default, high, max. The priority level influences how the operating system presents notifications to subscribers. The notification *channel importance* supersedes the priority feature in Android 8.0 and above [And20d]. A chi-square

Table 6.6 Observed frequencies of delivered notifications, with and without enticing text, and their assigned priority level

| Priority Level | isEnticing    |                 | <i>p</i> |
|----------------|---------------|-----------------|----------|
|                | False         | True            |          |
| default        | 4247<br>(71%) | 1321<br>(54.3%) | <.001*   |
| high           | 1374<br>(23%) | 652<br>(26.8%)  | <.001*   |
| max            | 359<br>(6%)   | 300<br>(12.3%)  | <.001*   |
| min            | 0<br>(0%)     | 160<br>(6.6%)   | <.001*   |

test of homogeneity was run, with an adequate sample size established. The two multinomial probability distributions were not equal in the population,  $\chi^2(3) = 565.223$ ,  $p < .001$ . Observed frequencies and percentages of notifications delivered per app for enticing and neutral notifications are presented in Table 6.6. Post-hoc analysis involved pairwise comparisons using multiple z-tests of two proportions with a Bonferroni correction. Statistical significance was accepted at  $p < .013$ . There was a number of statistically significant differences in the proportion of enticing notifications delivered compared to neutral notifications delivered for differing priority levels. Interestingly, only the *default* priority level had a higher proportion of neutral notifications delivered over enticing notifications. A larger proportion of notifications delivered, with *min*, *high* and *max* set as the priority level, were found to be enticing rather than neutral.

Finally, 629 WeAreUs notifications, classified as having enticing or neutral text, were analysed with respect to the vibration pattern used for alerting participants to the presence of a notification. As discussed previously in Chapter 3, and similar to the LED colour, the vibration pattern tends to be unique for differing applications. The vibration pattern is a sequence of values which indicate the duration of a pause or vibration executed when a notification is delivered [And20i]. The first value of the sequence denotes the duration of a delay before the first vibration. Thereafter, values are consecutively vibration/pause durations. For example, the sequence [0 100] indicates a 0ms delay before vibrating for 100ms. A chi-square test of homogeneity was run, with an adequate sample size established. The two multinomial probability distributions were not equal in the population,  $\chi^2(5) = 46.801$ ,  $p < .001$ . Observed frequencies and percentages of notifications



Table 6.7 Observed frequencies of delivered notifications, with and without enticing text, and their assigned vibration pattern

| Vibration Pattern | isEnticing     |               | <i>p</i> |
|-------------------|----------------|---------------|----------|
|                   | False          | True          |          |
| [0 100]           | 37<br>(7%)     | 16<br>(15.7%) | 0.829    |
| [0 200 200 200]   | 90<br>(17.1%)  | 29<br>(28.4%) | 0.277    |
| [0 250 200 250]   | 192<br>(36.4%) | 32<br>(31.4%) | <.001*   |
| [0]               | 23<br>(4.4%)   | 11<br>(10.8%) | 0.652    |
| [0 500]           | 139<br>(26.4%) | 1<br>(1%)     | <.001*   |
| other             | 46<br>(8.7%)   | 13<br>(12.7%) | 0.246    |

delivered per app for enticing and neutral notifications are presented in Table 6.7. Post-hoc analysis involved pairwise comparisons using multiple z-tests of two proportions with a Bonferroni correction. Statistical significance was accepted at  $p < .008$ . There were two statistically significant differences in proportion of enticing notifications compared to neutral notifications delivered using differing vibration patterns. A smaller proportion of notifications delivered, with vibration patterns of [0 500] and [0 250 200 250], were found to be enticing rather than neutral. This indicates that the vibration modality could help participants distinguish between notifications which are clickbait-like before even unlocking their device and reading the message. This is important as it could deter smartphone users from being manipulated into using their smartphone unnecessarily when a clickbait-like notification intrudes. A chi-square test for association was also conducted between enticement and whether or not a notification was updated multiple times, thus disrupting the participant more than once. All expected cell frequencies were greater than five. There was a small, statistically significant association between notification enticement and whether the notification was updated after delivery,  $\chi^2(1) = 226.831$ ,  $p < .001$ ,  $\phi = 0.164$ .

### 6.2.4 Notification Engagement Actions Toward Enticing Text

The actions taken toward notifications by the 11 participants were also analysed. Notifications with text content classified as enticing had higher CTRs among participants (67.51%) compared with the CTR of notifications classified as non-enticing (48.44%). Of the 11 analysed, the enticing notifications elicited an increased CTR in 10 participants, with 1 outstanding participant having a lower CTR for enticing notifications. An exact sign test was used to compare the differences in CTR between enticing and non-enticing notifications. The enticing notifications elicited a statistically significant median increase in CTR (15.59%) compared to the non-enticing notifications,  $p = .012$ .

The utility of the enticement feature in contributing to predictions of participant engagement actions toward notifications was also assessed. Permutation feature importance was used to identify the contribution made by the enticement feature compared to other notification features commonly used for predicting notification actions (open/dismiss). The permutation feature importance algorithm implemented in the Scikit-learn python library [Ped+11] was used to identify importance values. Feature importance values were extracted for each of the 11 participants and the mean was calculated to create a holistic ranking of features. The AdaBoost classifier [Has+09], implemented in Scikit-learn, was the estimator of choice for predicting open/dismiss actions. Figure 6.6 illustrates the final ranking of notification feature importance, in which the enticement feature ranks 4th. Interestingly, the enticement feature ranks above others which might be considered more indicative of notification engagement behaviour such as *category* and *timeOfDay*. Understandably, features such as the *appPackage* and *dayOfWeek* rank highest.

## 6.3 Training a NMS with Notification Text

As discussed in previous chapters, an idealistic Notification Management System (NMS) would leverage available information, such as notification content, context and the previous engagement history of subscribers, to intelligently deliver notifications, such that disruptions and distractions would be minimised. Chapter 3 briefly discussed notification engagement and smartphone etiquette currently adhered to *in-the-wild* and identified relationships between notification features,

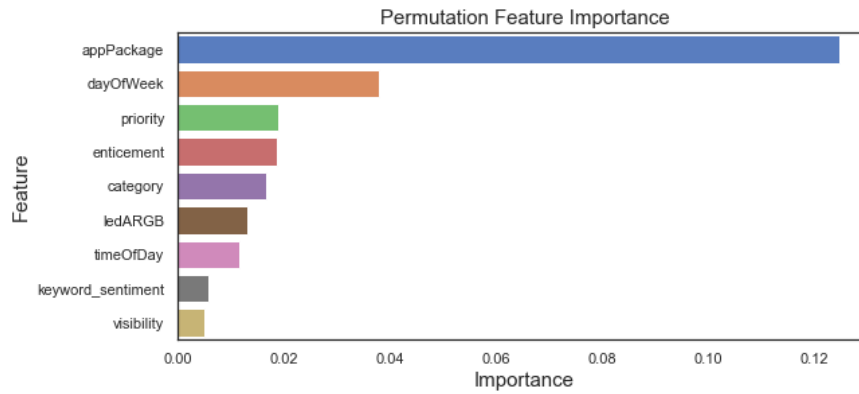


Fig. 6.6 Feature importance rankings when predicting notification actions

contexts and behaviours with respect to etiquette. Chapter 4 introduced methods of managing notifications using a variety of ML algorithms as well as extracting persuasive notification features to be used for informing subscribers of potential manipulations. Yet to be discussed however, are some of the risks associated with using a NMS. This section discusses notification delivery with respect to enticing notifications and the influence they may have on end-users, and subsequently, the NMSs trained upon their historical notification engagements.

In contrast to previous chapters which explored training a NMS using features extracted from the text fields of notifications, such as the *subject* or *contact significance*, this chapter describes the creation of a RL agent trained on the privacy-preserving raw notification text content. The goal of an RL agent acting as a NMS, in practice, would be to identify notifications which the subscriber is likely to open and deliver them, whilst caching or removing notifications which are predicted to be dismissed by the subscriber, thus reducing the cognitive load caused by an influx of irrelevant notifications. Using the notification text as input and the notification action as ground-truth, a RL agent was trained to predict whether a participant would open or dismiss a notification based on its textual content.

A Deep Q-learning (DQN) agent was chosen as it has shown previously to be effective at managing push-notifications on behalf of a notification subscriber [SFC19]. The notification text was transformed into a fixed-length (768 dimension) vector using the sentence-transformers [RG19] library. The pre-trained BERT-base model fine-tuned on *The Stanford Natural Language Inference Corpus* and *The Multi-Genre NLI Corpus* was used in this case. The DQN’s performance at predicting user actions toward notifications was evaluated using the following metrics: accuracy, precision, recall and F1-score. Recall is of particular interest in

this context as the cost of false negatives is quite high. For example, consider a scenario in which the agent dismissed a notification when it was actually important for the subscriber to see. The agent was then withholding information from the subscriber, which could foster feelings of mistrust and result in removal of the NMS.

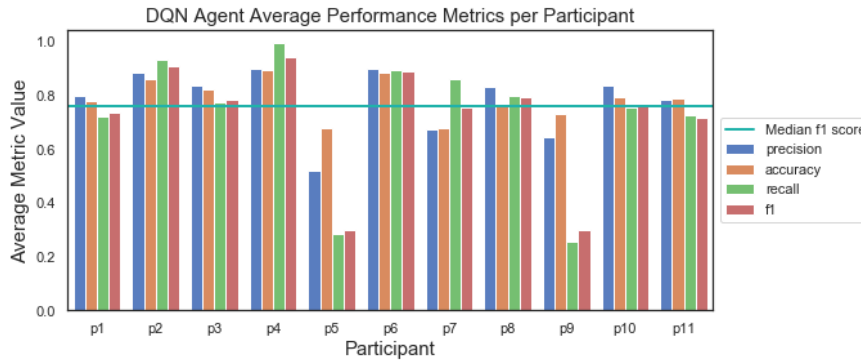


Fig. 6.7 Average performance metrics of DQN agents trained on notification text feature

The performance of the DQN agent was assessed using WeAreUs notification data of the 11 participants analysed in the previous section. The data set was split into train (80%) and test (20%) sets. The median accuracy, precision, recall and F1 scores obtained were: 83%, 79%, 77% and 76% respectively, demonstrating that the key phrases extracted from within the notification text fields were indicative of subsequent user actions and a RL agent showed potential for predicting notification dismissal and subsequently could be used for alleviating distractions borne of an influx of notifications.

Whilst the RL agents show potential for replicating participant actions taken toward notifications, this may not be a desired outcome. The agents could be learning poor engagement habits inherent in a participant's previous behaviour with respect to notifications. For example, a participant may be susceptible to enticing notification text and therefore, an agent, trained using their past notification engagements, would also be susceptible, thus reinforcing poorer habits as opposed to improving and encouraging better notification-engagement behaviour. To test this hypothesis, the trained agents were evaluated using two subsets of data from the WeAreUs data set. The first was comprised only of notifications classified as enticing and the second only of unenticing (neutral) notifications. The agents predicted actions of open or dismiss and the subsequent CTR for each agent was recorded. The results, illustrated in Figure 6.8, demonstrate that the median CTR for enticing notifications was much larger than that of neutral notifications.

Therefore the intelligent agents, trained on real participant notification data, were also susceptible to enticing notifications. It is important to note that there are, most likely, a number of confounding factors at play which contribute to subsequent notification action, as discussed in Chapter 3. Further work should explore identifying the notification types in which enticement, positive or neutral, is the cause of subsequent action as well as the impact the notification action has on the digital health of subscribers, such as anxiety or stress. Notifications which are enticing but show to negatively impact a subscriber’s health could be filtered to protect the subscriber. Conversely, notifications which positively impact a subscriber’s health, such as fitness reminders for example, but are inferred to be unenticing could be flagged at moments of creation thus encouraging the creator to craft a more enticing notification which would have a higher likelihood of being opened by subscribers (and not filtered by NMSs).

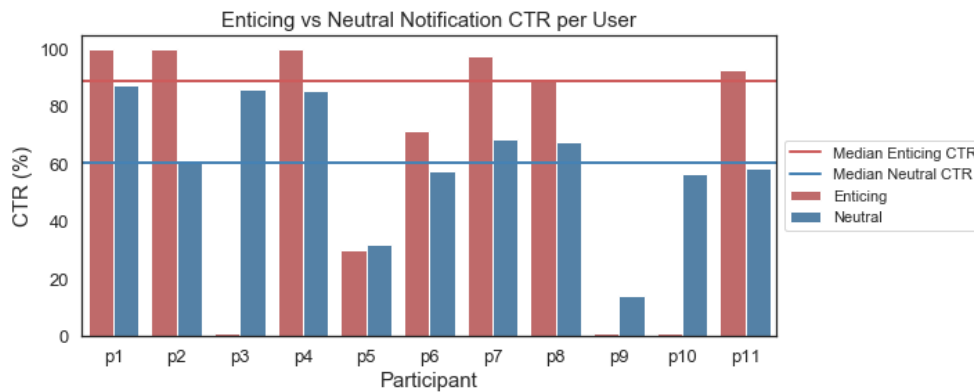


Fig. 6.8 Agent CTR for enticing vs neutral notifications

## 6.4 Enticing vs Summarised Notification Text

In the previous sections, push-notification text was analysed for enticement and its subsequent impact on notification engagement actions were discussed. However, due to the privacy-preserving process executed when gathering data during the WeAreUs study, the quality of the notification text available was a limiting factor. This section introduces a second notification data set, the *Alert* data set, which includes a richer source of text data to evaluate enticement. A summarisation method was also explored as a potential solution to reduce the influence of enticing notification text. Evaluations were performed using the RL agents, trained with

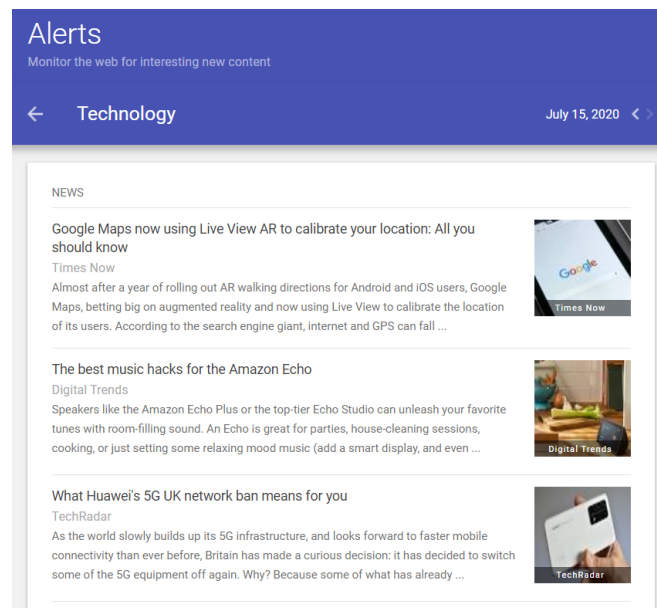


Fig. 6.9 Example of Google Alert notification feed

the WeAreUs notification data, for simulating user actions of opening or dismissing notifications.

### 6.4.1 Procuring Notifications & Summarising Text

*Google Alerts* is a content change detection and notification service [Wik20] enabling users to subscribe to and receive notifications of new stories published to the web based on search terms. For this study, a number of generic search terms were subscribed to and the generated notifications were recorded in real-time. The set of terms subscribed to were:  $\{breaking, entertainment, health, beauty, style, science, sport, tech, world\}$  and the subsequent data set generated was comprised of 35,841 notifications. The Alert notifications were logged between June 27th and September 26th 2019. Each notification was made up of *title* and *content* components. An example of three notifications generated from the *technology* subscription are illustrated in Figure 6.9. The *title* feature is generally a short headline such as "The best music hacks for the Amazon Echo" or "What Huawei's 5G UK network ban means for you". The *content* feature contains a larger amount of text pertaining to the actual story within. Notifications also contained a URL reference to images, as shown in Figure 6.9 however, additional analysis of the images was out of scope for this work.







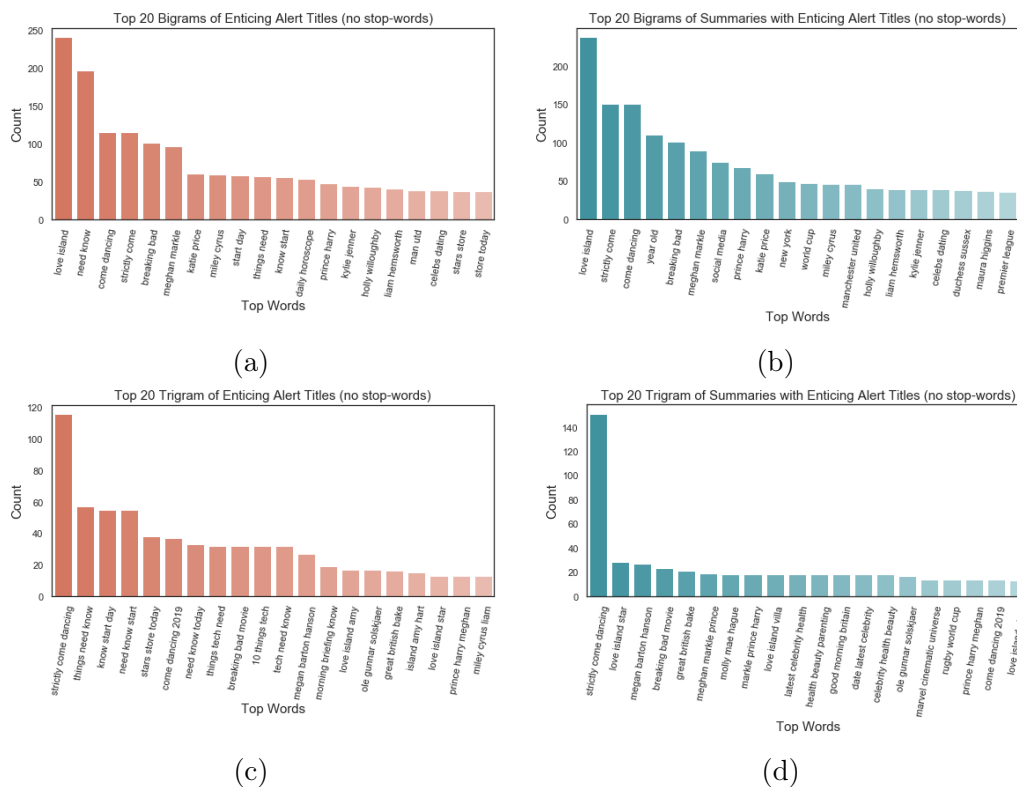


Fig. 6.12 Top bigrams and trigrams extracted from enticing notification texts of the Alert data set

correspond to clickbait-like titles such as *5 Things You Need to Know* or *Start Your Day with*. In contrast, the top summary bigrams did not include these terms and were much less sensational in appearance. Similar results were found in the extracted trigrams illustrated in Figures 6.12c and 6.12d. The enticing notification titles included sequences such as: *things need know*, *need know today*, *things tech need*, *10 things tech*, which were also similar to clickbait-like text exploiting the curiosity gap.

The summaries of enticing titles did not exhibit the same trend of enticing language. To investigate further, a Mann-Whitney U test was run to determine if there were differences in inferred enticement of summaries between notifications classified as enticing or neutral. Distributions of inferred enticement of summaries were not similar for notifications with enticing and neutral titles. The enticement level of summaries extracted from enticing notifications (mean rank = 23771) was statistically significantly higher than the enticement found in summaries extracted from neutral notifications (mean rank = 14872),  $U = 214050316.5$ ,  $z = 77.092$ ,  $p < .001$ . Thus illustrating that summarising the content of an enticing notification would not be a perfect alternative to presenting the enticing

Table 6.9 A sample of notification titles from topics generating largest proportion of enticing notifications

| Topic   | Title                                                              |
|---------|--------------------------------------------------------------------|
| Healthy | Should You Keep Taking Zantac for Your Heartburn?                  |
| Healthy | Can You Eat Orange Peels, and Should You?                          |
| Style   | What should you do if your crush is dating someone else            |
| Style   | 9 eyeshadow brushes to give you the eye makeup look of your dreams |
| Parents | How do you tell your Mom you're dating someone                     |
| Parents | What You Should Do If Someone "Steals" Your Baby Name              |
| Comedy  | We tested CBD dog treats so your hound didn't have to              |
| Comedy  | Before He Spammed You, this Sly Prince Stalked Your Mailbox        |

notification title. Although summarisation might not generate a fully neutral notification title, it could still be of benefit in cases where notification titles were found to be enticing. A sign test with continuity correction was conducted to determine the effect of summarisation on the enticement level of notifications. Of the 12,068 enticing notification in the Alert data set, the summarised text elicited a decrease in the enticement of 8,132 notifications compared to 3,905 notifications which saw an increase in enticement. There was no difference found in 31 notifications. Therefore, there was a statistically significant median decrease (7%) in enticement level when summaries of enticing notifications were used as opposed to their clickbait-like title,  $z = -38.519$ ,  $p < .001$ . A potential solution could subsequently involve classifying a notification as enticing or neutral and, in cases where a notification's text is found to be enticing, a summarised version of its content could be presented in place of the enticing text title.

In addition to uncovering common phrases appearing in enticing notifications, topics of the notification text were also used for identifying trends between enticing and unenticing (neutral) notifications. Figure 6.13 illustrates the distribution of enticement and sentiment across differing Alert notification topics. *Healthy living*, *style*, *parents* and *comedy* were among the top topics generating enticing notifications. Topics which generated fewest enticing notifications included *worldpost*, *sports* and *science*. Table 6.9 demonstrates a sample of enticing Alert notifications from the topics in which enticement was highest. A Kruskal-Wallis H test was run to determine if there were differences in enticement between the 33 topics extracted from the text fields of the Alert notification data set. Distributions of enticement were not similar for all topics, as illustrated by the boxplot

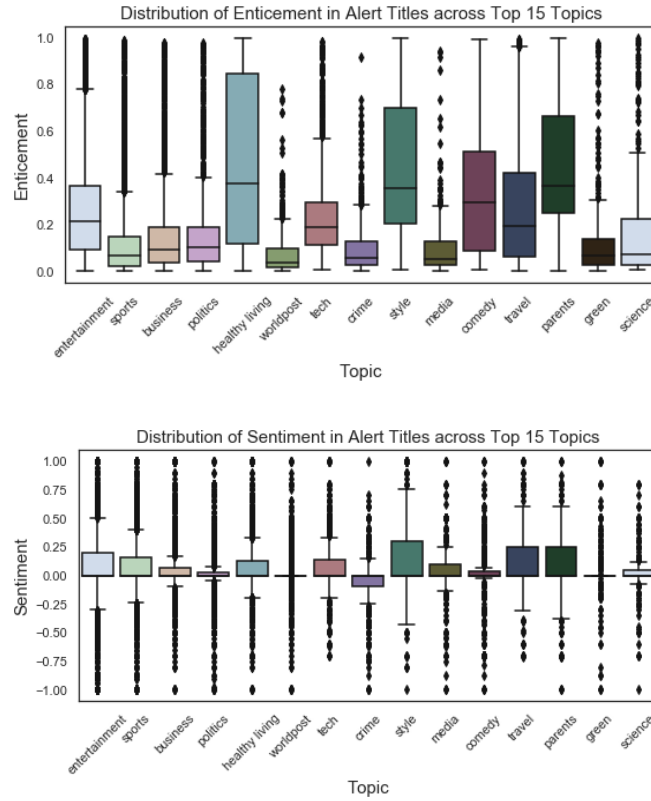


Fig. 6.13 Analysis of enticing notification text found in the Alert data set

in Figure 6.13. The distributions of enticement were statistically significantly different between topics,  $\chi^2(32) = 8139.750$ ,  $p < .001$ . The sentiment across all topics tended toward positive polarity, but for crime which, understandably, was negatively oriented. *Style*, *travel* and *parents* had the most positive notification titles. A Kruskal-Wallis H test was run to determine if there were differences in sentiment polarity between the 33 topics. Distributions of sentiment were not similar for all topics, as illustrated by the boxplot in Figure 6.13. The distributions of sentiment were statistically significantly different between topics,  $\chi^2(32) = 585.846$ ,  $p < .001$ .

### 6.4.3 Influence of Enticing vs Summarised Notification Titles using Simulated Users

To ascertain the effects of enticing *Alert* notifications on user engagement, the DQN agents trained in the previous section were used to simulate participant CTRs of notification titles and summaries, respectively. The *Alert* notification titles and summaries were converted to vectors using the pre-trained BERT-base

model of the sentence-transformers library [RG19]. This pre-processing step was identical to the one used during the training of the agents with the WeAreUs data. The agents were then given each notification text type (title and summary) in turn and tasked with predicting user actions (open or dismiss). Figure 6.14 illustrates the CTR of each participant’s DQN agent when given only enticing notifications. The median CTR for notifications when the agents were given notification text summaries was 48%. In contrast, the median CTR was 57% when the agents were given the notification titles. Therefore, summarising the content of notifications and using the summaries as a feature for managing notifications could reduce the influence of highly enticing notifications on both subscribers and agents trained using their data, thus improving notification management and reinforcing healthier engagement with notifications.

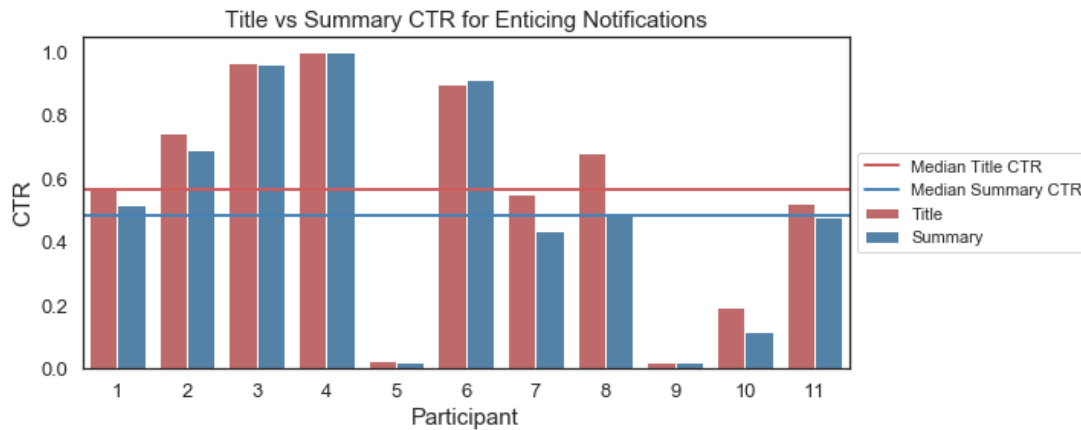


Fig. 6.14 DQN notification CTRs for enticing titles and summarised enticing notifications

## 6.5 Conditional Notification Text Generation

Chapter 4 demonstrated the use of a generative model for creating personalised and persuasive push-notifications. The features generated included a context, such as the *day of week* and *time of day* in which it was inferred a notification would be positively received, as well as notification features which were inferred to perform well in the chosen context. The generated notification features included alert modalities such as *vibration patterns* and *LED colours*, as well as features describing the content, such as *category* and *topic*. A limitation of the work was the absence of actual notification text content. This section addresses this

limitation by discussing the application of an additional generative model for autonomous and conditional creation of notification text titles. The Language Model (LM) chosen was used to generate enticing and neutral notification titles of particular topics, similar to those found in the Alert data set, for the purpose of again evaluating NMSs with respect to enticing text.

### 6.5.1 Language Model Background

*Language modeling is the task of assigning a probability to sentences in a language. Besides assigning a probability to each sequence of words, the language models also assigns a probability for the likelihood of a given word (or a sequence of words) to follow a sequence of words. [Gol17]*

The simplest form of LM learns the probability of individual words (unigram:  $N=1$ ) occurring while more complex models integrate larger sequences (bigrams:  $N=2$ , trigrams:  $N=3$ ) or probabilistic context-free grammars [Sak97]. Once trained on a corpus of text, a LM can be used to generate additional sequences of text, based on the vocabulary of words found within the corpus and their relative likelihood of occurring. Language models are also a key component of larger models applied to standard NLP tasks, such as *Question Answering*, *Summarisation*, *Topic Modeling*, *Machine Translation* and *Fake News Detection*, among others.

On February 14th 2019, the OpenAI group published their paper describing their GPT-2 Language Model. GPT-2, at the time of release, was shown to achieve "*state of the art results on 7 out of 8 tested language modeling datasets in a zero-shot setting*" [Rad+]. Two main features of the model which contributed to its superior performance was size and training data quality.

*"GPT-2 is a large transformer-based language model with 1.5 billion parameters, trained on a data set of 8 million web pages. GPT-2 is trained with a simple objective: predict the next word, given all of the previous words within some text."* [Rad+20]

For the purposes of training GPT-2, a large corpus of text was scraped from the web. Text quality was a priority, hence the scrape was limited to snippets of text verified by humans in-the-wild, ascertained using Reddits karma mechanism. The subsequent data set, *WebText*, contained 8 million documents and was 40GB in size. Four GPT-2 models of increasing size were released, illustrated in Table 6.10.

Table 6.10 The four released model sizes of GPT-2. Table source: [Rad+].

| Parameters | Layers | $d_{model}$ |
|------------|--------|-------------|
| 117M       | 12     | 768         |
| 345M       | 24     | 1024        |
| 762M       | 36     | 1280        |
| 1542M      | 48     | 1600        |

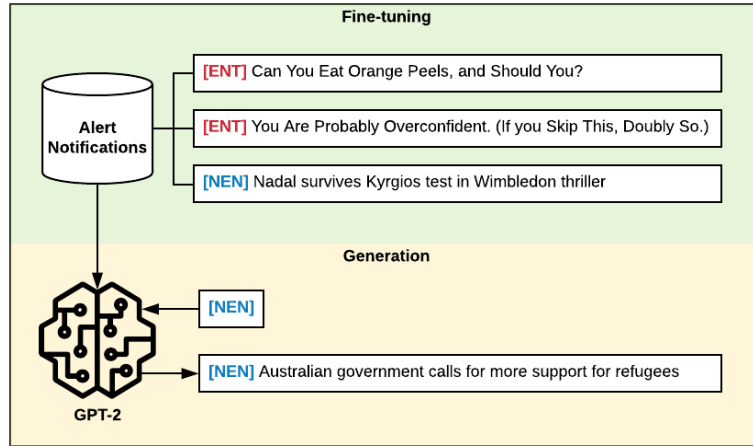


Fig. 6.15 Fine-tuning GPT-2 on Alert notifications

OpenAI decided to stage the release of the models due to ethical concerns over the potential use of GPT-2 [Sol+19].

### 6.5.2 Language Model Implementation

The HuggingFace implementation of GPT-2 was used in this study for generating notification titles [Wol+19]. The LM was fine-tuned on the Alert data set and was used to generate enticing and neutral notification titles on request. The smaller 12-layer 117M parameter model was configured and conditioned on the enticement feature of the Alert data set by inserting an enticement token at the beginning of each notification title. The token [ENT] was used to distinguish enticing notification titles and [NEN] for non-enticing (neutral) notifications. Once fine-tuned, the LM was then seeded with the enticement token indicating whether enticing or neutral notifications should be generated. Figure 6.15 illustrates the process. Model parameter values for *temperature* and *top\_k* were set at 0.7 and 40 respectively as they were found to strike the best balance between diverse and deterministic results.

Table 6.11 Enticing and neutral notification title examples extracted from the GPT-2 generated notification data set

| Entice | Title                                                              |
|--------|--------------------------------------------------------------------|
| Yes    | 17 Pictures That Will Make You Laugh Out Loud                      |
| Yes    | This Is What It's Like To Be A Woman And A Dad                     |
| Yes    | 19 Reasons You'll Love This Photo Of A Man In A Trousers           |
| Yes    | Which 'The Wire' Character Should You Be Based On Your Personality |
| No     | UK must decide next step on Brexit, says France's Europe minister  |
| No     | French footballers lose to United in Euro 2016                     |
| No     | Russian hackers attack 'cyberbullying' forum                       |
| No     | No. 2, Stanford, wins in NCAA Tournament                           |

### 6.5.3 Enticing Notification Text Analysis

25,000 notification titles were generated by the fine-tuned GPT-2 Language Model. The data set was comprised of 12,500 notification titles conditioned to be enticing and 12,500 conditioned to be unenticing (neutral). Table 6.11 illustrates a sample of enticing and neutral notifications generated by the LM. The enticement classifier, created and discussed in Section 6.2.2, was used to additionally infer the enticement of the generated GPT-2 notifications. This was done to evaluate the conditional property of the LM and verify that enticing and neutral notifications were generated correctly on request. Visual inspection of the distribution of inferred enticement suggests that the LM was able to generate enticing and neutral notifications distinctly, as demonstrated in Figure 6.16 where few notifications can be found between both extremes. Furthermore, a Mann-Whitney U test was run to determine if there were differences in inferred enticement between notifications conditionally generated as neutral or enticing. Inferred enticement values for notifications conditioned to be enticing (mean rank = 18473.49) were statistically significantly higher than for notifications conditioned to be neutral (mean rank = 6329.56),  $U = 152787413$ ,  $z = 133.381$ ,  $p < .001$ .

Figure 6.17 illustrates the most frequent words found in notification titles generated by GPT-2 which were conditioned to be neutral (left) and enticing (right). Interestingly, the LM tended toward generating enticing titles which involved questioning the notification reader. For example, phrases such as "Are you" (N=1445), "Do you" (N=1062) and "Can you" (N=648) appear prominently. Phrases associated with titles exploiting the curiosity gap were also discernible: "How to" (N=489), "Need to" (N=400) and "What Happens" (N=394). The

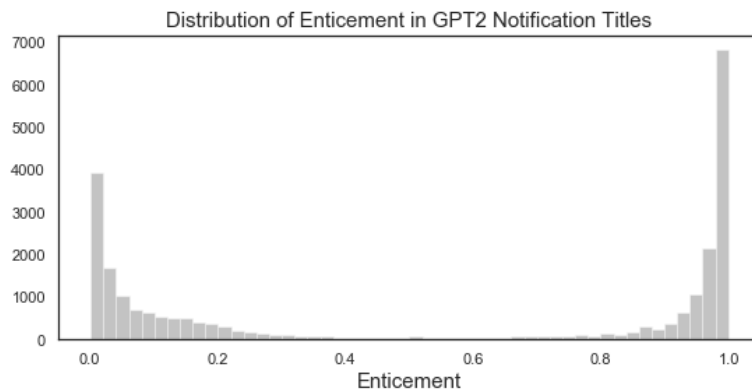


Fig. 6.16 Inferred enticement distribution of enticing and neutral notifications generated by GPT-2

neutral notification titles generated tended to be political in nature, which was unsurprising as the Alert data set assigned low enticement values to titles with topics of *politics* and *worldpost*. In particular, titles referencing "US" (N=2202), "Russia" (N=1188) and "China" (N=517) were amongst the most prominent neutral notification titles generated.

Figures 6.18a and 6.18b illustrate frequent trigrams present in the set of notification titles conditioned to be neutral and enticing, respectively. "To meet with", "New York City" and "court of appeals" were the most frequent trigrams found in neutral titles while "and your favorite", "based on your" and "your zodiac sign" were prominent in enticing titles. Other clickbait-like phrases, which tend to exploit the curiosity gap, were also noticeable in enticing titles:

- *Can You Guess The Best Movie Ever Made*
- *This Is What **Happens** When You Try To Break The Silence*
- *Which "Game of Thrones" **Character Are You** Based On Your Zodiac Sign*
- *We Know **What Your Favorite** Thing To Do On Instagram Is*

With regard to the most frequent topics appearing in neutral and enticing notification titles generated by GPT-2, trends similar to those found in the Alert data set, on which the model was trained, were identified. Figure 6.19 illustrates the distribution of enticement across differing title topics in GPT-2 generated notification titles. In particular, topics with higher enticement values in the Alert data set, such as *healthy living*, *style* and *parents*, were extremely enticing in the generated data set. Titles with topics such as *politics* and *worldpost* were created





Fig. 6.17 Word clouds of GPT-2 data set. Left: conditioned as neutral; Right: conditioned as enticing

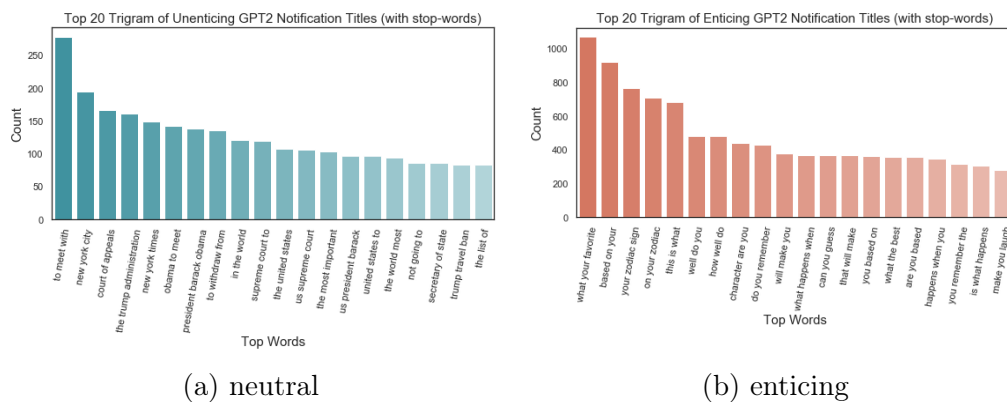


Fig. 6.18 Top trigrams extracted from neutral and enticing notification titles generated by GPT-2

less enticing, which was also consistent with the Alert data set. Interestingly, the titles with a *science* topic, which exhibited relatively low levels of enticement in the Alert data set, were quite high in the generated data set. Further exploration identified a large percentage (63%) of *science* titles referencing the bigram "zodiac sign", which appeared frequently in enticing notifications under different phrasing:

- *Which Character Are You Based On Your **Zodiac Sign***
- *How Well Do You Know Your **Zodiac Sign***
- *We Know Your Name Based On Your **Zodiac Sign***

Neutral notification titles with a *science* topic were also generated, but were a minority:

- *A Russian satellite captures images of the Earth in its orbit*
- *A group of scientists proposes the use of a laser to study how the brain works*
- *A 'small and light' asteroid slams into Earth*

Thus the distribution of enticement values for the *science* topic in the generated data set did not match the Alert data set on which the LM was trained. This demonstrates that further work must ensue to ensure a balanced data set, with respect to title topic, can be created. A Kruskal-Wallis H test was run to determine if there were differences in levels of enticement between the 30 notification title topics. Distributions of enticement were not similar for all topics, as illustrated in Figure 6.19. The distributions of enticement were statistically significantly different between topics,  $\chi^2(29) = 16542.518, p < .001$ .

An analysis of sentiment polarity was also carried out on the GPT-2 generated notification titles. Figure 6.20a illustrates that the notification titles generated were positively oriented. In particular, enticing notifications were shifted slightly more positive than notifications conditionally generated as neutral. Additionally, Figure 6.20b illustrates the distribution of sentiment across titles of differing topics. Consistent with the Alert data set, the GPT-2 generated title topics of *style*, *travel* and *entertainment* had high positive sentiment, while topics *worldpost* and *crime* had lower sentiment values. A Mann-Whitney U test was also run to determine if there were differences in sentiment polarity of titles between notifications conditionally generated as enticing or neutral. Distributions of sentiment polarity for enticing and neutral notification titles were not similar, as illustrated in Figure 6.20a. Title sentiment for enticing notifications (mean rank

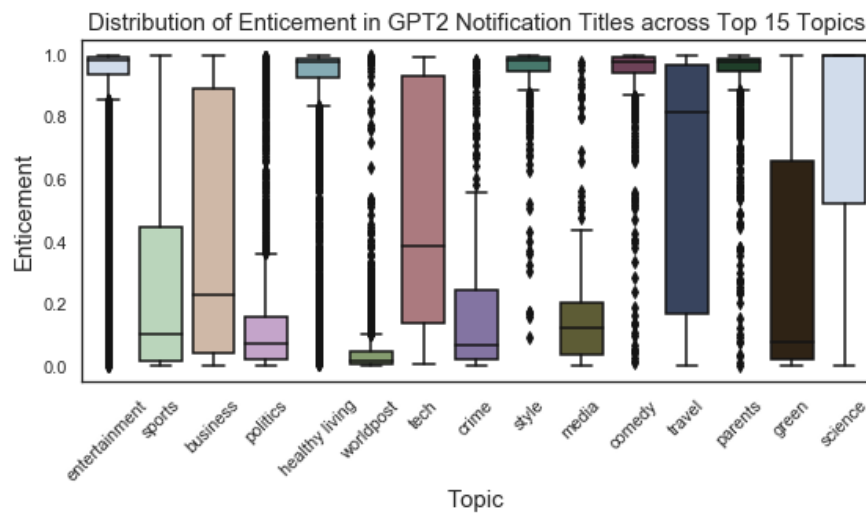


Fig. 6.19 Distribution of enticement over differing title topics generated by GPT-2

= 14155.90) were statistically significantly higher than neutral titles (mean rank = 10693.59),  $U = 98817461.5$ ,  $z = 39.715$ ,  $p < .001$ . This illustrates the potential value of the fine-tuned language model. If leveraged by notification authors at times notification campaigns are developed, the LM could aid in creating notification titles which deliberately do not over-entice subscribers. Alternatively, the LM could also be used to conditionally create notification titles which actively entice subscribers, assuming opening the notification would be of benefit to them, such as in behaviour change interventions. Additionally, the scope to further fine-tune the LM on notification text which the subscriber is most receptive to or finds agreeable, e.g. colloquial language, could further improve the quality of personalised notification messages.

#### 6.5.4 Influence of Conditionally Generated Notification Titles on Simulated Users

Following Section 6.4.3 in which enticing and neutral Alert notification titles were evaluated using simulated participants, this section discusses an identical evaluation process but for the notification titles conditionally generated using the GPT-2 Language Model, thus enabling comparisons to be drawn between both notification text data sets. Once again the DQN agents were used as proxy for simulating participant engagement actions. The GPT-2 notification titles were converted to vector representations using the pre-trained BERT-base model of the sentence-transformers library. The agents were given, in turn, notifications

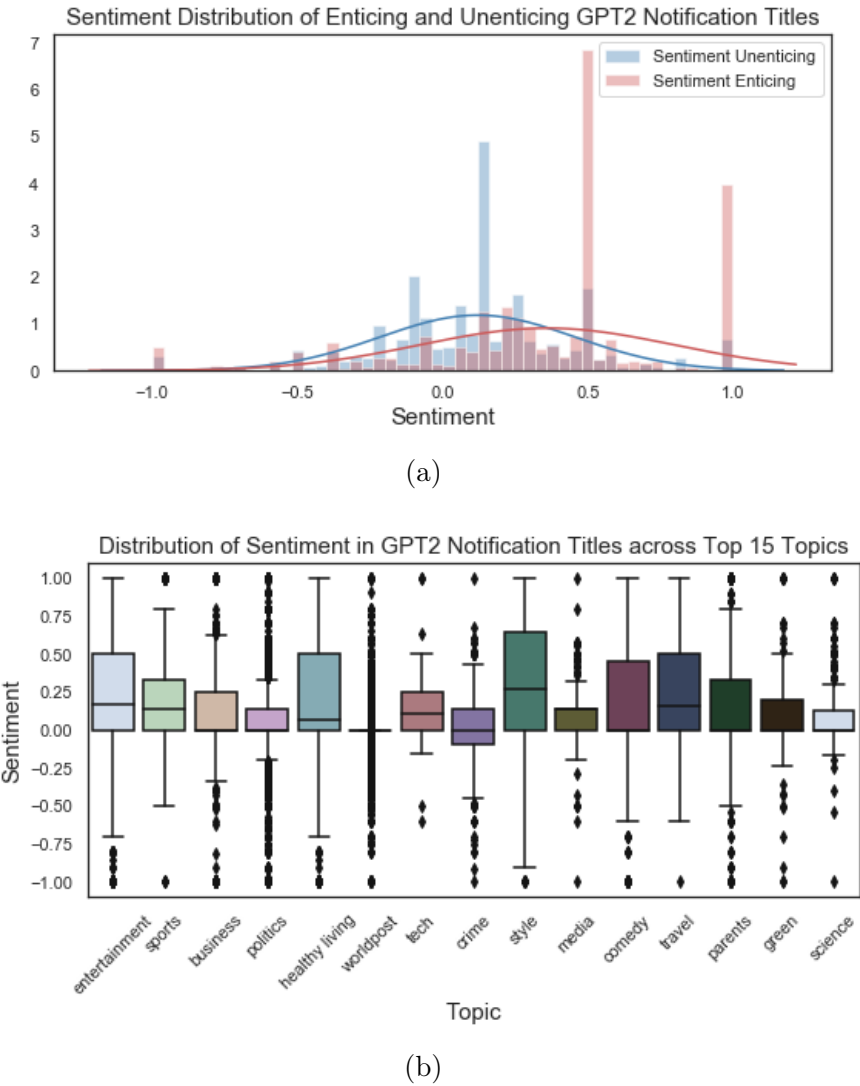


Fig. 6.20 Sentiment analysis of GPT-2 generated notification titles

conditionally generated as enticing or neutral and tasked with predicting user actions (open or dismiss). Figure 6.21 illustrates the CTR of each participant’s DQN agent on enticing and neutral notifications, respectively. The median CTR of conditionally generated enticing notification titles was 59.90%. In contrast, the median CTR of conditionally generated neutral notification titles was 27.40%.

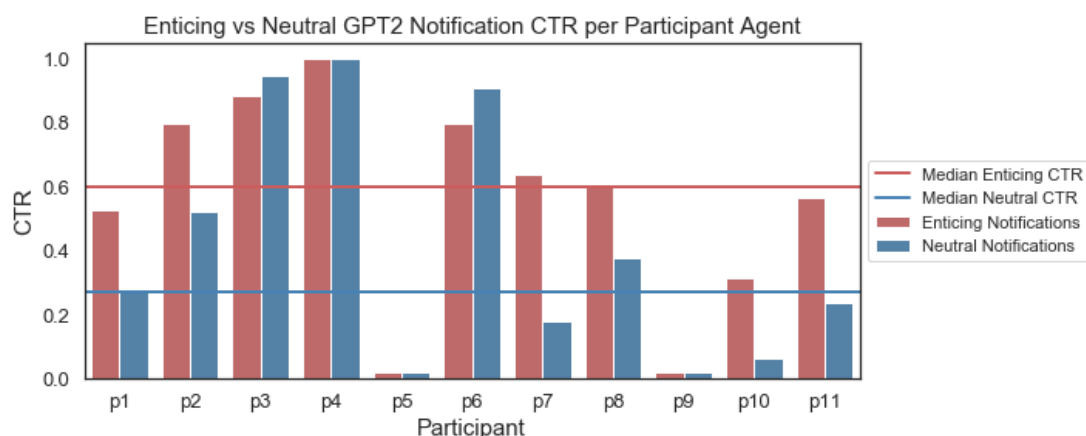


Fig. 6.21 DQN notification CTRs for enticing and neutral GPT-2 generated notification titles

## 6.6 Evaluation of Autonomous Notification Curation In-the-wild

Chapter 3 introduced a study of web push-notifications in-the-wild. Phase 1 of the *Pushd* study was designed to assess different methods of push-notification curation and how it may influence subsequent subscriber engagement. The experiment had a particular focus on examining the impact of enticing text and how it might be autonomously mediated at the point of notification creation as opposed to delivery when, for example, an agent (as described in the previous section) may be tasked with identifying and removing overly-enticing notification content. This is particularly relevant to the *curation* component of the ETHOS-Push framework which strives to ensure notifications are created with empathy for the subscriber i.e. in this case, not using clickbait-like language and attempting to entice clicks for the sake of a higher Click-through Rate. This section discusses an evaluation

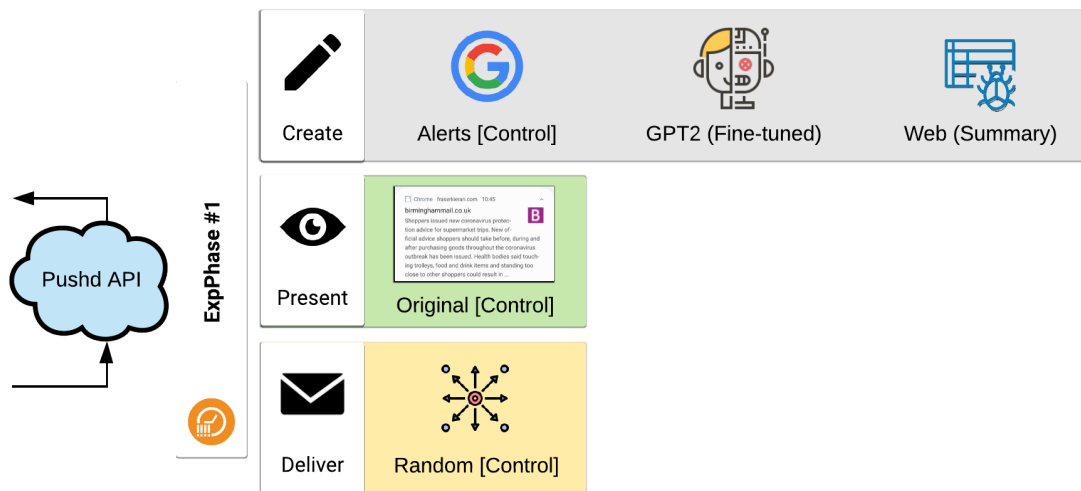


Fig. 6.22 Phase 1 of the Pushd web push-notification study

of various methods related to autonomous creation and empathetic mediation of web push-notifications in-the-wild.

### 6.6.1 Experiment Approach

During phase 1 of the Pushd study, participants were divided into three distinct groups, each with its own method of curating push-notifications, with all other variables, such as presentation format and delivery, fixed at control values to ensure consistent evaluation of the notification source. The three curation methods studied were:

- Alerts

The Google Alerts service, described in Section 6.4.1, was used as the benchmark control for this experiment. A number of categories were selected in the Google Alert console for subscription and the subsequent notifications generated by the service were recorded in real-time. Features such as enticement, topic and sentiment were extracted from the title of the notification and the source URL was the subsequent web page opened if a participant of the study clicked the delivered notification. The steps of curation are illustrated in Figure 6.23a.

- GPT-2

The method of conditional notification text generation described in Section 6.5 was implemented as the second method of notification curation. The GPT-2 language model was trained using the set of notifications collected

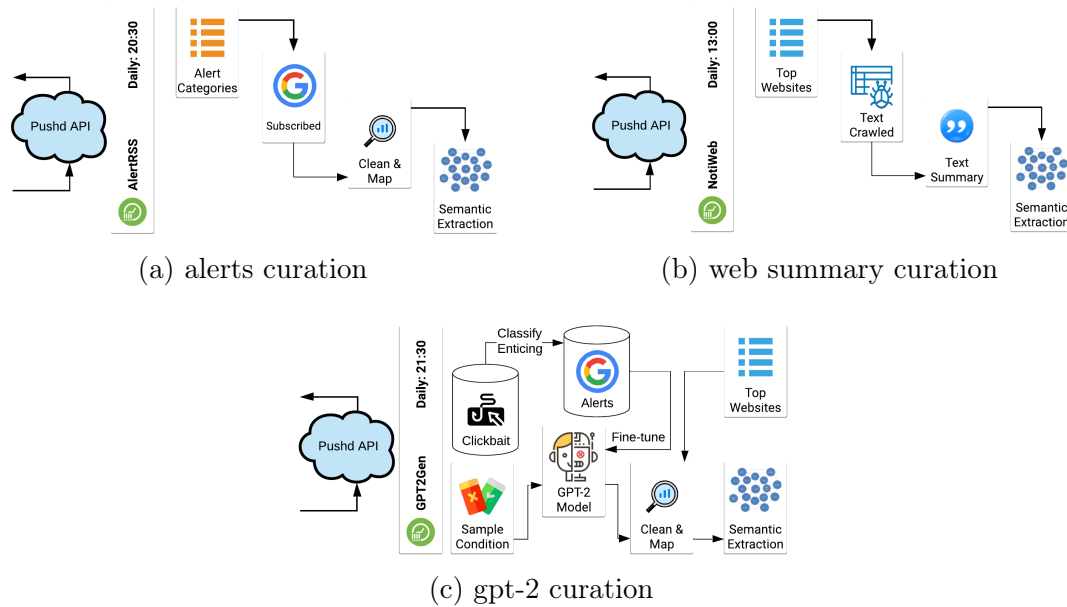


Fig. 6.23 Phase 1 of the Pushd web push-notification study

using the Google Alerts service and it was also conditioned on two features: enticement and topic. The enticement feature was inferred using the classifier introduced in Section 6.2.2. The purpose of this was to assess the competency of the LM at creating notification text content, but also to evaluate whether, by conditioning the LM, the influence of enticement could be controlled and subsequent engagement throttled autonomously. The implementation of this method also facilitated an evaluation of misleading notifications and the impact they have on quality of experience. As no accompanying true web page source was attached to the notifications generated by the language model, a source was randomly assigned from the list of web pages presented to participants during the pre-study questionnaire. The steps of curation are illustrated in Figure 6.23c.

- Web Summary (WebSum)

The final alternate method of curation was a simple web crawler tasked with extracting text from a predefined list of websites and summarizing into a notification format for consumption. This method follows the hypothesis, presented in section 6.4.3, that summarisation of the source may help reduce levels of enticement and improve the quality of experience for subscribers receiving push-notifications. The steps of curation are illustrated in Figure 6.23b.

### 6.6.2 Enticement and Impact on Engagement

A key reason for exploring methods of autonomous notification curation was to explore the feasibility of suppressing enticing text before it was packaged as a notification and pushed at subscribers. This section identifies how enticing notifications influenced subscribers in-the-wild and how by controlling this feature, engagement of subscribers could also be throttled on command. In addition, methods of mediating or suppressing the influence of enticing text was also explored as a key component of empathy within the ETHOS-Push framework.

There were 31, 30 and 25 participants in the Alert, GPT-2 and WebSum groups respectively. 5,121 notifications were pushed during the experiment of which 1,778 (34.7%) were delivered to participants in the Alert group, 1,892 (36.9%) were delivered to participants in the GPT-2 group and 1,451 (28.4%) were delivered to participants in the WebSum group.

A Kruskal-Wallis H test was run to determine if there were differences in levels of enticement between notifications curated by the GPT-2 method, in which the language model was conditioned to generate enticing text, and the WebSum and Alert groups. Distributions of enticement scores were not similar for all groups. The distributions of enticement were statistically significantly different between groups,  $\chi^2(2) = 62.285, p = < .001$ . Pairwise comparisons were performed using Dunn's (1964) procedure with a Bonferroni correction for multiple comparisons. Adjusted p-values are presented. Values are mean ranks unless otherwise stated. This post hoc analysis revealed statistically significant differences in enticement levels between: the Alert (mean rank = 1939.74) and WebSum (mean rank = 2087.29) ( $p = .002$ ) groups; the Alert and GPT-2 (mean rank = 2322.98) ( $p = < .001$ ) groups; and the Web and GPT-2 ( $p < .001$ ) groups. This demonstrates that the GPT-2 language model was successful in generating enticing notification text on command as levels of enticement are significantly higher than both the control (Alert notifications) and web summary groups.

Running the same analysis, but with notifications generated by GPT-2 conditioned to be non-enticing, illustrates that the language model was also successful at generating non-enticing or "neutral" notifications on command. The distributions of enticement were statistically significantly different between groups,  $\chi^2(2) = 296.318, p = < .001$ . Pairwise comparisons were once again performed using Dunn's (1964) procedure with a Bonferroni correction for multiple comparisons. This post hoc analysis revealed statistically significant differences in enticement levels between: the Alert (mean rank = 2237.87) and GPT-2 (mean



rank = 1515.35) ( $p = < .001$ ) groups; and the WebSum (mean rank = 2320.05) and GPT-2 ( $p < .001$ ) groups. In both instances the GPT-2 group notifications had significantly lower levels of enticing text content.

Participants who correctly identified enticing text during the pre-study questionnaire were also tasked with annotating a number of notifications received as: "enticing" or "neutral". A chi-square test for association was conducted between the inferred and annotated enticement levels. All expected cell frequencies were greater than five. There was a small, statistically significant association between both enticement features,  $\chi^2(1) = 7.19, p = .007$ , indicating that the inferred enticement feature aligned with participant perceptions of enticing notifications however, as it was a small effect, further work should be undertaken to improve identification of enticement and, in particular, the underlying features of text which may be indicative of enticement.

To determine the effect of enticing notifications on participants receiving push-notifications from the Alert (control) group, an exact sign test was conducted. The CTR of 18 participants in the Alert group was calculated for incoming notifications classified as enticing and neutral. Of the 18 participants, 10 had increased CTR when the notifications were annotated as enticing, 2 had a decrease in CTR and 6 had no difference in CTR. There was a statistically significant median increase in CTR (11.295%) when participants received notifications with enticing text (33.33%) compared to notifications with neutral text (5%),  $p = .039$ . This indicated that push-notifications containing clickbait-like enticing text were influencing participant CTRs in the control group. As a potential method of mediation, the push-notifications of the WebSum group contained autonomously generated summaries of website text within the push-notification message, instead of clickbait-like headlines, hoping to reduce the effect of enticement on CTRs of subscribers. An exact sign test was conducted for assessment. The CTRs of 23 participants of the WebSum group were calculated for notification text classified as enticing or neutral. Of the 23 participants, 16 had increased CTRs for enticing notifications, 3 had decreased CTRs and 4 had no change in CTR. There was a statistically significant median increase in CTR (16.67%) when participants received notifications with enticing text (25%) compared to notifications with neutral text (6.67%),  $p = .004$ . Although the median increase for the WebSum group was higher than that of the control group, the median CTR for enticing notifications was also much lower (25% compared to 33% in the control group), suggesting that although enticing notifications were created and still influenced

participant CTR in this group, it was at a lower threshold of enticement. This correlates with the offline simulation results discussed in the Section 6.4.3.

This was further validated by comparing participant CTRs across all groups for enticing notifications. A Kruskal-Wallis H test was run to determine if there were differences in CTR due to enticing notifications pushed from the Alert, GPT-2 and WebSum groups. The median CTR between groups was statistically significantly different,  $\chi^2(2) = 8.021, p = .018$ . Subsequently, pairwise comparisons were performed using Dunn's (1964) procedure with a Bonferroni correction for multiple comparisons. This post-hoc analysis revealed a statistically significant difference in median CTR between the GPT-2 (69.05%) and WebSum (25%) ( $p = .042$ ) groups but not quite (assuming a 95% confidence level) between the GPT-2 and Alert (33.3%) ( $p = .052$ ) groups or WebSum and Alert ( $p = 1.0$ ) groups, which will require further exploration.

Unsurprisingly, the competence at identifying enticing clickbait-like text within notifications also influenced subsequent engagement when such notifications appeared. A Mann-Whitney U test was run to determine if there were differences in CTRs for enticing notifications between participants who correctly identified enticing text in the pre-study questionnaire and those who did not. Distributions of CTRs for both groups were not similar. CTRs for participants who did not correctly identify clickbait-like text (mean rank = 37.85) were statistically significantly higher than for participants who did (mean rank = 23.52),  $U = 681.5, z = 2.922, p = .003$ . Thus illustrating how exploitative enticing clickbait-like text can be for those who may not be digitally literate with respect to clickbait, particularly paired with push-notifications. Again, further work should be done to fully assess subscriber competency at detecting enticing text within push-notifications.

### 6.6.3 Participant Assessment of Curation Methods

Participants were presented with a questionnaire for the purposes of assessing the performance of each curation method in a number of areas. Participants were asked submit scores rating how well their method of curation (i.e. the notifications they received) performed in the following areas: diversity, informative and credibility.

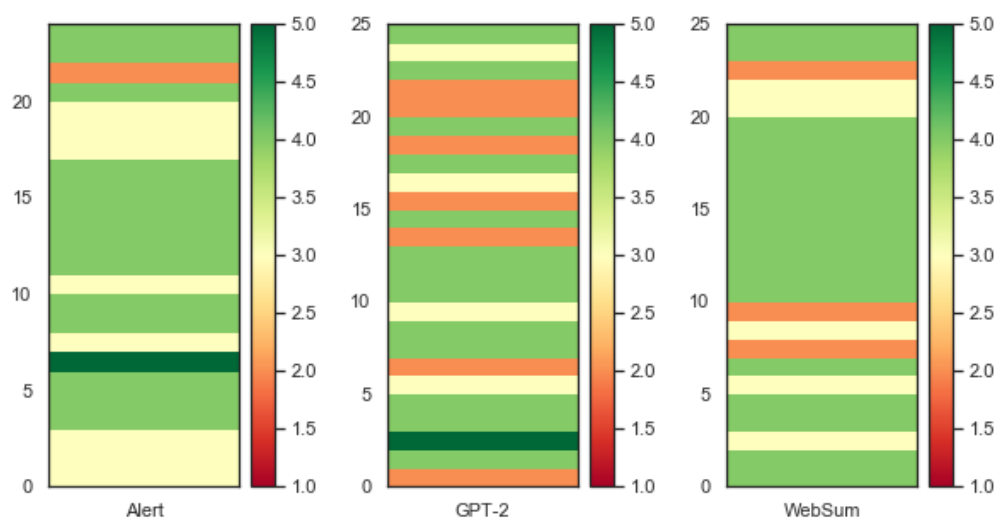


Fig. 6.24 Likert scores (5-point scale) of participants measuring the diversity of notifications from each curation method.

### 6.6.3.1 Diversity

This metric was to assess whether the range of topics and website sources pushed remained consistent and to identify whether alternate methods adversely duplicated notification content or overly favoured particular topics/sources (e.g. echo chambers). Diversity of the notifications was scored by participants on a 5-point Likert scale (1 = very indistinguishable; 5 = very diverse). Figure 6.24 illustrates the average scores recorded for each curation method. All groups fell between *Neutral* and *Somewhat diverse* on average. This demonstrates that the GPT-2 and WebSum methods were at least as competent as the control at generating diverse web push-notifications, albeit with no statistically significant difference. A Kruskal-Wallis H test was run to determine if there were differences in diversity score between the three different curation methods: Alerts, GPT-2 and WebSum. Median scores were not statistically significantly different between groups,  $\chi^2(2) = 3.539, p = .170$ .

### 6.6.3.2 Informative

This metric was to assess whether the GPT-2 and WebSum methods were as informative as the control or if, in the conditioning and mediation of enticing text, notifications became less informative. The level at which notifications were informative was also scored on a 5-point Likert scale (1 = very uninformative, 5 = very informative). Figure 6.25 illustrates the scores recorded for each curation

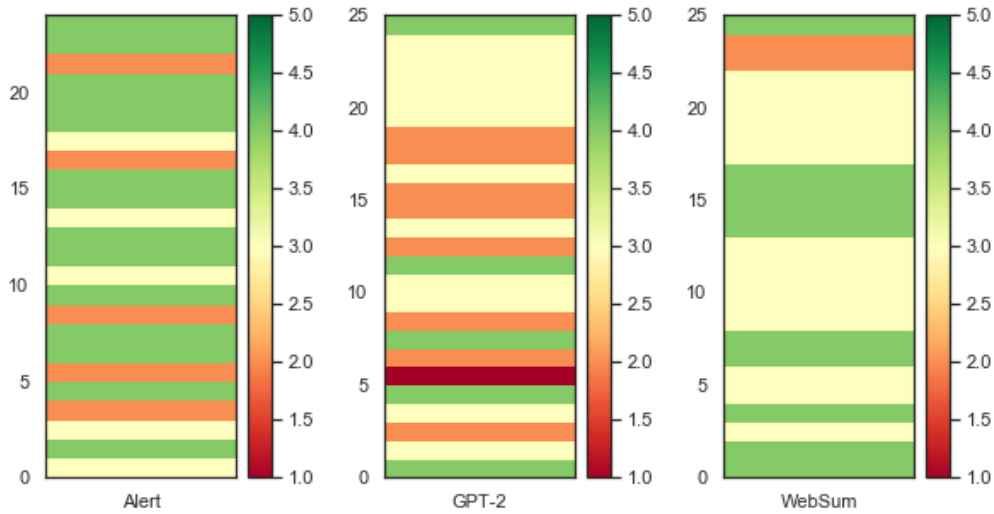


Fig. 6.25 Likert scores (5-point scale) of participants measuring the informativeness of notifications from each curation method.

method. A Kruskal-Wallis H test was run to determine if there were differences in diversity score between the three different curation methods: Alert, GPT-2 and WebSum. Distributions of scores were not similar for all groups. Informative scores were statistically significantly different between the different notification curation methods,  $\chi^2(2) = 9.140, p = .010$ . Subsequently, pairwise comparisons were performed using Dunn's (1964) procedure with a Bonferroni correction for multiple comparisons. This post-hoc analysis revealed statistically significant differences in informative scores between the Alert (mean rank = 42.83) and GPT-2 (mean rank = 27.38) ( $p = .026$ ) methods, and WebSum (42.50) and GPT-2 ( $p = .028$ ) methods, but not between the Alert and WebSum methods. Figure 6.25 illustrates that informative scores were similar for both Alert and WebSum indicating little informative power was lost whilst mediating notification enticement through summarisation. Again, further work should explore this point attempt to validate with statistically significant results.

### 6.6.3.3 Credibility

The credibility scores were used to assess the credibility of text generated/summarised by the GPT-2 and WebSum curation methods compared to the control (Alert) method. The measurement for this score was taken from evaluations carried out by Kreps and McCain [Sol+19; KMB20] when investigating the social risks and benefits of the GPT-2 language model. The credibility index comprises the scaled

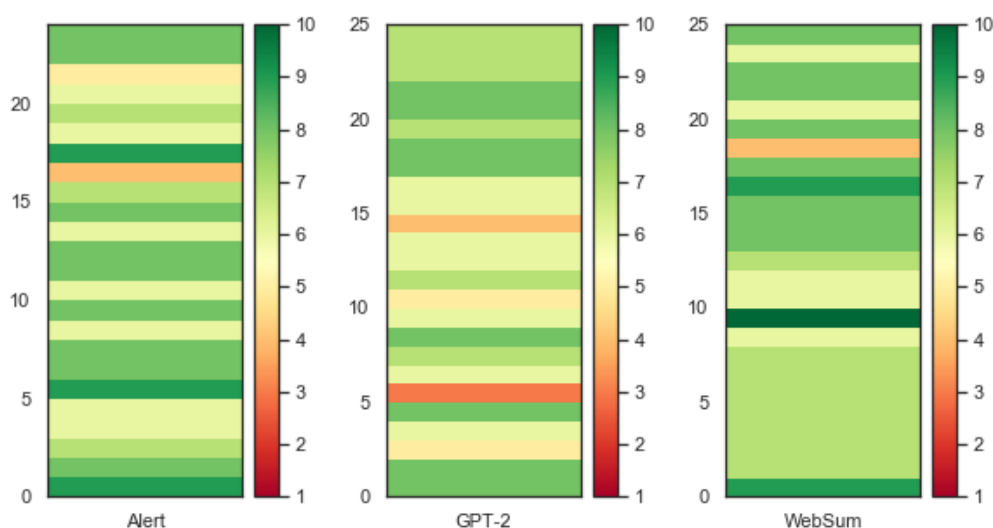


Fig. 6.26 Credibility scores (1-10) of notifications from each curation method.

(between 1-10) sum of three distinct scores: accuracy, clarity and believability (each score was measured on a scale of 1-4, 4 being the best rating e.g. most accurate, clear or believable). Figure 6.26 illustrates the average scores recorded per method. Based on the responses, the GPT-2 method was considered slightly less credible than the Alert or WebSum methods, but still performed well over the majority of participants signifying the competency of the language model at generating realistic notification content. The WebSum method was just as, if not more, credible than the Alert method in this study, demonstrating that autonomous summarisation does not adversely reduce the credibility of notifications. A Kruskal-Wallis H test was run to determine if there were differences in credibility score between the three different curation methods: Alert, GPT-2 and WebSum. Scores were not found to be statistically significantly different between the methods,  $\chi^2(2) = 3.475, p = .176$ . However, examining the individual attributes of credibility, using a Kruskal-Wallis H test for each feature, revealed that believability was statistically significantly different between curation methods  $\chi^2(2) = 6.262, p = .044$ . Figure 6.27 illustrates that believability for the GPT-2 method was poorer on average than WebSum or Alert methods. However, a subsequent post-hoc analysis of pairwise comparisons using Dunn's (1964) procedure with a Bonferroni correction did not reveal any statistically significant differences between pairs. Nonetheless, this study showed that GPT-2 generated notifications were significantly less believable than notifications curated by control (Alert) and summarisation (WebSum) methods, and click-through rates for this

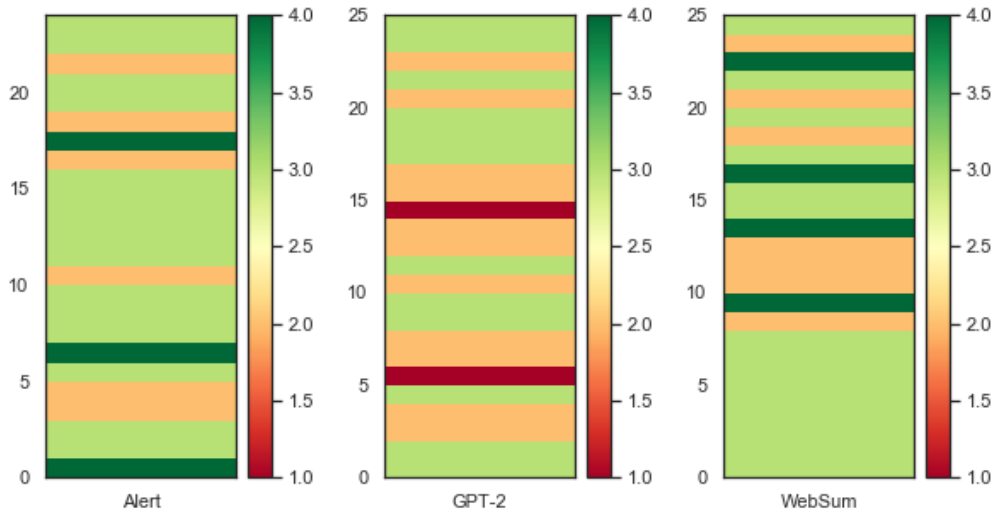


Fig. 6.27 Believability scores (1-4) of notifications from each curation method.

method, when conditioned to be enticing, remained startlingly higher than both alternate methods which did not fabricate content.

## 6.7 Limitations

A limitation of the work described in this chapter includes the granularity of notification text data available. The WeAreUs data set provided textual notification content, however due to disclosure restrictions the quality of the text was greatly reduced thus hindering the effectiveness of applying classification methods. The Alert data set provided a much richer source of text content, however, this data set did not include end-user notification engagement features indicating how a person would react (open or dismiss) to a notification. Simulated users, in the form of DQN agents, were instead used as proxy for evaluation purposes.

## 6.8 Summary

The goal of this chapter, motivated by **RO2** for satisfying **RQ1**, was to explore the prevalence of enticing text (exploiting the curiosity gap) in push-notifications and its impact on notification engagement. Through an analysis of notification data collected in-the-wild, the CTR was identified to be statistically significantly larger for notifications with text classified as enticing than for those classified as not

enticing. Additionally, text enticement was identified as an important feature when predicting notification engagement actions. A summarisation method was proposed and evaluated, in simulated and online environments, as a solution for reducing the influence of enticing notification text on subscribers. Results, providing empirical evidence for **RQ4**, showed that the CTR was reduced when enticing notification text content was summarised. In addition, a Language Model for conditional generation of enticing and neutral notification titles was implemented and fine-tuned upon an existing notification data set. The subsequent evaluation identified that notifications conditioned to be enticing or neutral did elicit accurate enticement levels for their respective category. The generated notification text titles were also used to demonstrate the influence of enticement on simulated users and the potential use of a similar LM for notification creation was presented. The influence of notifications generated by the LM was subsequently demonstrated in-the-wild and evaluated with respect to its influence on engagement and user experience metrics.





# Chapter 7

## Conclusions

Attention is a finite resource and is in growing demand as many of our daily tasks and interactions shift from the physical to virtual world. The push-notification is a prominent tool for capturing attention, therefore, as the number of digital services clamouring for attention have increased, so too have the number of interruptions caused by an unintelligent push. Marketing teams, tasked with creating notification content, implement strategies of persuasion to entice positive engagement to which subscribers of the technology are unaware, yet remain regularly nudged to satisfy other agendas. The studies of push-notifications within literature have observed how subscribers interact when interrupted and provide data-driven approaches for managing notification delivery in order to provide a more intelligent push experience.

This thesis contributes to the body of push-notification literature in the same vein by proposing methods which instill empathy for the subscriber, such that: (i) their data can be used minimally and sustainably; (ii) notification delivery and methods of presentation are contextually adaptive; and (iii) overtly enticing or persuasive elements of notifications are moderated and/or used to facilitate explanation and subscriber scrutiny.

As such, this chapter presents a conclusion for this thesis by discussing the extent to which:

- the Research Objectives, **RO1-RO4**, proposed in Chapter 1 were achieved
- the Research Questions, **RQ1-RQ4**, defined in Chapter 1 were satisfied

- the limitations identified within literature, outlined in Chapter 2, were addressed, thus advancing the state-of-the-art

## 7.1 Summary of Contributions

### 7.1.1 Research Question 1

*How do subscribers act toward, perceive and become influenced by notifications in-the-wild?*

Through a rigorous analysis of literature pertaining to push-notifications, which satisfied **RO1**, it was established notifications provide, through timely interventions, unquestionable value, however it was also shown they also negatively impact productivity and task completion. Their sudden absence evoke adverse feelings of anxiety and fear suggesting an unhealthy dependence. External contextual features of notifications, such as location and activity, have shown to be indicative of subscriber receptivity, however the internal context, such as the notification text and its propensity to be persuasive and enticing, and the transient contexts in which notifications undergo interactions, have not yet been fully explored. Two studies of push-notification engagement in-the-wild were introduced and discussed in Chapter 3 to further investigate how subscribers interact with notifications, perceive their behaviour and manage their relationship with their smartphone.

The contributions of the WeAreUs study, an observational study of Android push-notifications in-the-wild, included identification of a number of significant relationships between transient context features and notification engagement behaviours, such as those between positive engagement, contact relationships and varying environmental characteristics. The work also identified subscriber etiquette with respect to their engagement with, and receptivity to, their smartphone and showed how this changed across contexts. In addition, the work also explored the relationship between goal-directed thinking and smartphone use, identifying patterns of engagement and contexts in which it was strong or found to be lacking. An important contribution of the work was the creation of an abstract text feature which facilitated the gathering, and later use, of redacted notification text content for identifying potentially enticing and persuasive push-notifications.

The Pushd study, which evaluated web push-notifications in-the-wild, contributed to the state-of-the-art by providing an alternate perspective to notification

management, that of the notification *pusher*, and investigated the performance of applying empathy to the creation, delivery and presentation components of notification orchestration, compared to standard, unintelligent push methods currently in widespread use. The study identified the propensity of web users to subscribe to web push-notifications and the preferred parameters of those subscriptions, such as use of data and receptivity to topics and sources. Significantly, the study, whose relevant sections are also discussed in Chapters 4, 5 and 6, show that participant engagements are influenced by both enticing notification text and the contexts in which notifications are delivered, as later defined by a novel application of Cialidini’s principle’s with respect to push-notifications.

### 7.1.2 Research Question 2

*Can a generative approach to push-notifications help facilitate intelligent notification management?*

The NMSs proposed in literature depend on a vast array of rich features pertaining to: the content of the notification; the context in which it was delivered, seen or acted on; and the behaviour of subscribers engaging with it and their device. The data-driven approach exploits sensitive data of the subscriber, albeit to provide them with a better, more intelligent notification experience. This thesis argues that a truly empathetic notification service should rely on exploiting only the minimal amount of data necessary from the subscriber to achieve performance adequate to their needs. To support this ideal, and move the field toward a means of achieving it, Chapter 4 discusses work satisfying **RO3**, in which a generative approach was proposed for creating synthetic notification engagement data, similar by association to the real seed data from which it was generated, but also differentiable to support privacy and exploration. A comprehensive set of quality, utility and privacy metrics were defined to evaluate the synthetic notifications generated and offline evaluations suggested that synthetic notification engagements facilitated a good alternative to constantly harvesting subscriber data, which, as privacy regulations and consent requirements become more pervasive, becomes less sustainable. The generative approach was also deployed in-the-wild, contributing to **RO4**, which provided evidence of the utility of synthetic notifications in practice when applied to agents tasked with optimising the delivery and presentation of web push-notifications. The synthetic data was also used to seed an open-source, custom OpenAI gym environment, *gym-push*, which was

created to facilitate comparative evaluation of User Models and was deployed to the research community through the EvalUMAP workshop series [Con+19].

### 7.1.3 Research Question 3

*Do persuasive principles applied to push-notifications enhance scrutability?*

Few studies have provided subscribers with tools to better understand their push-notification engagement behaviour. Mehrotra et al. in their seminal work [MHM16a; MHM17] provided mobile users with a mechanism to view and control rules which governed their incoming push-notifications, thus keeping the "human-in-the-loop" and supporting explainable AI, which has become an important topic of research since the widespread adoption and pervasiveness of AI technology. To further address the absence of mechanisms by which subscribers can better understand their engagement behaviour, and additionally, to provide a mechanism which can enlighten subscribers as to how they may be persuaded or influenced into certain actions, an approach of extracting scrutable persuasive principles from push-notifications was defined and evaluated in Chapter 5. The work, which contributed to satisfying **RO4**, highlighted that the novel application of persuasive principles applied to push-notifications were indicative of future engagement actions, thus aiding NMSs anticipating subscriber preferences, and were also easily understandable. A tool to promote reflection of notification engagement, the SCRUPLE, built upon the explanations provided by the persuasive features, was deployed in-the-wild and shown to measurably reduce CTRs whilst empowering subscribers to better understand push-notification behaviour.

### 7.1.4 Research Question 4

*Can autonomous moderation, at the point of notification creation, reduce overtly influential notifications?*

Following evidence provided by **RO2** for **RQ1**, it was concluded that enticing notification text content influences the engagement actions of subscribers. Once again, relatively few studies have investigated the persuasive influence of text through notifications. Liu et al. showed in their recent paper how aggressive notifications are widespread throughout app market verticals [Liu+19], which paired with findings from this work, indicates it is a feature of notifications which needs to be tempered in order to facilitate an empathetic push experience for

subscribers. To address this gap, a method of autonomous moderation through summarisation was proposed and shown to reduce the extreme effects of enticing notification text in both simulated and in-the-wild environments, contributing, in part, to the fulfillment of **RO4**.

### 7.1.5 Public Resources

The following is a list of resource contributions, stemming from the work described in this thesis, released publicly to the research community to aid future work, foster collaboration and facilitate comparative evaluation. Unless otherwise stated, the public resources can be found at: <https://ethos-push.github.io/>.

- Mobile push-notification data - a data set comprised of both mobile and push-notification behaviours, created by the generative models described in this work.
- Mobile push-notification generative models - the fully trained generative models which can be used for creating new instances of mobile and push-notification behaviour.
- Web push-notification data - a data set comprised of web push-notification behaviour and user preferences, created by the generative models described in this work.
- Web push-notification generative models - the fully trained generative models which can be used for creating new instances of web push-notification behaviour and user preferences.
- Gym-push - a custom OpenAI Gym environment for improved push-notification management. This PyPI library includes synthetic notification data with which to train, metrics with which to evaluate results and benchmarks with which to compare performance with other researchers. The Gym-push repository is located at: <https://github.com/kieranfraser/gym-push>

## 7.2 Discussion

The primary thesis of this work is that *attributes describing notification engagement behaviour, derived and generated using minimal/synthetic data, can be used*

*to autonomously orchestrate the tasks of empathetic push-notification creation, delivery and presentation whilst facilitating scrutability.*

Borne from the evidence provided by **RO1** and **RO2**, four requirements (**R1-R4**) which align with this thesis were defined as key pillars of a novel approach toward implementing empathetic push-notifications, named ETHOS-Push (EmpaTHetic Orchestration of Scrutable Push-notifications). Each requirement focused on inserting empathetic intelligence at key moments of the push journey from *pusher* to *subscriber*. The methods proposed to satisfy the requirements for the ETHOS-Push approach were designed and evaluated throughout the work completed in this thesis:

- **R1:** *Notification content does not overtly entice engagement*

Chapters 5 and 6 both evaluate the persuasive and enticing influence of push-notifications on subscribers. Chapter 5 proposed an approach of mediating its effect through increased awareness using the SCRUPLE tool. Chapter 6 proposed an approach of identifying and autonomously moderating enticing text content. Both approaches were shown to reduce the impact of overtly persuasive and enticing push-notifications on subscribers.

- **R2:** *Past subscriber engagement data is used minimally to inform intelligent systems*

Chapter 4 evaluates features and algorithms used by intelligent agents to manage push-notifications on behalf of subscribers. A generative approach was proposed and evaluated as a means of decoupling the dependency of intelligent agents on real notification engagement data. The approach was shown to provide agents with quality data and offer a feasible alternative to the use of real subscriber data.

- **R3:** *Notification presentation is personalised and adaptive to varying contexts*

Chapter 5 proposed a method of intelligently adapting notification presentation templates to the context of the notification content in order to once again assuage persuasive strategies inherent in notification content. The approach showed that intelligently summarised content showed promise with respect to subscriber receptivity and in contexts in which privacy is a key concern.

- **R4:** *Notification influence is scrutable and understandable facilitating awareness*

Chapter 5 implemented an approach of presenting explainable features to subscribers to promote greater awareness of the persuasive elements of their incoming notifications. The SCRUPLE tool was found to be both useful and understandable to subscribers and had a measurable impact on their engagement with push-notifications.

## 7.3 Future Directions

The work presented in this thesis provides a number of interesting avenues and opportunity for future work. Due to the finite scope in which this work was undertaken, further exploration and investigation was not possible. Therefore they are listed here to potentially inform and direct future work in this area:

- Chapter 4 introduced a generative approach capable of producing notification engagement data for subscribers. Yet to be investigated is the potential for combining separate subscriber data sets, with differing weighted proportions, in order to generate push engagement data for a melded persona. This could also help focus exploration of the notification search space and improve the generation of realistic notifications. Additionally, with respect to behaviour intervention, the melded approach could be used to incrementally guide subscribers exhibiting poor behaviour with notifications toward another subscriber deemed to have a healthier relationship.
- Chapter 4 also touched on the creation of *gym-push*, a custom OpenAI gym environment in which NMSs can train in standard environments, seeded by the data created through the generative approach proposed in this thesis. Further work could be done to provide additional environments to this gym which support a wide range of notification, context and smartphone features enabling NMSs to train and improve in offline environments and facilitate comparative evaluation which would improve the progress of the field, both in academia and the in the wider community of mobile computing.
- This thesis also discussed an approach toward using *minimal* data from subscribers and Chapter 4 demonstrated the number of notifications necessary to achieve adequate utility, quality and privacy scores for the generated data. Missing from the discussion was the concept of a transient subscriber persona and the subsequent impact it has on the generation of notification data. As

subscriber preferences fluctuate day-to-day, week-to-week and year-to-year, how can the generative model adapt to these changes if it is to use *minimal* data. For instance, a small seed of real data to train the generative model might produce adequate notification quality for a week, but the subscriber's preferences might change to such a degree that the quality reduces after that period. Therefore, how can we ensure the generative model changes and grows in tandem with the subscriber, in such a way that also minimises the exploitation of data.

- Chapter 5 introduced the SCRUPLE tool for promoting awareness of notification engagement data to evaluate its impact on engagement. The tool was a basic representation of the persuasive principles, defined by Cialdini, and applied to push-notifications. Future work should consider a much more comprehensive scrutable mechanism with which to visualise, explain and support subscribers managing their notifications and provide them with intelligent tools which highlight potential dark interaction patterns applied against them, such as the use of disinformation, clickbait and deep fakes.
- Future work should also consider the dissemination of push-notifications in-the-wild. Push-notification systems are becoming increasingly automated. An interaction by a single person on a social media site might cause hundreds or thousands of push-notifications to be sent to users subscribed to the person interacting, the topic of interaction or simple due to being a member of the social media site. Interestingly, the person causing this massive *push* of notifications is mostly unaware of the disruption they may have caused others due to their interaction. Is it their fault? Should they feel responsible for the interruptions they inadvertently caused? We each have a *carbon footprint* which describes the total amount of greenhouse gas emissions caused by our actions. A push or interruption footprint could describe our push-notification emissions and the disruptions they have caused, which could create awareness and cultivate a culture of acting with increased empathy on digital platforms.

## 7.4 Coding Effort

The algorithms, apps and analyses described in this Thesis, unless otherwise explicitly stated, were coded and implemented solely by the candidate of this



thesis for the purpose of exploring the empathetic orchestration of scrutable push-notifications. Where applicable, the code has been uploaded to GitHub at <https://github.com/ethos-push/thesis-resources> for reference. The coding effort included:

- A Java program implementing a Mamdami Fuzzy Logic Controller and Particle Swarm Optimisation for transparent push-notification management
- An Android app (WeAreUs) implementing the Experience Sampling Method and background sensing for unobtrusive observation of mobile and push-notification behaviour
- A Java App, hosted on Heroku, implemented using the Spring Framework, and used to provide authentication, encryption and Natural Language Processing to the WeAreUs app
- A number of python scripts for cleaning and preparing push-notification features for Machine Learning (e.g. extracting NLP features from text, converting notification objects to vector representations)
- A number of modeling scripts implementing, evaluating and visualising ML algorithms applied to managing and generating push-notifications
- A custom OpenAI gym environment (Gym-push), implemented in Python and released on PyPI, for training intelligent agents to recommend optimal push-notification features for end-users
- A Progressive Web App (PWA), implementing web push-notifications and including service workers for tracking web push-notification behaviour
- A Flask REST API, hosted on Heroku, and used by the PWA for authentication and encryption. It was also used for logging notification behaviour and facilitating shared resources between the PWA and the back-end server hosting intelligent models.
- A Flask REST API, hosted on *Microsoft Azure*, which facilitated training of intelligent models for later inference when creating, scheduling and presenting personalised web push-notifications.
- A number of Batch jobs, hosted on Heroku, implemented in Python which handled the pushing of notifications to end-user devices, crawling websites

for new push content, subscribing to Google Alert RSS feeds and launching new training jobs via the back-end server hosted on Azure.

## 7.5 Final Remarks

It is hoped, by the author of this thesis, that future methods of push-notification orchestration, deployed in-the-wild, will adopt an approach which is empathetic toward push subscribers, such as that expressed by ETHOS-Push, and facilitates healthier communication channels between those pushing and those receiving notifications.

It is also hoped that the synthetic data generated as part of this work, and method of generation, further supports advances within the field of Notification Management by encouraging collaboration through freely sharing notification engagement data and facilitates comparative evaluation within the NM research community.

# References

- [ACM] ACM. *About the ACM Digital Library*. URL: <https://dl.acm.org/about> (visited on 02/15/2020).
- [AB18] A. Adadi and M. Berrada. “Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI)”. In: *IEEE Access* 6 (2018), pp. 52138–52160.
- [Aha+11] Nadav Aharony et al. “Social fMRI: Investigating and Shaping Social Mechanisms in the Real World”. In: *Pervasive Mob. Comput.* 7.6 (Dec. 2011), pp. 643–659. ISSN: 1574-1192. DOI: 10.1016/j.pmcj.2011.09.004. URL: <http://dx.doi.org/10.1016/j.pmcj.2011.09.004>.
- [And+19] Christoph Anderson et al. “The impact of private and work-related smartphone usage on interruptibility”. In: *Adjunct Proceedings of the 2019 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2019 ACM International Symposium on Wearable Computers*. 2019, pp. 1058–1063.
- [And20a] Android. *Activity Recognition API*. 2020. URL: <https://developers.google.com/location-context/activity-recognition> (visited on 06/15/2020).
- [And20b] Android. *Android 8.0 Behaviour Changes*. 2020. URL: <https://developer.android.com/about/versions/oreo/android-8.0-changes> (visited on 06/15/2020).
- [And20c] Android. *Android Geofences*. 2020. URL: <https://developer.android.com/training/location/geofencing> (visited on 06/15/2020).
- [And20d] Android. *Create and Manage Notification Channels*. 2020. URL: <https://developer.android.com/training/notify-user/channels#importance>.
- [And20e] Android. *Environment Sensors*. 2020. URL: <https://developer.android.com/reference/android/hardware/SensorManager> (visited on 06/15/2020).
- [And20f] Android. *Google APIs for Android: DetectedActivity*. 2020. URL: <https://developers.google.com/android/reference/com/google/android/gms/location/DetectedActivity#constants> (visited on 06/15/2020).
- [And20g] Android. *Notification Listener Service*. 2020. URL: <https://developer.android.com/reference/android/app/NotificationManager> (visited on 06/15/2020).

- [And20h] Android. *UsageStatsManager*. 2020. URL: <https://developer.android.com/reference/android/app/usage/UsageStatsManager> (visited on 06/15/2020).
- [And20i] Android. *VibrationEffect*. 2020. URL: <https://developer.android.com/reference/android/os/VibrationEffect>.
- [App20] Google Apple. *Privacy-Preserving Contact Tracing*. 2020. URL: <https://www.apple.com/covid19/contacttracing> (visited on 06/15/2020).
- [ACB17] Martin Arjovsky, Soumith Chintala, and Léon Bottou. “Wasserstein generative adversarial networks”. In: *International conference on machine learning*. PMLR. 2017, pp. 214–223.
- [Aus11] Peter C Austin. “An introduction to propensity score methods for reducing the effects of confounding in observational studies”. In: *Multivariate behavioral research* 46.3 (2011), pp. 399–424.
- [APT19] Ravit Avraham Bahir, Yisrael Parmet, and Noam Tractinsky. “Effects of visual enhancements and delivery time on receptivity of mobile push notifications”. In: *Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems*. 2019, pp. 1–6.
- [Bas+17] Sarah Bashirieh et al. “Nudge Your Workforce: A Study on the Effectiveness of Task Notification Strategies in Enterprise Mobile Crowdsourcing”. In: *Proceedings of the 25th Conference on User Modeling, Adaptation and Personalization*. 2017, pp. 4–12.
- [Bea+19] Brett K Beaulieu-Jones et al. “Privacy-preserving generative deep neural networks support clinical data sharing”. In: *Circulation: Cardiovascular Quality and Outcomes* 12.7 (2019), e005122.
- [BMI11] Stephen Bell, Alisdair McDiarmid, and James Irvine. “Nodobo: Mobile phone as a software sensor for social network research”. In: *Vehicular Technology Conference (VTC Spring), 2011 IEEE 73rd*. IEEE. 2011, pp. 1–5.
- [Bi+20] Qifang Bi et al. “Epidemiology and transmission of COVID-19 in 391 cases and 1286 of their close contacts in Shenzhen, China: a retrospective cohort study”. In: *The Lancet Infectious Diseases* (2020).
- [Bid+18] Niranjana Bidargaddi et al. “Predicting which type of push notification content motivates users to engage in a self-monitoring app”. In: *Preventive Medicine Reports* 11 (July 2018). DOI: 10.1016/j.pmedr.2018.07.004.
- [Blo53] Benjamin S Bloom. “Thought-processes in lectures and discussions”. In: *The Journal of General Education* 7.3 (1953), pp. 160–169.
- [Böh+14] Matthias Böhmer et al. “Interrupted by a phone call: exploring designs for lowering the impact of call notifications for smartphone users”. In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. 2014, pp. 3045–3054.
- [Bon20] Emily Bonnie. *What 300 Billion Push Notifications Say About Mobile User Engagement*. 2020. URL: <https://clevertap.com/blog/push-notification-report/> (visited on 01/20/2021).

- [Bor17] Nicolas Boring. *Library of Congress*. 2017. URL: <https://www.loc.gov/law/foreign-news/article/france-right-to-disconnect-takes-effect/> (visited on 07/10/2020).
- [Bre01] Leo Breiman. “Random forests”. In: *Machine learning* 45.1 (2001), pp. 5–32.
- [Bre+84] Leo Breiman et al. *Classification and regression trees*. CRC press, 1984.
- [Bri20] Karl Bridge. *Microsoft Docs: Understanding and interpreting lux values*. 2020. URL: <https://docs.microsoft.com/en-us/windows/win32/sensorsapi/understanding-and-interpreting-lux-values> (visited on 06/15/2020).
- [Bri] Harry Brignull. *Dark Patterns*. URL: <https://www.darkpatterns.org/> (visited on 07/10/2020).
- [Bro+16] Greg Brockman et al. *OpenAI Gym*. 2016. eprint: arXiv:1606.01540.
- [Car+19] Ana Caraban et al. “23 ways to nudge: A review of technology-mediated nudging in human-computer interaction”. In: *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. 2019, pp. 1–15.
- [Cha+16] Abhijnan Chakraborty et al. “Stop clickbait: Detecting and preventing clickbaits in online news media”. In: *2016 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*. IEEE. 2016, pp. 9–16.
- [Cha16] Anupam Chander. “The racist algorithm”. In: *Mich. L. Rev.* 115 (2016), p. 1023.
- [CCS19] Yung-Ju Chang, Yi-Ju Chung, and Yi-Hao Shih. “I Think It’s Her: Investigating Smartphone Users’ Speculation about Phone Notifications and Its Influence on Attendance”. In: *Proceedings of the 21st International Conference on Human-Computer Interaction with Mobile Devices and Services*. 2019, pp. 1–13.
- [Che06] Chia-Chong Chen. “A PSO-based method for extracting fuzzy rules directly from numerical data”. In: *Cybernetics and Systems: An International Journal* 37.7 (2006), pp. 707–723.
- [Che+16] Jiachen Chen et al. “CNS: content-oriented notification service for managing disasters”. In: *Proceedings of the 3rd ACM Conference on Information-Centric Networking*. 2016, pp. 122–131.
- [CG16] Tianqi Chen and Carlos Guestrin. “Xgboost: A scalable tree boosting system”. In: *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*. 2016, pp. 785–794.
- [CGM98] Hugh A Chipman, Edward I George, and Robert E McCulloch. “Bayesian CART model search”. In: *Journal of the American Statistical Association* 93.443 (1998), pp. 935–948.
- [Cho+16] Minsoo Choy et al. “Looking back on the current day: interruptibility prediction using daily behavioral features”. In: *Proceedings of the 2016 ACM international joint conference on pervasive and ubiquitous computing*. 2016, pp. 1004–1015.

- [Chu+15] Karen Church et al. “Understanding the challenges of mobile phone usage data”. In: *Proceedings of the 17th International Conference on Human-Computer Interaction with Mobile Devices and Services*. 2015, pp. 504–514.
- [Cia16] Robert Cialdini. *Pre-suasion: A revolutionary way to influence and persuade*. Simon and Schuster, 2016.
- [Cia87] Robert B Cialdini. *Influence*. Vol. 3. A. Michel Port Harcourt, 1987.
- [CC07] Robert B Cialdini and Robert B Cialdini. *Influence: The psychology of persuasion*. Vol. 55. Collins New York, 2007.
- [Clo20a] Google Cloud. *Content Categories*. 2020. URL: <https://cloud.google.com/natural-language/docs/categories> (visited on 06/15/2020).
- [Clo20b] Google Cloud. *Natural Language*. 2020. URL: <https://cloud.google.com/natural-language#natural-language-api-demo> (visited on 06/15/2020).
- [Clo20c] Google Cloud. *Place Types*. 2020. URL: [https://developers.google.com/places/supported\\_types](https://developers.google.com/places/supported_types) (visited on 06/15/2020).
- [Clo20d] Google Cloud. *Places*. 2020. URL: <https://cloud.google.com/maps-platform/places> (visited on 06/15/2020).
- [Coc07] William G Cochran. *Sampling techniques*. John Wiley & Sons, 2007.
- [Coc54] William G Cochran. “Some methods for strengthening the common  $\chi^2$  tests”. In: *Biometrics* 10.4 (1954), pp. 417–451.
- [Col+15] Patrick Colp et al. “Protecting data on smartphones and tablets from memory attacks”. In: *Proceedings of the Twentieth International Conference on Architectural Support for Programming Languages and Operating Systems*. 2015, pp. 177–189.
- [Com20] Apache Commons. *JaroWinklerSimilarity (Apache Commons Text 1.8 API)*. 2020. URL: <https://commons.apache.org/proper/commons-text/apidocs/org/apache/commons/text/similarity/JaroWinklerSimilarity.html> (visited on 06/15/2020).
- [Con+19] Owen Conlan et al. “A User Modeling Shared Challenge Proposal”. In: *Experimental IR Meets Multilinguality, Multimodality, and Interaction*. Ed. by Fabio Crestani et al. Cham: Springer International Publishing, 2019, pp. 171–177. ISBN: 978-3-030-28577-7.
- [CDS15] Charles Consel, Lucile Dupuy, and Hélène Sauzéon. “A unifying notification system to scale up assistive services”. In: *Proceedings of the 17th international ACM SIGACCESS conference on Computers & accessibility*. 2015, pp. 77–87.
- [CDM18] Fulvio Corno, Luigi De Russis, and Alberto Monge Roffarello. “AwareNotifications: Multi-device semantic notification handling with user-defined preferences”. In: *Journal of Ambient Intelligence and Smart Environments* 10.4 (2018), pp. 327–343.
- [CDM15] Fulvio Corno, Luigi De Russis, and Teodoro Montanaro. “A context and user aware smart notification system”. In: *2015 IEEE 2nd World Forum on Internet of Things (WF-IoT)*. IEEE. 2015, pp. 645–651.

- [CDM19] Fulvio Corno, Luigi De Russis, and Teodoro Montanaro. “XDN: cross-device framework for custom notifications management”. In: *Computing* 101.11 (2019), pp. 1735–1761.
- [Cov99] Thomas M Cover. *Elements of information theory*. John Wiley & Sons, 1999.
- [Cro17] Crowdsignals. *A massive new mobile data collection*. 2017. URL: <http://crowdsignals.io/> (visited on 04/19/2017).
- [CL14] Mihaly Csikszentmihalyi and Reed Larson. “Validity and reliability of the experience-sampling method”. In: *Flow and the foundations of positive psychology*. Springer, 2014, pp. 35–54.
- [Cze+16] Mary Czerwinski et al. “Challenges for designing notifications for affective computing systems”. In: *Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing: Adjunct*. 2016, pp. 1554–1559.
- [DB01] Thomas H Davenport and John C Beck. “The attention economy”. In: *Ubiquity* 2001.May (2001), 1–es.
- [Del17] Deloitte. *Global mobile consumer trends, 2nd edition*. 2017. URL: <https://www2.deloitte.com/content/dam/Deloitte/us/Documents/technology-media-telecommunications/us-global-mobile-consumer-survey-second-edition.pdf>.
- [Dev+18] Jacob Devlin et al. “Bert: Pre-training of deep bidirectional transformers for language understanding”. In: *arXiv preprint arXiv:1810.04805* (2018).
- [Dha20] Shivangi Dhawan. “Online learning: A panacea in the time of COVID-19 crisis”. In: *Journal of Educational Technology Systems* 49.1 (2020), pp. 5–22.
- [Doh+18] Kevin Doherty et al. “A mobile app for the self-report of psychological well-being during pregnancy (BrightSelf): qualitative design study”. In: *JMIR mental health* 5.4 (2018), e10007.
- [DR+14] Cynthia Dwork, Aaron Roth, et al. “The algorithmic foundations of differential privacy.” In: *Foundations and Trends in Theoretical Computer Science* 9.3-4 (2014), pp. 211–407.
- [Ela+17] Hanan Elazhary et al. “Context management for supporting context-aware Android applications development”. In: *International Journal of Interactive Mobile Technologies (iJIM)* 11.4 (2017), pp. 186–201.
- [Elh+16] Jon D Elhai et al. “Fear of missing out, need for touch, anxiety and depression are related to problematic smartphone use”. In: *Computers in Human Behavior* 63 (2016), pp. 509–516.
- [Eng07] Andries P Engelbrecht. *Computational intelligence: an introduction*. John Wiley & Sons, 2007.
- [EN14] Yaniv Erlich and Arvind Narayanan. “Routes for breaching and protecting genetic privacy”. In: *Nature Reviews Genetics* 15.6 (2014), pp. 409–421.

- [Esm07] Ahmed Ali Abdalla Esmine. “Generating fuzzy rules from examples using the particle swarm optimization algorithm”. In: *7th International Conference on Hybrid Intelligent Systems (HIS 2007)*. IEEE. 2007, pp. 340–343.
- [EHR17] Cristóbal Esteban, Stephanie L Hyland, and Gunnar Rätsch. “Real-valued (medical) time series generation with recurrent conditional GANs”. In: *arXiv preprint arXiv:1706.02633* (2017).
- [EGB19] Anja Exler, Zeynep Günes, and Michael Beigl. “Preferred notification modalities depending on the location and the location-based activity”. In: *Adjunct Proceedings of the 2019 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2019 ACM International Symposium on Wearable Computers*. 2019, pp. 1064–1069.
- [FGG20] Mohammad Navid Fekri, Ananda Mohon Ghosh, and Katarina Grolinger. “Generating energy data for machine learning with recurrent generative adversarial networks”. In: *Energies* 13.1 (2020), p. 130.
- [Fen05] Hsuan-Ming Feng. “Particle swarm optimization learning fuzzy systems design”. In: *Third International Conference on Information Technology and Applications (ICITA’05)*. Vol. 1. IEEE. 2005, pp. 363–366.
- [FGB11] Joel E. Fischer, Chris Greenhalgh, and Steve Benford. “Investigating Episodes of Mobile Phone Activity As Indicators of Opportune Moments to Deliver Notifications”. In: *Proceedings of the 13th International Conference on Human Computer Interaction with Mobile Devices and Services*. MobileHCI ’11. Stockholm, Sweden: ACM, 2011, pp. 181–190. ISBN: 978-1-4503-0541-9. DOI: 10.1145/2037373.2037402. URL: <http://doi.acm.org/10.1145/2037373.2037402>.
- [Fos19] David Foster. *Generative deep learning: teaching machines to paint, write, compose, and play*. O’Reilly Media, 2019.
- [FC20] Kieran Fraser and Owen Conlan. “Enticing Notification Text the Impact on Engagement”. In: *Adjunct Proceedings of the 2020 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2020 ACM International Symposium on Wearable Computers*. UbiComp-ISWC ’20. 2020.
- [FYC16] Kieran Fraser, Bilal Yousuf, and Owen Conlan. “A Context-aware, Info-bead and Fuzzy Inference Approach to Notification Management”. In: *2016 IEEE 7th Annual Ubiquitous Computing, Electronics & Mobile Communication Conference sponsored by IEEE (Published by IEEE Xplore) (IEEE UEMCON 2016)*. New York, USA, Oct. 2016.
- [FYC17a] Kieran Fraser, Bilal Yousuf, and Owen Conlan. “An in-the-wild and synthetic mobile notification dataset evaluation”. In: *2017 Intelligent Systems Conference (IntelliSys)*. IEEE. 2017, pp. 510–519.



- [FYC19a] Kieran Fraser, Bilal Yousuf, and Owen Conlan. “Generation and Evaluation of Personalised Push-Notifications”. In: *Adjunct Publication of the 27th Conference on User Modeling, Adaptation and Personalization*. 2019, pp. 223–224.
- [FYC19b] Kieran Fraser, Bilal Yousuf, and Owen Conlan. “Scrutable and persuasive push-notifications”. In: *International Conference on Persuasive Technology*. Springer. 2019, pp. 67–73.
- [FYC17b] Kieran Fraser, Bilal Yousuf, and Owen Conlan. “Synthesis & Evaluation of a Mobile Notification Dataset”. In: *Adjunct Publication of the 25th Conference on User Modeling, Adaptation and Personalization*. 2017, pp. 179–184.
- [FS97] Yoav Freund and Robert E Schapire. “A decision-theoretic generalization of on-line learning and an application to boosting”. In: *Journal of computer and system sciences* 55.1 (1997), pp. 119–139.
- [Gia+15] Wayne CW Giang et al. “Smartwatches vs. smartphones: A preliminary report of driver behavior and perceived risk while responding to notifications”. In: *Proceedings of the 7th international conference on automotive user interfaces and interactive vehicular applications*. 2015, pp. 154–161.
- [GLP18] Martin Gjoreski, Mitja Luštrek, and Veljko Pejović. “My watch says I’m busy: Inferring cognitive load with low-cost wearables”. In: *Proceedings of the 2018 ACM International Joint Conference and 2018 International Symposium on Pervasive and Ubiquitous Computing and Wearable Computers*. 2018, pp. 1234–1240.
- [Gol17] Yoav Goldberg. “Neural network methods for natural language processing”. In: *Synthesis Lectures on Human Language Technologies* 10.1 (2017), pp. 1–309.
- [GT09] Jacob Goldberger and Tamir Tassa. “Efficient anonymizations with enhanced utility”. In: *2009 IEEE International Conference on Data Mining Workshops*. IEEE. 2009, pp. 106–113.
- [Gol+04] Jacob Goldberger et al. “Neighbourhood components analysis”. In: *Advances in neural information processing systems* 17 (2004), pp. 513–520.
- [Gol97] Michael H Goldhaber. “The attention economy and the net”. In: (1997).
- [GM04] Victor M González and Gloria Mark. “"Constant, constant, multi-tasking craziness" managing multiple working spheres”. In: *Proceedings of the SIGCHI conference on Human factors in computing systems*. 2004, pp. 113–120.
- [Goo+14] Ian Goodfellow et al. “Generative adversarial nets”. In: *Advances in neural information processing systems*. 2014, pp. 2672–2680.
- [Goo20a] Google. *Chromium Blog: Introducing quieter permission UI for notifications*. Jan. 2020. URL: <https://blog.chromium.org/2020/01/introducing-quieter-permission-ui-for.html> (visited on 01/15/2021).

- [Goo20b] Google. *Exposure Notifications: Using technology to help public health authorities fight COVID-19*. 2020. URL: <https://www.google.com/covid19/exposurenotifications/> (visited on 01/23/2021).
- [Gow71] John C Gower. “A general coefficient of similarity and some of its properties”. In: *Biometrics* (1971), pp. 857–871.
- [Gul+17] Ishaan Gulrajani et al. “Improved training of wasserstein gans”. In: *Advances in Neural Information Processing Systems*. 2017, pp. 5767–5777.
- [HJ13] Pelle Guldberg Hansen and Andreas Maaløe Jespersen. “Nudge and the manipulation of choice: A framework for the responsible use of the nudge approach to behaviour change in public policy”. In: *European Journal of Risk Regulation* 4.1 (2013), pp. 3–28.
- [Has+09] Trevor Hastie et al. “Multi-class adaboost”. In: *Statistics and its Interface* 2.3 (2009), pp. 349–360.
- [Hes18] Pepita Hesselberth. “Discourses on disconnectivity and the right to disconnect”. In: *new media & society* 20.5 (2018), pp. 1994–2010.
- [Hin18] Avinash Hindupur. *The GAN Zoo*. 2018. URL: <https://github.com/hindupuravinash/the-gan-zoo> (visited on 02/18/2021).
- [Ho+18] Bo-Jhang Ho et al. “Nurture: notifying users at the right time using reinforcement learning”. In: *Proceedings of the 2018 ACM International Joint Conference and 2018 International Symposium on Pervasive and Ubiquitous Computing and Wearable Computers*. 2018, pp. 1194–1201.
- [HI05] Joyce Ho and Stephen S Intille. “Using context-aware computing to reduce the perceived burden of interruptions from mobile devices”. In: *Proceedings of the SIGCHI conference on Human factors in computing systems*. 2005, pp. 909–918.
- [HKA04] Eric Horvitz, Paul Koch, and Johnson Apacible. “BusyBody: Creating and Fielding Personalized Models of the Cost of Interruption”. In: *Proceedings of the 2004 ACM Conference on Computer Supported Cooperative Work*. CSCW '04. Chicago, Illinois, USA: ACM, 2004, pp. 507–510. ISBN: 1-58113-810-5. DOI: 10.1145/1031607.1031690. URL: <http://doi.acm.org/10.1145/1031607.1031690>.
- [How+17] Bill Howe et al. “Synthetic data for social good”. In: *arXiv preprint arXiv:1710.08874* (2017).
- [HK] TonTon Hsien-De Huang and Hung-Yu Kao. “C-3PO: Click-sequence-aware deeP neural network (DNN)-based Pop-uPs recOmmendation”. In: *Soft Computing* (), pp. 1–7.
- [Ire20] Central Statistics Office Ireland. *Employment and Life Effects of COVID-19*. 2020. URL: <https://www.cso.ie/en/releasesandpublications/er/elec19/employmentandlifeeffectsofcovid-19/> (visited on 07/10/2020).
- [JYS18] James Jordon, Jinsung Yoon, and Mihaela van der Schaar. “Measuring the quality of synthetic data for use in competitions”. In: *arXiv preprint arXiv:1806.11345* (2018).

- [JYV18] James Jordon, Jinsung Yoon, and Mihaela Van Der Schaar. “PATE-GAN: Generating synthetic data with differential privacy guarantees”. In: *International Conference on Learning Representations*. 2018.
- [JMO17] Rosemary Josekutty Thomas, Judith Masthoff, and Nir Oren. “Personalising Healthy Eating Messages to Age, Gender and Personality: Using Cialdini’s Principles and Framing”. In: *Proceedings of the 22nd International Conference on Intelligent User Interfaces Companion*. ACM. 2017, pp. 81–84.
- [Kal19] Ioannis Kaloskampis. *Synthetic data for public good*. 2019. URL: <https://datasciencecampus.ons.gov.uk/projects/synthetic-data-for-public-good/> (visited on 02/18/2021).
- [Kan+20] Thivya Kandappu et al. “PokeME: Applying context-driven notifications to increase worker engagement in mobile crowd-sourcing”. In: *Proceedings of the 2020 Conference on Human Information Interaction and Retrieval*. 2020, pp. 3–12.
- [Kan+09] Min Jeong Kang et al. “The wick in the candle of learning: Epistemic curiosity activates reward circuitry and enhances memory”. In: *Psychological science* 20.8 (2009), pp. 963–973.
- [Kar20] Evangelos Karapanos. “Technology-assisted reconstruction: a new alternative to the experience sampling method”. In: *Behaviour & Information Technology* 39.7 (2020), pp. 722–740.
- [Kem21] Simon Kemp. *DIGITAL 2021: GLOBAL OVERVIEW REPORT*. 2021. URL: <https://datareportal.com/reports/digital-2021-global-overview-report> (visited on 01/23/2021).
- [KE95] James Kennedy and Russell Eberhart. “Particle swarm optimization”. In: *Proceedings of ICNN’95-International Conference on Neural Networks*. Vol. 4. IEEE. 1995, pp. 1942–1948.
- [KS] Nicky Kern and Bernt Schiele. “Context-aware notification for wearable computing”. In: *Proceedings of the 7th IEEE International Symposium on Wearable Computers (ISWC’03)*, pp. 223–230.
- [Kin+13] Anna Lucia S King et al. “Nomophobia: Dependency on virtual environments or social phobia?” In: *Computers in Human Behavior* 29.1 (2013), pp. 140–144.
- [Kiu+] Niko Kiukkonen et al. “Towards rich mobile phone datasets: Lausanne data collection campaign”. In: ().
- [KFG18] Andreas Komninos, Elton Frengkou, and John Garofalakis. “Predicting user responsiveness to smartphone notifications for edge computing”. In: *European Conference on Ambient Intelligence*. Springer. 2018, pp. 3–19.
- [Kom+19] Andreas Komninos et al. “Discovering user location semantics using mobile notification handling behaviour”. In: *European Conference on Ambient Intelligence*. Springer. 2019, pp. 219–234.
- [KMB20] Sarah E Kreps, Miles McCain, and Miles Brundage. “All the News that’s Fit to Fabricate: AI-Generated Text as a Tool of Media Misinformation”. In: *Available at SSRN 3525002* (2020).

- [KKK17] Florian Künzler, Jan-Niklas Kramer, and Tobias Kowatsch. “Efficacy of mobile context-aware notification management systems: A systematic literature review and meta-analysis”. In: *2017 IEEE 13th International Conference on Wireless and Mobile Computing, Networking and Communications (WiMob)*. IEEE. 2017, pp. 131–138.
- [KPD16] Kostadin Kushlev, Jason Proulx, and Elizabeth W Dunn. “" Silence Your Phones" Smartphone Notifications Increase Inattention and Hyperactivity Symptoms”. In: *Proceedings of the 2016 CHI conference on human factors in computing systems*. 2016, pp. 1011–1020.
- [LR09] Peter E. Latham and Yasser Roudi. *Mutual Information*. 2009. URL: [http://www.scholarpedia.org/article/Mutual\\_information](http://www.scholarpedia.org/article/Mutual_information) (visited on 01/23/2021).
- [Lat+13a] Neal Lathia et al. “Contextual dissonance: Design bias in sensor-based experience sampling methods”. In: *Proceedings of the 2013 ACM international joint conference on Pervasive and ubiquitous computing*. ACM. 2013, pp. 183–192.
- [Lat+13b] Neal Lathia et al. “Open source smartphone libraries for computational social science”. In: *Proceedings of the 2013 ACM conference on Pervasive and ubiquitous computing adjunct publication*. 2013, pp. 911–920.
- [Lee+19] Hao-Ping Lee et al. “Does who matter? Studying the impact of relationship characteristics on receptivity to mobile IM messages”. In: *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. 2019, pp. 1–12.
- [Lee+14a] Yu-Kang Lee et al. “The dark side of smartphone usage: Psychological traits, compulsive behavior and technostress”. In: *Computers in human behavior* 31 (2014), pp. 373–383.
- [Lee+14b] Uichin Lee et al. “Hooked on smartphones: an exploratory study on smartphone overuse among college students”. In: *Proceedings of the 32nd annual ACM conference on Human factors in computing systems*. ACM. 2014, pp. 2327–2336.
- [Leo20] Paul M Leonardi. “COVID-19 and the new technologies of organizing: digital exhaust, digital footprints, and artificial intelligence in the wake of remote work”. In: *Journal of Management Studies* (2020).
- [Li+16] Frank Li et al. “Remedying web hijacking: Notification effectiveness and webmaster comprehension”. In: *Proceedings of the 25th International Conference on World Wide Web*. 2016, pp. 1009–1019.
- [Lim+15] Soo Ling Lim et al. “Investigating country differences in mobile app user behavior and challenges for software engineering”. In: *IEEE Transactions on Software Engineering* 41.1 (2015), pp. 40–64.
- [Lin+14] Yu-Hsuan Lin et al. “Development and validation of the Smartphone Addiction Inventory (SPAI)”. In: *PloS one* 9.6 (2014), e98312.

- [Liu+19] Tianming Liu et al. “Dapanda: Detecting aggressive push notifications in android apps”. In: *2019 34th IEEE/ACM International Conference on Automated Software Engineering (ASE)*. IEEE. 2019, pp. 66–78.
- [Loe94] George Loewenstein. “The psychology of curiosity: A review and reinterpretation.” In: *Psychological bulletin* 116.1 (1994), p. 75.
- [Lor20] Steven Loria. *TextBlob*. 2020. URL: <https://textblob.readthedocs.io/en/dev/index.html>.
- [Lou+13] Gilles Louppe et al. “Understanding variable importances in forests of randomized trees”. In: *Advances in neural information processing systems*. 2013, pp. 431–439.
- [LWY19] Pei-Hsuan Lu, Pang-Chieh Wang, and Chia-Mu Yu. “Empirical evaluation on synthetic data generation with generative adversarial network”. In: *Proceedings of the 9th International Conference on Web Intelligence, Mining and Semantics*. 2019, pp. 1–6.
- [Luc+17] Mario Lucic et al. “Are gans created equal? a large-scale study”. In: *arXiv preprint arXiv:1711.10337* (2017).
- [MN18] Sofiane Mahiou and Fintan Nagle. “Anonymity, Utility and Risk: an Overview”. In: (2018).
- [MS04] Bradley Malin and Latanya Sweeney. “How (not) to protect genomic data privacy in a distributed network: using trail re-identification to evaluate and design anonymity protection systems”. In: *Journal of biomedical informatics* 37.3 (2004), pp. 179–192.
- [MIC17] Gloria Mark, Shamsi Iqbal, and Mary Czerwinski. “How blocking distractions affects workplace focus and productivity”. In: *Proceedings of the 2017 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2017 ACM International Symposium on Wearable Computers*. 2017, pp. 928–934.
- [Mar+15] Gloria Mark et al. “Focused, aroused, but so distractible: Temporal perspectives on multitasking and communications”. In: *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing*. ACM. 2015, pp. 903–916.
- [MP43] Warren S McCulloch and Walter Pitts. “A logical calculus of the ideas immanent in nervous activity”. In: *The bulletin of mathematical biophysics* 5.4 (1943), pp. 115–133.
- [MHM17] Abhinav Mehrotra, Robert Hendley, and Mirco Musolesi. “Interpretable machine learning for mobile notification management: An overview of prefminer”. In: *GetMobile: Mobile Computing and Communications* 21.2 (2017), pp. 35–38.
- [MHM19] Abhinav Mehrotra, Robert Hendley, and Mirco Musolesi. “NotifyMeHere: Intelligent notification delivery in multi-device environments”. In: *Proceedings of the 2019 Conference on Human Information Interaction and Retrieval*. 2019, pp. 103–111.

- [MHM16a] Abhinav Mehrotra, Robert Hendley, and Mirco Musolesi. “PrefMiner: mining user’s preferences for intelligent mobile notification management”. In: *Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing*. 2016, pp. 1223–1234.
- [MHM16b] Abhinav Mehrotra, Robert Hendley, and Mirco Musolesi. “Towards multi-modal anticipatory monitoring of depressive states through the analysis of human-smartphone interaction”. In: *Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing: Adjunct*. 2016, pp. 1132–1138.
- [MM20a] Abhinav Mehrotra and Mirco Musolesi. “Intelligent Notification Systems”. In: *Synthesis Lectures on Mobile and Pervasive Computing* 11.1 (2020), pp. 1–75.
- [MM17] Abhinav Mehrotra and Mirco Musolesi. “Intelligent notification systems: A survey of the state of the art and research challenges”. In: *arXiv preprint arXiv:1711.10171* (2017).
- [Meh+15a] Abhinav Mehrotra et al. “Ask, but don’t interrupt: the case for interruptibility-aware mobile experience sampling”. In: *Adjunct Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2015 ACM International Symposium on Wearable Computers*. 2015, pp. 723–732.
- [Meh+15b] Abhinav Mehrotra et al. “Designing content-driven intelligent notification mechanisms for mobile applications”. In: *Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing*. 2015, pp. 813–824.
- [Meh+16] Abhinav Mehrotra et al. “My phone and me: understanding people’s receptivity to mobile notifications”. In: *Proceedings of the 2016 CHI conference on human factors in computing systems*. 2016, pp. 1021–1032.
- [Meh+17a] Abhinav Mehrotra et al. “MyTraces: Investigating correlation and causation between users’ emotional states and mobile phone interaction”. In: *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 1.3 (2017), pp. 1–21.
- [Meh+17b] Abhinav Mehrotra et al. “Understanding the role of places and activities on mobile phone interaction and usage patterns”. In: *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 1.3 (2017), pp. 1–22.
- [MZC12] Ravi Mehta, Rui Zhu, and Amar Cheema. “Is noise always bad? Exploring the effects of ambient noise on creative cognition”. In: *Journal of Consumer Research* 39.4 (2012), pp. 784–799.
- [Mera] Merriam-Webster. *Notification*. In: *Merriam-Webster.com dictionary*. URL: <https://www.merriam-webster.com/dictionary/notification> (visited on 01/01/2020).
- [Merb] Merriam-Webster. *Scruple*. In: *Merriam-Webster.com dictionary*. URL: <https://www.merriam-webster.com/dictionary/scruple> (visited on 01/01/2020).

- [MC06] Catherine A Middleton and Wendy Cukier. “Is mobile email functional or dysfunctional? Two perspectives on mobile email usage”. In: *European Journal of Information Systems* 15.3 (2006), pp. 252–260.
- [Mil19] Derek Miller. *bert-extractive-summarizer*. 2019. URL: <https://github.com/dmmiller612/bert-extractive-summarizer>.
- [MO14] Mehdi Mirza and Simon Osindero. “Conditional generative adversarial nets”. In: *arXiv preprint arXiv:1411.1784* (2014).
- [MN86] Yoshiro Miyata and Donald A Norman. “Psychological issues in support of multiple activities”. In: *User centered system design: New perspectives on human-computer interaction* (1986), pp. 265–284.
- [Moh+09] David Moher et al. “Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement”. In: *PLoS medicine* 6.7 (2009), e1000097.
- [MD19] Alberto Monge Roffarello and Luigi De Russis. “The race towards digital wellbeing: Issues and opportunities”. In: *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. 2019, pp. 1–14.
- [MM20b] Kathleen Anne Moore and Evita March. “Socially Connected during COVID-19: Online social connections mediate the relationship between loneliness and positive coping strategies”. In: (2020).
- [Mor+17] Leanne G Morrison et al. “The effect of timing and frequency of push notifications on usage of a smartphone-based stress management intervention: an exploratory trial”. In: *PloS one* 12.1 (2017), e0169162.
- [NS08] Arvind Narayanan and Vitaly Shmatikov. “Robust de-anonymization of large sparse datasets”. In: *2008 IEEE Symposium on Security and Privacy (sp 2008)*. IEEE. 2008, pp. 111–125.
- [Now15] Beata Nowok. “synthpop : An R package for generating synthetic versions of sensitive microdata for statistical disclosure control”. In: 2015.
- [NRD+16] Beata Nowok, Gillian M Raab, Chris Dibben, et al. “synthpop: Bespoke creation of synthetic data in R”. In: *Journal of statistical software* 74.11 (2016), pp. 1–26.
- [OTT19] Tadashi Okoshi, Kota Tsubouchi, and Hideyuki Tokuda. “Real-world product deployment of adaptive push notification scheduling on smartphones”. In: *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. 2019, pp. 2792–2800.
- [Oko+17] Tadashi Okoshi et al. “Attention and engagement-awareness in the wild: A large-scale study with adaptive notifications”. In: *2017 IEEE international conference on pervasive computing and communications (percom)*. IEEE. 2017, pp. 100–110.

- [Oko+15] Tadashi Okoshi et al. “Reducing Users’ Perceived Mental Effort Due to Interruptive Notifications in Multi-device Mobile Environments”. In: *Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing*. UbiComp ’15. Osaka, Japan: ACM, 2015, pp. 475–486. ISBN: 978-1-4503-3574-4. DOI: 10.1145/2750858.2807517. URL: <http://doi.acm.org/10.1145/2750858.2807517>.
- [OM18] Rita Orji and Karyn Moffatt. “Persuasive technology for health and wellness: State-of-the-art and emerging trends”. In: *Health informatics journal* 24.1 (2018), pp. 66–91.
- [Oul+12] Antti Oulasvirta et al. “Habits make smartphone use more pervasive”. In: *Personal and Ubiquitous Computing* 16.1 (2012), pp. 105–114.
- [Pal+15] Shekhar Pal et al. “Mobile phones: Reservoirs for the transmission of nosocomial pathogens”. In: *Advanced biomedical research* 4 (2015).
- [Par21a] AJ Park. *15 Customer Engagement Stats for 2021*. 2021. URL: <https://www.airship.com/blog/2021-customer-engagement-stats/> (visited on 01/20/2021).
- [Par21b] AJ Park. *15 Customer Engagement Stats for 2021*. 2021. URL: <https://www.airship.com/blog/2021-customer-engagement-stats/> (visited on 02/18/2021).
- [Par+18] Noseong Park et al. “Data Synthesis Based on Generative Adversarial Networks”. In: *Proc. VLDB Endow.* 11.10 (June 2018), pp. 1071–1083.
- [PW16] Christopher J Parker and Huchen Wang. “Examining hedonic and utilitarian motivations for m-commerce fashion retail app engagement”. In: *Journal of Fashion Marketing and Management: An International Journal* (2016).
- [Par10] Konstantinos E Parsopoulos. *Particle swarm optimization and intelligence: advances and applications: advances and applications*. IGI Global, 2010.
- [PWV16] N. Patki, R. Wedge, and K. Veeramachaneni. “The Synthetic Data Vault”. In: *2016 IEEE International Conference on Data Science and Advanced Analytics (DSAA)*. 2016, pp. 399–410. DOI: 10.1109/DSAA.2016.49.
- [PKA14] David Paulk, Lisa Kim, and Diogo Adrados. “TigerBites: an assistive notification system for local dining”. In: *Proceedings of the 7th International Conference on Pervasive Technologies Related to Assistive Environments*. 2014, pp. 1–3.
- [Pea+17] Jennifer Pearson et al. “Chameleon devices: investigating more secure and discreet mobile interactions via active camouflaging”. In: *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*. 2017, pp. 5184–5196.
- [Ped+11] F. Pedregosa et al. “Scikit-learn: Machine Learning in Python”. In: *Journal of Machine Learning Research* 12 (2011), pp. 2825–2830.



- [PM14] Veljko Pejovic and Mirco Musolesi. “InterruptMe: Designing Intelligent Prompting Mechanisms for Pervasive Applications”. In: *Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing*. UbiComp ’14. Seattle, Washington: ACM, 2014, pp. 897–908. ISBN: 978-1-4503-2968-2. DOI: 10.1145/2632048.2632062. URL: <http://doi.acm.org/10.1145/2632048.2632062>.
- [PSM14] Jeffrey Pennington, Richard Socher, and Christopher D. Manning. “GloVe: Global Vectors for Word Representation”. In: *Empirical Methods in Natural Language Processing (EMNLP)*. 2014, pp. 1532–1543.
- [PH18] Erik Peper and Richard Harvey. “Digital addiction: Increased loneliness, anxiety, and depression”. In: *NeuroRegulation* 5.1 (2018), pp. 3–3.
- [Pie19] Martin Pielot. *CrawDad - MobilePhoneUse Data*. 2019. URL: <https://crawdad.org/telefonica/mobilephoneuse/20190429/MobilePhoneUse/index.html> (visited on 02/18/2021).
- [PCD14] Martin Pielot, Karen Church, and Rodrigo De Oliveira. “An in-situ study of mobile phone notifications”. In: *Proceedings of the 16th international conference on Human-computer interaction with mobile devices & services*. 2014, pp. 233–242.
- [PR16] Martin Pielot and Luz Rello. “Productive, anxious, lonely-24 hours without push notifications”. In: *arXiv preprint arXiv:1612.02314* (2016).
- [PR15] Martin Pielot and Luz Rello. “The do not disturb challenge: a day without notifications”. In: *Proceedings of the 33rd annual ACM conference extended abstracts on human factors in computing systems*. 2015, pp. 1761–1766.
- [PVP18] Martin Pielot, Amalia Vradi, and Souneil Park. “Dismissed! a detailed exploration of how mobile phone users handle push notifications”. In: *Proceedings of the 20th International Conference on Human-Computer Interaction with Mobile Devices and Services*. 2018, pp. 1–11.
- [Pie+17] Martin Pielot et al. “Beyond interruptibility: Predicting opportune moments to engage mobile phone users”. In: *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 1.3 (2017), pp. 1–25.
- [Pie+14] Martin Pielot et al. “Didn’t you see my message? predicting attentiveness to mobile instant messages”. In: *Proceedings of the SIGCHI conference on human factors in computing systems*. 2014, pp. 3319–3328.
- [Pie+15] Martin Pielot et al. “When attention is not scarce-detecting boredom from mobile phone usage”. In: *Proceedings of the 2015 ACM international joint conference on pervasive and ubiquitous computing*. 2015, pp. 825–836.

- [PSH17] Haoyue Ping, Julia Stoyanovich, and Bill Howe. “Datasyntesizer: Privacy-preserving synthetic datasets”. In: *Proceedings of the 29th International Conference on Scientific and Statistical Database Management*. 2017, pp. 1–5.
- [Pot+16] Martin Potthast et al. “Clickbait detection”. In: *European Conference on Information Retrieval*. Springer. 2016, pp. 810–817.
- [Pot+17] Martin Potthast et al. *The clickbait challenge 2017*. 2017. URL: <https://www.clickbait-challenge.org/>.
- [Prz+13] Andrew K Przybylski et al. “Motivational, emotional, and behavioral correlates of fear of missing out”. In: *Computers in Human Behavior* 29.4 (2013), pp. 1841–1848.
- [Pus19] PushWoosh. *Emoji increase push notifications CTR by over 100%*. 2019. URL: <https://www.pushwoosh.com/blog/emojis-increase-ctr/> (visited on 01/23/2021).
- [QH17] Mohiuddin Abdul Qader and Vagelis Hristidis. “Dualdb: An efficient lsm-based publish/subscribe storage system”. In: *Proceedings of the 29th International Conference on Scientific and Statistical Database Management*. 2017, pp. 1–6.
- [Qin+14] Yujue Qin et al. “A Context-aware Do-not-disturb Service for Mobile Devices”. In: *Proceedings of the 13th International Conference on Mobile and Ubiquitous Multimedia*. MUM ’14. Melbourne, Victoria, Australia: ACM, 2014, pp. 236–239. ISBN: 978-1-4503-3304-7. DOI: 10.1145/2677972.2678003. URL: <http://doi.acm.org/10.1145/2677972.2678003>.
- [RND16] Gillian M Raab, Beata Nowok, and Chris Dibben. “Practical data synthesis for large samples”. In: *Journal of Privacy and Confidentiality* 7.3 (2016), pp. 67–97.
- [Rac+10] Kiran K Rachuri et al. “EmotionSense: a mobile phones based adaptive platform for experimental social psychology research”. In: *Proceedings of the 12th ACM international conference on Ubiquitous computing*. 2010, pp. 281–290.
- [Rad+20] Alec Radford et al. *Better Language Models and Their Implications*. 2020. URL: <https://openai.com/blog/better-language-models/>.
- [Rad+] Alec Radford et al. “Language models are unsupervised multitask learners”. In: ().
- [RSL16] Rahul Rajan, Ted Selker, and Ian Lane. “Effects of mediating notifications based on task load”. In: *Proceedings of the 8th International Conference on Automotive User Interfaces and Interactive Vehicular Applications*. 2016, pp. 145–152.
- [Reg16] General Data Protection Regulation. “Regulation EU 2016/679 of the European Parliament and of the Council of 27 April 2016”. In: *Official Journal of the European Union*. Available at: [http://ec.europa.eu/justice/data-protection/reform/files/regulation\\_oj\\_en.pdf](http://ec.europa.eu/justice/data-protection/reform/files/regulation_oj_en.pdf) (accessed 20 September 2017) (2016).

- [RG19] Nils Reimers and Iryna Gurevych. “Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks”. In: *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Nov. 2019.
- [RD20] ALBERTO MONGE ROFFARELLO and LUIGI DE RUSSIS. “Understanding, Discovering, and Mitigating Habitual Smartphone Use in Young Adults”. In: (2020).
- [Rog+20] Yvonne Rogers et al. “The Dark Side of Interaction Design.” In: *CHI Extended Abstracts*. 2020, pp. 1–4.
- [Ros+20] Lucas Rosenblatt et al. “Differentially Private Synthetic Data: Applied Evaluations and Enhancements”. In: *arXiv preprint arXiv:2011.05537* (2020).
- [Sae+20] Ahmed Saeed et al. “Annulus: A Dual Congestion Control Loop for Datacenter and WAN Traffic Aggregates”. In: *Proceedings of the Annual conference of the ACM Special Interest Group on Data Communication on the applications, technologies, architectures, and protocols for computer communication*. 2020, pp. 735–749.
- [Sah+14] Alireza Sahami Shirazi et al. “Large-scale assessment of mobile notifications”. In: *Proceedings of the SIGCHI conference on Human factors in computing systems*. 2014, pp. 3055–3064.
- [Sak97] Yasubumi Sakakibara. “Recent advances of grammatical inference”. In: *Theoretical Computer Science* 185.1 (1997), pp. 15–45.
- [SB10] Dario D Salvucci and Peter Bogunovich. “Multitasking and multitasking: The effects of mental workload on deferred task interruptions”. In: *Proceedings of the SIGCHI conference on human factors in computing systems*. 2010, pp. 85–88.
- [Sam+18] Mary Jane C Samonte et al. “Development of Online Hospital Document Management with SMS Notification System”. In: *Proceedings of the 2nd International conference on E-Society, E-Education and E-Technology*. 2018, pp. 150–154.
- [Sch+00] Bernhard Schölkopf et al. “New support vector algorithms”. In: *Neural computation* 12.5 (2000), pp. 1207–1245.
- [SG16] Florian Schulze and Georg Groh. “Conversational Context Helps Improve Mobile Notification Management”. In: *Proceedings of the 18th International Conference on Human-Computer Interaction with Mobile Devices and Services*. MobileHCI ’16. Florence, Italy: ACM, 2016, pp. 518–528. ISBN: 978-1-4503-4408-1. DOI: 10.1145/2935334.2935347. URL: <http://doi.acm.org/10.1145/2935334.2935347>.
- [Sen68] Pranab Kumar Sen. “Estimates of the regression coefficient based on Kendall’s tau”. In: *Journal of the American statistical association* 63.324 (1968), pp. 1379–1389.
- [SSH08] Saul Shiffman, Arthur A Stone, and Michael R Hufford. “Ecological momentary assessment”. In: *Annu. Rev. Clin. Psychol.* 4 (2008), pp. 1–32.

- [SSL19] In-geon Shin, Jin-min Seok, and Youn-kyung Lim. “Ten-minute silence: a new notification UX of mobile instant messenger”. In: *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. 2019, pp. 1–13.
- [SBV18] Alberto Vianna Dias da Silva, Lucas Borges, and Vaninha Vieira. “CDNA: A Context-Aware Notification System for Driver Interruption”. In: *Proceedings of the 17th Brazilian Symposium on Human Factors in Computing Systems*. 2018, pp. 1–8.
- [SSL18] Mark Silvis, Anthony Sicilia, and Alexandros Labrinidis. “PittGrub: a frustration-free system to reduce food waste by notifying hungry college students”. In: *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. 2018, pp. 754–763.
- [Smi+14] Jeremiah Smith et al. “Learning to recognise disruptive smartphone notifications”. In: *Proceedings of the 16th international conference on Human-computer interaction with mobile devices & services*. ACM. 2014, pp. 121–124.
- [SDM16] Kirsten A Smith, Matt Dennis, and Judith Masthoff. “Personalizing reminders to personality for melanoma self-checking”. In: *Proceedings of the 2016 Conference on User Modeling Adaptation and Personalization*. ACM. 2016, pp. 85–93.
- [Sno+16] Joshua Snoke et al. “General and specific utility measures for synthetic data”. In: *arXiv preprint arXiv:1604.06651* (2016).
- [Sny+96] C Rick Snyder et al. “Development and validation of the State Hope Scale.” In: *Journal of personality and social psychology* 70.2 (1996), p. 321.
- [Sol+19] Irene Solaiman et al. “Release strategies and the social impacts of language models”. In: *arXiv preprint arXiv:1908.09203* (2019).
- [SS17] M Sudha and K Sheena. “Impact of influencers in consumer decision process: the fashion industry”. In: *SCMS Journal of Indian Management* 14.3 (2017), pp. 14–30.
- [SFC19] Rowan Sutton, Kieran Fraser, and Owen Conlan. “A Reinforcement Learning and Synthetic Data Approach to Mobile Notification Management”. In: *Proceedings of the 17th International Conference on Advances in Mobile Computing Multimedia*. MoMM2019. Munich, Germany: Association for Computing Machinery, 2019, pp. 155–164. ISBN: 9781450371780. DOI: 10.1145/3365921.3365932.
- [Tan+18] Sarah Tan et al. “Distill-and-compare: Auditing black-box models using transparent model distillation”. In: *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*. 2018, pp. 303–310.
- [TK16] Channary Tauch and Eiman Kanjo. “The roles of emojis in mobile phone notifications”. In: *Proceedings of the 2016 acm international joint conference on pervasive and ubiquitous computing: Adjunct*. 2016, pp. 1560–1565.

- [Ter07] G Henri Ter Hofte. “Xensible interruptions from your mobile phone”. In: *Proceedings of the 9th international conference on Human computer interaction with mobile devices and services*. ACM. 2007, pp. 178–181.
- [Tho+14] Bill Thornton et al. “The mere presence of a cell phone may be distracting: Implications for attention and task performance.” In: *Social Psychology* 45.6 (2014), p. 479.
- [TAW17] Liam D Turner, Stuart M Allen, and Roger M Whitaker. “Reachable but not receptive: Enhancing smartphone interruptibility prediction by modelling the extent of user engagement with notifications”. In: *Pervasive and Mobile Computing* 40 (2017), pp. 480–494.
- [Ulg+09] Fatma Ulger et al. “Are we aware how contaminated our mobile phones with nosocomial pathogens?” In: *Annals of clinical microbiology and antimicrobials* 8.1 (2009), p. 7.
- [UP16] Gašper Urh and Veljko Pejović. “TaskyApp: inferring task engagement via smartphone sensing”. In: *Proceedings of the 2016 ACM international joint conference on pervasive and ubiquitous computing: Adjunct*. 2016, pp. 1548–1553.
- [VFK17] Niels Van Berkel, Denzil Ferreira, and Vassilis Kostakos. “The experience sampling method on mobile devices”. In: *ACM Computing Surveys (CSUR)* 50.6 (2017), pp. 1–40.
- [Vie21] Armando Vieira. *Hazy synthetic data quality metrics explained*. 2021. URL: <https://hazy.com/blog/2021/01/15/hazy-synthetic-data-quality-metrics-explained> (visited on 01/23/2021).
- [Vis+19] Aku Visuri et al. “Understanding smartphone notifications’ user interactions and content importance”. In: *International Journal of Human-Computer Studies* 128 (2019), pp. 72–85.
- [Voi+16] Alexandra Voit et al. “UbiTtention: smart & ambient notification and attention management”. In: *Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing: Adjunct*. 2016, pp. 1520–1523.
- [WRB14] Daniel T Wagner, Andrew Rice, and Alastair R Beresford. “Device Analyzer: Large-scale mobile data collection”. In: *ACM SIGMETRICS Performance Evaluation Review* 41.4 (2014), pp. 53–56.
- [Wal+17] Jason Walonoski et al. “Synthea: An approach, method, and software mechanism for generating synthetic patients and the synthetic electronic health care record”. In: *Journal of the American Medical Informatics Association* 25.3 (2017), pp. 230–238.
- [War+17] Adrian F Ward et al. “Brain drain: the mere presence of one’s own smartphone reduces available cognitive capacity”. In: *Journal of the Association for Consumer Research* 2.2 (2017), pp. 140–154.
- [WMZ10] William Webber, Alistair Moffat, and Justin Zobel. “A similarity measure for indefinite rankings”. In: *ACM Transactions on Information Systems (TOIS)* 28.4 (2010), pp. 1–38.

- [WVH19] Dominik Weber, Alexandra Voit, and Niels Henze. “Clear all: A large-scale observational study on mobile notification drawers”. In: *Proceedings of Mensch und Computer 2019*. 2019, pp. 361–372.
- [Web+18] Dominik Weber et al. “Snooze! investigating the user-defined deferral of mobile notifications”. In: *Proceedings of the 20th International Conference on Human-Computer Interaction with Mobile Devices and Services*. 2018, pp. 1–13.
- [WSB13] Jason Wiese, T Scott Saponas, and AJ Brush. “Phoneprioception: enabling mobile phones to infer where they are kept”. In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM. 2013, pp. 2157–2166.
- [Wij+17] Sanjaya Wijeratne et al. “Emojinet: An open service and api for emoji sense discovery”. In: *Proceedings of the International AAAI Conference on Web and Social Media*. Vol. 11. 1. 2017.
- [Wik20] Wikipedia. *Google Alerts*. 2020. URL: [https://en.wikipedia.org/wiki/Google\\_Alerts](https://en.wikipedia.org/wiki/Google_Alerts).
- [Woh14] Claes Wohlin. “Guidelines for snowballing in systematic literature studies and a replication in software engineering”. In: *Proceedings of the 18th international conference on evaluation and assessment in software engineering*. 2014, pp. 1–10.
- [Wol+19] Thomas Wolf et al. “HuggingFace’s Transformers: State-of-the-art Natural Language Processing”. In: *ArXiv* abs/1910.03771 (2019).
- [Won+17] Min Hao Wong et al. “Nitroaromatic detection and infrared communication from wild-type plants using plant nanobionics”. In: *Nature materials* 16.2 (2017), pp. 264–272.
- [Woo+09] Mi-Ja Woo et al. “Global measures of data utility for microdata masked for disclosure limitation”. In: *Journal of Privacy and Confidentiality* 1.1 (2009).
- [WC] Jesse H Wright and Robert Caudill. “Remote treatment delivery in response to the COVID-19 pandemic”. In: *Psychotherapy and Psychosomatics* 89.3 (), p. 1.
- [Yam+17] Sezin Yaman et al. “Notifying and involving users in experimentation: ethical perceptions of software practitioners”. In: *2017 ACM/IEEE International Symposium on Empirical Software Engineering and Measurement (ESEM)*. IEEE. 2017, pp. 199–204.
- [YS20] Kevin P Yancey and Burr Settles. “A Sleeping, Recovering Bandit Algorithm for Optimizing Recurring Notifications”. In: *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. 2020, pp. 3008–3016.
- [YGL17] Fengpeng Yuan, Xianyi Gao, and Janne Lindqvist. “How busy are you? Predicting the interruptibility intensity of mobile users”. In: *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*. 2017, pp. 5346–5360.
- [Zad83] Lotfi Asker Zadeh. “The role of fuzzy logic in the management of uncertainty in expert systems”. In: *Fuzzy sets and systems* 11.1-3 (1983), pp. 199–227.

- 
- [Zha04] Harry Zhang. “The Optimality of Naïve Bayes”. In: *In FLAIRS2004 conference*. 2004.
- [Zha+18] Bo Zhao et al. “Notification Volume Control and Optimization System at Pinterest”. In: *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. 2018, pp. 1012–1020.





# Appendix A

## Preliminary: Exploring Open Data for Optimal Push

### A.1 Overview

Manually managing the vast number of notifications pushed through mobile devices can be challenging, particularly as the number of connected devices and apps with push capabilities continues to grow [PCD14]. As a solution to stem the constant inundation, research has explored autonomous management of incoming push-notifications [Meh+15b; KS; Oko+15; Qin+14; HI05]. A Notification Management System (NMS) uses data to decide optimal parameters for how and when a notification will be received by a subscriber. To achieve this, explicit data contained within the notification, and contextual features describing a subscriber's immediate environment, are both desirable, and arguably necessary, data points for aiding a NMS in accurately inferring a subscriber's preferred delivery time. However, as both notification and context features tend to contain sensitive data, such as Personally Identifiable Information (PII), it can be difficult to acquire these data points. This creates two major barriers for improving the experience of push-notifications. The first can be found in research. Data which contains explicit notification and context information is rarely shared due to ethical, privacy and consent constraints. This poses a problem for prospective research in this domain as mobile development and participant recruitment are then necessary to obtain data before new NMSs can be trained and evaluated. Evidence of this can be found

in the many studies on push-notifications which first include the development of a mobile app for capturing notification and context features, many of which overlap with other studies [Meh+15b; PM14; HI05; HKA04; FGB11; Meh+16]. The second barrier to improved experiences of push-notifications is the availability of data which can be used for autonomously managing notifications. This barrier is relevant to applied NMSs where subscribers may not grant, or subsequently withdraw, consent to track granular notification or context behaviours, reducing the amount of rich data available for identifying trends and training effective NMSs. This problem is not unique to notification management however, as many domains are confronted with sparse data and cold start problems. The ability to adapt on-the-fly to different levels of data availability, or privacy settings, is new and now necessary for NMSs as subscribers are given greater control over their data and how it is used by machines.

To address the first barrier outlined above, a number of open smartphone usage data sets were evaluated with respect to their applicability and usefulness for notification engagement research. Of the data sets explored, one was chosen for creating a synthetic mobile notification data set for the purpose of improving an existing NMS. The process of creation was used as an opportunity to identify applicable features for effective notification engagement as well as identify methods for abstracting sensitive notification and context information. With respect to the second barrier, a method of augmentation was used to scale the synthetic data set to provide NMSs with greater amounts of data with which to train. The synthetic data set was evaluated using an existing NMS and a notification data set previously collected in-the-wild.

The key contributions of this work are:

1. An analysis of existing open data applicable to notification engagement
2. A process of constructing and scaling a notification data set using open data
3. A method of optimising a NMS using abstract, synthetic, notification data
4. Comparative evaluation between a fixed rule NMS and a NMS trained using synthetic data

## A.2 Open-source Notifications

This section discusses the challenges of data availability and comparative evaluation in the notification engagement domain and explores current features used in SOTA studies as well as applicable open data sets.

Previous research in the domain of personalized notification management has utilized both incoming notification content and user context to infer preferred delivery times [Meh+15b; KS; Oko+15; Qin+14; HI05]. Rich data sets, describing real-world trends of notification engagement behaviours, were a vital component for enabling the training and evaluation of these systems. However, few open data sets exist which contain explicit notification content, user context at moments of interest and qualitative or quantitative user feedback on preferential notification delivery. This creates an initial barrier to entry for researchers wishing to develop and evaluate a new NMS algorithm and subsequently means allocating time to develop a notification logging app for collecting notification and context attributes *in-the-wild* [Meh+15b; PM14; HI05; HKA04; FGB11; Meh+16].

Time is also devoted to acquiring participants and encouraging engagements over prolonged periods of time, which is difficult to achieve at scale without significant resources or industry partnerships. In addition, due to the ethical implications of collecting sensitive information contained within participant notifications, the use and subsequent sharing of data is constrained. Therefore, the existing evaluations of NMSs occur in segmented partitions defined by individual study conditions. For example, definitions of when a notification is seen and how it is engaged with have varying interpretations and the set of features, collection methods and sampling strategies used differ from study to study. Therefore, the conclusions drawn of NMSs are applicable only to narrow subsets of the population. Comparative evaluations of NMSs would require a large-scale open data set accompanied by a standard set of engagement metrics with which performance could be evaluated and compared to benchmark methods. This has been achieved in Computer Vision, Natural Language Processing and Information Retrieval domains in which standard tasks are well defined and benchmark data sets freely available. As a first step toward contributing to a notification engagement benchmark data set, a set of standard features, already in use across existing notification engagement studies, had to be identified.

Through an analysis carried out on a number of state-of-the-art studies (see Tables 2.4, 2.5, 2.6), a set of common and most frequently used features pertaining to smartphone usage and notification engagement was created. The majority

contained features which overlapped with other studies, however, there was no discernible consensus on a best practice for collecting the most common features. Tables 2.4, 2.6 and 2.5 illustrate the consolidated features prevalent in notification engagement studies. [Meh+15b; Meh+16; Lat+13a; WSB13; PM14; Ter07].

In addition to identifying a standard set of notification engagement features, a brief analysis of some existing open data sets, or sources of data, pertaining to smartphone usage was also carried out for the purpose of identifying suitable, freely available data which could be used to improve, and comparatively assess, a NMS without requiring costly app development and participant acquisition. The sources analysed (ca. September 2016) were as follows:

#### **A.2.0.1 Nodobo**

[BMI11] This data set was comprised of smartphone usage data collected in 2010/11 by the University of Strathclyde. The features included were calls (13035 records), messages (83542 records), presence (5292103 records), Wifi and cell towers. All data points were anonymised. The data set was freely available for use with the requirement that researchers cite the relevant publication of the Nodobo authors. The data set was available for download in the form of multiple CSV files.

#### **A.2.0.2 Worldwide Mobile App User Behavior**

[Lim+15] This was a worldwide survey of mobile app user behavior encompassing 15 countries and 10,000 participants in 2012. Survey questions included demographics, age, education level, personality (Big-Five), app usage, triggers for downloading apps and reasons for disengaging with them. This data set was also freely available on condition of citing the relevant publication of the authors if used. The data was available for download in the form of an Excel file.

#### **A.2.0.3 Device Analyzer**

[WRB14] This was a growing data set harvested through an Android app which logged statistics of ongoing smartphone usage in the background. It was developed at the University of Cambridge and had 31,031 contributors with over 100 billion records of smartphone usage from over 17,000 devices around the world. The smartphone usage data collected included events such as device charging, picture

taken or sensor values. The data set was freely available to academic researchers and was obtained via email, once a brief description of use was given.

#### A.2.0.4 CrowdSignals.io

[Cro17] CrowdSignals was a community which aimed to generate ongoing labeled smartphone and sensor data. The data could be freely used for research purposes 12-18 months after collection, before which it was only accessible if a fee was paid. The data sets advertised included features such as geo-location, communications, user interactions, social connections, user-provided ground-truth labels and survey feedback. At the point of this analysis, there were only sample data sets available as ongoing data collection periods had yet to take place.

#### A.2.0.5 Mobile Data Challenge (MDC)

[Kiu+] This was a data collection campaign run by Idiap and NRC-Lausanne from 2009 to 2011 whereby 200 participants had their mobile device behavior and social network details collected. The data set was restricted to non-profit organizations and required a process of establishing a site manager to act as a legal representative for an eligible institution before the data was passed on. The data was also considered accessible only to permitted staff within the eligible institution.

#### A.2.0.6 Friends & Family (F&F)

[Aha+11] This was an experiment studying how people make decisions in varying social situations with a goal of identifying methods for improved decision making. The data set included over 130 participants. It was run in 2010/11 and included features such as call/SMS logs, apps installed, relationships, activities, events, mood, personality (Big-Five) and others. The data was collected *in-the-wild* via an *Android* app by the MIT Human Dynamics Lab and was freely available with the condition that relevant publications by the authors were cited, the data was not redistributed to others and re-identification of the participants was not attempted. The data set was available for download as a set of CSV files.

The Friends & Family (F&F) dataset [Aha+11] was collected for a mobile-phone-based social and behavioral study undertaken by the MIT Media Lab. The data was collected in 2011 within a residential community with ties to a

Table A.1 Features recorded in the Friends &amp; Family data set

| Collection | Category       | Description                                        |
|------------|----------------|----------------------------------------------------|
| Sensor     | Bluetooth      | Scan of devices in proximity                       |
|            | Call Log       | Incoming and outgoing calls                        |
|            | SMS Log        | Incoming and outgoing messages                     |
|            | Location       | Affine-transformed coordinates (every 30 mins)     |
|            | Apps           | State of apps installed on device (every 3 hours)  |
|            | Running Apps   | Name of apps running (every 30 seconds)            |
|            | Acceleration   | Accelerometer readings                             |
| Survey     | Battery        | Charge level, connection type, health, temperature |
|            | Monthly        | Personality, demographics, dynamic social network  |
|            | Weekly         | Sleep quality, dynamic social network              |
|            | Social Network | Couples and friendship surveys                     |
|            | FunFit         | Daily activity stats, weight report, buddy network |

nearby university. The data was used to investigate social involvements in decision making processes and to explore how those decisions could be improved. The data was collected *in-the-wild* through the *Android* mobile phone devices of the participants. The data includes quantitative sensory data such as accelerometer readings and proximity networks, as well as qualitative data in the form of weekly and monthly surveys.

#### A.2.0.7 Additional Update

[WVH19; KFG18; Kom+19; Pie19] Since this preliminary study into available notification engagement data sets was completed (Dec 2016), there has been an effort by prominent researchers in the field to release ESM and mobile sensory data, including notification data (albeit, not to the extent of full granularity) within the community, in addition to open-source tools empowering researchers to conduct their own studies.

## A.3 Synthetic & Augmented Notifications

This section details the process used for manipulating open-source smartphone data into a form consistent with notification engagement behaviour such that it could be used for improved training of an existing NMS. The process is documented to

enable other researchers recreate the same freely available notification engagement data set.

### A.3.1 Challenge

As discussed in the previous sections, the availability of open-source notification data sets is negligible. The majority of data sets collected are fit-for-purpose and can't be freely shared by researchers for ethical reasons. The lack of data available makes it difficult to train a NMS without first developing software to capture both notification and context data. As well as the added time it takes to develop the mechanism to capture the necessary notification features, the process also requires finding a group of participants willing to have their notifications monitored on an ongoing basis. Paired with this are multiple ethical concerns. For instance, a participant may be willing to share their explicit notification data with researchers for the purposes of a smartphone usage study, but they are not the sole contributors to these notifications. The profile of the notification sender is also captured in this process. Subsequently, this raises the following question: should consent also be sought from the contacts of participants, as they are just as much owners of the notifications sent and received as the participants who have granted consent? Do the participants, ethically, have the power to grant consent over data which is not solely theirs? These questions make it difficult to acquire open source, granular, notification data sets which accurately model real-world engagements.

### A.3.2 Proposed Solution

An augmented, or synthetic, data set is proposed as a possible solution to the challenges outlined. If the underlying real-world behaviours and scenarios of notification engagements could be effectively mapped to a synthetic data set, ethical constraints governing the use of real-world data sets could be circumnavigated and freely available notification engagement data sets could be shared for improving and evaluating NMSs. For the purposes of this work, an existing NMS, named NAbs [FYC16], was chosen as a tool for evaluating the utility of a created synthetic data set. This dictated the features chosen for the synthetic data set, illustrated in Table A.2. The sender, app, title and content features were collected from notifications delivered to two participants over a period of 68 days.

Table A.2 Features used for design and evaluation of existing NMS

| Feature     | Description                                                    |
|-------------|----------------------------------------------------------------|
| sender      | The relationship type between receiver and sender              |
| senderRank  | The importance rank given to this relationship by the receiver |
| app         | The app through which the notification was delivered           |
| appRank     | The importance rank given to this app by the receiver          |
| title       | The topic extracted from the title of the notification         |
| titleRank   | The importance rank given to the title topic by the receiver   |
| content     | The topic extracted from the content of the notification       |
| contentRank | The importance rank given to the content topic by the receiver |

The sensitive notification features were uplifted to abstract terms to preserve the privacy of the participants. For example, the name of the sender was replaced by a term categorising the relationship between the sender and the receiver, such as *family*, *friend* or *colleague*. The title and content values were uplifted to topics representing the subject of the combined text. Additionally, the participants also ranked the fixed terms which were used for uplift. For example, the importance of sender, app and content were scored on a scale of 1 to 10. Using a combination of these features, the NAbs NMS identified opportune moments for delivery.

None of the open-source data sets discussed in the previous section contained notification engagement data and none were found referenced in literature at the point of analysis. Therefore, regardless of the smartphone data chosen, some assumptions would have to be made for creating notification engagements. The source chosen for creating synthetic notifications was the Friends & Family (F&F) data set. The F&F data included call and SMS logs, which could be interpreted as incoming notifications. The data set was also comprised of app usage behaviour, contextual data (such as calendar events), responses to mood questionnaires, personality types and perceptions of relationship. No other data set gave such a granular and holistic view of participants with reference to their smartphone. The data provided also aligned well with the existing NMS, NAbs. The limitation of the data, and all open-source data sets assessed, was the absence of an engagement action, such as opening or dismissing a notification. While the F&F data set provided call and SMS events at differing times and days of week, it did not provide the action taken once delivered. An alternative solution was therefore applied and will be detailed in Section A.4.1. Table A.1 lists the data recorded in the F&F data set. Table A.3 further illustrates the features chosen from the F&F data set which were used for inferring notification features applicable to the NAbs NMS. The inference process is detailed in the following section.



Table A.3 F&amp;F data features used for creating NMS compatible features

| <b>F&amp;F Data</b>               | <b>NMS Data</b>                                     |
|-----------------------------------|-----------------------------------------------------|
| Apps installed                    | Notification app                                    |
| Call log                          | Notification date/time                              |
| SMS log                           | Notification date/time                              |
| Couples & Friendship              | Sender rank                                         |
| Monthly, weekly,<br>daily surveys | Subject, Sender, App,<br>App rank, Context activity |

### A.3.3 Engineering Notification Features

As the F&F data set did not contain all the necessary notification features required for the NAbs NMS, some assumptions and inferences were necessary to build a synthetic data set. This section details the assumptions made and methods of inference.

#### A.3.3.1 Notification Date & Time

The time and date of incoming calls and SMS messages were logged during the F&F study. As SMS messages and calls create notifications on a device, it could be naturally inferred that these times were also consistent with incoming notifications. Unfortunately, other types of notification, such as updates from installed apps, were not recorded. Figures A.1a and A.1b illustrate the incoming notifications and calls for participants through the whole study period and for a single month, respectively. On average,  $\approx 40$  notifications were delivered per day, interrupting participants as they went about their daily activities. The majority of notifications were delivered during the afternoon and evening, illustrated in Figure A.1d. There was little discrepancy between days of the week. The proportion of notifications delivered on Fridays were marginally larger than other days and Sunday had the smallest proportion of notification deliveries, illustrated in Figure A.1c.

#### A.3.3.2 Notification Sender & Ranking

The call and SMS data of the F&F data set also included anonymised participant ids for the sender as well as the recipient. The NAbs NMS required a notification to have a sender and for the notification recipient to have a relationship ranking for that sender. Using the provided ids, the relationship between the receiver

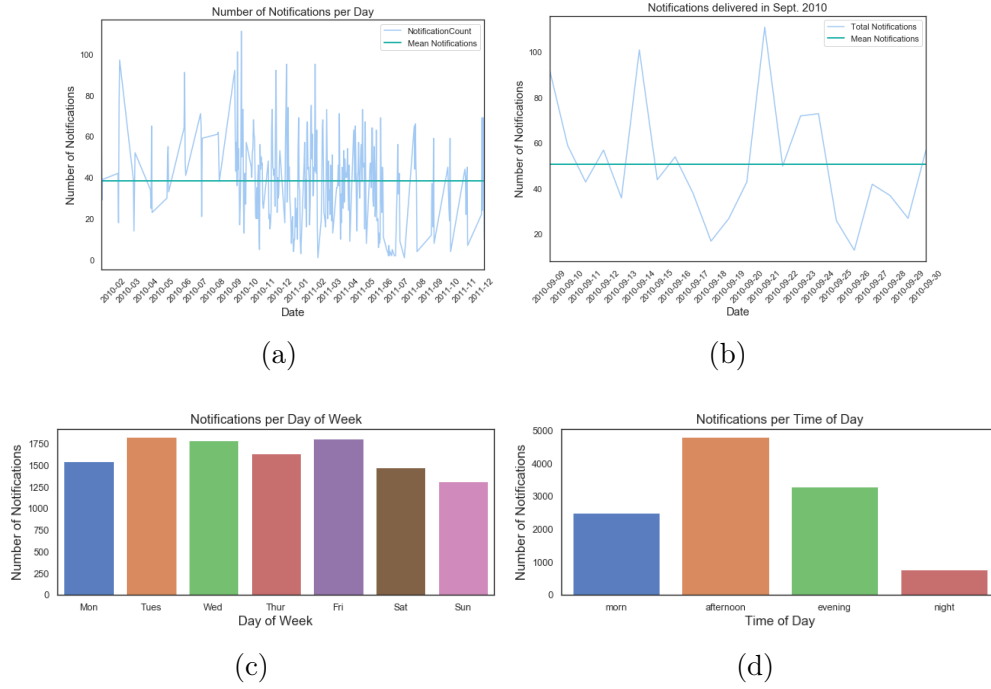


Fig. A.1 Timings of notification delivery in the synthetic data set drawn from real-world events

and sender of the notification could be found using the responses to the *Couples* or *Friendship* surveys included in the F&F data set. The *Friendship* data was comprised of relationship ratings given by each participant to other participants. The *Couples* data was comprised of a mapping of one participant id to another to signify a couple. Therefore, the assumptions made for inferring the sender importance ranking were as follows:

- If a sender could not be found in the *Friendship* or *Couple* survey responses, the sender was categorised as a stranger and was assigned an importance ranking of 0.
- If a sender was found in any *Friendship* responses and had a ranking greater than 0, the sender was classified randomly as either a *colleague* or a *friend*. This was done simply to add context to the notification and aid in inferring the subject (discussed next). Interestingly, some participants gave differing rankings to the same sender at different times of the study, suggesting the relationship importance varied over time, place or activity. This insight inspired further experimentation, as detailed in Section 3.2.5.1 of Chapter 3. For the purposes of this work, the most recent previous ranking given to the sender was used.

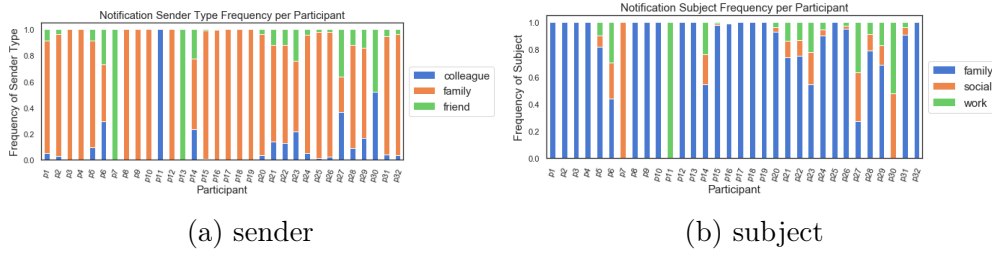


Fig. A.2 Frequency of differing subject and sender types present in synthetic notifications

- If a sender could be found in a *Couples* survey response then the sender was classified as *family* and the importance ranking was set at 10.

Figure A.2a illustrates the diversity of sender types in the subsequent synthetic data set created. The *family* sender type was found to dominate in most cases, although a few edge cases emerged such as a participant only receiving communication from work colleagues (p11) or friends (p7). A number of participants also displayed balanced communication proportions between family, friends and colleagues (p6 and p17). This diversity is positive as it enables NMSs be evaluated under different scenarios and for different user profiles.

### A.3.3.3 Notification Subject & Ranking

It was also necessary for each notification to have a subject and an associated ranking of importance for that subject, assigned by the participant. As no information on the subject was recorded in the *Call Log* or *SMS Log* datasets, the subject had to be inferred through other means. A fixed set of subjects were therefore assumed and the sender of the notification was used to infer which subject of the set was chosen. The subject set was comprised of {family, work, social, interest}. This set of abstract subjects was used by the NAbs NMS and also aligned with the type of sender relationships found in the F&F survey data. The assumptions for subject were as follows:

- If the sender of the notification was a family member, the subject was family.
- If the sender of the notification was a colleague, the subject was work.
- If the sender of the notification was a friend, the subject was social.
- If the sender of the notification was a stranger, the subject was interest.

Table A.4 Mapping of app categories to notification subjects.

| App Category          | Subjects                       |
|-----------------------|--------------------------------|
| Games                 | Family, Interest               |
| Lifestyle             | Family, Interest               |
| Shopping              | Family, Interest               |
| Communication         | Family, Interest, Social, Work |
| Entertainment         | Interest, Social               |
| Phone Personalization | Interest                       |
| Productivity          | Work                           |
| Social                | Social                         |
| Other                 | Interest                       |

These assumptions were quite restrictive as, in a real-world scenario, a family member might also send a notification with a subject categorised as social or work. It must also be noted that the subject ranking was not derived from the F&F data set. The subject ranking was instead used as a dynamic value which could be tuned depending on the type of participant wishing to be modeled. For instance, the NMS could be evaluated on career driven user profiles by adjusting the subject ranking for *work* higher than the subject ranking for *social*. Figure A.2b illustrates the frequency of notification subject types for participants. Naturally, as the majority of notifications were sent by a family member and due to the assumptions outlined above, *family* was also the dominant subject within notifications delivered. However, interesting scenarios also occur whereby the only subject received by some participants was *social* (p7) or *work* (p11).

#### A.3.3.4 Notification Application & Ranking

The NAbs NMS required a notification to have an associated app through which it was delivered and to have that app assigned an importance value provided by the participant. In this case, the app for a particular notification was chosen based on the inferred subject of the notification. A manual mapping of subject to app category, illustrated in Table A.4, was assumed and used to infer which app category was chosen for a given notification. Once the app category was chosen, participant apps were retrieved from the apps and survey responses recorded in the F&F data set. An app for the notification was then chosen from the set of apps which matched the category. The user ranking of importance was inferred using the *Survey* data which asked participants to rank their most used and favorite apps.

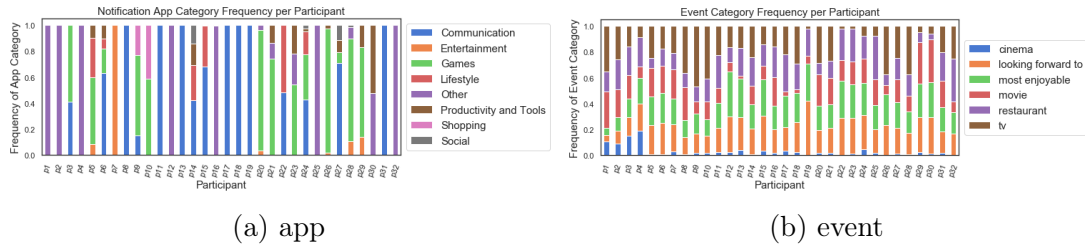


Fig. A.3 Frequency of differing apps present in synthetic notifications and differing event types occurring per participant

Figure A.3a illustrates the diversity of app categories through which notifications were delivered during the study. Many participants received notifications from a number of differing categories, however there were some cases in which participants received notifications from one category alone (e.g. p11, *Communication*), thus providing interesting scenarios on which to train and evaluate a NMS.

#### A.3.3.5 User Events/Activities

The NAbs NMS also required contextual data for making notification delivery inferences. There was no explicit information regarding a participant's current activity at moments of incoming calls or SMS messages in the F&F data set hence, a number of assumptions were made regarding participant activities. The survey responses gathered information regarding events and activities that a participant participated in during the weeks of the study. For example, it recorded visits to the cinema, restaurants and nights in front of the TV. It also recorded which groups of people the activity or event was associated with. For example, participants specified whether events were attended by family members, colleagues, friends or whether they were alone. The responses were used to create a sparse calendar for participants. Figure A.3b illustrates the frequency of event types per participant throughout the study period. Additional, predictable events were also created to fill out participant calendars based on their profile, such as work, lectures or lunch. The event assumptions were as follows:

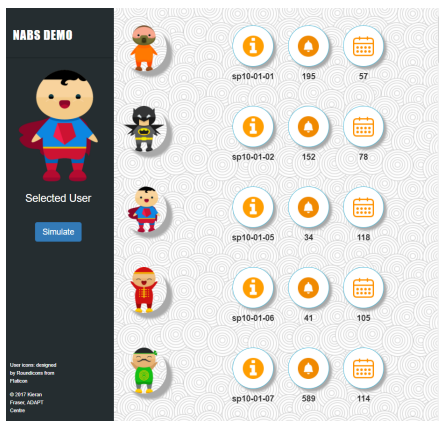
- Participants were given work activities from 9am to 5pm each weekday with a break of 1 hour for lunch at 12pm.
- For events found in the *Survey* responses, the day of the week was chosen randomly and the time and duration were chosen based on the type of event, illustrated in Table A.5.

Table A.5 Mapping of event types to assumed relevant times.

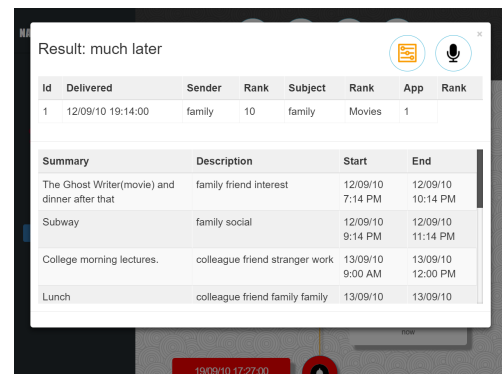
| <b>Event Type</b>    | <b>Time</b> |
|----------------------|-------------|
| Most look forward to | 5pm - 8pm   |
| Most enjoyed         | 5pm - 7pm   |
| Cinema               | 8pm-9pm     |
| Movie                | 9pm-10pm    |
| TV                   | 5pm-8pm     |
| Restaurant           | 7pm-9pm     |

The synthetic data set was modeled using real-world behavioral data found in the F&F data set, albeit with limiting assumptions. However, the assumptions and limitations provided a mechanism for controlling the type of data set, and underlying participant profiles, generated, such that differing scenarios could be created and used for evaluating the NMSs under different conditions and participant profiles. For example, the frequency of events could be increased or decreased to model participants who had dense or sparse schedules. Notification frequency could be increased or notification subject values switched, depending on the personality type of participants wishing to be modeled. This synthetic data set was purposely built to match a data set which could be used by an existing NMS, however, many more inferences could have been made to generate a richer data set. For example, personality traits, bluetooth proximity, location and accelerometer data were all recorded in the F&F study, but were out of scope for the creation of this synthetic data set. The task of creating the synthetic data set unearthed possible features which could prove important for future work, such as the dynamic sender relationship. The final data set was comprised of 32 participants, 11,395 notifications and 3,148 calendar events.

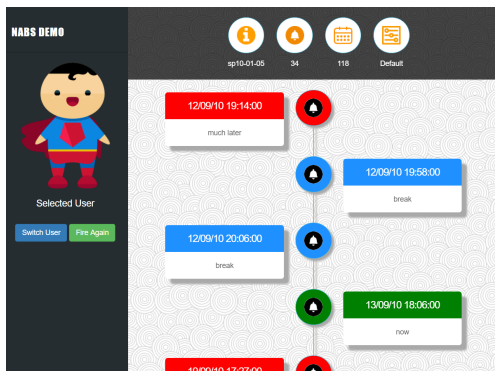
A web app, illustrated in Figure A.4, was created for exploring the newly created synthetic notification data set. A novel user interface, including an interactive chat bot, was later added for visualising, controlling and understanding the inferences made by the NAbs NMS.



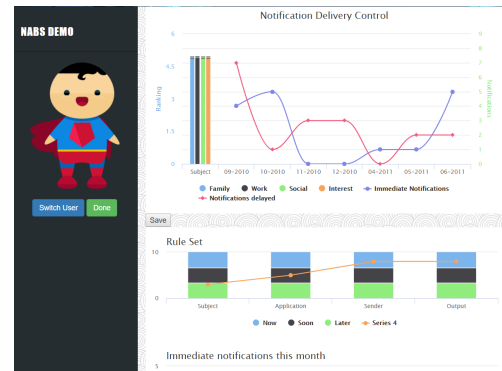
(a) Participant info



(b) Inferred delivery within context



(c) Preferred delivery inferred by NABS



(d) Visualisation, control &amp; understanding

Fig. A.4 Web app created for exploring the synthetic notification data set

## A.4 Evaluating a Notification Management System

The purpose of the work was to identify a set of common features for NMSs, create a synthetic notification data set using open data and use the newly generated data to improve and evaluate an existing NMS. The previous sections have discussed relevant notification engagement features and the steps for generating a synthetic notification data set using open data. This section explores the functionality of the NAbs NMS, the process of optimising inference using synthetic data and the subsequent evaluation results afforded by the synthetic notification engagements created.

### A.4.1 NAbs Notification Management System

The design of the NAbs NMS was motivated by a number of gaps identified in notification management literature: trust, transparency and control. At the time, few NMSs took into account a notification recipient's trust of autonomous systems. For the NMS to be of use in the real world, the recipient had to relinquish control over what notifications they would and would not see. Research has shown that a portion of the population are wary of relinquishing power to a NMS and prefer transparency paired with some level of personal input to the decision making process with regard their notifications [SG16]. Moreover, "black-box" predictions provided by AI systems have been a focal point of discussion in the research community in recent years [AB18], as issues, such as gender or racial bias [Cha16; Tan+18] for example, have surfaced in systems which are incapable of explaining their decision process. Mehrotra et al. were some of the first to address transparency and intelligent control of notification engagements [MHM16a].

The NAbs NMS applied Fuzzy Logic [Zad83] as a solution for notification management in the form of a Fuzzy Inference System (FIS), made up of two key components: Fuzzy Membership Functions and a Fuzzy Knowledge Base. The Fuzzy Knowledge Base (FKB) contained heuristic rules governing notification delivery decisions and provided insight into the decisions made by the NMS which could subsequently be presented to end-users to improve transparency and explainability. An example of the rules contained within the FKB were as follows:

- If *SenderInput* is *NIP* and *SubjectInput* is *NIP* and *AppInput* is *NIP* then *AlertOutput* is *MUCHLATER*.



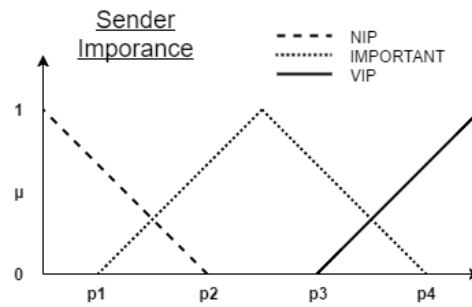


Fig. A.5 Fuzzy membership function inferring importance of the relationship between notification recipient and sender

- If *SenderInput* is *IMPORTANT* and *SubjectInput* is *IMPORTANT* and *AppInput* is *NIP* then *AlertOutput* is *VERYSOON*.

The values *NIP* (not important), and *IMPORTANT*, were inferred using the rankings given by participants for particular apps, subjects and sender relationships. The rankings were used as input to Fuzzy Membership Functions (FMF) whose role was to indicate if an app, subject or sender were relevant in a given context, using contextual data such as events logged in participant calendars. Depending on the parameters set for the FMF, values such as: Not Important (NIP), Important or Very Important (VIP) were inferred for notifications in a given context. Figure A.5 illustrates a simple membership function for sender importance. As participants applied their own rankings, they had some input and control over the NMS's notification delivery decisions.

The two key components of the FIS used by NABs for inferring optimal moments of notification delivery were created with fixed parameters. For example, the lower and upper limits (e.g. p1 and p4 in Figure A.5) of the FMFs remained fixed across all situations and participants. Additionally, the set of heuristic rules found in the FKB also remained unchanged. This was identified as a limitation of the work as the FIS did not adapt to any dynamic conditions. The synthetic data set was therefore used to optimise the parameters of these two components.

As previously mentioned, the synthetic data set was limited by the absence of a preferred delivery time or ground truth action (open/dismiss) taken on a notification by a participant after delivery. As this feature was unavailable to the synthetic data set, an alternative approach was taken toward managing notification delivery. NABs would be optimised to favour quicker delivery of work related notifications during working hours and would favour deferring the delivery of social notifications until the working day ended, with a goal of reinforcing better work-life

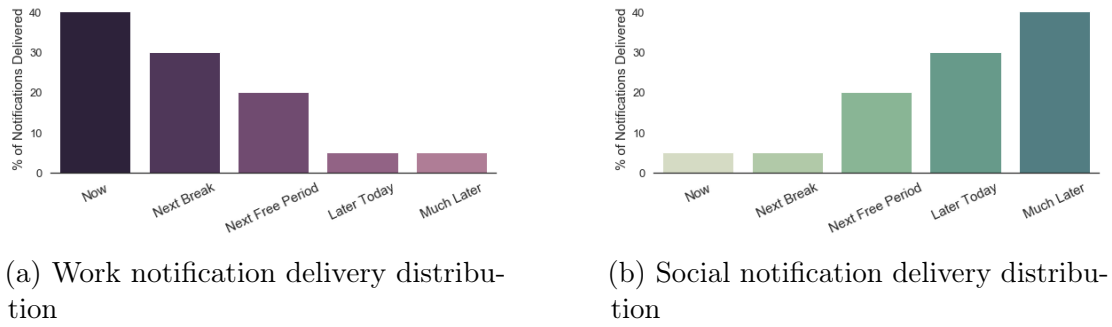


Fig. A.6 Preferred delivery distribution of work and social notifications during working hours (assumed 9am to 5pm)

balance with respect to notification engagement. The preferred distribution of notifications delivered during working hours for notifications with a work topic is illustrated in Figure A.6a and the distribution for social notifications delivered during those same hours is illustrated in Figure A.6b. The NMS would continue delivering notifications at contextual moments during these times, however, the majority of work notifications would be delivered during working hours and social notifications delivered during social hours. Whilst the parameters were optimised to reflect this behaviour, as the participants had input to the FIS, in the form of their importance rankings, any notifications with higher levels of importance, or relevancy, could still break through and be delivered, regardless of subject. A justification for distinguishing clearer boundaries between work and social notification delivery hours is the *Right to Disconnect* which "refers to the right of employees to not have to take calls or read emails related to work during their time off" [Bor17; Hes18]. This has become a prominent point of discussion, in recent times particularly, as large numbers of employees work from home during the COVID-19 pandemic. For example, in Ireland, data released by the Central Statistics Office [Ire20] indicated: "of the 47% of the population who have had their employment impacted by COVID-19, just over a third (34%) have started remote working from home and 23% of these employment affected persons have seen a change to their work hours". Therefore, applying rules which encourage better notification engagements, dependent on the subject and relevance of the activity, could be used to help remote-workers disconnect, take back their free time and promote healthier digital engagements.

### A.4.2 Parameter Optimisation

As previously noted, the synthetic data set generated was used to optimise the parameters of the FIS implemented in NAbs such that work notifications would favour immediate contextual delivery times during work hours and social notifications would be deferred, respecting the distinction between work and social life. Figure A.6 illustrates the target distribution of notification delivery for work and social notifications. It is important to note that a small number of social notifications were still encouraged to be contextually delivered during working hours, as anxiety has been shown to occur when notification recipients feel they are missing out on notifications [PR16].

Particle Swarm Optimisation (PSO) [KE95] was the algorithm chosen for identifying the optimal permutation of parameters to be used in the FIS. It was chosen due to its popularity and success in a wide range of applications, including that of neural network training and fuzzy logic control [Par10; Fen05; Esm07; Che06]. PSO is a population based search algorithm bio-inspired by the movement taken by flocks of birds when searching for food. The algorithm begins by initialising a number of particles with a random position in the search space and a random velocity. The position represents a solution, in this case, a permutation of parameters for the FMFs or FKB. Each particle continuously updates its own position and velocity based on its own personal best position (*pBest*) and a global best position (*gBest*) shared between all particles. In this manner, particles fly through the search space and converge towards areas of the search space with optimal solutions. The algorithm uses a fitness function to evaluate the performance of each particle's position/solution. Based on this value, *pBest* and *gBest* are updated. There is no generic fitness function to apply in this instance as it is unique to the problem at hand. The fitness function implemented in this case sought to minimize the differences between the distributions of *inferred* notification delivery times and *preferred* notification delivery times for work and social notifications, illustrated in Figure A.6.

The PSO algorithm was used in two subsequent experiments designed to optimise NAbs using synthetic data. The first attempted to find the optimal parameters for the FMFs, while keeping the FKB fixed. The second experiment attempted to find the optimal set of heuristic rules in the FKB while keeping the FMFs fixed. The two components were optimised in isolation in order to decrease the search space to be traversed by the PSO algorithm. Each experiment implemented a generic PSO algorithm, illustrated in Figure A.7. Equations (A.1)

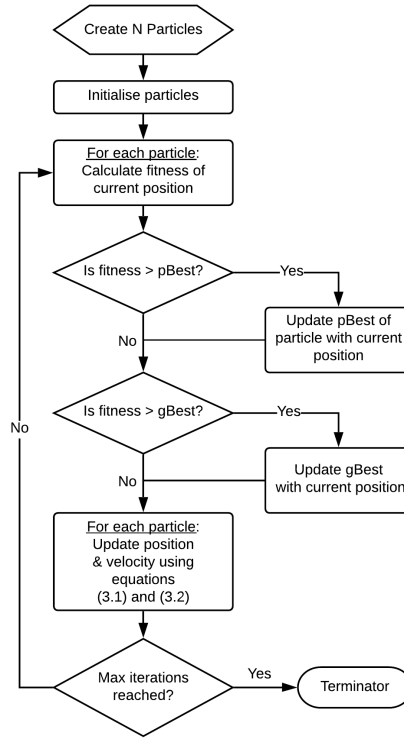


Fig. A.7 Particle Swarm Optimisation Algorithm

and (A.2), referenced in Figure A.7, were used to update the velocity and position of each particle. Equation (A.1) updated the velocity, i.e. the rate at which a particle moved towards the *pBest* and *gBest* solution. The *rand()* function generated a uniform random number between 0 and 1. The *c1* and *c2* variables were both learning factors. Equation (A.2) was used to update the position of particles using the new velocity value.

$$v_{t+1} = v_t + c1 * rand() * (pBest - position_t) + c2 * rand() * (gBest - position_t) \quad (A.1)$$

$$position_{t+1} = position_t + v_{t+1} \quad (A.2)$$

#### A.4.2.1 Tuning FKB parameters

The first experiment tuned the parameters of the FKB using the synthetic data. The parameters were linguistic variables which represented preferred delivery times for notifications:  $\{NOW, SOON, VERYSOON, LATER, MUCHLATER\}$ . By tuning these parameters the NMS was learning what permutation of rules would result in a delivery distribution of notifications similar to Figure A.6a.

#### A.4.2.2 Tuning FMF parameters

The second experiment tuned the FMFs. These functions mapped crisp input values representing the contextual importance of the sender, subject and app, to fuzzy linguistic terms of *NIP*, *IMPORTANT*, *VIP*. The fuzzy linguistic terms were then subsequently used produce crisp outputs which were fed to the FKB to influence the preferred delivery times of notifications. An example of the FMF for *Sender* importance is illustrated in Figure A.5 and the tunable parameters are denoted by  $p_n$ . Through tuning the FMF parameters, the NMS could learn to contextually prioritize the notification features such that predicted delivery times for notifications would match the preferred distribution illustrated in Figure A.6a. Previously, the static parameters of the FMFs were a limitation of the NMS, potentially introducing bias to the predictions, as stated in the introductory study [FYC16].

## A.5 Results

The PSO algorithm was executed on notification data of every user in the synthetic data set who received both *work* and *social* notifications, 18 users in total. The number of particles, or swarm size, used in the PSO algorithm was 15. Values between 10 and 30 were shown, in literature [Eng07], to be successful at finding optimal solutions. Additionally, the learning factors, C1 and C2, were set at 2, as per the original basic PSO algorithm [KE95] and the number of iterations was set at 10, although the algorithm would end before this if the fitness function determined the predicted delivery distributions matched the preferred delivery distributions perfectly. The notification *subject* importance rankings were set at 7 for *work* and 8 for *social* (derived from the rankings previously used for the

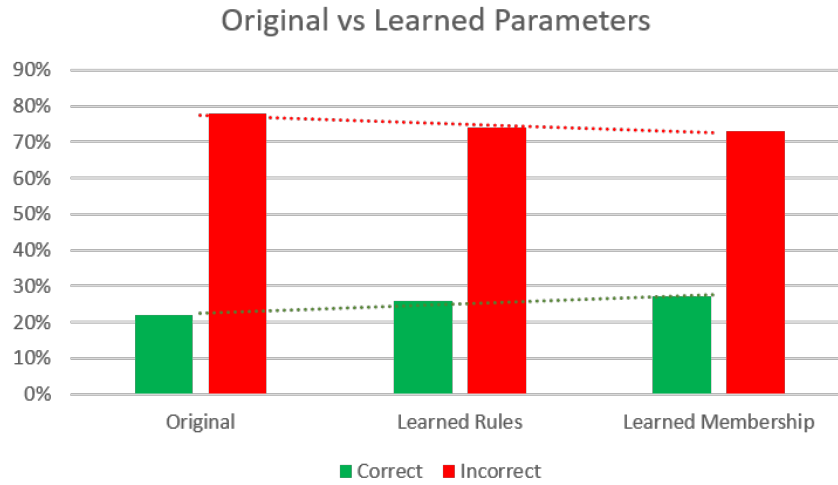


Fig. A.8 Performance comparison between using original parameters and parameters learned using synthetic data set and PSO.

NAbs NMS) as the F&F data set did not provide data for this feature of the NMS. These values remained fixed while tuning both the FKB and the FMFs.

The original NAbs NMS and real-world notification data set were used for evaluating the optimised parameters learned while training on synthetic data. The optimised parameters were applied to NAbs and used to predict preferred delivery times. The ground truth values of the real notification data set were used to identify correct or incorrect preferred delivery time predictions. Figure A.8 illustrates that the optimised FMFs had a marginal increase in performance over the optimised FKB, which in turn had a marginal performance improvement over the original, static rule, NMS. While the overall performance of the NMS remains quite low, the increased performance does demonstrate the utility of the synthetic data set in optimising the NMS.

## A.6 Limitations

There were a number of limitations associated with this work. One such limitation was the absence of a ground truth notification engagement feature within the synthetic data set. Subsequently, a preferred distribution of notification delivery times was used to optimise a NMS using synthetic data. Whilst this method did provide a measure of improved performance, it was dependent on the assumption that participants would prefer work notifications delivered immediately during

working hours and social notifications delivered outside of working hours. Thus the final NMS was using a mix of both general and personalised rules to deliver notifications. Additionally, the process of evaluation was limited by the number of participants ( $n=2$ ) in the real notification data set.

## A.7 Summary

This work identified and assessed state-of-the-art data sets with respect to mobile usage and *in-the-wild* notification-engagement studies. As an alternative to collecting real notification data, the process of creating a large-scale synthetic notification data set, derived from a mobile usage study carried out by the MIT Media lab, was evaluated. The synthetic data set was then used to optimize a previously created set of Fuzzy Knowledge Base rules and Fuzzy Membership Functions used within the Fuzzy Inference System (FIS) of an existing NMS. The purpose of this was to ascertain the utility of a synthetic data set created using existing contextual data, as opposed to collecting further data *in-the-wild*. Results illustrated a slight increased performance when the FKB and FMF components were optimised using synthetic data, however it was concluded that the optimisation process was limited by the absence of a ground-truth action feature. Contributions of this work included an analysis of open notification data, the creation of a scalable synthetic data set comprised of notification and context features and a method for improving a NMS through Particle Swarm Optimisation.





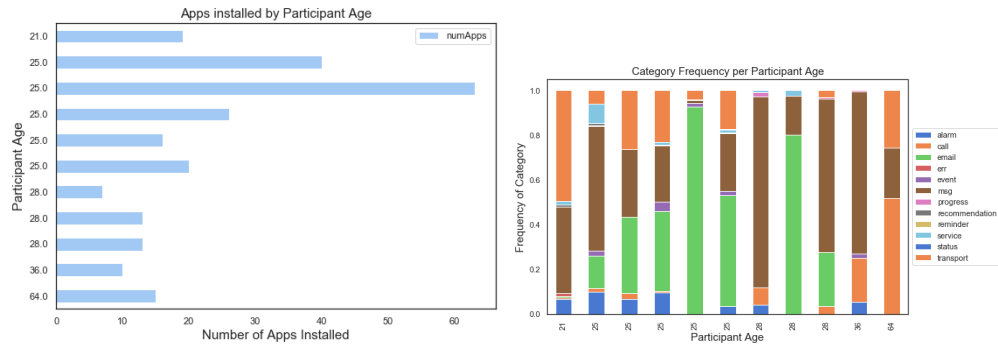


Fig. B.1 Number of apps installed and frequency of notification categories received for participants of different age

# Appendix B

## Supplementary Material for Chapter 3

### B.1 Smartphone Usage and Notification Engagement

In this section we investigate the relationships between participant characteristics and notification engagement behaviour as well as general smartphone usage. Key findings are summarised in B.1.5.

### B.1.1 Participant Age, Apps and Notifications

Participants that engaged with the study ranged in age from 21 to 64 allowing for an analysis of smartphone usage between young and old. On sign-up to the WeAreUs application, the number of unique apps installed on participant devices was recorded. On average, participants had a selection of 21 apps on their device ( $M=21.31$ ,  $SD=14.99$ ). A Spearman's rank-order correlation was run to assess the relationship between participant age and the number of apps installed. There was a statistically significant, strong negative correlation between age and the number of apps installed,  $r_s = -.714$ ,  $p = .014$ . That is, the number of apps installed decreased as age increased. This is an important relationship, as the number of apps installed on the device can give insight into the diversity of notifications expected. The hypothesis being that a larger array of apps allows for a greater range of categories and topics to be pushed at participants. This was also measured through the WeAreUs application as the topic and category of incoming notifications were recorded. On average, participants received notifications spanning 7 different categories ( $M=7.15$ ,  $SD=2.27$ ) and 18 different topics ( $M=18.39$ ,  $SD=6.55$ ). A Spearman's rank-order correlation was run to assess the relationships between the number of apps installed and unique topics and categories received in notifications. There was a statistically significant, strong positive correlation between the number of apps installed and unique categories received,  $r_s = .793$ ,  $p = .001$ , but not with unique topics. This illustrates that younger participants with larger selections of apps received notifications which were more diverse in terms of notification category, but not necessarily in terms of notification topic. Older participants, with a smaller selection of apps, received notifications which were less diverse in terms of notification category. This suggests that models developed for predicting notification usage behaviour might be simpler for older users. Figure B.1 illustrates the number of installed apps and frequency of categories received per participant age.

### B.1.2 Notification Topics & Places of Engagement

Spearman's rank-order correlation was also indicated a statistically significant, moderate positive correlation between the average number of notifications received per day by a participant ( $M=44.50$ ,  $SD=26.14$ ) and the number of unique topics received,  $r_s = .606$ ,  $p < .028$ . Figure B.2 illustrates the average number of notifications received per day and the notification topic frequencies per participant.

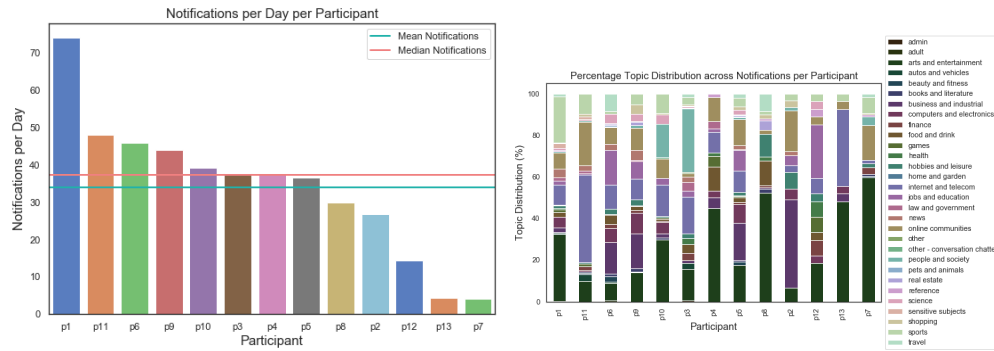


Fig. B.2 Average number of notifications received per day and frequency of notification topics received per participant

As such, the diversity of topics increased as the number of average daily incoming notifications increased, suggesting that participants overloaded with notifications also had to deal with a wider spectrum of topics when deciding to open or dismiss notifications, potentially adding to the cognitive load. This relationship was explored by running a Spearman's rank-order correlation between the average time participants spent deciding on whether to open or dismiss a notification (measured in seconds) and the number of unique topics received. There was a statistically significant, moderate positive correlation between the average decision time in minutes ( $M=1.96$ ,  $SD=1.16$ ) and the number of unique topics received,  $r_s = .645$ ,  $p = .017$ , such that as the number of unique topics increased, the average time participants spent deciding what action to take on a notification also increased. It is important to note that we do not assume a causation effect, but simply identify an association between these two variables, as there are many, potentially confounding, factors which may contribute to the decision time. For example, the places in which participants engaged with their notifications was also recorded by the WeAreUs app, and the relationship between the average decision time and the number of unique place categories where participants dismissed notifications was also found to be statistically significant,  $r_s = .724$ ,  $p = .018$ . The strong positive correlation shows participants who dismissed notifications in more types of place took longer, on average, to act upon notifications.

Over the course of the study, participants received notifications in 51 different place types and made an action on a notification (open/dismiss) in 52 different place types. Figures B.3a and B.3b illustrate the top place types where notifications were received and acted on, respectively. Notification place types such as *cafe*, *night club* and *lodging* had much higher CTRs than those of *university*, *gym* and *supermarket*, the former places associated with relaxed leisure while the

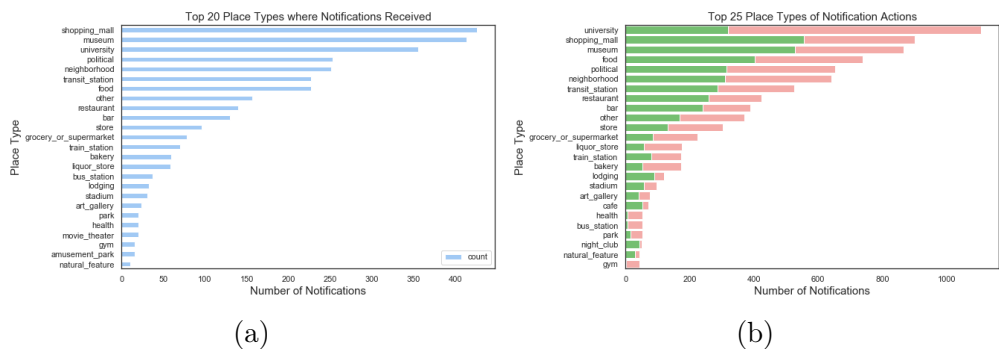


Fig. B.3 Notifications delivered and engaged with over differing place types

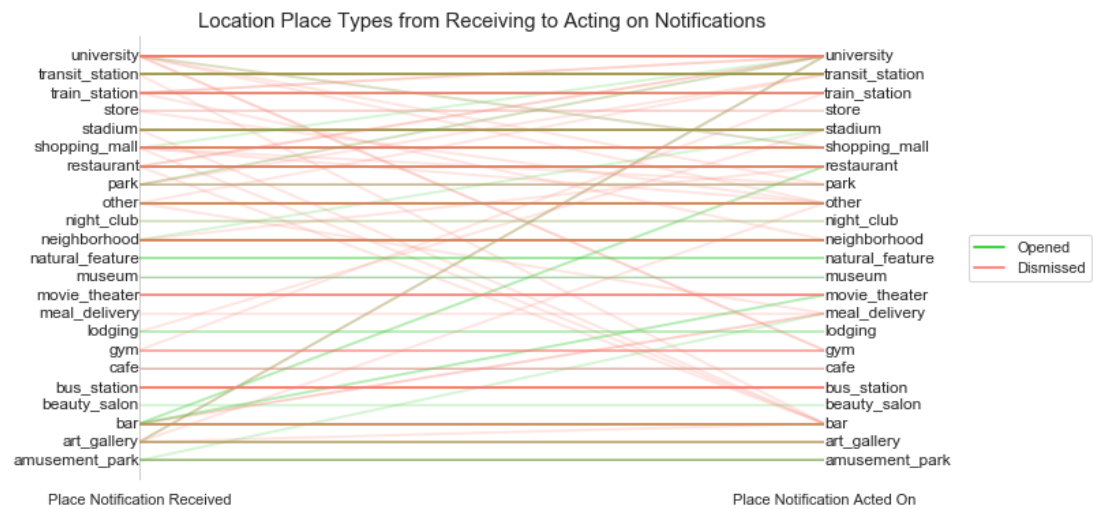


Fig. B.4 Changes of place type between notification delivery and notification action

latter places suggest more focused activities occurring (e.g. studying, exercising or grocery shopping). Figure B.4 also visualises notification engagements in which a notification was delivered in one place type and acted on in another. The edge colours represent the action taken on a notification and the strength of the edge reflects the number of occurrences. The strongest edges are those in which no change occurs between place types indicating that a large portion of notifications were acted on in the same place they were originally delivered. However, some interesting relationships can be identified in which a participant's context changed. For example, notifications delivered in bars, and acted on in restaurants or movie theaters, had high open rates. In contrast, notifications delivered in restaurants, shopping malls or stadiums were subsequently dismissed in bars. Section 3.2.5.5 further analyses changing contexts and the implications for notification engagement.

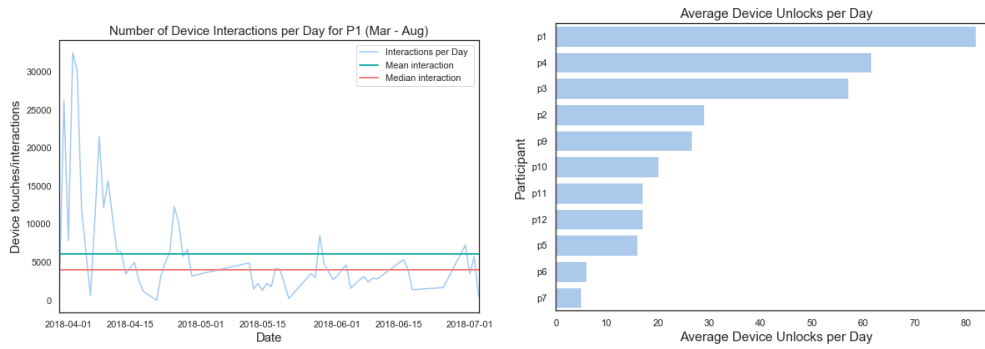


Fig. B.5 Number of device interactions per day (single participant) and the average unlock count per day (all participants)

### B.1.3 Notification Receptivity & Device Interactions

Participant actions toward notifications were observed throughout the study. The CTRs recorded showed that, on average, participants were opening slightly less than half of all notifications delivered ( $M=48.35\%$ ,  $SD=13.57$ ). This illustrates that the delivery method for notifications was sub-optimal and resulted in participants receiving 23 notifications on average per day, which they subsequently dismissed. Unsurprisingly, there was a statistically significant, strong negative correlation between participant CTR and average device battery level when receiving notifications,  $r_s = -.740$ ,  $p = .004$ . Participants who opened more notifications also had lower average battery levels at moments of receiving notifications. Conversely, the relationship between CTR and average smartphone usage time per day was significant and strongly positively correlated,  $r_s = .855$ ,  $p < .001$ . Participants who opened more notifications also spent more time on their smartphone throughout the day. This was also apparent by analysing the number of times participants touched their smartphone per day ( $M=1571.45$ ,  $SD=2754.08$ ).

The WeAreUs app tracked each time a participant interactively touched their device while it was unlocked. A Spearman's rank-order correlation was run to assess the relationship between CTR and the average number of touches per day. There was a statistically significant, strong positive correlation,  $r_s = .770$ ,  $p = .009$ , such that participants who accepted more notifications also interactively touched their smartphone more throughout the day. Understandably, there was also a statistically significant, strong positive correlation, between the average number of touches per day and the average number of times a participant unlocked their device during the day,  $r_s = .729$ ,  $p = .011$ . Figure B.5 demonstrates the daily interactions of a participant with their device over the study period as well as the average number of times participants unlocked their devices during the day. Participants

who unlocked their device frequently also physically touched their device more throughout the day. This is important to note as incoming notifications persuade mobile users to interact with their device when they otherwise, may not have done. Particularly in our current climate in which Covid-19 has been shown to spread rapidly through communities [Bi+20], hand hygiene and limiting unnecessary interactions with smartphones has become increasingly important. In a study of 200 health care worker devices, 94.5% were identified as being contaminated with bacteria [Ulg+09]. Another study involving 316 smartphones owned by doctors, hospital staff, college faculty, medical students and a control group identified that 81.8% of the smartphones sampled "showed growth of bacterial pathogens" and suggested that "strict adherence to infection control policies such as proper hand hygiene practices should be followed" [Pal+15]. In addition, as notifications are regularly dismissed, mobile users might be interacting with their device more frequently but for little added value, which may not be worth the risk. This was perhaps illustrated in the relationship between average unlocks per day and average notifications received. There was a statistically significant, moderate positive correlation,  $r_s = .647$ ,  $p = .031$  between them demonstrating that participants who received a higher number of notifications on average per day also had a higher average unlock rate per day. Similarly, the average time taken for a participant to notice a notification had been delivered to their device, defined as seen time, was statistically significantly lower when their average number of unlocks per day was higher,  $r_s = -.656$ ,  $p = .028$ . These relationships illustrate that push-notifications are intrinsically linked to how participants engage with, and exert energy into, their smartphones.

#### B.1.4 Elapsed Engagement Time & App Significance

The WeAreUs app measured the elapsed time between three notification engagement points of interest to enable further understanding of the behaviour of participants toward incoming notifications. We have already touched upon the significance of these relationships in the previous sections e.g. the relationship between average decision time and the number of unique topics; the relationship between average seen time and average unlocks per day. Participants, on average, noticed a notification on their device within 14 minutes ( $M=13.96$ ,  $SD=4.72$ ) of arrival. This is relevant only to notifications which were delivered when a participant was not using their device (i.e. it was locked), which accounts for

31.21% of notifications. The remaining 68.79% of notifications were delivered when devices were unlocked and we assume participants noticed these notifications immediately.

It subsequently took, on average, under 2 minutes ( $M=1.96$ ,  $SD=1.16$ ) for participants to decide how to act (open/dismiss) toward notifications, once they unlocked their device, although decision times differ greatly depending on whether the device was unlocked or not (i.e. immediately seen by the participant) at the time of notification delivery (see Figure B.6). Breaking the average decision time into these two categories yields an average of just over 3 and a half minutes ( $M=3.37$ ,  $SD=4.73$ ) for notifications seen immediately and an average decision time of under a minute ( $M=0.88$ ,  $SD=0.37$ ) for notifications delivered when the device was locked.

The average total response time, from notification delivery to a participant action, was just over 5 and a half minutes ( $M=5.63$ ,  $SD=4.67$ ), but again response times differ depending on whether devices are unlocked or not at the time of delivery. Breaking the average response time into these two categories yields an average of just over 3 minutes ( $M=3.35$ ,  $SD=3.37$ ) for notifications seen immediately and an average response time of just over 18 minutes ( $M=18.13$ ,  $SD=7.24$ ) for notifications delivered when the device was locked. From this analysis, it is clear that notifications delivered to participants when their device was unlocked had a quicker engagement turn-around than those arriving to locked devices. However, the increase in decision time for delivery at unlocked states may indicate that these notifications were interrupting participants focused on other device tasks.

The response time, similar to decision time, also has a statistically significant, moderate positive correlation,  $r_s = .678$ ,  $p = .011$ , with the number of unique topics received from notifications, illustrating that participants who received a greater array of topics through notifications also took longer on average to respond to their notifications.

The WeAreUs app also measured the amount of time elapsed between when a participant used an app and when the app sent them a notification. There is a statistically significant, strong positive correlation between the average notification response time and the average elapsed time between using an app and receiving a notification from it,  $r_s = .797$ ,  $p = .001$ . This illustrates that apps that were used by participants more recently were also responded to quicker. Similarly, statistically significant positive moderate correlations were also found for the average seen,  $r_s = .692$ ,  $p = .009$ , and decision times,  $r_s = .670$ ,  $p = .012$ ,

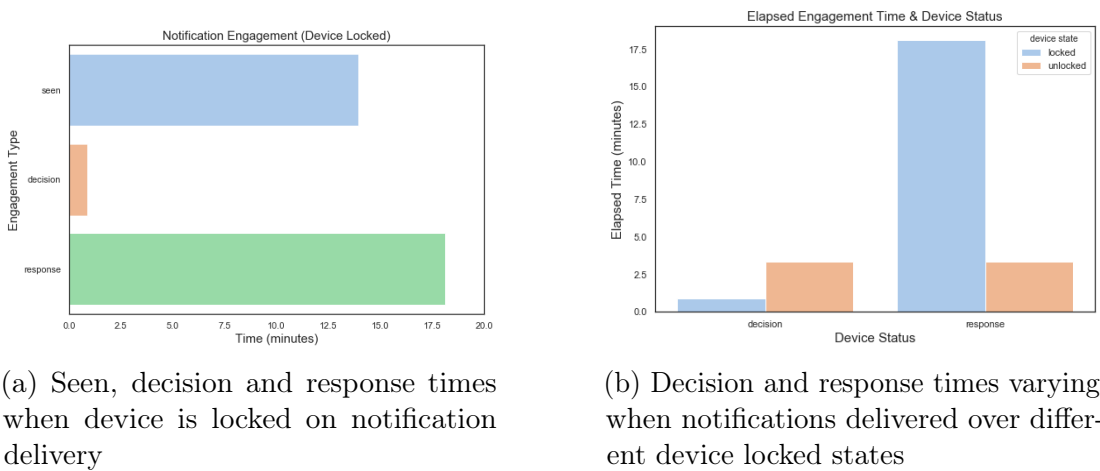


Fig. B.6 Elapsed time between notification engagement points of interest: seen, decision & response.

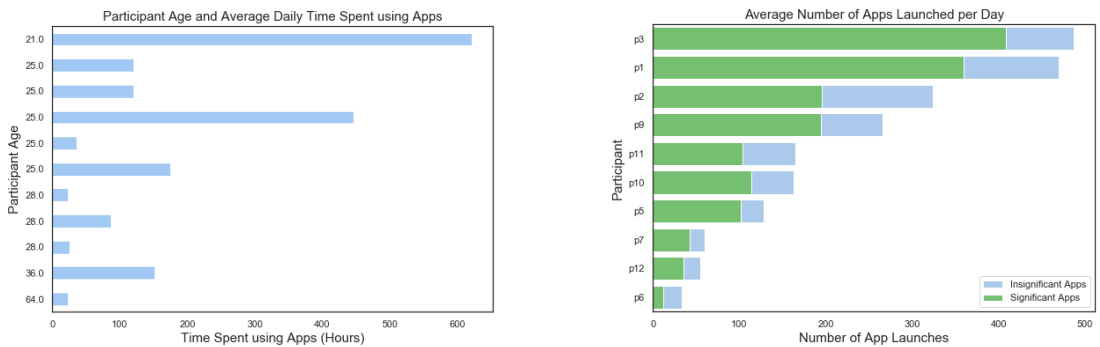


Fig. B.7 Participant app usage metrics

indicating that apps recently used also had their notifications noticed and decided upon quicker.

The WeAreUs app also recorded the interactions participants had with their installed apps (e.g. number of times apps were opened, time spent within apps) over the period of the study. App significance was then determined, in real-time, by calculating the average number of times apps were opened for a participant and categorising apps which were launched over above the average number of times as significant and below the average as insignificant, with respect to a particular participant. App significance was a dynamic feature which was recalculated at hourly intervals with respect to participant smartphone usage over the course of the study. Participants launched, on average, 344 apps per day ( $M=344.36$ ,  $SD=456.91$ ). This number includes multiple launches of the same app,  $\approx 6.5$  per app per day. Participants launched, on average, 53 ( $M=53.16$ ,  $SD=41.57$ ) distinct apps per day. Comparing the average number of app launches per day with the



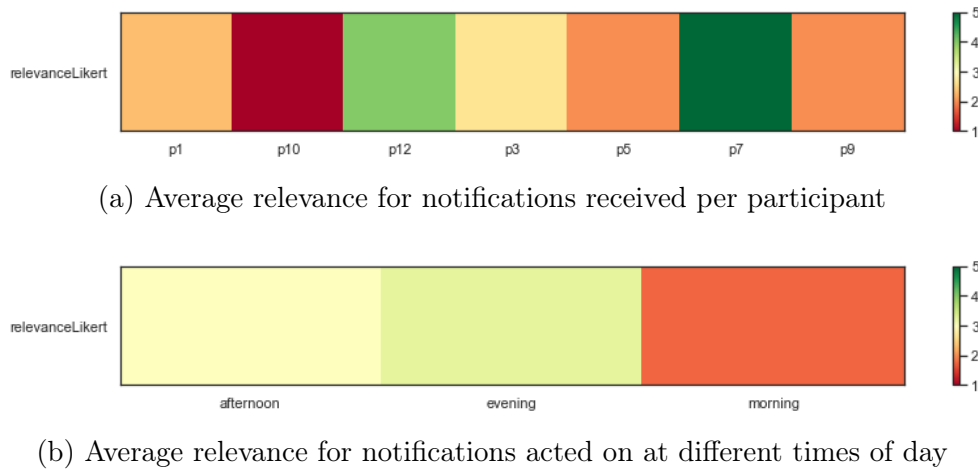


Fig. B.8 Average relevance scores for notifications received by participants (1:VeryIrrelevant; 2:Irrelevant; 3:Neutral; 4:Relevant; 5:VeryRelevant)

number of times participants touched their smartphone per day interestingly suggests an app launch occurred once every 10 times a participant touched/interacted with their unlocked smartphones. Figure B.7 illustrates the average number of app launches for significant and insignificant apps as well as the average time participants spent using apps per day. A Spearman's rank-order correlation was run to assess the relationship between distinct app launches per day and the number of notifications received per day. There was a statistically significant, strong positive correlation between them,  $r_s = .764$ ,  $p = .006$ , illustrating that participants opening a higher number of unique apps per day also received more notifications per day. Understandably, there were also statistically significant strong positive correlations between distinct app launches per day and the number of times participants unlocked their device during the day,  $r_s = .870$ ,  $p < .001$ , as well as the the number of touches/interactions they had with their device during the day,  $r_s = .755$ ,  $p = .007$ . Overall, the analysis shows that participants who engaged with a greater number of apps on their device, also received more notifications and unlocked their device more frequently.

Breaking down participant app launch records further, using the inferred significance flag, the average number of significant app launches per day ( $M=250.41$ ,  $SD=336.15$ ) was four times the number of insignificant app launches ( $M=64.64$ ,  $SD=59.19$ ). Participants opened 32 ( $M=32.27$ ,  $SD=26.78$ ) distinct significant apps per day on average and only 21 ( $M=20.86$ ,  $SD=15.23$ ) distinct insignificant apps per day. The average amount of time participants spent in apps throughout the day was over 2 and a half hours ( $M=2.77$ ,  $SD=3.22$ ). A statistically significant

negative moderate correlation was found between the average amount of time spent on apps and the age of participants,  $r_s = -.645$ ,  $p = .032$ , demonstrating that older participants spent less time in apps than younger participants. The WeAreUs app also recorded qualitative feedback from participants at moments they acted upon notifications. One such question in the questionnaire presented to participants queried the relevance of the notification to the given moment. Using Spearman's rank-order correlation, a statistically significant strong negative correlation was found between the average amount of time spent on apps and the average relevance of notifications,  $r_s = -.823$ ,  $p = .023$ , such that participants who spent greater amount of time using apps also found the notifications they received less relevant. Figure B.8 illustrates the average relevance score given by participants receiving notifications throughout the WeAreUs study. Just two of the seven participants who responded to the relevance questions of the ESM questionnaires believed their notifications were relevant. Furthermore, notifications received during the morning received the worst scores for relevancy. The relevance of notifications and other qualitative features recorded by the WeAreUs app are explored further in the following sections.

### B.1.5 Key Findings

- Number of unique apps installed reduced as participant age increased
- Participants with a higher number of installed apps received more distinct notification categories
- Participants receiving higher numbers of notifications also dealt with higher distinct notification topics and spent more time deciding how to act (open/dismiss) toward notifications
- Participants engaged with notifications across  $\approx 52$  distinct place types
- Participants dismissing notifications at more distinct place types typically spent longer deciding how to act (open/dismiss)
- Less than 50% of notifications pushed at participants were opened
- Participants were physically dismissing 23 notifications, on average, per day

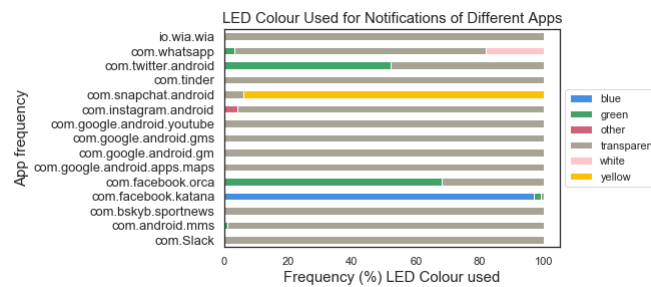


Fig. B.9 Frequency of LED colours implemented in notifications pushed by different apps

- Participants receiving higher numbers of notifications also unlocked their device more frequently throughout the day and spent more time on their smartphone
- 14 minutes for participants to unlock their device after a notification was delivered to their locked device
- Over 3 and a half minutes for participants to decide how to act (open/dismiss) toward a notification when their device was unlocked. Under a minute when it arrived when the device was locked
- 3 minutes for participants to respond to notifications when delivered while device unlocked. 18 minutes when it was locked
- Participants who were sent a larger array of topics typically took longer to decide to act on a notification
- Apps used recently by participants, had notifications seen and acted upon quicker
- App launches occurred, on average, once in every ten touches/interactions of the smartphone
- Participants spent, on average, over 2 and a half hours using apps per day
- Participants spending more time in apps on average found push-notifications they received less relevant

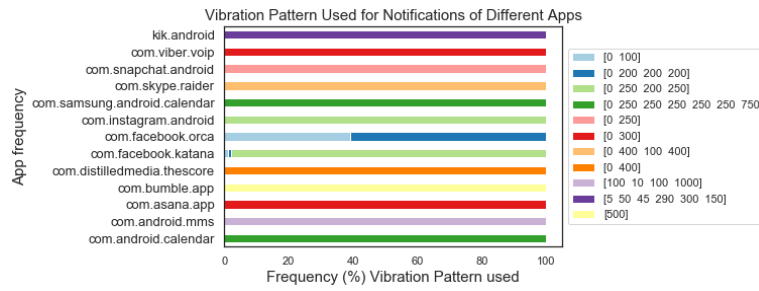


Fig. B.10 Frequency of vibration patterns implemented by different apps when pushing notifications

### B.1.6 The Role of the Sender

A chi-square test of independence was conducted between the LED colour of a notification and whether it was from a significant contact. 7.1% of cells had an expected count less than 5. There was a statistically significant association between the LED colour and contact significance,  $\chi^2(6) = 615.27, p < .001$ . The association was small to moderately strong, Cramer's  $V = .222$ . The majority of notifications with and without a significant contact associated had *transparent* set as the LED colour (78% and 84.4% respectively), meaning the LED did not visibly alert the user on delivery of the notification. Following this, white was the next most frequently seen colour for notifications with a significant contact (17.1%) and green was the next most frequently seen colour for notifications without a significant contact (5.3%). LED colours of notifications pushed from differing apps can be seen in Figure B.9.

In addition, a chi-square test of independence was conducted between the vibration pattern of a notification and whether it was from a significant contact. 16.7% of cells had an expected count less than 5. There was a statistically significant association between the vibration pattern and contact significance,  $\chi^2(8) = 207.257, p < .001$ . The association was moderate to large, Cramer's  $V = .421$ . Similar to the LED colour, the majority of notifications with and without a significant contact associated were set to not vibrate (72.7% and 21.4% respectively). Of those that did, the vibration pattern *[100 10 100 1000]* was most frequently used when a significant contact was associated (13.3%) and the pattern *[0 250 200 250]* was most frequently found in notifications without a significant contact associated (29.2%). Vibration patterns present in notifications pushed from differing apps can be seen in Figure B.10.

The results of these tests illustrate that a significant relationship exists between alert modalities and the presence of an important contact relevant to the context

Table B.1 Observed frequencies and percentages of notifications responded to at different times of day with respect to contact significance

| Time of Day   | isContactSignificant |                | <i>p</i> |
|---------------|----------------------|----------------|----------|
|               | False                | True           |          |
| afternoon     | 3289<br>(31.6%)      | 661<br>(31.8%) | 0.809    |
| early morning | 593<br>(5.7%)        | 87<br>(4.2%)   | 0.006*   |
| evening       | 2606<br>(25.0%)      | 667<br>(32.1%) | <.001*   |
| morning       | 3291<br>(31.6%)      | 549<br>(26.4%) | <.001*   |
| night         | 639<br>(6.1%)        | 112<br>(5.4%)  | 0.196    |

of a participant. Additionally, the vast majority of notifications delivered to participants did not make use of these modalities, which clearly would have aided participants distinguishing between important and irrelevant notification interruptions.

In addition to contextual features, the relationship between other notification features and the presence of a significant contact were also explored. From 5,888 notifications, the WeAreUs app succeeded in identifying both the category of notification and the significance of a contact. A chi-square test of homogeneity was run, with an adequate sample size established. The two multinomial probability distributions were not equal in the population,  $\chi^2(3) = 437.064$ ,  $p < .001$ . Observed frequencies and percentages of categories of notifications, and whether the notification was from a significant contact, are presented in Table B.2. Post-hoc analysis involved pairwise comparisons using multiple z-tests of two proportions with a Bonferroni correction. Statistical significance was accepted at  $p < .013$ . There were statistically significant differences in the proportion of *email*, *msg* and *other* categories of notification received from significant and insignificant contacts. The highest proportion of notifications were found to be in the *msg* category for notifications with and without a significant contact present. However, the proportion of notifications received and categorised as email and other were much higher, and closer to the *msg* category, for insignificant contacts. The notification categories are set by the app pushing the notification and each app has the ability to push any type of category. In this case, the results illustrate that notifications received by participants during the WeAreUs study, associated with a significant contact, were mainly categorised as messages.

Table B.2 Observed frequencies and percentages of notification categories with respect to contact significance

| Category     | isContactSignificant |                | <i>p</i> |
|--------------|----------------------|----------------|----------|
|              | False                | True           |          |
| <b>email</b> | 1568<br>(33.8% )     | 305<br>(24.5%) | <.001*   |
| <b>msg</b>   | 1994<br>(42.9%)      | 911<br>(73.2%) | <.001*   |
| <b>other</b> | 1059<br>(22.8%)      | 29<br>(2.3%)   | <.001*   |
| recommend    | 22<br>(0.5%)         | 0<br>(0.0%)    | 0.015    |

In 4,719 notifications, the WeAreUs app succeeded in identifying both the topic of notification and the significance of a contact. A chi-square test of homogeneity identified that the two multinomial probability distributions were not equal in the population,  $\chi^2(10) = 469.848$ ,  $p < .001$ . Observed frequencies and percentages of topics of notifications, and whether the notification was from a significant contact, are presented in Table B.3. Post-hoc analysis involved pairwise comparisons using multiple z-tests of two proportions with a Bonferroni correction. Statistical significance was accepted at  $p < .005$ . There were statistically significant differences in the proportions of many topics of notification received between significant and insignificant contacts (denoted by an asterisk in Table B.3). Of particular interest was the difference of proportions found in notifications of the *internet telecoms* and *online communities* topics. The proportion of notifications was low for these topics when no significant contact was identified and, conversely, much higher when a significant contact was found.

The WeAreUs app succeeded in identifying both the app of the notification and the significance of a contact (open/dismissal) in 12,494 notifications. A chi-square test of homogeneity found that the two multinomial probability distributions were not equal in the population,  $\chi^2(9) = 1957.036$   $p < .001$ . Observed frequencies and percentages of notification app, and whether the notification was from a significant contact, are presented in Table B.4. Post-hoc analysis involved pairwise comparisons using multiple z-tests of two proportions with a Bonferroni correction. Statistical significance was accepted at  $p < .005$ . There were statistically significant differences in the proportions of notifications, with and without a significant contact, for all but one app: *Gmail*. Of particular interest were *Snapchat* and

Table B.3 Observed frequencies and percentages of notification topics with respect to contact significance

| Topic                          | isContactSignificant |                | <i>p</i> |
|--------------------------------|----------------------|----------------|----------|
|                                | False                | True           |          |
| arts & entertainment           | 851<br>(22.0%)       | 212<br>(25.1%) | 0.044    |
| business & industry            | 483<br>(12.5%)       | 18<br>(2.1%)   | <.001*   |
| computers & electronics        | 293<br>(7.6%)        | 27<br>(3.2%)   | <.001*   |
| food & drink                   | 101<br>(2.6%)        | 15<br>(1.8%)   | 0.160    |
| <b>internet &amp; telecoms</b> | 318<br>(8.2%)        | 209<br>(24.8%) | <.001*   |
| jobs & education               | 254<br>(6.6%)        | 54<br>(6.4%)   | 0.875    |
| news                           | 133<br>(3.4%)        | 5<br>(0.6%)    | <.001*   |
| <b>online communities</b>      | 308<br>(7.9%)        | 181<br>(21.5%) | <.001*   |
| other                          | 576<br>(14.9%)       | 81<br>(9.6%)   | <.001*   |
| people & society               | 88<br>(2.3%)         | 9<br>(1.1%)    | 0.026    |
| sports                         | 471<br>(12.2%)       | 32<br>(3.8%)   | <.001*   |

Table B.4 Observed frequencies and percentages of notification **app** with respect to contact significance

| <b>App</b>                  | <b>isContactSignificant</b> |                 | <b><i>p</i></b> |
|-----------------------------|-----------------------------|-----------------|-----------------|
|                             | <b>False</b>                | <b>True</b>     |                 |
| com.android.chrome          | 162<br>(1.6%)               | 0<br>(0.0%)     | <.001*          |
| com.facebook.katana         | 300<br>(2.9%)               | 0<br>(0.0%)     | <.001*          |
| com.facebook.orca           | 462<br>(4.4%)               | 3<br>(0.1%)     | <.001*          |
| com.google.android.gm       | 1556<br>(14.9%)             | 305<br>(14.7%)  | 0.776           |
| <b>com.snapchat.android</b> | 174<br>(1.7%)               | 93<br>(4.5%)    | <.001*          |
| com.spotify.music           | 350<br>(3.4%)               | 11<br>(0.5%)    | <.001*          |
| com.tinder                  | 623<br>(6.0%)               | 12<br>(0.6%)    | <.001*          |
| com.twitter.android         | 492<br>(4.7%)               | 11<br>(0.5%)    | <.001*          |
| <b>com.whatsapp</b>         | 2680<br>(25.7%)             | 1479<br>(71.2%) | <.001*          |
| other                       | 3619<br>(34.7%)             | 162<br>(7.8%)   | <.001*          |

*WhatsApp*, as both apps held a larger proportion of notifications with a significant contact than without.

### B.1.7 Pushing with the Right App

Observed frequencies and percentages of LED colours present in notifications, and whether the notifications were pushed by a significant app, are presented in Table B.5. Post-hoc analysis involved pairwise comparisons using multiple z-tests of two proportions with a Bonferroni correction. Statistical significance was accepted at  $p < .007$ . There were statistically significant differences in the proportions of many notification LED colours received from significant and insignificant apps. Of particular interest were the difference of proportions found in notifications with *white* as its LED colour. The proportion of notifications with this type of alert modality was higher across apps inferred as significant, 11.9%, as opposed to proportions found across insignificant apps, 0.1%. Studying Figure B.9



Table B.5 Crosstabulation of app significance and LED colour of notification (significance accepted at  $p < .007$ ).

| LED Colour   | isAppSignificant |                 | <i>p</i> |
|--------------|------------------|-----------------|----------|
|              | False            | True            |          |
| blue         | 187<br>(2.9%)    | 228<br>(3.1%)   | 0.570    |
| green        | 538<br>(8.3%)    | 288<br>(3.9%)   | <.001*   |
| magenta      | 14<br>(0.2%)     | 0<br>(0%)       | <.001*   |
| yellow       | 268<br>(4.1%)    | 85<br>(1.1%)    | <.001*   |
| other        | 28<br>(0.4%)     | 14<br>(0.2%)    | .008     |
| <b>white</b> | 5<br>(0.1%)      | 887<br>(11.9%)  | <.001*   |
| transparent  | 5429<br>(83.9%)  | 5962<br>(79.9%) | <.001*   |

reveals *WhatsApp* to be the only app pushing notifications with *white* set as the LED colour, thus illustrating its significance.

Observed frequencies and percentages of notification vibration pattern, and whether the notification was pushed by a significant app, are presented in Table B.6. Post-hoc analysis involved pairwise comparisons using multiple z-tests of two proportions with a Bonferroni correction. Statistical significance was accepted at  $p < .007$ . There were statistically significant differences in the proportions of all but one vibration pattern found in notifications received from significant and insignificant apps: *[0]* (which indicates no vibration was set for the notification). Of particular interest were the difference of proportions found in notifications with *[0 250 200 250]* and *[0 500]* set as vibration patterns as the proportion of notifications with these patterns were higher across apps inferred as significant as opposed to proportions found across insignificant apps. The *[0 500]* vibration pattern was not used by any app inferred as insignificant.

Observed frequencies and percentages of notification priority level, and whether the notification was pushed by a significant app, are presented in Table B.7. Post-hoc analysis involved pairwise comparisons using multiple z-tests of two proportions with a Bonferroni correction. Statistical significance was accepted at  $p < .008$ . There were statistically significant differences in the proportions all priority levels found in notifications received from significant and insignificant apps.

Table B.6 Crosstabulation of app significance and vibration patterns of notification (significance accepted at  $p < .007$ ).

| Vibration Pattern      | isAppSignificant |                | <i>p</i> |
|------------------------|------------------|----------------|----------|
|                        | False            | True           |          |
| [0 100]                | 90<br>(16.4%)    | 53<br>(9.3%)   | <.001*   |
| [0 200 200 200]        | 163<br>(29.7%)   | 62<br>(10.9%)  | <.001*   |
| <b>[0 250 200 250]</b> | 135<br>(24.6%)   | 214<br>(37.5%) | <.001*   |
| [0]                    | 43<br>(7.8%)     | 35<br>(6.1%)   | 0.266    |
| <b>[0 500]</b>         | 0<br>(0%)        | 156<br>(27.4%) | <.001*   |
| other                  | 118<br>(21.5%)   | 50<br>(8.8%)   | <.001*   |

Of particular interest was the difference of proportions found in notifications with *high* set as the priority level as the proportion of notifications with this level was higher across apps inferred as significant, 28.8%, as opposed to proportions found across insignificant apps, 13.8%. Figure 3.7b illustrates that *default* was the most common priority level chosen for notifications delivered during the study and it appeared just as frequently in notifications delivered by apps inferred as significant as insignificant. The *high* priority level was the next most frequent level used for notification priority and it appeared more frequently in notifications pushed by significant apps. Note also, the low number of notifications that appeared with *low* and *min* priority levels, perhaps indicating no app willing to yield in the war over subscriber attention.

## B.2 Additional Figures & Tables

Table B.7 Crosstabulation of app significance and priority level of notification (significance accepted at  $p < .008$ ).

| Priority Level | isAppSignificant |                 | <i>p</i> |
|----------------|------------------|-----------------|----------|
|                | False            | True            |          |
| min            | 353<br>(5.5%)    | 0<br>(0%)       | <.001*   |
| low            | 116<br>(1.8%)    | 0<br>(0%)       | <.001*   |
| default        | 4438<br>(68.9%)  | 4719<br>(63.5%) | <.001*   |
| <b>high</b>    | 887<br>(13.8%)   | 2141<br>(28.8%) | <.001*   |
| max            | 643<br>(10%)     | 572<br>(7.7%)   | <.001*   |

Table B.8 A list of notification engagement features recorded by the WeAreUs app

| Feature      | Description                                                 |
|--------------|-------------------------------------------------------------|
| appPackage   | the app posting the notification                            |
| deliveryTime | datetime the notification was delivered                     |
| isClearable  | whether notification can be removed by the user             |
| isOngoing    | whether notification is ongoing e.g. download               |
| text         | text included in notification (includes all text fields)    |
| category     | category of notification                                    |
| ledARGB      | led color                                                   |
| priority     | priority level of notification                              |
| vibrate      | vibration pattern to be used on arrival of notification     |
| visibility   | when and how sytem UI reveals notification                  |
| actionTime   | datetime when the notification is removed                   |
| action       | whether user opened the notification                        |
| numUpdates   | total number of times a notification was updated            |
| appLastUsed  | time posting app was last used by the user                  |
| seenTime     | the time the notification was seen by user                  |
| decisionTime | time from seen time to when user acts on it                 |
| responseTime | total time from posting to removal                          |
| latitude     | latitude of device on delivery/action of notification       |
| longitude    | latitude of device on delivery/action of notification       |
| keywords     | entities extracted from notification text with Google's NLP |
| topic        | topic as defined by Google NLP                              |
| isContactSig | is notification associated with significant contact         |
| activity     | activity of user (Activity Recognition API)                 |
| isAppSig     | is the app posting the notification significant             |
| placeType    | category of place as defined by Google Places API           |
| ambientLight | ambient light levels around the device                      |
| ambientNoise | ambient noise levels around the device                      |

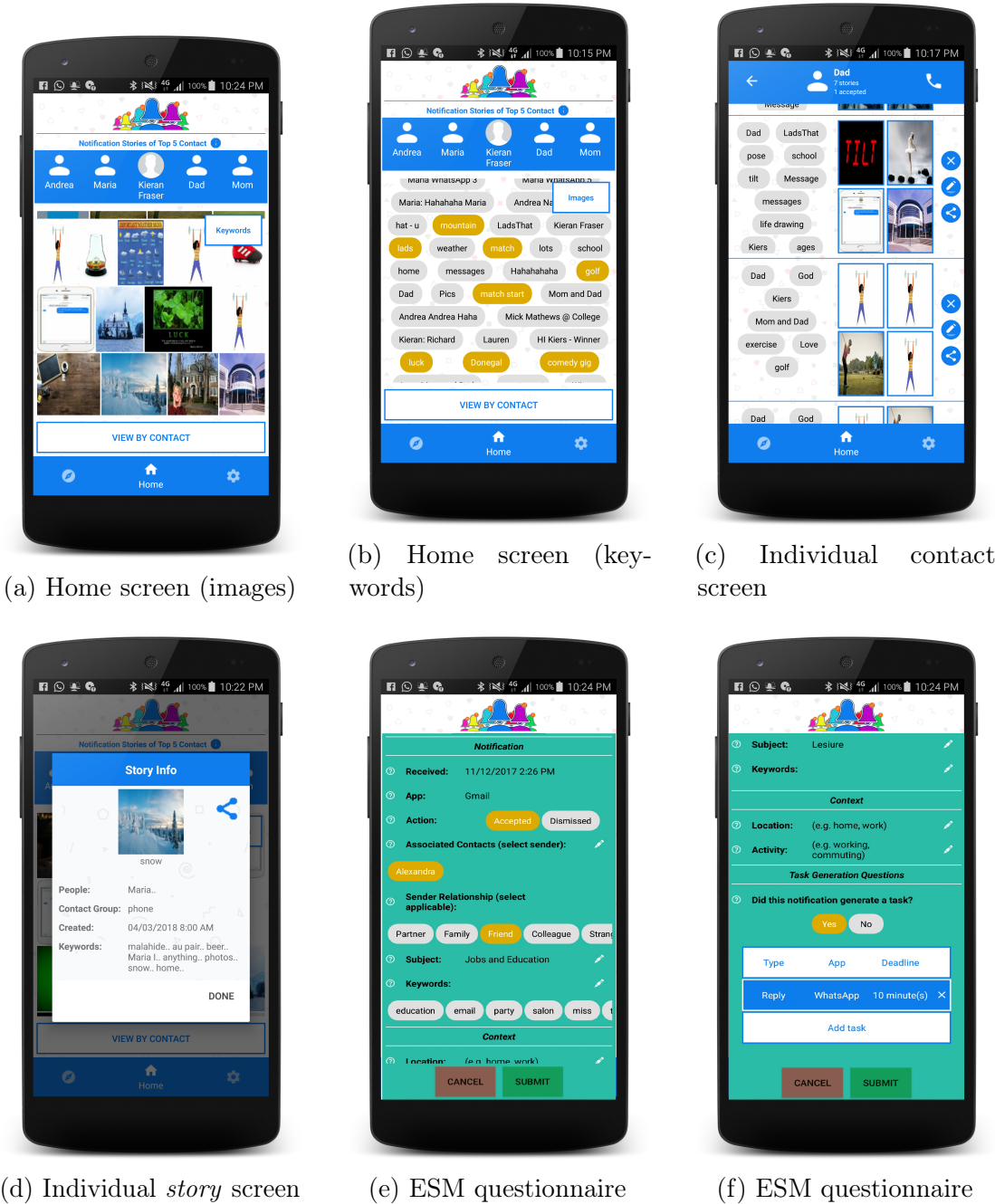


Fig. B.11 Screenshots of the WeAreUs mobile application used for capturing notification engagement and mobile usage data

Table B.9 A list of aggregated smartphone usage metrics recorded by the WeAreUs app at hourly intervals

| Feature                          | Description                                  |
|----------------------------------|----------------------------------------------|
| notificationCount                | Number of notifications received             |
| notificationAcceptedCount        | Average number of notifications accepted     |
| notificationSeenTimeAvg          | Average notification seen time               |
| notificationDecisionTimeAvg      | Average notification decision time           |
| notificationResponseTimeAvg      | Average notification response time           |
| unlockCount                      | Number of times phone unlocked               |
| clickCount                       | Number of times user interacted with device  |
| phoneUsageTime                   | Time user spends interacting with phone      |
| totalTimeInApps                  | Total time user spends in apps               |
| totalAppLaunchCount              | Total number of times apps launched          |
| uniqueAppsLaunched               | Number of unique apps launched               |
| totalSignificantAppsLaunchCount  | Number of significant apps launched          |
| uniqueSignificantAppsLaunched    | Number of unique significant apps launched   |
| totalInsignificantAppLaunchCount | Number of insignificant app launches         |
| uniqueInsignificantAppsLaunched  | Number of unique insignificant apps launched |
| callCount                        | Number of calls received                     |
| callUniqueCount                  | Number of calls from unique callers          |
| callDurationAvg                  | Average call duration                        |
| smsCount                         | Number of sms received                       |
| smsUniqueCount                   | Number of sms received from unique senders   |
| smsAverageLength                 | Average length of sms body                   |
| smsSentReceivedRatio             | Ratio of sent sms to received sms            |

Table B.10 A list of qualitative and quantitative features collected through the ESM questionnaire presented at moments of notification action

| Feature            | Description                                                      |
|--------------------|------------------------------------------------------------------|
| received           | Time notification received                                       |
| app                | App which posted the notification                                |
| action             | Whether or not notification clicked or dismissed                 |
| senderRelationship | Relationship user has with sender of notification                |
| topic              | Topic of notification (Google's NLP or user specified)           |
| place              | Participant current place (Google's Places API)                  |
| keywords           | Entities found in text using Google NLP or user specified        |
| behaviourScale     | Action taken is considered good/bad behaviour (Likert)           |
| behaviourReason    | Reason for behaviour e.g. 'aids goal', 'distracted'              |
| relevanceScale     | Notification was relevant to current context (Likert)            |
| relevanceEvidence  | Feature which lead user to believe this e.g. 'sender', 'subject' |
| modeCorrect        | Delivered through correct device e.g. computer, smartwatch       |
| modeAlternative    | Device the notification should have arrived through              |
| modeEvidence       | Feature which prompts this response e.g. 'subject', 'location'   |

Table B.11 A list of qualitative and quantitative features collected through the ESM questionnaire presented at moments participants unlocked their smartphone

| Feature   | Description                                                   |
|-----------|---------------------------------------------------------------|
| activity  | current activity inferred by Activity Recognition API         |
| company   | participant's current company e.g. alone, friends, colleagues |
| necessity | necessity of using device in the current context (Likert)     |
| place     | current place type (Google's Places API)                      |
| hope      | answers to the Adult State Hope Scale questionnaire           |

Table B.12 A list of demographic features gathered from participants when signing-up to the app/study

| Feature            | Description                                  |
|--------------------|----------------------------------------------|
| gender             | gender of user                               |
| age                | age of user                                  |
| totalContacts      | number of total contacts synced              |
| appBlacklist       | list of apps whose notifications are ignored |
| distinctLocations  | unique locations/contexts of the participant |
| distinctActivities | unique activities of the participant         |

Table B.13 Observed frequencies and percentages of notifications responded to at different place types with respect to contact significance

| Place Type               | Contact Significant |                | <i>p</i> |
|--------------------------|---------------------|----------------|----------|
|                          | False               | True           |          |
| art gallery              | 65<br>(0.7%)        | 4<br>(0.2%)    | 0.006    |
| bar                      | 92<br>(1.0%)        | 34<br>(1.6%)   | 0.014    |
| <b>cafe</b>              | 40<br>(0.4%)        | 26<br>(1.3%)   | <.001*   |
| lodging                  | 70<br>(0.8%)        | 20<br>(1.0%)   | 0.363    |
| natural feature          | 35<br>(0.4%)        | 6<br>(0.3%)    | 0.522    |
| neighborhood             | 483<br>(5.3%)       | 88<br>(4.2%)   | 0.050    |
| other                    | 414<br>(4.5%)       | 121<br>(5.8%)  | 0.012    |
| park                     | 34<br>(0.4%)        | 11<br>(0.5%)   | 0.304    |
| <b>point of interest</b> | 1563<br>(17.1%)     | 439<br>(21.1%) | <.001*   |
| restaurant               | 225<br>(2.5%)       | 69<br>(3.3%)   | 0.027    |
| shop                     | 790<br>(8.6%)       | 159<br>(7.7%)  | 0.145    |
| stadium                  | 82<br>(0.9%)        | 13<br>(0.6%)   | 0.224    |
| street address           | 4020<br>(44.0%)     | 856<br>(41.2%) | 0.022    |
| transit                  | 370<br>(4.0%)       | 67<br>(3.2%)   | 0.081    |
| university               | 856<br>(9.4%)       | 163<br>(7.9%)  | 0.030    |

Table B.14 Crosstabulation of app significance and notification topic (significance accepted at  $p < .005$ ).

| Topic                          | isAppSignificant |                | <i>p</i> |
|--------------------------------|------------------|----------------|----------|
|                                | False            | True           |          |
| arts & entertainment           | 576<br>(19.8%)   | 644<br>(28.5%) | 0.566    |
| <b>business &amp; industry</b> | 169<br>(5.8%)    | 351<br>(15.5%) | <.001*   |
| computers & electronics        | 309<br>(10.6%)   | 26<br>(1.2%)   | <.001*   |
| food & drink                   | 80<br>(2.7%)     | 50<br>(2.2%)   | 0.001*   |
| internet & telecoms            | 299<br>(10.3%)   | 284<br>(12.6%) | 0.016    |
| jobs & education               | 272<br>(9.3%)    | 52<br>(2.3%)   | <.001*   |
| news                           | 114<br>(3.9%)    | 30<br>(1.3%)   | <.001*   |
| <b>online communities</b>      | 124<br>(4.3%)    | 403<br>(17.8%) | <.001*   |
| other                          | 522<br>(17.9%)   | 206<br>(9.1%)  | <.001*   |
| <b>people &amp; society</b>    | 33<br>(1.1%)     | 101<br>(4.5%)  | <.001*   |
| sports                         | 415<br>(14.2%)   | 112<br>(5%)    | <.001*   |



Table B.15 Pre-study questionnaire presented to participants of the Pushd study.

| Question                                                                                                                                                                           | Purpose                                                                                                           |
|------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-------------------------------------------------------------------------------------------------------------------|
| Select applicable gender                                                                                                                                                           | <i>Used to distinguish web push engagement between distinct genders</i>                                           |
| Select applicable age group                                                                                                                                                        | <i>Used to distinguish web push engagement between distinct age groups</i>                                        |
| Have you altered any notification settings on your device?                                                                                                                         | <i>Used to identify participants implementing custom notification settings</i>                                    |
| Which phrase, a or b, contains an emoji:<br>a = I am 😊 to be taking part<br>b = I am happy to be taking part                                                                       | <i>Used to identify whether participants understand and recognize emojis</i>                                      |
| How likely are you to subscribe to push-notifications on the web?                                                                                                                  | <i>Used to understand receptivity to web push subscriptions</i>                                                   |
| To what degree are you comfortable with your data being used for automated decision making?                                                                                        | <i>Used to ascertain participant comfort with sharing data for automated decisions</i>                            |
| Which phrase, a or b, do you think is deliberately enticing:<br>a = Can You Eat Orange Peels, and Should You?<br>b = Intel launches Centrino 2 aiming for performance and mobility | <i>Used to distinguish between participant ability to recognize enticing / persuasive text, such as Clickbait</i> |
| Select multiple topics which you would like to receive notifications on                                                                                                            | <i>Used to understand participant preferences toward web push subscription</i>                                    |
| Select multiple websites which you would like to receive notifications from                                                                                                        | <i>Used to understand participant preferences toward web push subscription</i>                                    |
| Select multiple websites which you would not like to receive notifications from                                                                                                    | <i>Used to understand participant preferences toward web push subscription</i>                                    |
| Mindful Scale                                                                                                                                                                      | <i>Used to distinguish web push engagement between distinct mindful types</i>                                     |
| Curiosity Scale                                                                                                                                                                    | <i>Used to distinguish web push engagement between distinct curiosity types</i>                                   |
| Initial notifications                                                                                                                                                              | <i>Used to identify initial participant receptivity to web push notifications</i>                                 |

Table B.16 The Mindful Attention Awareness Scale (MAAS) presented as part of the Pushd pre-study questionnaire.

| Below is a collection of statements about your everyday experience.<br>Using the 1-6 scale below, please indicate how frequently or<br>infrequently you currently have each experience.<br>Please answer according to what really reflects your experience rather<br>than what you think your experience should be.<br>Please treat each item separately from every other item. |                                                                                                                  |                      |                        |                    |                 |
|---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|------------------------------------------------------------------------------------------------------------------|----------------------|------------------------|--------------------|-----------------|
| 1                                                                                                                                                                                                                                                                                                                                                                               | 2                                                                                                                | 3                    | 4                      | 5                  | 6               |
| Almost<br>Always                                                                                                                                                                                                                                                                                                                                                                | Very<br>Frequent                                                                                                 | Somewhat<br>Frequent | Somewhat<br>Infrequent | Very<br>Infrequent | Almost<br>Never |
| 1                                                                                                                                                                                                                                                                                                                                                                               | I could be experiencing some emotion and not be conscious of it until<br>some time later                         |                      |                        |                    |                 |
| 2                                                                                                                                                                                                                                                                                                                                                                               | I break or spill things because of carelessness, not paying attention,<br>or thinking of something else.         |                      |                        |                    |                 |
| 3                                                                                                                                                                                                                                                                                                                                                                               | I find it difficult to stay focused on what is happening in the present.                                         |                      |                        |                    |                 |
| 4                                                                                                                                                                                                                                                                                                                                                                               | I tend to walk quickly to get where I am going without paying<br>attention to what I experience along the way.   |                      |                        |                    |                 |
| 5                                                                                                                                                                                                                                                                                                                                                                               | I tend not to notice feelings of physical tension or discomfort<br>until they really grab my attention.          |                      |                        |                    |                 |
| 6                                                                                                                                                                                                                                                                                                                                                                               | I forget a persons name almost as soon as I have been told it<br>for the first time.                             |                      |                        |                    |                 |
| 7                                                                                                                                                                                                                                                                                                                                                                               | It seems I am running on automatic, without much awareness<br>of what I am doing.                                |                      |                        |                    |                 |
| 8                                                                                                                                                                                                                                                                                                                                                                               | I rush through activities without being really attentive to them.                                                |                      |                        |                    |                 |
| 9                                                                                                                                                                                                                                                                                                                                                                               | I get so focused on the goal I want to achieve that I lose touch with<br>what I am doing right now to get there. |                      |                        |                    |                 |
| 10                                                                                                                                                                                                                                                                                                                                                                              | I do jobs or tasks automatically, without being aware of<br>what I am doing.                                     |                      |                        |                    |                 |
| 11                                                                                                                                                                                                                                                                                                                                                                              | I find myself listening to someone with one ear,<br>doing something else at the same time.                       |                      |                        |                    |                 |
| 12                                                                                                                                                                                                                                                                                                                                                                              | I drive places on automatic pilot and then wonder why I went there.                                              |                      |                        |                    |                 |
| 13                                                                                                                                                                                                                                                                                                                                                                              | I find myself preoccupied with the future or the past.                                                           |                      |                        |                    |                 |
| 14                                                                                                                                                                                                                                                                                                                                                                              | I find myself doing things without paying attention.                                                             |                      |                        |                    |                 |
| 15                                                                                                                                                                                                                                                                                                                                                                              | I snack without being aware that I am eating.                                                                    |                      |                        |                    |                 |

Table B.17 The Curiosity and Exploration Inventory (CEI-II) presented as part of the Pushd pre-study questionnaire.

| Rate the statements below for how accurately they reflect<br>the way you generally feel and behave. Do not rate what you think<br>you should do, or wish you do, or things<br>you no longer do. Please be as honest as possible. |                                                                                              |            |             |           |
|----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|----------------------------------------------------------------------------------------------|------------|-------------|-----------|
| 1                                                                                                                                                                                                                                | 2                                                                                            | 3          | 4           | 5         |
| Very Slightly or Not at All                                                                                                                                                                                                      | A little                                                                                     | Moderately | Quite a Bit | Extremely |
| 1                                                                                                                                                                                                                                | I actively seek as much information as I can in new situations.                              |            |             |           |
| 2                                                                                                                                                                                                                                | I am the type of person who really enjoys the uncertainty of everyday life.                  |            |             |           |
| 3                                                                                                                                                                                                                                | I am at my best when doing something that is complex or challenging.                         |            |             |           |
| 4                                                                                                                                                                                                                                | Everywhere I go, I am out looking for new things or experiences.                             |            |             |           |
| 5                                                                                                                                                                                                                                | I view challenging situations as an opportunity to grow and learn.                           |            |             |           |
| 6                                                                                                                                                                                                                                | I like to do things that are a little frightening.                                           |            |             |           |
| 7                                                                                                                                                                                                                                | I am always looking for experiences that challenge how I think<br>about myself and the world |            |             |           |
| 8                                                                                                                                                                                                                                | I prefer jobs that are excitingly unpredictable.                                             |            |             |           |
| 9                                                                                                                                                                                                                                | I frequently seek out opportunities to challenge<br>myself and grow as a person              |            |             |           |
| 10                                                                                                                                                                                                                               | I am the kind of person who embraces unfamiliar<br>people, events and places.                |            |             |           |

Table B.18 Mid-study questionnaire presented to participants of the Pushd study.

| Feat.       | Question                                                                                                                                             | Purpose                                                                                                         |
|-------------|------------------------------------------------------------------------------------------------------------------------------------------------------|-----------------------------------------------------------------------------------------------------------------|
| Enticement  | Study the push notifications you received. Consider whether the notification text is deliberately enticing or not. Mark each as enticing or neutral. | To determine which notifications participants find enticing for comparison across differing participant groups. |
| Delivery    | How satisfied were you with the delivery time of notifications received                                                                              | To determine participant sentiment toward delivery and evaluate inferred times.                                 |
| Relevance   | How relevant (with regard to your interests and context) were the notifications which you received?                                                  | To determine whether notifications chosen were delivered at relevant moments.                                   |
| Diversity   | How similar or diverse were the notifications you received?                                                                                          | To determine whether notifications chosen, or synthetically generated, were diverse                             |
| Informative | How informative was the content of the notifications which you received?                                                                             | Used to identify notification content and presentation attributes which provide most clarity                    |
| Credibility | On a scale of 1-4, how believable were the notifications you received?<br>(4 = very believable)                                                      | Used to determine the credibility of the notification text content                                              |
|             | On a scale of 1-4, how accurate were the notifications you received?<br>(4 = very accurate)                                                          | Used to determine the credibility of the notification text content                                              |
|             | On a scale of 1-4, how clear were the notifications you received?<br>(4 = very clear)                                                                | Used to determine the credibility of the notification text content                                              |

Table B.19 Mid-study questionnaire presented to participants of the Pushd study.

|                 |                                                                                       |                                                                                                         |
|-----------------|---------------------------------------------------------------------------------------|---------------------------------------------------------------------------------------------------------|
| Engag-<br>ement | On average, how many notifications do you receive per day?                            | Used to determine if discrepancies exist between perceived number of notifications received and reality |
|                 | On average, how many notifications do you open per day?                               | Used to determine if discrepancies exist between perceived number of notifications opened and reality   |
|                 | Of the notifications received, how many did you read, find interesting, but not open? | Used to identify percentage of notifications which were positively received, but no engagement          |
|                 | Which notification topics do you believe you were most receptive to?                  | Used to identify changes in receptivity throughout the study                                            |

Table B.20 Post-study questionnaire presented to participants of the Pushd study.

| Feature            | Question                                                                                                                            | Purpose                                                                                                                                                                                   |
|--------------------|-------------------------------------------------------------------------------------------------------------------------------------|-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| Synth.<br>Data     | Which set of data, A or B, most accurately describes your notification engagements?                                                 | To determine if participants could discern between their own notification engagement behaviour and synthetically generated engagement behaviour (asked for topic, websites, day and time) |
|                    | Which notification content, A or B, is more informative at a glance?                                                                | Used to evaluate unexpanded notification formats with respect to imparting information through fleeting impressions                                                                       |
| Present.<br>format | Which notification content, A or B, would enable you to make a faster decision (with regard opening or dismissing the notification) | Used to evaluate unexpanded notification formats with respect to engagement decisions                                                                                                     |
|                    | Which notification content, A or B, would better protect sensitive/private information, if present, in a notification?              | Used to evaluate expanded notification formats with respect to privacy                                                                                                                    |
|                    | Which notification content, A or B, would you prefer to be used by a machine making automated decisions on your behalf?             | Used to evaluate expanded notification formats with respect to sharing personal data with automated decisions making algorithms                                                           |
|                    | Which notification content, A or B, would you prefer to be used by a machine making automated decisions on your behalf?             | Used to evaluate expanded notification formats with respect to sharing personal data with automated decisions making algorithms                                                           |
| Emojis             | Which notification content, A or B or C, is more visually enticing?                                                                 | Used to determine sentiment between use and presentation of emojis in notification content                                                                                                |
|                    | Which notification content, A or B or C, makes best use of emojis?                                                                  | Used to determine sentiment between use and presentation of emojis in notification content                                                                                                |
|                    | Which notification content, A or B or C, has the most relevant emojis for the content?                                              | Used to determine relevancy of emojis with respect to text content                                                                                                                        |

| Feature | Question                                                                                                                                        | Purpose                                                                             |
|---------|-------------------------------------------------------------------------------------------------------------------------------------------------|-------------------------------------------------------------------------------------|
| Misc.   | For notifications which you opened during the study, how much time, on average, do you think you spent reading the web page that was presented? | Used to evaluate the user experience of the various notification service components |
|         | For notifications which you opened during the study, did you often browse other stories on Chrome after opening and reading the notification?   | Used to evaluate the user experience of the various notification service components |
|         | Would you subscribe to web notifications again in the future?                                                                                   | Used to evaluate sentiment toward web push after the study                          |





# Appendix C

## Supplementary Material for Chapter 5

### C.1 SCRUPLE Persuasive Feature Definitions

The following were the definitions and explanations presented to subscribers when using the SCRUPLE tool to study their notification engagements. Depending on the persuasive values of the notification in question, a different persuasive principle would appear as the most influential for the notification selected. All features were listed below, regardless for full enlightenment.

- Unity

*This is a notification created by a source that shares and promotes topics and values you identify with most.*

**Example:** The only topics included in notifications pushed by this source might be ones you identify strongly with. Thus you identify strongly with this brand and you are more open to being interrupted by them.

- Scarcity

*This is a notification that is not delivered to you often and thus may spark curiosity and entice engagement.*

**Example:** It could reference a topic never seen before or be delivered from a source that does not regularly push notifications to you.

- Liking

*This is a notification that contains content you are known to like and thus may entice further engagement.*

**Example:** You are known to enjoy reading sports content. Therefore notification with topics of sports are pushed at you.

- Social Proof

*This is a notification that your peers usually engages positively with and thus may entice positive engagement from you.*

**Example:** A story is trending about a household name all over the world. The vast majority of people will read the story if sent via notification, so it is sent to you.

- Consistency

*This is a notification that contains elements which you have acted positively toward in the past thus you may be enticed to behave similarly out of habit.*

**Example:** It can be identified that you always engage positively with a certain topic. Therefore you are pushed more and more of this topic.

- Reciprocity

*This is a notification created by a source which previously gave you valuable information. Even if the content of this notification is not useful, you may be enticed to engage due to the value obtained previously.*

**Example:** You have recently opened a notification from an app/website. Therefore, you may be more willing to open more notifications from this app/website as your last experience of the app/website was valuable to you. Therefore you might assume that the next experience will be too.