



Trinity College Dublin

Coláiste na Tríonóide, Baile Átha Cliath

The University of Dublin

Clinical Sentiment Classification on Intensive Care Unit Clinical Notes

SHAHAD HUSSAIN NAGOOR

A THESIS SUBMITTED TO

THE UNIVERSITY OF DUBLIN, TRINITY COLLEGE

IN FULFILLMENT OF THE REQUIREMENTS OF THE DEGREE OF

DOCTOR OF PHILOSOPHY

SCHOOL OF COMPUTER SCIENCE AND STATISTICS

SUPERVISOR: PROF. LUCY HEDERMAN

CO-SUPERVISOR: PROF. KEVIN KOIDL

2026

Declaration

I declare that this thesis has not been submitted as part of a degree at this or any other university and that it is entirely my own work.

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

Shahad Hussain Nagoor

Abstract

Doctors' notes reflect their impressions, clinical sense, and opinions on patients' progress. This information is not documented elsewhere in patients' health records and is essential for doctors' daily clinical decisions. Decisions in critical care require an overall perception of the patient's clinical progress. Sentiment analysis tools are valuable for capturing judgments or perceptions from text. Despite this value, the literature indicates a lack of sentiment analysis methods that can function effectively on intensive care unit (ICU) clinical texts. Patients' privacy and confidentiality requirements have restricted the availability of open clinical text, contributing to this issue. Moreover, it remains unclear which of the several available approaches should be adopted to develop an effective sentiment model for ICU notes. Moreover, developing automated tools or AI models for analysing clinical notes requires involving clinicians in the process for validation by comparing machine output against clinicians' sentiment labels. Because of the difficulty of labeling, there are no public datasets of ICU notes labeled with sentiment classes, which is a significant issue hampering the development and evaluation of machine learning methods. The absence of sentiment classifiers which are able to distinguish polarity or more formally 'clinical sentiment' of ICU clinical notes has led to missing the value of doctors' impressions in these notes.

The aim of this thesis is to investigate which sentiment analysis method is most effective in recognizing clinical sentiment in ICU clinical notes. Clinical sentiment is clearly defined and specified for the task, avoiding the ambiguity and inconsistency found in related work. The thesis provides a comparison framework by which the effectiveness of different sentiment methods are evaluated, exploring which approach to sentiment and which model provides the best

classification. In response to lack of labeled data, the thesis contributes an automatic labeling strategy, based on clinical events, for creating training samples, and a ground truth dataset labeled by clinicians for evaluation. In experimentation, different methods, models, domains, notes' content, and machine learning techniques are investigated.

The major contributions of this thesis are a sentiment ICU model, which demonstrates excellent capability in performing clinical sentiment classification; and the first ground truth clinical sentiment dataset to support future work on this problem. This dissertation addresses a research gap that has been highly under-researched. Its contributions support clinical linguistics, NLP applications and ultimately clinicians' decisions and healthcare practices.

Acknowledgements

I would like to express sincere gratitude to my research supervisor Dr. Lucy Hederman, for her insightful guidance and expectational supervision throughout my PhD journey. I thank her for believing in me and for her continuous support and inspiration. I am thankful for her patience, motivation, sincerity, and kindness. Her valuable insights, knowledge, and advice have constantly refined my research, helped me overcome challenges, and grow as an independent researcher.

I am also thankful to Dr. Kevin Koidl for his guidance, advice and valuable input on the research. A special thanks to Dr. Annalina Caputo for providing valuable input and insights at early phase in the PhD before her absence due to illness. Though I mourn her passing, I am grateful for her kindness and the impact she left on my research. I also thank Dr. Ignacio Martin-Loeches for providing clinical insights on the research.

My deepest gratitude, endless love, and appreciation to my dear husband Foad for his pure love, compassion, and continuous support for my dreams, and for being there in my difficult times. I am grateful, blessed, and proud to have my precious children, my son Muhammed and my daughter Danah who are my greatest blessing and who have been my strength and joy along the way.

I am honoured to be a daughter of my parents, my mother Amal Alhindi, my father Hussain Nagoor. I express my profound appreciation for their unconditional love and belief in me, their kind words, and for all of their sincere prayers in my long absence. I am blessed with sisters: Faigah, Omniyah, Afnan, Mawadah and brothers: Muhammed and Ali who were supportive by their genuine love, and all their capacities, making this PhD journey easier for me. I am also

grateful for my parents in-law Sabriah, and Assa'ad and my sisters in-law for their kindness, support, and prayers.

I am grateful for how this journey has turned me into a different person and a researcher. I finally express my obedience and gratitude to God, for his providence, grants and grace in my dreams and path.

Shahad Hussain Nagoor

University of Dublin, Trinity College, 2025

Associated Publications

0.1 Journal Papers

Nagoor, Shahad, Lucy Hederman, Kevin Koidl, Foad Bukhari, Sandra Chiaka Amasike, Antonie Bianchi, Chris Yen-Chen Lo, Hiral Patel, Ignacio Martin-Loeches. 2025. "Clinical Sentiment Analysis of ICU Notes: A Dataset and a Comparison of Lexicon and Clinical Large Language Model Methods." *Submitted to BMC Medical Informatics and Decision Making*.

0.2 Conference Papers

S. Nagoor, L. Hederman, K. Koidl and I. Martin-Loeches, "Distinguishing Clinical Sentiment in Intensive Care Unit Clinical Notes," 2024 IEEE 37th International Symposium on Computer-Based Medical Systems (CBMS), Guadalajara, Mexico, 2024, pp. 249-259, doi: 10.1109/CBMS61543.2024.00049.

0.3 Awards

Best presentation award at ICBNLP 2023: XVII. International Conference on Biomedical Natural Language Processing. Nagoor, S. , Hederman, L. , Koidl, K. , Caputo, A. (2023), 'COVID_ICU_BERT: A Fine-tuned Language Model for COVID-19 Intensive Care Unit Clinical Notes', World Academy of Science, Engineering and Technology, Open Science Index 200, International Journal of Cognitive and Language Sciences, 17(8), 445 - 452.

Contents

Declaration	i
Abstract	iii
Acknowledgements	v
Associated Publications	vii
0.1 Journal Papers	vii
0.2 Conference Papers	vii
0.3 Awards	vii
List of Figures	xiii
List of Tables	xv
Glossary	xvii
1 Introduction	1
1.1 Introduction	1
1.2 Motivation	3
1.3 Research Scope	4
1.4 Sentiment Analysis Definitions	5
1.4.1 Clinical sentiment	5
1.5 Research Question	7
1.6 Research Objectives	7
1.7 Research Overview	9
1.8 Research Contributions	11
1.8.1 Major Contributions	11
1.8.2 Minor Contributions	11
1.9 Thesis Structure	11

2	Background and Related Work	15
2.1	Introduction	15
2.2	An Overview of ICU Data and Clinical Environment	15
2.3	Critical Care Notes	16
2.3.1	Structure of ICU Notes	16
2.3.2	Type of ICU Notes	17
2.3.3	Challenges in Clinical Notes	17
2.4	Natural Language Processing in Healthcare	18
2.5	NLP Recent Paradigm Shifts	20
2.5.1	Pre-trained Language Models (PLMs) and Large Language Models (LLMs)	21
2.5.2	BERT	22
2.5.3	Generative Prompt Language Models	23
2.5.4	Clinical Pre-trained Language Models (PLMs) and Large Language Models (LLMs)	24
2.6	Current Sentiment analysis Techniques	25
2.6.1	Sentiment Analysis Approaches Based on Text Level	25
2.6.2	Sentiment Analysis Approaches Based on Methods	27
2.6.3	Lexicon-Based Approach	27
2.6.4	Machine Learning Approach	28
2.6.5	Hybrid Approach: Lexicon-based and ML	32
2.7	Overview of Sentiment Analysis in Clinical Domain and Research Gaps	33
2.8	Conclusion	35
3	Methodology	37
3.1	Introduction	37
3.2	Dataset	38
3.2.1	Dataset Descriptive Analysis	39
3.2.2	Dataset Labeling	40
3.2.3	Pseudo-labeling Strategy	41
3.2.4	Dataset Samples	43
3.3	Method	44
3.3.1	Hypothesis	46
3.3.2	Models	46
3.3.3	Evaluation	47
3.3.4	Limitations	48
3.4	Summary	48
4	Initial Investigations of Sentiment Classification Techniques	49

4.1	Introduction	49
4.2	Experiment Overview	49
4.3	SOAP: Clinical Notes Structure and Components	50
4.4	Length of Clinical Notes	51
4.5	Hypothesis	51
4.6	Experiment Design	52
4.7	Pre-processing	53
4.8	Models and Metrics	54
4.9	Results and Discussion	56
4.10	Summary	60
5	Manual Annotation and Construction of Ground Truth	63
5.1	Introduction	63
5.2	Manual Annotation Project	64
5.2.1	Motivation for manual labeling	64
5.2.2	Data Collection	65
5.2.3	Annotators Recruitment	66
5.2.4	Annotation Tool	66
5.2.5	Task Description	67
5.2.6	Annotation File	68
5.2.7	Annotation Challenges and Takeaways	69
5.3	Preparation of Ground Truth	72
5.3.1	Inter-Annotator Agreement	73
5.4	Analysis of Extracts Associated with Manual Labeling	78
5.5	Validation of Pseudo-Labeling	80
5.6	Conclusion	83
6	Final Evaluation on Ground Truth Dataset	85
6.1	Experiment Method	85
6.2	Datasets	86
6.2.1	Evaluation Data	86
6.2.2	Training Data	87
6.3	Models	88
6.4	Results and Discussion	88
6.4.1	Results and insights	89
6.4.2	Statistical Significance	92
6.4.3	Error Analysis	92

6.5	Implementation and Reproducibility	93
6.6	Technical Challenges	94
6.7	Summary	94
7	Conclusion	97
7.1	Introduction	97
7.2	Fulfillment of Research Objectives	97
7.2.1	Fulfillment of RO1	97
7.2.2	Fulfillment of RO2	98
7.2.3	Fulfillment of RO3	98
7.2.4	Fulfillment of RO4	99
7.2.5	Fulfillment of RO5	99
7.3	Research Contributions	99
7.3.1	Major Contributions	100
7.3.2	Minor Contributions	100
7.4	Final Findings	101
7.5	Limitations	102
7.6	Future Work	104
7.7	Recent Large Language Models	104
7.8	Final Remarks	105
	Bibliography	107
A	Annotation Files	129
B	Submitted Manuscript	135

List of Figures

1.1	Scope of the research	5
1.2	Research Phases.	10
1.3	Overall methodology of research combining chapters, objectives (ROs), and research contributions (RCs).	12
2.1	Types of ICU notes, based on information of intensive care notes in Shapiro & Melendez (2010).	18
2.2	Common issues that are found in notes, from Leaman et al. (2015).	19
2.3	NLP Development in 1990-2023 (from Wang et al. (2025)).	21
2.4	Sentiment analysis by text level, from (Wankhade et al. 2022)	26
2.5	Difference between machine learning and lexicons approaches to sentiment analysis (from Taboada et al. (2011)).	29
2.6	Differences between machine learning and deep learning classification approaches of sentiment polarity (from Dang et al. (2020))	31
3.1	Pseudo-labeling strategy applied to MIMIC-III.	43
3.2	Pseudo-labeling strategy: labels for a deceased patient (on top), and a patient being discharged to home. Both patients length of stay is 10 days. 'Improving' is used here as shorthand for 'is already improving or will get better'.	44
3.3	Phases of investigation composing research methodology	46
3.4	Overall method of research	47
4.1	Description of SOAP structure of clinical notes	50
4.2	Overview of the design of experiments on the pseudo-labeled dataset. Parts related to Chapter 4 are highlighted	52
5.1	Research Phase 3, parts related to Chapter 5 are highlighted.	64
5.2	Correlation heatmap between annotators of samples of notes in 3 days, and 1 day prior death. . . .	76

5.3	Correlation heatmap between annotators of samples of notes in 3 days, and 1 day after admission.	76
5.4	Correlation heatmap between annotators of samples of notes in 3 days, and 1 day prior discharge.	77
5.5	Correlation heatmap between annotators of samples of notes that don't belong to admission, death, or discharge groups.	77
5.6	The whole path of annotation and construction of ground truth datasets.	78
5.7	WordClouds for notes' extracts for positive notes (left) and negative notes (right), based on majority votes.	79
5.8	Validation with pseudo-labels.	81
6.1	Research Phase 3, parts related to Chapter 6 are highlighted.	87
7.1	Research Contributions	100
7.2	Research contributions addressing the identified research gaps.	101
7.3	Summary of the key findings	102
A.1	Annotation File shared with annotators	133

List of Tables

3.1	Statistics of MIMIC-III patients, physician and nursing notes.	39
3.2	Proportion of notes of Kennedy et al. (2023) samples and sentiment keywords frequencies based on category.	40
3.3	Sizes of datasets related to pseudo-labeling strategy. Final pseudo-labeled data is obtained after all pre-processing and segmentation presented in Chapter 4.	43
3.4	Research datasets sizes	44
3.5	Definitions of TP, FP, TN and FN in sentiment analysis	47
4.1	Sizes of training, validation and test sets for each of physician and nursing notes	53
4.2	Sizes of datasets in used in extracting training, validation and testing datasets. "pt" refers to "patients".	54
4.3	Experiment details for both frozen and non-frozen models	56
4.4	Evaluation of three lexicon models: AFINN, Textblob, and VADER using overall accuracy, and recall, precision, f1-score for positive and negative classes and applied to whole notes with default truncation (whole/truncated) vs. assessment and plans sections onwards (AssPlan).	57
4.5	Evaluation of three BERT models: BERT, clinicalBERT, and PubMedBERT using overall accuracy, and recall, precision, F1-score for positive and negative classes, and applied to whole notes with default truncation (whole/truncated) vs. assessment and plans sections onwards (AssPlan). Top half is for frozen models and lower half is for unfrozen models.	58
4.6	Evaluation for three prompt models: T5-base, ClinicalT5, and SciFive using overall accuracy, and recall, precision, F1-score for positive and negative classes, and applied to whole notes with default truncation (whole/truncated) vs. assessment and plans sections onwards (AssPlan). Top half is frozen models and lower half is unfrozen models.	59
4.7	Summary average for each type of model, category, notes segment across all metrics	60

5.1	Statistics for the manually annotated dataset. ICYSTAYS refer to ICU stays. LOS refers to length of stay.	66
5.2	Structure of Annotation File	69
5.3	Counts matrix for annotators labeling on the manual annotation data. Original labels are: Is deteriorating?, Is already improving or will get better?, Is neither improving nor deteriorating?, OR that there is no indication of the note writer’s opinion., respectively. ‘ICU-R’ refers to ICU Registrar, ‘Resp-C’ refers to Respiratory consultant, and ‘ICU-C’ refers to ICU consultant.	74
5.4	Two synthesized examples of conflicting opinions about an individual note. ‘ICU-R’ refers to ICU Registrar, ‘Resp-C’ refers to Respiratory consultant, and ‘ICU-C’ refers to ICU consultant.	75
5.5	Note extract similarity: shows overall ROUGE-F1, BLEU (implemented by sentence_bleu), cosine similarity scores comparing texts from annotators’ extracts by pairwise analysis. 5V_R indicates ROUGE-F1 for 5-votes dataset, 5V_B indicates Bleu score for 5-votes dataset. 5V_C indicates cosine similarity for 5-votes (5V) dataset. The remaining columns indicate the same metrics for 4-votes (4V) dataset.	79
5.6	Statistics for matching labels between pseudo-labels and gold ground truth (5V) in middle, and silver ground truth (4V) on right. Two classes are samples labeled as ‘Is deteriorating’, or ‘Is already improving or will get better’ by annotators. ‘%’ is percentage of correct labels compared to ground truth.	82
6.1	Sizes of of physician and nursing notes training dataset.	88
6.2	Training details for learning models	88
6.3	Evaluation on physicians’ notes of ClinicalT5, SciFive, clinicalBERT, and Kennedy’s-lexicon using overall accuracy, and recall, precision, F1-score for positive and negative classes, and applied to assessment and plans sections onwards. Best result for each dataset is shown in bold.	89
6.4	Evaluation on nursing notes of ClinicalT5, SciFive, clinicalBERT, and Kennedy’s lexicon using overall accuracy, and recall, precision, F1-score for positive and negative classes, and applied to assessment and plans sections onwards. Best result for each dataset is shown in bold.	90
6.5	A summary table averaging clinical LLMs accuracies against keyword method. 5V indicates the gold ground truth with 5 annotators’ votes, and 4V indicates the silver ground truth with 4 annotators’ votes.	91
6.6	Examples of samples misclassified by ClinicalT5	95

Glossary

AI	Artificial Intelligence
AssPlan	Assessment and Plan
BERT	Bidirectional Encoder Representations from Transformers
ClinicalBERT	Clinical Bidirectional Encoder Representations from Transformers
ClinicalT5	Clinical Text-to-Text Transfer Transformer
DL	Deep Learning
ICU	Intensive Care Unit
ICUSTAYS	ICU Stays
LOS	Length of Stay
LLMs	Large Language Models
ML	Machine Learning
MIMIC	Medical Information Mart for Intensive Care
NLP	Natural Language Processing
PLMs	Pre-trained Language Models
PubMedBERT	PubMed Bidirectional Encoder Representations from Transformers
SA	Sentiment Analysis
SciFive	Scientific Text-to-Text Transfer Transformer
SentimentICUModel	Sentiment ICU Model
SOAP	Subjective, Objective, Assessment, and Plan
T5	Text-to-Text Transfer Transformer
5V	5 Votes
4V	4 Votes

Chapter 1

Introduction

1.1 Introduction

Clinical notes comprise, along with unstructured data such as images, around 80% of documentation of electronic health records (EHR) ([Goh et al. 2021](#)). They include information that is not recorded elsewhere in patients' health record such as doctors' impressions and plans, patients' medication reactions, patients' attitude, or doctors' social interactions with patients' families. This information is essential for doctors to make their daily clinical decisions ([Murdoch & Detsky 2013](#)). Monitoring patients' attitude, their reaction to medications, and their daily progress can be especially valuable in the context of critical care (a.k.a Intensive Care Unit or ICU and used interchangeably hereafter) as patients need close and frequent monitoring, timely interventions and their conditions can change or deteriorate quickly ([Pinsky et al. 2024](#)). Therefore, notes, especially daily progress notes in critical care help to recognize the patient's condition accurately and comprehensively and help to avoid inaccurate medical plans, mistakes or false alarms that incur high costs ([Lynch et al. 2017](#), [Johnson et al. 1992](#), [Thomas et al. 1999](#), [Leape et al. 1991](#), [Churpek et al. 2016](#)). Unintentionally inaccurate decision-making or medical mistakes can occur as doctors have narrow time to review patients' chart, busy work environment, and exchange shifts and documentation in treating patients. Automated tools using natural language processing techniques can help in assisting clinicians in comprehending patient journeys in critical care from clinical notes and reducing the potential risks.

Despite their value, clinical notes are insufficiently researched ([Kong 2019](#), [Huang et al. 2020](#)). This is because notes are unstructured as opposed to structured information such as physiological vital signs, and laboratory results. Most clinical studies utilize structured health data and few target clinical notes ([Goh et al. 2021](#), [Zeng et al. 2019](#)). In addition, providing NLP experts access to clinical notes is often infeasible given the challenges posed by the need for data protection. Moreover, variations in language expressions, diversity in style of note takers,

as well as the normal ambiguity that characterises natural language, make the processing and representation of clinical notes a complex task ([Huang et al. 2019](#)).

A basic yet significant piece of information is knowing if a patient is overall progressing well or not. This can be referred to as capturing positivity or negativity of a patient's status or polarity. Identification of polarity from clinical notes can provide such an overall assessment by examining the language and the impression written in these notes. Automated tools, specifically sentiment analysis tools, can assist in this task. However, extracting polarity from clinical text using regular sentiment tools is difficult as this type of text contains less subjectivity and fewer sentimental words than other types of text, and the language used can be highly implicit. Extracting sentiment from such implicit language is known to be a difficult task ([Wankhade et al. 2022](#)). Additionally, the language of ICU clinical notes is very specialized i.e., only experts with clinical knowledge can comprehend the language.

Recent development of artificial intelligence (AI) is remarkable. Among AI fields, natural language processing is advancing at unprecedented speed. In particular, deep learning large language models are becoming more capable and have achieved significant success in many language tasks. However, language tasks advancements vary depending on languages and domain. This is because languages differ significantly based on the domain they belong to and the target they are used for. It has been challenging to develop models that can function equally effectively in all tasks and all domains. Therefore, efforts have been made by researchers to develop and expand model capabilities while considering the demands of specific domains.

Despite the advancements, slow adoption of NLP in clinical notes results from difficulty accessing adequate sample data which is a key component for artificial intelligence, machine learning, or any data driven analysis. Additionally, as mentioned earlier, handling unstructured data including text is challenging since unstructured data doesn't have consistent format or structure, making tracing or extracting information difficult ([Kong 2019](#)). Text often requires intense pre-processing as well as manual feature engineering and sometimes mapping to ontologies for semantic interpretation ([Kasthurirathne et al. 2016](#), [Nuthakki et al. 2019](#)). The above challenges have led to a research gap and a lack of this type of analysis for the clinical domain.

Based on this, this thesis aims to investigate methods of identifying clinical sentiment polarity from ICU clinical notes, addressing this gap in the field of NLP in support of ICU clinicians' decisions. This chapter next discusses the motivation behind this research (Section [1.2](#)), the research scope (Section [1.3](#)), the definition of clinical sentiment (Section [1.4](#)), the research question (Section [1.5](#)), the research objectives (Section [1.6](#)), an overview of the research (Section [1.7](#)), and the research contributions (Section [1.8](#)). The chapter then concludes with the thesis structure (Section [1.9](#)).

1.2 Motivation

Automatic identification of clinical sentiment polarity from clinical notes can support the predictability of risk factors in intensive care unit (ICU) such as length of stay, readmission, or mortality and can aid in the decision-making of clinicians in ICU (McCoy et al. 2015, Gao et al. 2021, Waudby-Smith et al. 2018, Zou et al. 2021, Tran & Lee 2018). In critical care, improving accuracy of prediction even by a small margin can make a great impact considering mortality rate linked to quick deterioration and high costs linked to false alarms (Churpek et al. 2016). Despite this value, the related literature indicates a lack of clinical sentiment analysis models that can function effectively in ICU clinical settings. This has led to loss of significant clinical impressions exclusively contained in notes and to having incomplete view of a patient’s health status (Huang et al. 2021, Denecke & Deng 2015, Seinen et al. 2025). Patients’ privacy and confidentiality requirements have restricted the availability of open clinical text, contributing to this lag. Additionally, there is no clear answer for which of the now multiple available approaches to follow in order to get a well-functioning sentiment model for ICU notes. In addition, developing automated models, or AI models, requires involving clinicians in the validation process, while clinicians are burdened by hospital duties.

For example, having clinical experts manually annotate textual data is essential for AI supervised learning and automated analysis for comparing machine vs. human performance. Specifically, a set of notes whose “sentiment” has been assigned manually by expert annotators is required. No such labeled datasets of ICU notes exist. Many studies attempt to use general domain models for identifying sentiment of clinical text and notes. However, because clinical language is so specialized, the existing general domain models have not succeeded in recognizing clinical polarity (Weissman et al. 2019, Holderness et al. 2019). Based on the indicated issues, very little work and attention has been paid to sentiment in this specific domain, and the development of clinical-specific sentiment approaches, including clinical lexicons and AI models, have not been progressed. Therefore, their capabilities and performance have not been explored and clinical sentiment analysis remains significantly under-researched.

The following summarizes the problems found in the literature (discussed thoroughly in Chapter 2) in relation to investigating clinical sentiment: (1) The lack of effective sentiment classifiers has led to missing the potential benefit of this type of analysis. (2) The related literature shows no clear answer for which approach to follow in order to get a well-functioning sentiment model for ICU notes. (3) Related work shows no existence of public and labeled datasets of clinical sentiment for ICU notes, which is a significant issue for the development of machine learning methods and evaluation for other methods e.g., keyword-based methods. (4) Unavailability of clinicians who can do the labeling or extract texts for sentiment analysis hamper the development of specialized lexicons and creating ground truth data for training and evaluation of automated models. The above motivates this thesis to address those challenges in support of natural language processing research.

The significance of the research has many aspects. The following exemplifies some of the potential uses of clinical

sentiment in ICU settings. Firstly, it supports the predictability of ICU risk factors that have a great impact on hospital management and costs. For example, scores or labels of sentiment models can be used independently in creating prediction models of mortality or readmission or can be fused, using data fusion approaches, with other features such as patient vitals and laboratory results for this purpose. Secondly, this research has the potential to expand doctors' comprehension of events in patient journeys. Polarity can help identify the onset of progress or regress in patients' symptomatological changes. For instance, for a patient ICU stay of 60 days, it would be useful for doctors to know when the patient's condition begins to deteriorate. Clinical sentiment can be used to recognise such signals by investigating the polarity in notes and providing guidance to doctors into where changes occurred. Thirdly, clinical sentiment can answer inquiries in relation to the appropriateness of medical interventions like operations and prescribed drug side effects by analysing notes related to those events (Deng & Wiebe 2015). Fourthly, examining patients outcomes manually can be a time-consuming task especially for complicated cases with long stay in ICU, and additionally because multiple clinicians shifts between cases resulting in a collaborative documentation of notes, with pieces of information recorded in multi-documents and with different writing styles. This nature of documentation might include contradictions or inconsistent opinions or outcomes. Therefore, automatic identification of clinicians' documented opinions, and polarity of events, interventions, treatments, and outcomes in patient notes can assist clinicians, especially if presented with visual aid, in comprehending the patient journey in ICU, and can enable studying consistency, and/or detecting contradictory opinions or plans if any exists.

Finally, obtaining the polarity of the clinical notes after a medical event or intervention can also enable clustering of patient notes and studying of sub-population tendency (if any) represented in text. For example, clustering clinical notes for patients who possibly had drug reactions or complication after the operation versus clinical notes for patients who did not.

Uses of sentiment analysis are not limited to the mentioned examples; further uses are discussed in Chapter 2. Since the investigation of sentiment approaches can span to a wide range of methods, the scope of the thesis coverage is specified next.

1.3 Research Scope

The scope of the research is at the intersection of two areas, sentiment analysis in clinical domain including classic approaches, and ML approaches using recent deep learning LLMs as visualised in Figure 1.1. Hybrid approaches, which combine methods discussed in Chapter 2, are out of the scope of the research.

Within the scope of the research and related studies, there have been ambiguity and inconsistency in defining the two terms: sentiment analysis and clinical sentiment. This thesis describes a clear definition of clinical sentiment with brief background on the issue as presented next in Section 1.4.

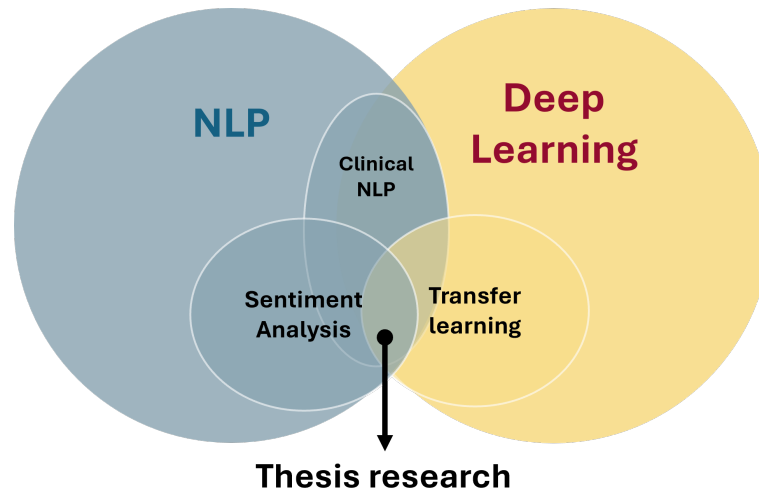


Figure 1.1: Scope of the research

1.4 Sentiment Analysis Definitions

Despite the use of sentiment analysis technique in many studies, there has been no consistent or standard definition for it. In fact, there has been ambiguity and inconsistency in defining the term sentiment analysis in the literature. [Cambria et al. \(2017\)](#) mentioned that the term "sentiment analysis" is a big suitcase covering (up to 15) NLP sub-tasks. Sentiment analysis has four categorizations as defined by [Vij & Pruthi \(2018\)](#) which are: polarity analysis, subjectivity analysis, intensity analysis, and emotion analysis. Some research refers to polarity analysis as sentiment analysis. [Cambria et al. \(2017\)](#) and others use opinion mining interchangeably with sentiment analysis ([Hoang et al. 2019](#), [Weissman et al. 2019](#), [Md Suhaimin et al. 2023](#)). Therefore, the interpretation of the term sentiment analysis depends on the domain, the text to which it is applied, and the selection of the task that is applicable and of interest to researchers.

As described in Section 1.1, this research investigates the application of sentiment analysis to clinical notes. Therefore, the thesis provides a specific definition and interpretation of sentiment which will be referred to as clinical sentiment as introduced in the next section.

1.4.1 Clinical sentiment

Applying sentiment analysis to clinical notes poses a computational challenge given the complexity of the text and the domain-specific nature of clinical text ([Thapa et al. 2022](#)). For example, assessing subjectivity is challenging as clinical text is more objective than many other types of text, and composed of concrete language, including clinical terms, jargon words, and abbreviations with very few sentimental, subjective words. In addition, assessing the polarity, that is, positivity or negativity, in clinical narratives is highly dependent on changes or events occurring for the patient. Therefore, there have been arguments for altering the interpretation of sentiment analysis for medical context and introducing the concept of clinical sentiment ([Deng et al. 2016](#), [Denecke & Deng 2015](#),

[Holderness et al. 2019](#)). Unlike regular sentiment analysis, which most commonly targets explicit sentimental words in text (e.g. happy, sad, joyful, etc.), clinical sentiment is driven more by clinical events, interventions, treatments, and outcomes. For example, having a drug reaction or organ failure indicates negativity, while stopping treatment or being discharged can indicate positivity. These are expressed by clinicians and are based on their personal impression of patients' assessments. This is why some research includes the term "opinion" of a clinician in the definition of clinical or medical sentiment. For example, [Liu \(2017\)](#) defines medical sentiment as the attitude or the judgment for something noticed in relation to a patient's health.

Understanding the concept of sentiment in clinical notes is not as straightforward if using the classic definition of sentiment analysis. For instance, with respect to the polarity analysis, one can think of sentiment as judging doctors' opinion about patient status, similar to the definition by [Liu \(2017\)](#). In this view, doctors' thoughts would have been imagined as something like: *I think that the patient's liver is unfortunately not functioning well*. However, that opinion in reality is typically implicit and untold, transforming the sentence to something like: *the patient has CLD (referring to chronic liver disease) or cirrhosis* or even *Pt has CLD*. This expression is referred to as a rational sentiment which contains no emotion ([Liu 2017](#)). It would be more straightforward to detect the polarity, by the regular sentiment, of the sentence if the text contains more sentimental words like *unfortunately*, and easier to detect subjectivity with more opinion words like *think*. Such high objectivity in clinical text contributes to the failure of existing sentiment analysis tools in analyzing clinical notes and this is the reason why clinical-specific sentiment tools are required. A clear definition for each sentiment analysis task in the clinical domain is also required to eliminate the existing ambiguity around the term ([Denecke & Reichenpfader 2023](#)).

In this research, and throughout this dissertation, clinical sentiment is defined as notes writers' opinion about the patient's condition and possibly prognosis, whether it is good or bad. On that basis, the sentiment task targets doctors' impression about the polarity of a patient's condition expressed in a patient clinical note. Therefore, a positive note may reflect successful intervention, improvement or progress in patient health status, good impact of a medication, and also desirable outcomes such as discharge. On the other hand, a negative note may reflect failed intervention, regression in patient health status, development of a disease, adverse medication reaction, and undesirable outcomes such as mortality or readmission.

The thesis may also refer to clinical sentiment as sentiment polarity as the task being explored is polarity analysis. Although the term clinical sentiment is the term used for consistency with other work, impression analysis, or clinical sentiment polarity are also valid. The thesis definition of clinical sentiment is applied to complete notes, as overall opinion is intuitively identified by considering overall narratives, especially due to the objective tone in notes ([Denecke & Reichenpfader 2023](#)). In subsequent experiments in the thesis, due to restrictions enforced by computational models on length of input texts, the length of notes required an adjustment. This involves utilizing the maximum permissible amount of text, with comparisons made between subsets of upper and lower sections

of notes, employing an informed segmentation based on a standardized note structure (see Section 4.3).

The clinical sentiment analysis conducted throughout the thesis experiments is a classification problem with binary labels (two-classes) reflecting **improvement** or **deterioration** of a patient's status and mapped into **positive** and **negative** labels, respectively.

The next sections present the research question, objectives, and contributions.

1.5 Research Question

This PhD research addresses the following question:

Which sentiment analysis method is most effective in recognizing clinical sentiment in ICU clinical notes?

Research sub-questions are:

RQ1: To what extent do the previous studies of sentiment approaches target clinical sentiment, and how effective are existing methods at identifying clinical sentiment? Addressed in Chapter 2.

RQ2: To what extent do the domain of an approach, category of notes, segments of notes, or training strategies affect the classification of clinical sentiment of ICU clinical notes? The conceptualization is introduced in Chapter 3, and implementation is in experiments of Chapter 4 and publication (Nagoor et al. 2024).

RQ3: Which among the methods of sentiment analysis approaches can offer the best capture of clinical sentiment based on human expert evaluation? Addressed in Chapter 6.

The research question results in five research objectives (ROs), as presented in Section 1.6, and four contributions (RCs) as presented in Section 1.8.

1.6 Research Objectives

The aim of this research is to identify the best approach and model to distinguish clinical sentiment in ICU clinical notes.

The following objectives collectively address the research question.

RO1: To review the literature comprehensively and reflect on the existing sentiment analysis methods that are applicable to the clinical domain.

This includes four sub-tasks as follows:

- (a) To review lexicon-based approaches to sentiment analysis.
- (b) To review machine learning, especially deep learning, approaches to sentiment analysis.
- (c) To review hybrid approaches (combined machine learning and lexicon-based) to sentiment analysis.
- (d) To review recent approaches to sentiment analysis using recent large language models.

This objective is addressed in Chapter 2.

RO2: To develop a pseudo-labeling strategy for automatic labeling of patients' ICU notes from patient outcomes and attributes of ICU stays. This pseudo-labeled dataset is used for training and initial experimentation while human labeled (ground truth) data is being constructed. The labeling is introduced in Chapter 3 (RO2-A), an initial experimentation of sentiment classification methods on the pseudo-labeled data are in Chapter 4, and validation of the labeling against human-labeled ground truth data (RO2-B) is presented in Chapter 5.

RO3: To investigate the effect of different domains, notes' category, notes segments, and transfer learning strategies on classification of clinical sentiment of ICU clinical notes. This objective was addressed with pseudo-labeled data while a manually labeled ground truth dataset was being constructed. The conceptualization is introduced in Chapter 3, and implementation is in experiments of Chapter 4 and publication (Nagoor et al. 2024).

RO4: To develop a ground truth dataset by specialized clinicians for human evaluation on models performance. Addressed in Chapter 5 and our recent submission (see 0.1 in Associated Publications and Appendix B).

RO5: To investigate the performance of the specialized models in clinical sentiment of ICU clinical notes on the ground truth dataset. This is addressed in Chapter 6 and our recent submission (see 0.1 in Associated Publications and Appendix B).

An explanation for each objective is presented below.

Research Objective 1:

Considering that sentiment analysis has been oriented to non clinical domains, the thesis begins by first exploring the existing techniques for their applicability and adoption in clinical domain.

Research Objective 2:

In the absence of a large clinical sentiment labeled training dataset, which is necessary for training supervised machine learning models but is resource-intensive to create, a cost effective alternative approach for training is desirable. One strategic method is to automate the labeling for training samples. In the context of ICU, this can be achieved by utilizing some of the attributes of patients' ICU stays which are intuitively related to positivity and negativity of events. Therefore, the thesis explores the use of automatic labeling (also called pseudo-labeling interchangeably) for labeling training datasets. **Research Objective 3:**

The existing methods in the state of the art are mostly built for general domains based on data from Wikipedia, books, tweets, and other sources not related to clinical domain. Few approaches target clinical language, yet they require further training for the task of clinical sentiment classification. Therefore, this thesis explores the performance of some of the most popular general and clinical-specific methods.

The nature of clinical notes depends on the note’s writer and note’s purpose. For example, nursing notes are more detailed and narrative-like than physician notes. It is also unknown which segment of notes conveys more sentiment than others, and which specific technique or strategy is able to detect that sentiment more than others. For example, LLMs can be trained fully (all-layers) or (few-layers) for the same task. Many studies in NLP investigate different portions of text, and different learning strategies while seeking optimized prediction. This work consider different aspects, while exploring the best method in predicting clinical sentiment.

Research Objective 4:

While preparing ground-truth data by experts in the field is essential in model development and evaluation, no such labeled datasets of ICU notes exist due to the cost and the resources required as mentioned in Section 1.2. This leads to having no human evaluation for models, and making it unclear to what extent they are aligned. Therefore, this objective is focused on addressing the lack of ground-truth.

Research Objective 5:

Taken the outcomes of research objectives RO2, RO3, RO4, and RO5, the analysis is finalized by determining the best method and the best performing model using the ground truth dataset. Past studies as well as the results of experiments of the previous objective RO3, has shown that specialized models are more accurate than general ones. Thus, this objectives is focused on specialized models exclusively.

Taken together, the results of the objectives answer the research question and result in the contributions introduced in Section 1.8. The next section describes an overview of the research phases.

1.7 Research Overview

The thesis research has been through a sequence of exploratory experiences. A main challenge, described as motivation for the work in Section 1.2, was the difficulty in accessing relevant data and the difficulty in forming ground truth data labeled by experts to validate the results of the research. These challenges led to the research being built in an iterative manner through three major phases, discussed in detail in Chapter 3, and listed here briefly. The phases are iterations of investigations, in which each phase makes use of the findings of the previous. The phases are summarised in Figure 1.2 and outlined as follows:

- Research Phase 1 (RP1): investigating the type of sentiment classification methods in the literature and applicable to the clinical notes. This is discussed in Chapter 2.

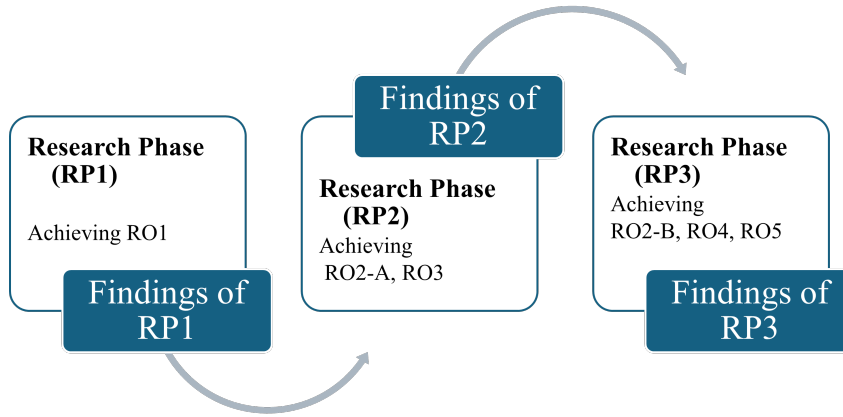


Figure 1.2: Research Phases.

- Research Phase 2 (RP2): taking the finding from the first stage and starting experimentation with automatically labeled data. The labeling strategy is proposed and used in this stage for training and evaluation due to unavailability of labeled data. The labeling strategy is introduced in Chapter 3, and the experimentation is in Chapter 4. Concurrently with the experimentation of this stage, a project to recruit clinicians and building the ground truth was initiated.
- Research Phase 3 (RP3): taking the finding from the second phase and replicate the experimentation using ground-truth data. This phase is when the ground truth became prepared. In this phase, the training continues to rely on automatic labeling; however, the evaluation was transitioned to using ground-truth data for human assessment. This stage is presented in Chapter 5.

In summary, automatic labeling were utilized consistently throughout the experimentation, and ground-truth data has been incorporated at the final stage to providing (1) evaluation for automatic labeling vs. human labeling; and (2) evaluation of models performance vs. human performance.

The challenge and path of obtaining access to appropriate data is discussed in Section 3.2 in Chapter 3. Overall, this research makes use of two datasets: (1) MIMIC dataset (The primary, and referred to hereinafter as MIMIC dataset or the MIMIC dataset). (2) Covid dataset (The secondary, and referred to hereinafter as the Covid dataset).

The MIMIC dataset was used for the second and third phases for training and evaluation including automatic and ground-truth samples. The Covid dataset is only used for secondary evaluation targeting model generalisability in the third phase. The key outcome of the research is the identification of which method and, more specifically, which model provides the best classification for clinical sentiment, evaluated on human validated ground truth data. Details on ground truth data are given in Chapter 5, and the identification of the best method and model

is given in Chapter 6. The next section lists the contributions of the research.

1.8 Research Contributions

This thesis makes two major and two minor contributions to the discipline of clinical NLP.

1.8.1 Major Contributions

RC1: SentimentICUModel, A model demonstrating the best performance in identifying clinical sentiment in ICU notes as fulfillment of RO3, and RO5.

RC2: MIMIC-III-Ext-Notes-Sentiment¹ a ground-truth ICU-specific clinical sentiment dataset labeled by clinicians as fulfillment of RO4. This dataset can be used in developing methods either by complete notes or their associated fragments and validation by comparing machine output against clinicians' sentiment labels.

1.8.2 Minor Contributions

RC3: An Automatic Clinical Sentiment Labeling Strategy based on attributes of patients' ICU stays that enables training on large dataset as fulfillment of RO2.

RC4: A Comparative Analysis for evaluating the effectiveness of several popular models in clinical sentiment classification of ICU notes as fulfillment of RO3, and RO5.

1.9 Thesis Structure

This thesis consists of six chapters.

Chapter 1 has introduced the research including motivation, research question, objectives, scope, overview of research phases, contributions, and thesis structure.

Chapter 2 is the review of the literature. First, it provides brief background information on the critical care environment and a description of the documentation in the ICU. The rest of this chapter presents recent advances in NLP, especially for healthcare, followed by a review of sentiment approaches in general and in the health domain. The chapter concludes by identifying the research gaps which this thesis addresses. This chapter fulfills RO1.

Chapter 3 presents an overview of the research methodology and discusses the evaluation method. The chapter presents first the training and evaluation datasets used throughout the research. Afterwards, it introduces the

¹Currently under-review in Physionet, titled: 'MIMIC-III-Ext-Notes-Sentiment: Labelled ICU Clinical Notes Dataset for Clinical Sentiment Polarity Identification'.

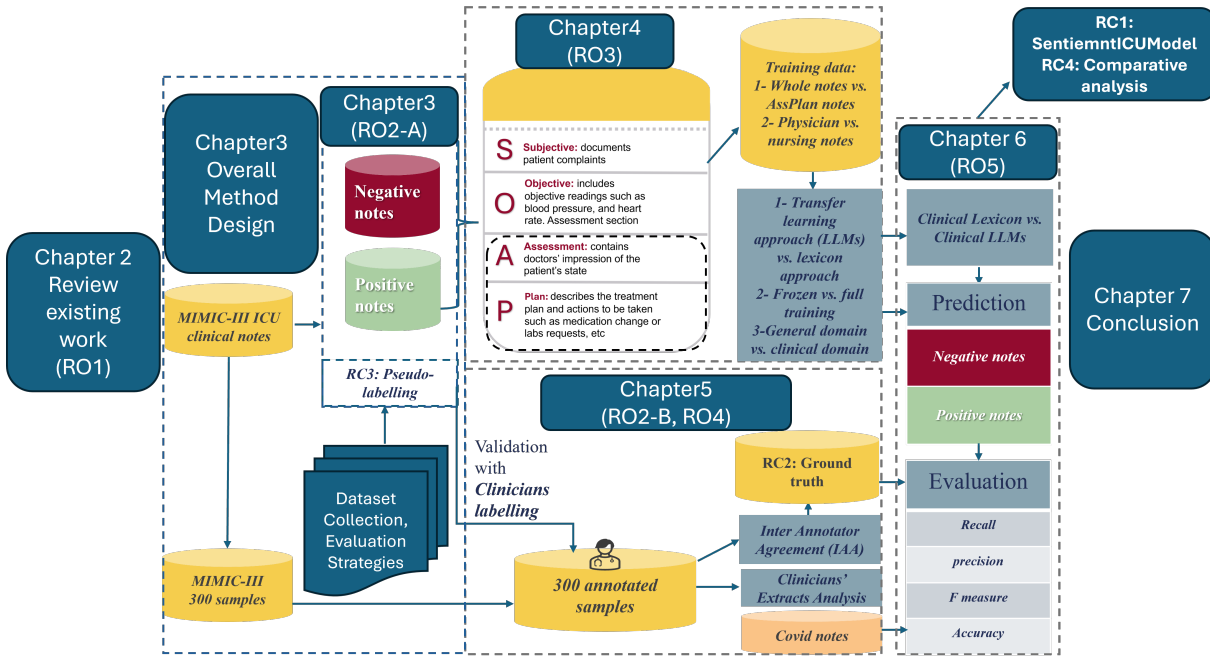


Figure 1.3: Overall methodology of research combining chapters, objectives (ROs), and research contributions (RCs).

method designed to achieve the research objectives, the main hypothesis, the models, and the evaluation strategies. This chapter presents the conceptualization of the research and fulfills RO2-A.

Chapter 4 presents an initial investigations of sentiment classification techniques conducted using the suggested labeling scheme while considering segments of notes, different categories of notes, different methods to clinical sentiment, and different domains and learning strategies. This involves comparing several general domain lexicons and LLMs and specialized ones using the mentioned settings. The chapter concludes by a discussion of the results and findings. This chapter fulfills RO3.

Chapter 5 presents the development of a manually labeled data by clinicians including the collection process, the recruitment process, the tools and files shared during the project. Then, it presents inter-annotator study, construction of ground truth, and validates the pseudo-labeling scheme using ground truth. This chapter fulfills RO2-B and RO4.

Chapter 6 presents the final evaluation of clinical sentiment approaches on the ground truth dataset obtained from Chapter 5. It compares the top performing models identified in initial experimentation in Chapter 4, with another clinical-specific method (a clinical lexicon), and identifies the most successful method and model in clinical sentiment classification. The results are supported by statistical significance and error analysis. The chapter concludes by specifying the implementation and technical details, for reproducibility of the results. This chapter fulfills RO5.

Chapter 7 presents the summary of the hypotheses, experiments conducted, the fulfillment of the research objectives, research limitations and future work.

Figure 1.3 combines the thesis structure, chapters, objectives, and contributions in an overall diagram. Later chapters use divisions and derivatives of the diagram.

Chapter 2

Background and Related Work

2.1 Introduction

This chapter reviews the literature of sentiment analysis techniques and their applicability in clinical settings. Section 2.2 first introduces an overview of ICU data and clinical environment with a brief discussion on demanding costs in ICU as well as the general workflow there. Then, Section 2.3 describes ICU notes and discusses the challenges in analysing clinical notes. Section 2.4 presents an overview of natural language processing in healthcare. Section 2.5 describes recent paradigm shifts in NLP, describing the development of NLP over time. After that, Section 2.6 reviews the existing approaches of sentiment analysis, and provides examples of studies for each approach. The conclusion of this chapter in Section 2.7 identifies the challenges and gaps observed in the related work.

2.2 An Overview of ICU Data and Clinical Environment

As described in Section 1.2, predicting risk factors in ICU such as length of stay, readmission, or mortality is important for ICU decisions and management. For example, predicting mortality in ICU can help in reducing mortality. Rates of these factors, such as readmission rate within 30 days, are also used as quality indicators of hospitals' level of care (Zubillaga et al. 2022). Therefore, research has long studied the predictability of ICU risk factors to reduce potential risks for patients with critical illness, increase the quality of care and reduce potential costs in ICU.

Costs in ICU are significant. They can become worse depending on patients' status, outcomes, and also due to spreads of infectious diseases such as the recent Covid-19 pandemic in 2020. Kramer et al. (2017) observed an increase of 12.4% in ICU associated costs for patients deceased before discharge compared to survivors. In their

study they emphasize that mortality is important in evaluating ICU related costs. With respect to disease-specific burden, Covid-19 has significantly increased the cost of hospitals and ICUs, at least at the beginning of the pandemic. In a study conducted in the U.S. by [Ohsfeldt et al. \(2021\)](#), the median costs for hospitalization in ICU for 11-16 days were estimated in the range of \$47,454 to \$41,510 for patients infected by Covid and mechanically ventilated in the period April to December 2020. The effect on cost has extended in other countries and periods of time. A study by [Popescu et al. \(2022\)](#) conducted in low-vaccinated European countries during the fourth wave of Covid-19 observed a 70.8% increase in the median daily ICU-associated cost for patients in 2021 compared to patients hospitalized in 2020, and a 63.2% increase compared to those in 2019. The above shows how crucial ICU costs are. If early prediction of risks in ICU was improved, doctors could be alerted, minimize possible risks, minimize costs, and hospitals could improve health services and plan capacities wisely ([Van Aken et al. 2021](#)).

ICU is an environment with intense workflow and rich documentation ([Sodhi & Chanchalani 2022](#), [Huang et al. 2021](#)). In ICU practice, devices record vital signs continuously ([Simpson & Lyndon 2019](#)). Patients are monitored by teams of clinicians with different backgrounds, different roles and levels of expertise. All members communicate and cooperate and effort is made to stabilise patients' status. This dynamic nature of ICU requires efficient management of workflow and data-flow ([Sodhi & Chanchalani 2022](#)). Timing is very sensitive in critical care and can contribute to successful or harmful outcomes ([Churpek et al. 2016](#)). The main concern in the critical care environment is to enhance patient progress to stability and discharge and avoid deterioration to death. Nurses visit patients frequently, assess them, provide their treatments, review their vitals and document them with their own reflections and suggested plan ([Lewandowska et al. 2020](#)). Doctors, on the other hand, visit patients less frequently, perhaps once or twice a day, depending on their rounds and consultation plans. Their role is to do assessment, prescribe treatments, and suggest efficient plans for the patients ([Sodhi & Chanchalani 2022](#)). The nature of work in ICU introduces information load and workload on clinicians making the environment difficult and therefore making prediction systems and decision support systems valuable ([Sodhi & Chanchalani 2022](#), [McCoy et al. 2015](#), [Gao et al. 2021](#), [Waudby-Smith et al. 2018](#), [Zou et al. 2021](#), [Tran & Lee 2018](#)). Decision support, especially AI based systems is going to be integrated in ICU in the near future ([Pinsky et al. 2024](#)). With the ongoing advancements in AI, the focus on analyzing and aiding healthcare professionals in their documentation tasks is growing. To optimise such aid, understanding the type, the structure, and the role of notes is important. The next section describes the structure and the types of notes in ICU.

2.3 Critical Care Notes

2.3.1 Structure of ICU Notes

ICU notes are typically written in SOAP ([Podder et al. 2023](#)) structure. SOAP is a clinical documentation methodology in which S refers to Subjective, O refers to Objective, A refers to Assessment, and P refers to

Plan. Subjective section describes a patient's main complaint and symptoms. Objective section includes readings such as breathing rate. Assessment section contains doctors' judgments of the patient's state, and Plan section describes the plan, such as operations, or medication, etc. Further details on SOAP is provided in Section 4.3. This structure is recommended for daily progress notes in ICU since it includes the required and up-to-date information for daily decision making and communication for medical teams (Brown et al. 2014).

2.3.2 Type of ICU Notes

Medical records or electronic health records (EHR) contain both structured and unstructured data. It has been observed that the proportion of unstructured data in EHRs is larger than the proportion of the structured fields. In particular, unstructured data holds 80% of the EHR (Goh et al. 2021). Unstructured data is valuable but more complex (Hultman et al. 2019, Leaman et al. 2015). Clinical notes are examples of unstructured documentation as they are composed of free text and typically have no limitation in length. They are used to describe personalized features about a patient's condition and progress. In ICU, many types of notes are used to keep track of patient's situation and facilitate exchange of patient data between clinicians and medical teams. Figure 2.1 shows multiple notes that are documented during patient stays in ICU. Details in the figure are based on information about intensive care notes in Shapiro & Melendez (2010). All of these notes capture important information about patients at different point of time in their courses of hospitalization in ICU. Nevertheless, each plays a different role in the documentation of patient's medical record and in the management of workflow and information exchange between medical team members (Huang et al. 2021). Doctors' notes contain their daily assessment, impression and plan for the patient. Their notes are written during rounds and are also key components of their decision making in daily consultations (Brown et al. 2014). Nursing notes are more detailed and frequent compared to doctors' notes as nurses visit and examine patients more frequently and monitor them closely (Huang et al. 2021). Discharge summaries describe an overall yet short summary from patient admission to discharge point. Daily progress notes are meant to describe patient progress on a daily basis and mostly reflect events on one day. Other notes may include radiology reports, pre- and post- operative notes, and rehabilitation. Despite the value of notes, they are not easy to digest and have some challenges, requiring attention or improvements as discussed in the next section.

2.3.3 Challenges in Clinical Notes

According to Hultman et al. (2019) notes are "data-rich and information-poor". In their study, they focus on how clinicians read different variants of notes structure: SOAP (Podder et al. 2023), APSO (Lin et al. 2013), SAPO. In the findings, clinicians prioritise examining the assessment first, and prioritise the chief complaint with the greatest importance, followed by assessment and plan. With respect to writing style, Leaman et al. (2015) reports a number of problems with processing clinicians writing, as presented in Figure 2.2.

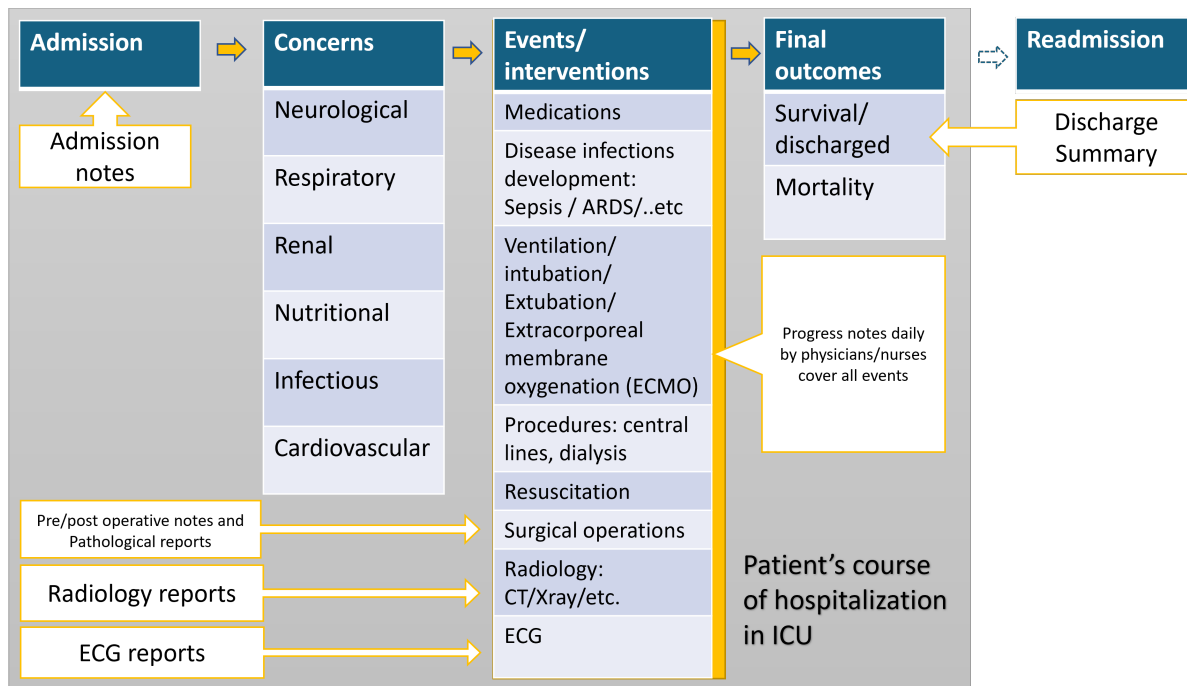


Figure 2.1: Types of ICU notes, based on information of intensive care notes in [Shapiro & Melendez \(2010\)](#).

[Perera & Sheth \(2015\)](#) study implicit entity extraction in clinical notes, extracting untold entity names from the text. Implicit language is commonly used by providers in clinical notes and leads to difficulty in interpretation of text or information extraction.

Being specialized and containing ambiguous and inconsistent linguistic components, has made it complicated even for analysis by automated models such as NLP models ([Leaman et al. 2015](#)). The field of natural language processing includes many studies related to healthcare and clinical notes analysis. Section 2.4 explore studies for NLP applications and uses on healthcare. To understand the integration of NLP in clinical domain more, it is essential to have a background on the history of the development of NLP as each development phase has its advantages and weakness. Section 2.5 sheds the light on the development of NLP and its recent paradigm shifts.

2.4 Natural Language Processing in Healthcare

This section aims to offer an overview of studies employing NLP techniques within the healthcare sector. Applications of NLP in medicine vary and range from educational purposes and research support to clinical diagnosis, decision support and patient care ([Locke et al. 2021](#)). [Chary et al. \(2019\)](#) utilizes NLP techniques for medical education purposes, including tracking resident and trainee performance and medical decision-making through their documentation in hospitals and exams.

Illustrative examples of common challenges in processing text from clinical narratives.

Category	Detail	Example
Flexible formatting	Variable formatting semantics	• Section header: "Admitting Diagnosis: SPLENOMEGALIA" • Inseparable phrase: "Neuro: nonfocal"
	Structure without sentences	• "Height: (in) 75 Weight (lb): 245 BSA (m2): 2.39 m2 BP (mm Hg): 92/52 HR (bpm): 120" • "Digoxin 250 mcg Tablet Sig: One (1) Tablet PO DAILY (Daily)."
	Missing punctuation	• Commas: "no m/r/g b/l coarse rhonchi w/assoc upper airway sounds." • Periods: "Well appearing male in no acute distress Chest is clear No hernias"
	Parenthetical expressions	• "Severe AS (AoVA < 0.8cm2)." • "Mild (1+) aortic regurgitation is seen."
Atypical grammar	Missing expected words	• Verb: "No pneumothorax." • Object: "Lopressor beta blockade was given." • Articles: "Ultrasound showed no evidence of [a] pseudoaneurysm." • Adjective without noun modified: "Head, eyes, ears, nose, and throat examination revealed normocephalic and atraumatic."
	Unusual part-of-speech combinations	
Rich descriptions	Variety of textual subjects	• Patient: "Awake, alert and oriented times three." • Anatomy: "The left atrium is mildly dilated." • Test or procedure: "Suboptimal image quality - poor echo windows." • Family: "mother had CABG grandmother had CAD" • Diagnosis: "He was also tachycardic" • Evidence: "He had a fingerstick of 142." • Jargon: "septic picture likely aspiration pneumonia secondary to dobhoff placement." • Ad-hoc abbreviations: "His iron studies, LDH/Bili/Hapto, B12/Folate were normal." • Acronyms: "recent tension PTX at OSH" • "Mitral stenosis is not present and definate [sic] mitral regurgitation is not seen." • "s/p gsw now with fevers r/o abcess [sic]" • "Sclarea [sic] anicteric." • "ventilator dependent respiratoy [sic] failure"
	Variety of communication styles	
	Language specific to medical context	
	Misspellings	

Figure 2.2: Common issues that are found in notes, from [Leaman et al. \(2015\)](#).

[Pool et al. \(2024\)](#) reviews the use of large language models in telehealth using multiple dimensions related to responsible use of AI. The study identified three directions for LLMs in telehealth focusing on (1) adaptability to patients' need and benefits to healthcare providers, (2) maintaining ethics and patient rights, and (3) ensuring safety and reliability of information and systems. Adoption of LLMs and AI in telehealth is increasing and should be accompanied by responsible use.

Another application of NLP in health care is detection of diseases and conditions. Early detection of sudden changes is important and can help in early interventions and management of patients' status and resources. For example, early detection of deterioration can help in sepsis cases where early intervention is essential to prevent death. Therefore, many studies target detection of sepsis and more recent studies involve NLP techniques and features from notes in the sepsis detection algorithm ([Goh et al. 2021](#), [Liu et al. 2019](#), [Jensen et al. 2017](#)). For example, [Goh et al. \(2021\)](#) were able to detect sepsis up to 48 hours prior to its onset when they added clinicians' notes to patients' structured data in the prediction algorithm. The prediction utilizes latent Dirichlet allocation (LDA), a text mining technique by which a major topic is extracted from text.

Recently, with the emergence of LLMs, many studies use language models to detect diagnosis such as dementia ([Liu et al. 2023](#)) and orthodontic ([Ohtsuka et al. 2023](#)).

Recent studies target the use of LLMs in support of patients and clinicians. In support of patients, [Chowdhury et al. \(2023\)](#) investigate the use of ChatGPT in a teleclinic to answer routine patient questions in cases of routine surgeries. The model demonstrates satisfactory capability in addressing the questions with no further training.

Many LLMs have shown excellence in clinical related tasks like answering questions and generation of documentation. Some even outperform human in certain task. For example, [Workum et al. \(2025\)](#) found that GPT-4o provides top accuracy in medical practice exam. However, the study emphasizes the need for further investigations seeking responsible use in clinical environments. Similarly, [Kung et al. \(2023\)](#) found that ChatGPT achieved around the passing score, with no further training, on the MedQA dataset ([Jin et al. 2021](#)) which has questions in the style of the United States Medical Licensing Examination (USMLE). Med-PaLM2 ([Singhal et al. 2023](#)) was the first model which passed the exam. All of these studies indicate the potential of LLMs in assistance in medical education.

In support of clinicians, [Zhou et al. \(2023\)](#) surveyed a number of studies for using LLMs in assisting clinicians to write medical reports. [Singh et al. \(2023\)](#), [Mishra & Desetty \(2023\)](#) and [Ben Abacha et al. \(2023\)](#) study LLM generation of clinical notes from dialogues between patients and doctors.

[De Brito et al. \(2024\)](#) conducted a systematic review for NLP in cancer related research. The study raises the question about the major challenges of NLP in health care. Their answers include:

- The need for large annotated data.
- Privacy and security of data.
- Specialty, diversity, and complexity of clinical text.
- Understanding ambiguous language, and handling inconsistent structure.
- Adjusting LLMs models to specific tasks and demands.
- Generalizability of models for different text types and institutional changes.

Language processing issues has been studied in NLP for many years, and involving different techniques. The following section presents an overview of the development of language models and techniques over years.

2.5 NLP Recent Paradigm Shifts

Figure 2.3 by [Wang et al. \(2025\)](#) shows visually some of the key advancements in NLP the timeline 1990-2023. The language models have evolved as follows: statistical language models (SLMs), neural language models (NLMs), pre-trained language models (PLMs), and large language models (LLMs). SLMs are traditional and based on statistical methods such as n-grams. They predict next words based on previous ones. Therefore, they are limited in handling context. NLMs are based on neural networks which are better at learning language patterns and complexities, but should be trained on large data. PLMs are pre-trained on large datasets, and are adaptable to specific tasks, without the need to repeat the training. They also have dynamic embeddings while they process words allowing for context-aware analysis. LLMs are large-scale PLMs with much increased scale up to billions of parameters. They have shown excellence in diverse language tasks. Recent NLP models consider words'

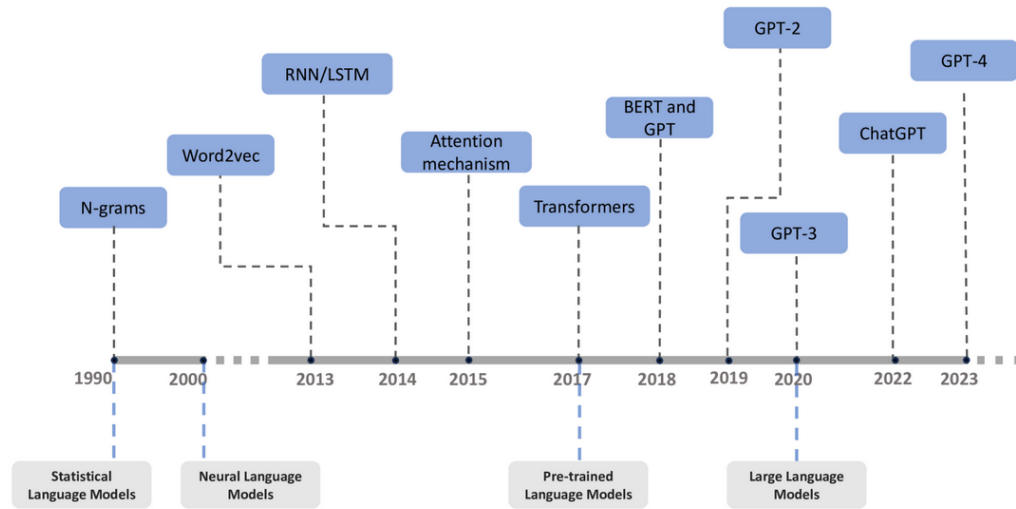


Figure 2.3: NLP Development in 1990-2023 (from Wang et al. (2025)).

contexts and so are context-aware language models. Context-aware language models were first proposed by Dai & Le (2015). Unlike previous models with static word embeddings (word2vec and GloVe), these models have the ability to distinguish the meaning of words based on the context surrounding each word and therefore introduce different representation (embedding) for the same word in different contexts. The emergence of the concept of contextualized word vectors in 2017 has led to unprecedented improvement in NLP tasks. Among others, sentiment analysis has significantly been improved by contextualized processing of text as sentiment components such as polarity or subjectivity highly depend on the whole context of text. However, the language models back then were less advantageous as they were pre-trained on small datasets and their performance varies in many NLP tasks (Wang et al. 2023). Recently, many NLP models are trained on very large data and can function effectively in many NLP tasks. Using pre-trained models is called transfer learning and has become the standard in NLP in the recent years as it allows for easier, faster, and more efficient training compared to training from scratch with relatively small labeled data and few steps for tasks' adaptation. A further elaboration on this approach with a focus on BERT models is described in Section 2.5.2. However, there is also an evolving paradigm where the approach of NLP tasks will be to use a unified model that fits all NLP tasks in a text-to-text fashion and few or no labeled samples. This approach is explained in Section 2.5.3. Finally this section ends with an overview of PLMs and LLMs being used for clinical tasks in Section 2.5.4.

2.5.1 Pre-trained Language Models (PLMs) and Large Language Models (LLMs)

As mentioned in Section 2.5, transfer learning using pre-trained models including PLMs and LLMs has recently become the standard in NLP, achieving remarkable performance that has not been feasible with smaller models

(Szep et al. 2024). They become capable in comprehending, and generating human-like texts (Khurana et al. 2023). The prevalence of transformer models (since 2017) is a key element to NLP advancements. Among the popular models are BERT models (developed in 2018) which were the state of the art for a while. Afterwards, more advanced models with encoder-decoder structure, and prompt-based generation of text has outperformed BERT in many tasks, becoming the state of the art in the field of NLP in many aspects.

The next two sections discuss BERT model and prompt LLMs, due to their relevance to the research.

2.5.2 BERT

The transformer model BERT was state-of-the-art when it was released in 2018 and has achieved excellent results in many language tasks using transfer learning. The concept of pre-trained word embeddings (Mikolov et al. 2013, Pennington et al. 2014) has been an important component of recent NLP models. BERT has been trained on massive data before being released to use. Therefore, NLP researchers can access and make use of BERT, and its pre-trained word embeddings for NLP research without the significant computational and time costs required in the pre-training stage. BERT-large was trained on 3.3 billion words collected from Wikipedia and Google's BooksCorpus using 64 TPU chips for a period of four days costing around \$7,000. Based on that, many studies have used pre-trained BERT transformer models through transfer learning and fine-tuning. Transfer learning is basically transferring knowledge obtained from learning to solve a task into solving another new task (Esterle 2022), and fine-tuning adjusts the transferable model's internal parameters which often results in a much more accurate model with relatively little data. Fine-tuning also changes the representation of text in the underlying embedding space of the pre-trained model in a way that serves a specific downstream task (Zhou & Srikumar 2022). Tasks can be any of NLP tasks such as text classification or question answering.

The strength of the BERT model is its ability to process text bidirectionally at the same time, enhancing its comprehension of language. Therefore, BERT has contextual embeddings for every word in the text. This bidirectional analysis and context-aware embedding was not provided in earlier models like word2vec and GloVe. Earlier models rely on one direction to generate the next word in a sentence. The studies described below exemplify the use of BERT for different purposes.

Su & Vijay-Shanker (2022) fine-tune BERT for extracting biomedical relations. Their analysis suggests two methods for improved performance: further pre-training with sub-domain data, and fine-tuning the model with complete knowledge in BERT's last layer. Peng et al. (2019) explore the performance of ELMO, Embeddings from Language Models by Peters et al. (2018), and BERT in biomedical domain, and based on their analysis, BERT models showed better results when pre-trained on PubMed abstracts and clinical notes than ELMO and generally than most of the state of the art models.

Despite the advantages gained by transfer learning, one challenge in this technique is that there is no standard fine-tuning strategy; the process of tuning pre-trained models can be performed in multiple ways (Sun, Qiu, Xu

& Huang 2019). Sun, Qiu, Xu & Huang (2019) explores different fine-tuning strategies on BERT and carried out multiple trials on text classification task. As a result they propose an effective multi-step solution for fine-tuning BERT: first, continue to pre-train on data of the same task or data from the same domain. Then, an optional step for fine-tuning BERT on multiple related tasks. Finally, perform the fine-tuning for the target task. The study also shows the effect of different choices for preprocessing long text, selecting layers and their learning rates, and discusses a few potential problems in transfer learning such as catastrophic forgetting. The authors indicate that despite the effective performance of BERT in many NLP tasks, its full capabilities have been partially unexplored. There has been little research studying the performance of BERT on target tasks or investigating factors for improving performance on them, e.g. Zhou & Srikumar (2022). Therefore, fine-tuning methods are still subject to experiments, investigations, and development.

2.5.3 Generative Prompt Language Models

Prompt-based learning is the most recent paradigm shift in NLP state of the art (Taylor et al. 2024). In prompt-based learning methods, a text is transformed into a specific format that includes slots to be filled. This transformation is performed by a prompting function or templates, and a pre-trained model is asked to fill in the blanks based on probabilities generated with suggested outputs. The output can be a token, a phrase, or a sentence (Liu et al. 2021). This style of NLP has enabled having a unified language model where the same model can accept different task requests with only a prefix to input and no required task-specific structural modification, or hyperparameter changes. This is why it is called prompt-tuning as the prompt is the piece being modified and the complete task is done in text-to-text fashion. In *pre-training, fine-tuning paradigm*, pre-training is unsupervised and performed by masked language modeling and next sentence prediction while fine-tuning is performed with additional classification heads or layers on top of the pre-trained model and training the combined model with supervised learning and labeled dataset. In contrast, in prompt-based learning, there is no difference between pre-training and fine-tuning and thus they become consistent and convenient. Moreover, prompt-based learning can also function with few-shot learning (i.e., few examples) or zero-shot learning (ZSL) which classifies samples from new classes unseen previously in training. The general frame for the paradigm is *pre-train, prompt, predict* (Liu et al. 2021), and task-based templates are used for NLP tasks. An example is the attempt to predict a disorder from a symptom, as described by Sivarajkumar & Wang (2023). The template/prompting function is chosen as: *[text]. This is a symptom of [Z] disorder*. In this example, [text] is the text to be classified: *The patient has a cough and expanded chest that does not deflate when he exhales*. The pre-trained model will look at the text in the template to generate [Z], which can be a word, phrase or sentence. The model will predict the probabilities of texts used to fill [Z]. Potential outputs are “lung”, “asthma”, or “respiratory”. Sivarajkumar & Wang (2023) study and other studies have explored prompt learning LLMs for clinical context and many reported promising results in information extraction. The subsequent section centers on research utilizing clinical PLMs and LLMs for clinical applications.

2.5.4 Clinical Pre-trained Language Models (PLMs) and Large Language Models (LLMs)

Existing language models are trained on generic knowledge, and their performance in medical language is not optimal. Therefore, clinical domain language models have been developed, e.g., ([Van Aken et al. 2021](#)), ([Hao et al. 2020](#)), and ClinicalBERT and DischargeBERT by [Alsentzer et al. \(2019\)](#). ClinicalBERT and Bio_Discharge_Summary_BERT are among the first attempts in LLMs to address this problem, and create models that are clinical-friendly, but they are not general enough being trained on only one dataset with a very specific time span, and being trained for masked language and next sentence prediction and a few specific target prediction tasks. As a result, these models capture the language characteristics of clinical narratives but not directly the medical knowledge that is critical for inference or prediction related tasks. Therefore, some recent studies attempt to enhance the medical knowledge for these models by further training or improving word embeddings. An example, [Huang et al. \(2019\)](#) further train BERT for predicting clinical outcome using admission notes and advance the models' capabilities by integrating knowledge about relations between symptoms, risk factors and outcomes derived from public patients' notes, articles and wikis.

[Hao et al. \(2020\)](#) enhanced the embedding of clinicalBERT by integrating information from Unified Medical Language System (UMLS) in the pre-training stage. This was achieved by transforming the relation between concepts in UMLS into a binary classification by which the model aims to predict the existence of the relation. The resulting model outperformed other models in NLP inference tasks and clinical named entity recognition.

The above examples have shown that research has been increasingly inclined to enhance clinical models like clinicalBERT for clinical NLP tasks since they have both specialized and contextualized representation of the vocabularies in the clinical texts.

Despite their enhanced representation for clinical applications, these models lack the ability to recognize numbers in text. In the study by [Van Aken et al. \(2021\)](#), ClinicalBERT was unable to detect values representing life-threatening risks. For example, it fails to recognize a temperature over 105°F as indication of mortality risk.

More recent studies explored prompt learning in clinical context. Examples are [Taylor et al. \(2024\)](#), [Sivarajkumar & Wang \(2023\)](#), and [Moradi et al. \(2021\)](#). [Taylor et al. \(2024\)](#) find that the prompt method outperforms fine-tuning when freezing pre-trained models in training for target tasks. They point out that the prompt method, requiring only few trainable parameters, can offer cost-effective solutions (even if not achieving the best performance) in clinical support tools. However, the existing general prompting models lack domain-knowledge and as such they miss an important component for explainability. The final thoughts of the authors suggest training and use of smaller PLM like BERT on domain-specific data with prompt-based learning.

[Sivarajkumar & Wang \(2023\)](#) developed HealthPrompt on six pre-trained language models using prompt-based learning on clinical texts. In particular, they apply zero-shot learning using six different pre-trained models with

no data. According to their findings, prompting methods perceive the clinical texts well and show excellent performance in phenotype classification task, where the proposed prompt model was asked to provide a phenotype category for a sample text extracted from MIMIC-III phenotype dataset. Their results show that using clinical-BERT for the prompt-based learning achieved the best results. This aligns with the conclusion by [Taylor et al. \(2024\)](#) that domain knowledge for models is important and should not be ignored.

[Moradi et al. \(2021\)](#) use few shot learning with GPT-3 and BioBERT for multiple NLP tasks (completion, answer, and classification) on biomedical text. Their findings show that GPT-3 performed poorly compared to BioBERT, regardless of the great difference of the models' sizes. They advise that domain-specific pre-training is beneficial but not enough and that new strategies for domain-specific pre-training and few-shot learning are necessary for task-specific few-shot learning in the domain.

[Peng et al. \(2023\)](#) developed GatorTronGPT, a generative clinical LLM trained on clinical text with GPT-3 structure and explored its uses in healthcare. The model provided excellent results on Turing test tasks; the paragraphs generated by the model was undistinguished from human performance based on readability and clinical consistency.

Prompt based approach are increasingly getting attention and emerging in NLP. However, their potential has not been fully explored as this paradigm is evolving, and many variants are expected to be introduced in the near future ([Taylor et al. 2024](#)). Despite the advancements and studies, many NLP models are not developed exclusively for specific tasks. Therefore, though many methods exist, there is no certainty about which method provides the best performance on language tasks. This is true for sentiment analysis, the task being investigated in this thesis. In fact, the use of the above recent LLMs for sentiment analysis in clinical domain specifically has been identified as an open research gap in the literature ([Denecke & Reichenpfader 2023](#)). The next section presents sentiment analysis techniques, before the chapter wraps up with the identification of the research gaps this thesis addresses (Section [2.7](#)).

2.6 Current Sentiment analysis Techniques

This section of the literature review presents the existing techniques and studies for sentiment analysis in the following order: first, sentiment analysis based on the level of application; second, sentiment analysis based on the used approaches or algorithms; third, sentiment analysis in the health and clinical domain. This then leads to Section [2.7](#) where the research gaps identified by this review are presented.

2.6.1 Sentiment Analysis Approaches Based on Text Level

Sentiment analysis can be applied to: document level, sentence or phrase level, and aspect level. Figure [2.4](#) from the study by [Wankhade et al. \(2022\)](#) depicts those levels. Document-level sentiment analysis is the highest

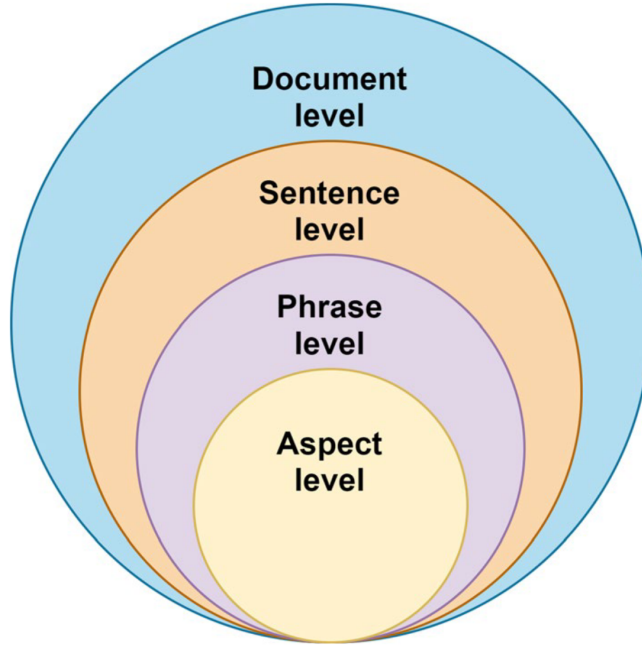


Figure 2.4: Sentiment analysis by text level, from (Wankhade et al. 2022)

level aiming for a single sentiment label or score for a complete document. Examples of studies using this scope are Taboada et al. (2011), Spinczyk et al. (2018), Sarker et al. (2011) and Choi et al. (2020). Document-level sentiment analysis is being improved with the advancement in deep learning as will be discussed later with some example studies in Section 2.6.4.1. Sentence-level sentiment has narrower focus, aiming to provide a sentiment score or label for individual sentences. The same applies for phrase level. Examples of research targeting sentence-level sentiment analysis are Holderness et al. (2019), Nguyen et al. (2017), and Carrillo-de Albornoz et al. (2018). Aspect-based sentiment analysis is the most fine-grained but also the most complicated. Aspect-based sentiment analysis (ASBA) aims to identify sentiments associated with specific aspects within the text, rather than identifying the overall sentiment of a chunk of text. Typically, this requires intensive annotation at the word/sentence level because each sentence may include multiple aspects. Therefore, although ASBA aims to provide a fine-grained and more accurate sentiment, it is more complex, requiring additional steps of prediction for the aspects themselves and the sentiment expressed for them, and increasing the burden with respect to manual annotation. As mentioned in Wankhade et al. (2022), ASBA requires three steps: finding out the aspect, finding out the polarity, and aggregating the results. Hoang et al. (2019) and Sun, Huang & Qiu (2019) are recent studies that perform ASBA using BERT model (BERT models are described in Section 2.5.2). The methods treat ASBA as a classification using a pair of sentences e.g., question-answer style and demonstrate improvement over previous studies with single-sentence classification. Salas-Zárate et al. (2017) use a semantic annotation technique which involves Stanford CoreNLP and diabetes diagnosis ontology to annotate aspects in tweets about diabetes. The polarity of the aspect is assigned depending on the polarity of the surrounding words using SentiWordNet. The results then were compared to human annotation and show that involving domain ontology

in identifying aspects was positive, though less accurate than human expert, suggesting demand for improvement in terminologies. The study also emphasizes the weakness of non-specialized lexicons in understanding health texts. [Guo et al. \(2021\)](#) provides a through review on datasets and issues identified in state of the art in relation to aspect sentiment analysis. Next sections describe the existing sentiment models, their methodologies, advantages, and limitations.

2.6.2 Sentiment Analysis Approaches Based on Methods

Existing sentiment models are either lexicon-based with specific built-in dictionaries or machine learning models which learn from pre-labeled data. Some studies use a hybrid approach where both methods are combined. The next sub-sections discuss each approach and its advantages and limitations.

2.6.3 Lexicon-Based Approach

Lexicon-based approaches to sentiment analysis have been the classic for sentiment analysis methods ([Ojeda-Hernández et al. 2023](#), [Holderness et al. 2019](#)). Lexicon-based approaches are also known as rule-based models. A number of rules or calculations are used to arrive to a final sentiment from text. Lexicons have dictionaries where words (typically adjectives and adverbs) are assigned to scores or weights in advance. Sentiment is calculated by examining the text, looking for specific words and assigning the corresponding score for each of the identified words. The overall sentiment is then decided by an aggregation equation ([Taboada et al. 2011](#)). Lexicons can be classified into two categories: dictionary-based and corpus-based ([Liu & Hosseini 2025](#), [Darwich et al. 2019](#)). Dictionary-based approaches contain a predefined dictionary extracted either manually or automatically from vocabularies in SentiWordNet and WordNet or online thesaurus. Although the dictionaries can be extended, this approach is still limited, as covering unlimited range of words is not feasible ([Liu & Hosseini 2025](#)). Corpus methods typically rely on words surrounding the desired word to determine its polarity ([Darwich et al. 2019](#)). In particular, corpus-based sentiment analysis develops a lexicon by having an initial set of words, and increasingly adding more by specific techniques. For example, one technique is to decide sentiment of words based on their distance to predefined positive and negative existing words ([Darwich et al. 2019](#)). Techniques can be based on statistical or semantic methods, or both. The statistical methods observe e.g, co-occurrence frequency of words in texts and study words behaviors to determine their polarity. i.e., words appearing (with specific criteria) in positive text are considered positive while words appearing frequently in negative text are considered negative, and equal occurrences are considered neutral. Some of the techniques used in statistical-based sentiment are k-nearest neighbors (k-NN), and hidden Markov models (HMM). The semantic approach finds a word's synonyms and antonyms and provides scores to these words by adding their semantic similarity to the loop ([Darwich et al. 2019](#)).

In general, corpus based methods demand a massive corpus to produce acceptable sized vocabularies. In many

cases, the lexicon may include noisy information compared to manually structured dictionaries, and may contain conflicting assignments for scores as statistics of words may not be reliable. For example, it was observed that only around 60% of co-occurrences share the same sentiment orientation, which reduces the quality of the identification of sentiment (Kanayama & Nasukawa 2006, Darwich et al. 2019).

Taboada et al. (2011) is an example of a lexicon method relying on dictionaries and rules including handling of intensity and negation. On the other hand, Velikovich et al. (2010) and Kanayama & Nasukawa (2006) are examples of corpus-based methods. Velikovich et al. (2010) utilizes label propagation technique for assigning polarity using lexical graphs from web content. Kanayama & Nasukawa (2006)’s lexicon is built based on the assumption that adjacent sentences often convey similar sentiment polarity. Darwich et al. (2019) summarises many techniques and studies with a focus on corpus-based lexicon generation.

Some of the problems that lexicons cannot resolve are negated content, ambiguous or ironic language, modern terms or vocabularies (Ojeda-Hernández et al. 2023). Another drawback of lexicon-based methods, mainly dictionary-based ones, is their ignorance of context of words (Liu & Hosseini 2025). In particular, they assign a score to a word in advance using a dictionary without consideration of its context (Nguyen et al. 2017). In addition, the performance of the lexicon based methods is usually consistent when applied to the same domain but typically fails in cross-domain tasks (Grissette & Nfaoui 2021) as they are vocabulary-dependent and they lack the ability to learn.

Weissman et al. (2019) evaluated the validity of six existing, general sentiment models on clinical notes derived from MIMIC datasets. Five of these models are lexicon based. Based on the results, the authors conclude that domain-specific sentiment analysis is required to improve the task on clinical texts. Holderness et al. (2019) evaluated Pattern, a general domain lexicon against ML models. Pattern was assessed as a baseline, and provided the weakest performance due to lack of specialization. Weissman et al. (2019), Holderness et al. (2019) are further discussed in Section 2.7. The next section presents the machine learning approach to sentiment analysis.

2.6.4 Machine Learning Approach

The machine learning approach to sentiment analysis is focused on finding features that distinguish the classes of sentiment into e.g., positive, negative or neutral. The model (a classifier) is built from a large volume of samples (either a sentence or a document) and their corresponding pre-assigned labels (Taboada et al. 2011). Examples of ML classifiers which can be used for sentiment classification are Support Vector Machine (SVM), Naive Bayes (NB), Decision Tree (DT), Logistic Regression (LR) and k-nearest neighbors (KNN) (Mao et al. 2024).

Figure 2.5 shows the differences between the lexicon-based methods described in the previous section and the ML approaches to sentiment analysis. As shown, lexicon-based method does not require training but depends on pre-prepared dictionaries and rules. In contrast, machine learning classifiers require training on labeled samples before they are ready to function and classify new samples.

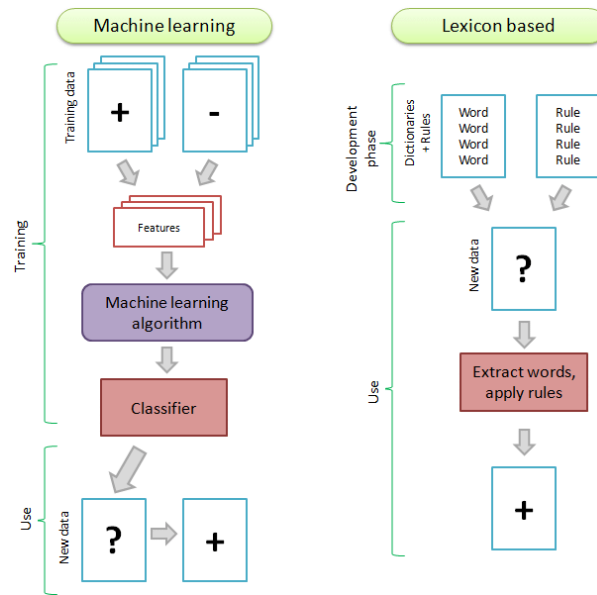


Figure 2.5: Difference between machine learning and lexicons approaches to sentiment analysis (from Taboada et al. (2011)).

Three main challenges in ML approaches are collecting large data, annotating this data, and selecting the best feature set. These steps require multiple experiments and trials before achieving satisfactory results. For example, Sarker et al. (2011) use four feature sets and four machine learning classifiers including Naive Bayes, Bayes Net, SVMs and C4.5 Decision Tree for sentiment on medical publications. The classifiers were trained using multiple feature sets, e.g., N-grams, and two proposed features: Relative Average Negation Count, and Sentence Signature. Relative Average Negation Count is an aggregate measure of negations that takes into account the position of negation relative to the document being processed. Sentence Signatures are patterns composed of two or three terms that represent sentences in a way that describe changes occurring in a sentence e.g., *less-bad* means a bad thing was decreased. The suggested features improved the accuracy of the sentiment classification. The findings motivate the inclusion of carefully selected features in machine learning approaches. Another example of this approach is the study by Holderness et al. (2019) who performed sentence-level sentiment analysis using three machine learning classifiers on psychiatric EHR texts. They annotate sentences based on selected risk factors for readmission from psychiatric point of view such as appearance, mood, and interpersonal relations. Their analysis included Pattern lexicon (a well-known general sentiment analysis model) as a baseline for comparison. They found out that it does not capture the sentiment of many of the significant words such as ‘depressed’ as these words were not covered by its dictionary. Furthermore, their analysis concludes that semi-supervised models using Self-Training and K-Nearest Neighbors (KNN) achieve comparable results to the fully-supervised learning.

ML approaches have shown improvements over lexicon-method in many applications due to their learnability. However, as mentioned they require annotated data and features extraction, which can be resource-intensive. The next subsection introduces the more recent deep learning method, which addresses some of these limitations.

2.6.4.1 Deep Learning Approach

Section 2.6.4 discusses traditional machine learning methods and this section introduces deep learning methods which, although a subset of ML methods, obtained more popularity and adaption in recent years due to advancement and access to computational resources. Deep learning approaches utilize neural networks with multiple hidden layers. Deep learning networks learn features automatically, making them more convenient and more powerful compared to traditional machine learning (Dang et al. 2020). This holds true for text data as deep learning has shown capabilities in extracting syntactic and semantic features with no need for manual textual feature engineering (Alshuwaier et al. 2022).

Some of the common deep learning models are convolutional neural networks (CNN), Deep Neural Networks (DNN), Recurrent Neural Networks (RNN), Generative Adversarial Networks (GANs), and long short-term memory (LSTM). Choi et al. (2020) proposed a methodology of document sentiment prediction which considers and distinguishes the degree of importance of sentences in the input document by applying a gate mechanism in their deep learning model. A gate mechanism is a mechanism that is added to models to allow for selective actions e.g., update, forget, or pass for data within hidden internal states in cells and allow for easier backpropagation through time stamps. The study results outperformed the state of the art models reflecting the positive impact of considering sentences' importance in document sentiment analysis. Shin et al. (2017) shows reliable accuracy for classifying radiology reports using Convolutional Neural Network (CNN) models with an attention mechanism for highlighting important features. Alshuwaier et al. (2022) provide a recent (up to 2022) comprehensive systematic reviews for deep learning methods in document level sentiment analysis for multiple domains. Despite the large number of studies and the ranges that the study covers, there was no mention of any studies targeting clinical domain for document sentiment level, confirming the research gap in the field.

Dang et al. (2020) provide a comparative study for three different deep learning methods including DNN, RNN, CNN as well as two input transformation methods including TF-IDF and word embedding. The analysis shows three findings: the best deep learning models should be selected based on best trade-off between time complexity and accuracy; word embedding is more robust than TF-IDF in transforming text to numeric representation; sentiment models built for a specific topic have better performance than the ones built on generic topics. It is important to note that explainability for deep learning models might be challenging as they contain less transparent methodologies compared to lexicons models with manual or pre-defined dictionaries (Weissman et al. 2019). This is because as mentioned earlier, deep learning methods skip explicit feature extraction steps and rely on model layers to identify hidden features from text and therefore those features might not be disclosed clearly to researchers. The difference between machine learning approach and deep learning approach to sentiment analysis is described in Figure 2.6 by Dang et al. (2020).

Zhang et al. (2018) offers a survey of the use of deep learning in sentiment analysis. A key advantage in deep learning approach is also the attention mechanism that allows the model to keep track for long dependency links

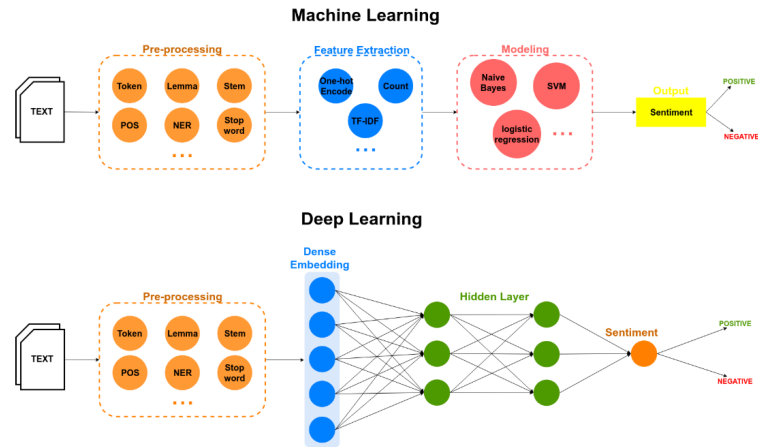


Figure 2.6: Differences between machine learning and deep learning classification approaches of sentiment polarity (from [Dang et al. \(2020\)](#))

in words and context ([Zhang et al. 2018](#)).

In more recent years, since around 2018 to the present, pre-trained language models (PLMs) and large language deep learning models (LLMs) are increasingly being developed, advanced, and adapted, with unprecedented and significant improvement in NLP tasks. The next subsection provides background of PLMs and LLMS in relation to sentiment analysis.

2.6.4.2 Pre-trained Language Models (PLMs) and Large Language Models (LLMs)

Sections 2.5.2 and 2.5.3 provided background about BERT and subsequent prompt models. This section discusses the application of PLMs and LLMs to sentiment analysis.

Based on BERT excellence on NLP tasks, a number of studies use BERT based models for general sentiment analysis and aspect based sentiment analysis. Examples are SentiBERT ([Yin et al. 2020](#)), BERT-PT ([Xu et al. 2019](#)), DomBERT ([Xu et al. 2020](#)), SKEP ([Tian et al. 2020](#)), TransBERT ([Li et al. 2022](#)), SentiLARE ([Ke et al. 2020](#)). Those studies target sentiment analysis tasks but have not studied their performance in clinical domain. When describing a patient state, the physician may make use of a language that bears some sentiment, or implicit polarity, reflection of the untold opinion of the doctor, something that if captured by an algorithm can be used to detect early sings of specific patient complications or disease. Therefore, a clinical based model can enhance the understanding of the language used in clinical text.

However, as indicated earlier, many of clinical models from both BERT-models and prompt-models were used to predict clinical outcomes such as mortality, length of stay or readmission, but none were specifically designed or used for clinical sentiment. In fact, training models for such task requires a labeled dataset for transfer learning, which is unavailable.

A number of BERT-Style models target outcomes such as [Huang et al. \(2019\)](#) and [Van Aken et al. \(2021\)](#). Likewise,

more recent prompt clinical models are mostly trained for tasks unrelated to sentiment, such as ICD-9 triage task, in-hospital mortality, length of stay in ICU in [Taylor et al. \(2024\)](#), and disease Named Entity Recognition, 30-day readmission risk, 30-day and 1-year mortality risk in [Lu, Dou & Nguyen \(2022\)](#).

In general, as mentioned above, the use of recent LLMs for sentiment analysis in the clinical domain has been identified specifically as an open research gap in the literature ([Denecke & Reichenpfader 2023](#)). The topic is still underexplored and this motivates the thesis. The next section presents hybrid approach to sentiment analysis.

2.6.5 Hybrid Approach: Lexicon-based and ML

As mentioned before, hybrid approaches combine lexicon-based methods with machine learning methods. The purpose of this combination is to bring together the advantages of both by making the ML classifier learn from features extracted by the lexicon model. In other words, the results of the lexicon methods are used as input for the machine learning methods. This has shown improvement reported in the literature, but did not eliminate the challenges of machine learning methods including the demand for feature engineering, large data and manual labeling. [Nguyen et al. \(2017\)](#) propose a hybrid method that combines machine learning and lexicon-based methods applied to medical text from articles. In particular, semantic orientation calculator (SO-CAL) was used to calculate a polarity score for each sentence by scores of its words. The scores of the words are assigned by a predefined dictionary with consideration of word form, negation, and intensification. Their experiments show that the combination of few linguistic baseline features and SO-CAL scores to train a support vector machine (SVM) classifier improves the classification precision, recall, and F1-score to up to 70%. The study by [Neshan & Akbari \(2020\)](#) applies a hybrid approach which utilises four lexicons and four machine learning models. The lexicons' output scores for the text are combined and passed as input for the machine learning models. The lexicons scores are extracted by SentiWordNet Score, Liu's lexicon score, SentiStrength score, and ratios of positive and negative words are calculated for every opinion extracted by opinion mining algorithm in a preprocessing step. The final decision is determined by a meta-classifier which combines the four classifiers results (classifiers are Naïve Bayes, Support Vector Machine, Multi-layer Perceptron classifier, and Decision Tree) for final classification. The results of the hybrid method show slight improvement over the individual classifiers and individual lexicons.

In addition to the above categorization, multi-modal sentiment analysis is a combined approach, which includes fusion techniques for different data types e.g., audio, video, and text. [Gandhi et al. \(2023\)](#) provides an overview and a discussion of multi-modal sentiment analysis. [Mao et al. \(2024\)](#) provides a recent survey on all approaches to sentiment analysis to guide in choosing the appropriate method for analysis. Hybrid approach and multi-modality sentiment are out of the scope of the thesis. The next section describes an overview of sentiment analysis in clinical domain followed by research gaps.

2.7 Overview of Sentiment Analysis in Clinical Domain and Research Gaps

This section provides a brief overview of both lexicon-based approaches and ML-based approaches in relation to sentiment analysis for clinical-related data. Based on the literature survey, sentiment analysis research in the medical domain has been mostly focused on either biomedical literature or patient-generated text (Deng et al. 2014), with little focus on clinical texts or notes. Moreover, most existing work applying NLP to clinical notes focuses on predicting risks of re-admission or outcomes (Kumar et al. 2022, Liu et al. 2022, Thapa et al. 2022). There has been minimal attention to adapting and improving sentiment analysis for clinical notes. Therefore, more studies are required to address this gap and enable the integration of this analysis in clinical decision support systems (Denecke & Reichenpfader 2023).

Weissman et al. (2019) evaluated the validity of six existing, general sentiment models on clinical notes derived from MIMIC datasets. Five of these models are lexicon-based including Pattern (De Smedt & Daelemans 2012), sentimentr (Rinker 2021), AFINN (Nielsen 2011), EmoLex (Mohammad & Turney 2013), and Opinion (Hu & Liu 2004) and only one (i.e. StanfordCoreNLP (Manning et al. 2014)) deployed a pre-trained model. The analysis showed that, although a few of the selected models have shown minor validity, (i.e. inverse correlation between positive sentiment and notes on the day of death), they have little potential to succeed in clinical text sentiment analysis due to their weak lexical coverage for the terminologies used in the clinical notes and the variation in their agreement on sentiment prediction resulting in inconsistency between them. Based on the results, the authors conclude that domain-specific sentiment analysis is required to improve the task on clinical texts. Holderness et al. (2019) came to the same conclusion emphasizing the importance of domain adaptation of sentiment analysis in clinical text. A 2023 review (Denecke & Reichenpfader 2023) found most clinical sentiment studies use general domain lexicons, including Denecke & Deng (2015), Cuenca-Zaldívar et al. (2022), Zubillaga et al. (2022), and Deng et al. (2014).

SentiHealth (Asghar et al. 2016) is among the very few health-specific lexicons and was constructed from online forums of health-related discussions about drug reviews. The authors conclude that clinical-specific sentiment lexicons are needed due to linguistics differences in the languages. Similarly, Elbers et al. (2023) employs a data-based method to re-calibrate the vocabulary scores and improves the coverage of words for the oncology domain in the LabMT sentiment dictionary. Waheeb et al. (2020) identified some issues that should be addressed in relation to sentiment analysis during their analysis for clinical discharge summaries. They are context-dependent polarity, difference in words meanings, models' poor lexical coverage, and contexts that share semantics but have conflicting polarity.

Ntiamoah et al. (2024) applies sentiment on posts extracted from a public blog regarding sepsis events provided by patients and providers to observe the emotion expressed. The analysis uses National Research Council (NRC)

Canada Emotion Lexicon, and reveals that emotional support is required for the patients and families based on the prevalence, timing, and links of the negative emotions such as fear or sadness.

[Chintalapudi et al. \(2021\)](#) applies both lexicon-based analysis and ML-based sentiment classification on seafarers' medical documents to explore medical issues in their trips. The analysis was successful with 80% accuracy using the domain specific pre-assembled lexicon and the trained Naïve Bayes classifier.

[Deng et al. \(2016\)](#) was among the very few studies that target the generation of a clinical specific corpus for sentiment analysis using a particular definition for sentiment. The definition includes three components including event about patients status, and care provider attitude towards clinical, pharmaceutical, and social aspects. However, the dataset is not yet released for public use.

Several studies use Naive Bayes and other ML models with linguistic features, such as part of speech or n-gram, in predicting outcome polarity from medical publications ([Sarker et al. 2011](#)), posts in medical forums ([Ali et al. 2013](#)), online drug surveillance ([Chee et al. 2011](#)), and psychiatric patients' electronic health records narratives ([Holderness et al. 2019](#)). [Spinczyk et al. \(2018\)](#) analyze document sentiment. First, they use a bag of words approach to represent documents of two individual groups including healthy individuals and individuals with anorexia nervosa. After obtaining documents' representation, they use dictionary and rules to identify sentimental keywords from the text, and classify the document into positive, negative, or neutral based on counts of keywords.

Two studies, by [Gupta et al. \(2022\)](#) and [Naseem et al. \(2021\)](#), studied sentiment on tweets about Covid-19. A recent study by [Cuenca-Zaldívar et al. \(2022\)](#) explored sentiment analysis in nursing clinical notes in the early stage of COVID-19 pandemic in Spain. The study uses dictionaries (Afinn and Bing) and ML models (Support Vector Machine and neural network) to identify polarity from two groups (a group of patients with COVID-19 vs. patients without it). The study showed that polarity in notes varies significantly and more positive sentiment was found for the group of subacute COVID-19.

In summary, general sentiment models demonstrate variability and different levels of validity in performance but are generally poor in capturing sentiment from clinical notes ([Kennedy et al. 2023](#), [Weissman et al. 2019](#)). A recent scoping survey by [Denecke & Reichenpfader \(2023\)](#) revealed the following in relation to clinical sentiment:

- Analyzing clinical sentiment can enhance the forecasting of clinical outcomes.
- General domain methods are still the most prevalent for clinical sentiment analysis.
- Recent approaches such as LLMs are still underutilized for clinical sentiment analysis.
- Improvements are required for effective capture of clinical sentiment in the clinical domain.

These findings align with the following, which are identified gaps based on the literature review of this thesis:

- The lack of effective sentiment classifiers has led to missing the potential benefit of clinician impressions in clinical notes.
- The literature shows no clear answer for which approach to follow in order to get a well-functioning sentiment model for ICU notes.
- Related work shows no existence of public and labeled datasets of clinical sentiment for ICU notes, which is a significant issue for the development of machine learning methods and evaluation for other methods e.g., keyword-based methods.
- Unavailability of clinicians who can do the labeling or extract texts for sentiment analysis hamper the development of specialized lexicons and creating ground truth data for training and evaluation of automated models.

The above motivates this thesis to address those challenges in support of clinical sentiment analysis. Chapter 3 presents the methodology used to do so. Future studies should address the incorporation of clinical sentiment into clinical practice, by addressing the following:

- Compare the prediction of risk factors in ICU using the structured data in electronic health records including readings such as heart rate and blood pressure, and labs results with prediction using unstructured clinical notes (Denecke & Reichenpfader 2023).
- Investigate the fusion of both structured and unstructured data into clinical prediction frameworks.
- Investigate the integration of clinical sentiment analysis in clinical decision support systems and the implications.

2.8 Conclusion

In conclusion, this chapter has reviewed the available methods in the literature for sentiment analysis in general domain and clinical domain. The literature survey revealed a number of research gaps which require attention by researchers. With the above, RO1 has been achieved by providing the necessary background and overview of traditional and recent methods for sentiment analysis in the literature. Chapter 3 introduces the methodology of the research addressing the research question.

Chapter 3

Methodology

3.1 Introduction

This chapter presents an overview of the research methodology. Section 3.2 presents the research dataset. Section 3.2.1 presents background and statistics of the research data. Section 3.2.2 discusses dataset labeling issue. Section 3.2.3 introduces the pseudo-labeling scheme. Section 3.2.4 presents the samples of the datasets. Section 3.3 presents the overall methodology of the research, and the phases that the research underwent. Section 3.3.1 presents the hypothesis of the research, and Section 3.3.2 discusses the models used in the study. Section 3.3.3 presents the evaluation strategies. Section 3.4 concludes the chapter with a brief summary.

As discussed in Chapter 2, different approaches can be implemented to distinguish sentiment in ICU clinical notes. However, as mentioned earlier, the existing techniques demand specific requirements. For example, in supervised ML methods, both training and evaluation require labeled datasets. In fact, training usually requires a large set of labeled data. On the other hand, non-learning methods such as dictionary or keyword-based lexicons do not require any training data, but they require domain experts for development and labeled data for evaluation. Based on these requirements, training and evaluating sentiment models on clinical data is challenging as no such datasets are available. This is the motivation behind the labeling strategy proposed in this thesis. Another challenge in analyzing text data is deciding the length of a text for a desired task (Su et al. 2021). No standard amount of data is known to be sufficient, and that is particularly true for textual data, as the context is a key component to understand text and can be affected if text was segmented. Even the recent large language models limit the number of words (or particularly sub-words known as tokens) they process at once.

Additionally, it is also unknown if capturing clinical sentiment is more accurate using clinical notes written by different providers e.g., physicians vs. nurses. While different approaches exist, each has different advantages, disadvantages and capabilities. For instance, LLMs require high computational cost (often large data, memory and

GPUs) for training while keyword-based lexicon are cost-effective in terms of computational resources. However, keyword lexicons are built with explicit coding of domain knowledge obtained by experts (often a team of 3 or more) and lack learning ability, so they cannot generalize their functionality to new or evolving issues. As mentioned earlier, model performance also varies depending on the domain used during their development. All these aspects leave a large room for exploration and demand through experiments. These aspects and challenges are addressed while designing the thesis methodology which are presented in this chapter.

3.2 Dataset

At the early stage of the PhD, the intended research primary dataset was clinical notes from Intensive Care Medicine at St. James's Hospital, which is a large teaching and research hospital in Dublin. However, despite repeated promises to deliver this dataset, the dataset was never delivered, due to disruption caused by the Covid-19 pandemic and difficulties obtaining ethical approval from the hospital. Therefore, an alternative was required. Another trial to obtain hospital data was attempted with King Faisal Specialist Hospital & Research Centre (KFSHRC) in Riyadh. However, difficulties arose in relation to gaining approval in time. Searching for suitable online data, MIMIC-III (Medical Information Mart for Intensive Care) is the only free and publicly available dataset which contains progress notes of patients in critical care. Being appropriate for the research purpose, it became the primary dataset for the thesis experiments. MIMIC-III is available under protected access through *Physionet*¹. Access to the data is restricted to credentialed users who are required to do specific training called "CITI Data or Specimens Only Research" as described on the website. The MIMIC-III database² contains de-identified information about patients of the critical care units at Beth Israel Deaconess Medical Center in Boston, Massachusetts, USA in the period 2001-2012 (Johnson et al. 2016, 2015, Goldberger et al. 2000). Clinical notes are in English. The dataset contains categories of notes such as nursing and physician notes, electrocardiogram (ECG) reports, radiology reports, and discharge summaries. Other types are General, Case Management, Consult, Pharmacy, Respiratory, Social Work, Echo (echoencephalogram), Nursing/other, Nutrition, and Rehab Services. This research considers patients in ICU represented in ICUSTAYS, and collects the physicians and nursing notes from clinical notes recorded within the duration of their stay. Physician notes and nursing notes are extracted for the research as previous studies have proven their importance and their strong connection with prediction in ICU (Cioffi 2000, Collins & Vawdrey 2012, Collins et al. 2013). Table 3.1 provides statistics of patients, ICU stays, and clinical notes in MIMIC-III.

In addition to MIMIC data, another dataset has been utilized as a secondary data in evaluation of the research. This data contains clinical notes associated with chest X-ray and CT images of patients infected with or suspected of COVID-19 or other pneumonia causes (Cohen et al. 2020). It was mainly collected from publications, but some samples were collected independently from hospitals and physicians. The notes in this dataset do not have specific

¹<https://physionet.org/content/mimiciii/1.4/>

²Dataset license is available at: <https://physicianionet.org/content/mimiciii/view-license/1.4/>.

Table 3.1: Statistics of MIMIC-III patients, physician and nursing notes.

Description	Statistic
All patients	46,520
ICU stays	61,532
Physician notes	266,520
Nursing notes	169,642
Average length of stay (LOS)	11.75
Median length of stay (LOS)	3.72
Average counts of notes per patient	71
Median counts of notes per patient	29

standard structure. Notes are mostly brief and some focus on observations made in relation to the chest X-ray. Other information provided with data includes sex, age, survival, extenuation information and some physiological readings such as temperature, and oxygen saturation. These notes were written by physicians. Only samples of patients who are located in ICU are extracted (158). Of those, 23 samples were annotated by a clinician.

3.2.1 Dataset Descriptive Analysis

MIMIC-III contains 26 tables which has distributed patients details including admissions, patients demographics, and ICU stays details. All tables and database schema are documented in MIT Laboratory for Computational Physiology (MIT-LCP)³.

As discussed earlier in 2.3.2 and RO3, notes differ based on the provider and their purpose. Depending on the category of the note, clinicians tailor a suitable language to effectively communicate the information to the medical team. Therefore, it is not clear which category of notes may contain information that convey more clinical sentiment. Kennedy et al. (2023) study counts of sentiment words in different notes' categories. Table 3.2 shows the prevalence of sentiment words keywords from their study. As can be seen from the table, physician and nursing notes are the most frequent note categories. Additionally, physician notes have the greatest percentage of occurrence of any (i.e., positive or negative) sentiment words, but have slight increase in negative terms. On the other hand, positive terms are more frequent in nursing notes. Taking into account the prevalence of sentiment words in physician and nursing notes along with their associations with predictions in ICU as mentioned in 3.2, these two categories are suitable for clinical sentiment analysis.

The following text shows a sample of a nursing note. (MIMIC-III notes are protected. Hence, the following samples have been synthesized by a clinician.)

Pt p/w acute lower abd pain, with marked hypotension, in ed received 4l iv fluid with sbp maintained on 70's. admitted to sicu for bp monitoring and abd exam. o/e abdominal tenderness. pt bo x 3 abd soft positive bowel sounds + flatus, low bp this am bp as low as 70

³<https://lcp.mit.edu/mimic-schema-spy/index.html>

Table 3.2: Proportion of notes of Kennedy et al. (2023) samples and sentiment keywords frequencies based on category.

Category	Notes proportion	Positive word	key-	Negative word	key-	Any keyword
Nursing	57,683 (28.8%)	67.6%		53.0%		81.1%
Physician	52,033 (26.0%)	89.0%		92.3%		97.8%
Respiratory	18,478 (9.2%)	17.7%		80.3%		65.0%
General	5,927 (3.0%)	50.3%		60.1%		65.0%
Nutrition	4,561 (2.3%)	33.2%		63.1%		60.5%
Rehab Services	4,581 (2.3%)	78.5%		53.1%		85.9%
Consult	89 (0.0%)	68.1%		82.6%		91.3%

s/40 s, pain level 6 out of 10 chest clear, urine normal. 2l iv blouse for severe low bp then improve 100s/50s, dilaudid for pain improve pain to 2 out of 10. plan: cont. monitor and report new s+s to sicu, assess abd, painkillers, iv fluid, support prn.

A synthesised example of a physician note is shown here:

65-year-old female patient presented to the ED with severe sub-sternal chest pain, known case of CAD with PCI to LCx in 2019. This pain started 30 minutes ago, patient then shocked for suddenly collapsed with VT/VE, then, ECG showed STEMI with significant trop leak. PCI was done and showed 90% lesion in LAD with patent LCx. DES to LCx. Patient recovers well with Trop trending down. TTE: LV with HK, EF is 45-40%. HD is stable during CCU. Ongoing trop trending down.

Assessment: HR: 70 bpm. NSR. Chest discomfort post CPR improving with Tylenol.

Action: Cardiac meds started.

Response: Meds well tolerated.

Plan: Cont. STEMI protocol For discharge today.

3.2.2 Dataset Labeling

As noted earlier, supervised ML requires a labeled dataset i.e. a dataset of clinical notes labeled individually by clinicians. Manual labeling requires time and effort by busy clinicians. Also privacy concerns limit availability of clinical data (Taylor et al. 2024). The lack of labeled data has obstructed developing and evaluating models. In the early stage of the PhD, no manually annotated sentiment dataset was available. Recruiting clinical annotators to create one was a very slow and laborious process (see Chapter 5). To proceed with sentiment analysis experimentation in the first phase of investigation, an alternative labeling was adopted. This is a sentiment labeling strategy based on the timing of the note relative to key events such as admission to ICU (at which point patient status is assumed to be poor), death (prior to which status is assumed to be poor) and discharge from

ICU (prior to which status is assumed to be more positive). This approach is inspired by [Weissman et al. \(2019\)](#) as explained in the next subsection.

3.2.3 Pseudo-labeling Strategy

Recent studies by [McCoy et al. \(2015\)](#), [Gao et al. \(2021\)](#), and [Waudby-Smith et al. \(2018\)](#) studied the association between sentiment applied to ICU clinicians' notes, mortality and readmission and show an association between sentiment and risk of readmission and mortality. However, they did not assess the accuracy of the sentiment analysis itself since they did not have a labeled dataset. On the other hand, [Weissman et al. \(2019\)](#) focused on evaluating sentiment models, but again did not use labeled data to test the sentiment in their study. They rather depended on an equation calculating positive sentiment in relation to daily aggregated score of positivity and negativity in aggregated notes. They used this score to study how sentiment varies in relation to events in ICU.

This thesis adopts a labeling strategy inspired by the measures suggested in [Weissman et al. \(2019\)](#). It is a patient-dependent, time-and-survival-based method to provide positive and negative pseudo-labels for notes rather than the manual method. No reasonable strategy was found to identify notes for neutral class and neutral class has been discussed as a difficult class and less-dominant than positive and negative classes in other studies. [Kennedy et al. \(2023\)](#), and [Holderness et al. \(2019\)](#) discuss the difficulty of the neutral class. The key point is that distinguishing neutral class is not easy even for experts and so for models. Additionally, it was observed that class distribution in clinical documents is uneven and biased to positive or negative classes. The majority class tend to be negative. This is because clinicians keep mentioning and documenting ongoing issues for patients until the patient shows significant improvement. Being focused on documenting noticeable outcomes or changes, clinicians' notes tend not to have much of neutral language.

The details and logic of the labeling involve three patients' timing attributes in relation to the following:

- Patients' admission.
- Patients' discharge.
- Patients' death.

Specifically, three subgroups of notes are extracted including:

- Group1: admission to ICU note and first 3 days of progress notes for patients regardless of their discharge outcomes.
- Group2: notes for last 3 days for discharged patients.
- Group3: notes for the last 3 days for deceased patients.

The labeling involves two assumptions as follows:

- Patients are likely to be in very critical health status when admitted to ICU and near to mortality time. Based on this fact, progress notes close to admission time and close to death time are labeled as "deteriorating" or "negative" polarity.
- On the other hand, progress notes close to discharge time are likely to reflect improvement in patients, thus are labeled as "improving or will get better" or "positive" polarity.

The names of the pseudo-labels were derived in discussions with two physicians, an ICU consultant and a respiratory consultant. Throughout the thesis, the two terms "improving or will get better", and "deteriorating" may be referred to as "positive" and "negative", or binary labels collectively. These assumptions are not only intuitive, but align with two validity measures used to evaluate the extent of effectiveness of sentiment analysis models as proposed in [Weissman et al. \(2019\)](#), namely concurrent validity and predictive validity. A brief explanation is provided below. These measures are used to assess validity of sentiment analysis models:

1. **Concurrent validity:** checks if model temporal sensitivity to events, interventions, and outcomes show evidence of validity or reliability for the clinical sentiment model, i.e. a model with concurrent validity should have strong association between the sentiment measure and the outcomes in the same period of time, e.g. providing negative sentiment in the same day of death.
2. **Predictive validity:** checks if correlation between sentiment and prediction of outcomes like mortality show evidence of validity or reliability for the clinical sentiment model, i.e. model with predictive validity should have strong association between the sentiment measure and the future outcomes, e.g., association between negative sentiment and death.

Also, [Holderness et al. \(2019\)](#) pointed out that involving timing of events in studying patient status enables noticing gradual changes over time and is desirable to be part of sentiment definition. On that basis, the two physicians recommended a time window of 3 days pre or post events. However, the number of days can be reassessed depending on the analysis in further experiments. In practice, this pseudo-labeling is done as follows: ICU notes in MIMIC-III dataset are divided into 5 sub-groups. Group1: admission to ICU note and first 3 days of progress notes for patients who were discharged from ICU. Group2: admission to ICU note and first 3 days of progress notes for deceased patients. Group3: discharge from ICU note and last 3 days of progress notes for discharged patients. Group4: discharge from ICU note and last 3 days of progress notes for deceased patients. Group5: all other notes. Based on the assumption stated above, Group1, Group2 and Group4 are labeled as negative; and Group3 is labeled as positive. Because no method can be found to infer the sentiment for the remaining notes, Group5 was withdrawn from further analysis. Then, all notes with labels were collected, and the final pseudo-labeled dataset was constructed. Table 3.3 shows the sizes of datasets related to pseudo-labeling strategy. Figure 3.1 visualizes the labeling, and Figure 3.2 shows visually the labeling for a survived patient and a deceased patient. Notes are extracted using 'CATEGORY': 'physician', 'Nursing' and 'DESCRIPTION': 'Intensivist Note'. In labeling notes close to discharge time, patients discharged to HOSPICE services including

Table 3.3: Sizes of datasets related to pseudo-labeling strategy. Final pseudo-labeled data is obtained after all pre-processing and segmentation presented in Chapter 4.

Category	Physician	Nursing
Size of notes 3 days after admission	17,021	25,026
Size of survived pt notes 3 days prior discharge	6,534	11,559
Size of died pt Notes 3 days prior death	3,412	4,706
Total of 3 days	26,967	41,291
Size of final pseudo-labeled data	12,666	12,666

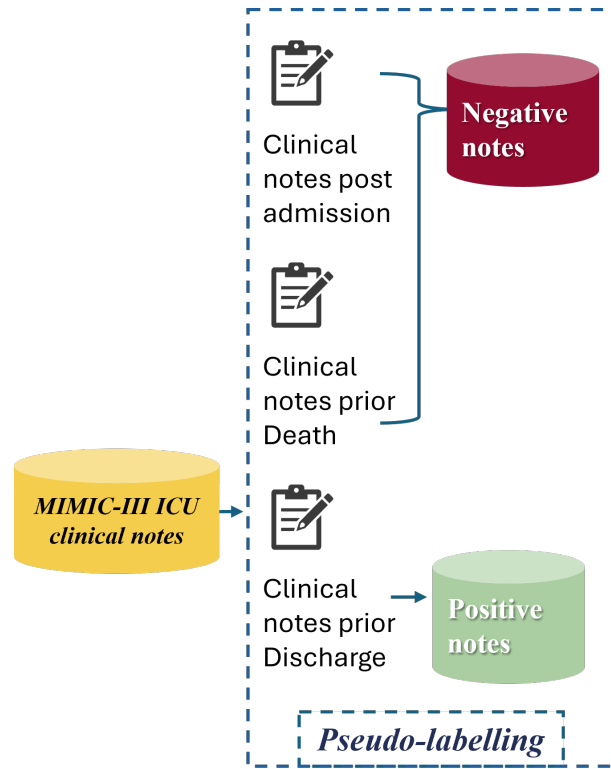


Figure 3.1: Pseudo-labeling strategy applied to MIMIC-III.

'HOSPICE-MEDICAL FACILITY' or 'HOSPICE-HOME', and patients who left against medical advice ('LEFT AGAINST MEDICAL ADVI') were removed from positive samples as they do not belong to positive outcomes. CHARTTIME, INTIME, OUTTIME, DISCHTIME (chart time, ICU in time, ICU out time, discharge time) were used to determine the timing window (1-3 days) and DOD_HOSP (date of death reported by hospital) for flagging death using the reported death time.

3.2.4 Dataset Samples

Overall, this research use two categories of notes: physician and nursing notes, which will be further segmented in two segments depending on notes' structure as explained in Section 3.3. The initial phase of the research utilizes only MIMIC-based datasets for training and evaluation which are automatically labeled using the strategy. The final phase utilizes three datasets:

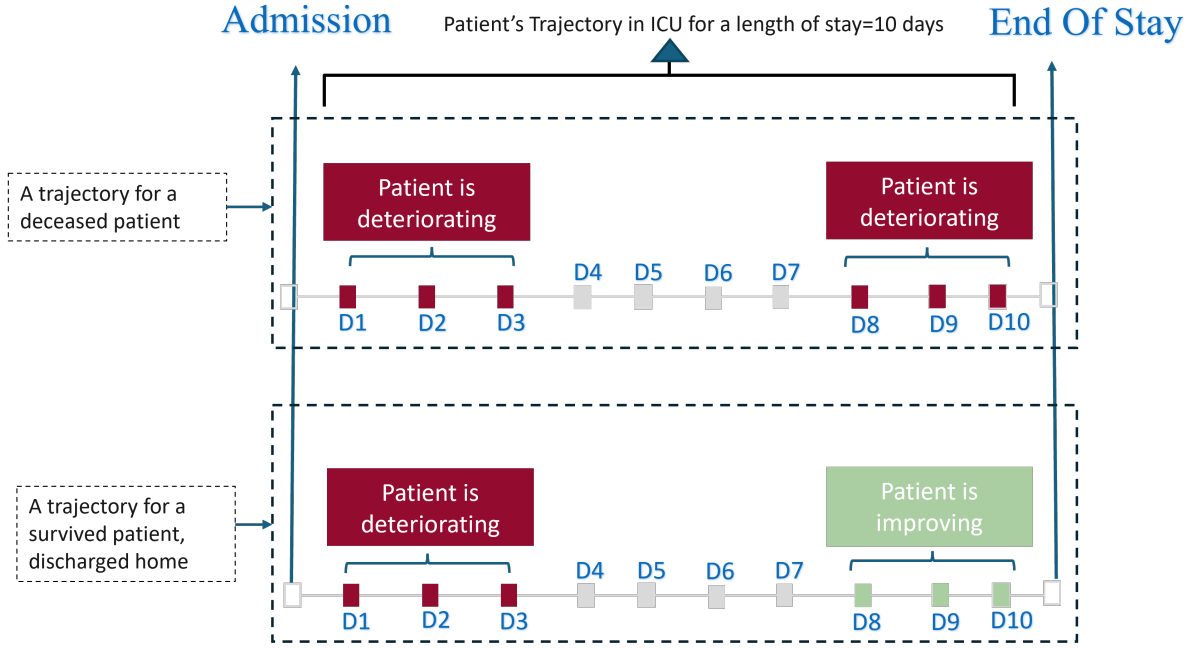


Figure 3.2: Pseudo-labeling strategy: labels for a deceased patient (on top), and a patient being discharged to home. Both patients length of stay is 10 days. 'Improving' is used here as shorthand for 'is already improving or will get better'.

- The MIMIC-based training data automatically labeled.
- The MIMIC-based evaluation data or 'ground truth' labeled by clinicians.
- The Covid dataset, which is a non-MIMIC data based on Covid-19 patients and annotated by a clinician.

Table 3.4: Research datasets sizes

Category	Physician	Nursing
Size of pseudo-labeled training data	10766	10766
Size of pseudo-labeled Testing data	1900	1900
Manually annotated data by clinicians	156	144
MIMIC Ground Truth (5V)	17	15
MIMIC Ground Truth (4V)	46	64
The Covid dataset	23	-

All datasets are described in Table 3.4. More details about research phases is explained next in Section 3.3. Details about datasets are documented in Chapter 4 for initial experimentation, and Chapters 5 and 6, for generation and evaluation on ground truth.

3.3 Method

The thesis methodology employs a comparative research with an iterative design used to optimize the selection of the most effective method and model for identifying clinical sentiment in ICU clinical notes. This process is realized through a sequence of experiments conducted on ICU clinical notes. The methodology is progressively

refined through sequential and chronological investigative phases, as mentioned briefly in Chapter 1 and outlined below. Because the manually annotated dataset has taken 15 months to obtain (details in Chapter 5) it was not available until the third (final) phase of the research. Hence, the research, particularly the third phase, reflects an amendment and improvement in evaluation over the second phase, while utilizing its findings. The research phases are as follows:

- **Research Phase 1 (RP1)** reviews sentiment techniques of previous work and state of the art and their adoption to ICU clinical texts. This is presented in Chapter 2, and achieves RO1.
- **Research Phase 2 (RP2)** explores the performance of different clinical sentiment techniques on various aspects of clinical notes using the pseudo-labeled datasets. The investigation involves:
 - Creating a pseudo-labeled clinical sentiment dataset as described above in Section 3.2.3 (achieving RO2-A). Validation of pseudo-labeling with humans (RO2-B is in RP3).
 - Investigating the performance on the pseudo-labeled dataset of different approaches identified in the State of Art review (in RP1), including general domain and clinical-specific (addressing RO3). Presented in Chapter 4.
 - Investigating different learning strategies for the machine learning techniques, including frozen models vs. full training (addressing RO3). Presented in Chapter 4.
 - Investigating the performance of sentiment classification methods on different types of notes including physician and nursing notes (addressing RO3). Presented in Chapter 4.
 - Investigating the performance of sentiment classification methods on different segments of notes using the SOAP structure (addressing RO3). SOAP is introduced in Section 4.3 and all is presented in Chapter 4.

This phase overall achieves RO3, and is presented in Section 3.2.3, as well as Chapter 4. It also produces the contribution RC3 and establishes RC4.

- **Research Phase 3 (RP3)** compares the best performing models from RP2 and a recently published clinical lexicon on a manually annotated ground truth MIMIC-based dataset and a separate (Covid) dataset. The investigation involves:
 - Creating manually annotated data for human evaluation by a team of clinicians. This involves sampling from clinical notes, studying annotator agreement and creating ground truth data. This achieves RO4, and is presented in Chapter 5.
 - Validating the pseudo-labeling introduced in Section 3.2.3 on the ground truth data. This achieves RO2-B and is presented in Chapter 5.
 - Investigating the top performing specialized models (from RP2) with a recent specialized lexicon and evaluating all models on clinician's ground truth data. This addresses (RO5) and is presented in Chapter 5.

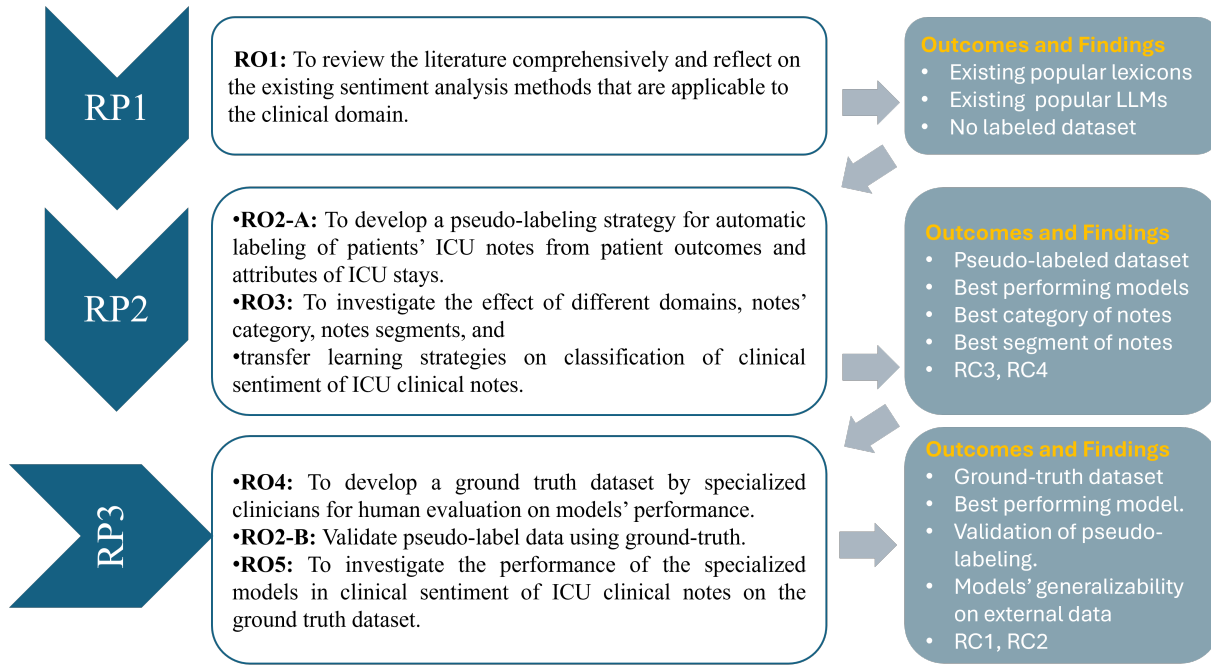


Figure 3.3: Phases of investigation composing research methodology

- Evaluating all models on external data (the Covid dataset) for generalization. This is presented in Chapter 5.

Overall, this phase achieves RO2-B, RO4, and RO5 and produces the contributions RC1, RC2, and RC4.

Figure 3.3, visually illustrates the three phases, and Figure 3.4, presents the overall methodology.

3.3.1 Hypothesis

The key hypothesis for the research is that a clinical-specific sentiment model from the recent generative AI models can provide a more accurate detection of clinical sentiment polarity from ICU clinical notes than clinical-specific sentiment lexicons, the classic non-specific sentiment lexicons, and non-specific LLMs.

3.3.2 Models

Considering the sensitive nature of the research data, particularly the primary dataset MIMIC-III, the owners imposed certain restrictions that permit only open-sourced models, ensuring data and processing are kept within a local processing environment. This limitation reduced the range of models available for the study. For instance, the use of ChatGPT with MIMIC-III is not allowed as indicated on the dataset's website.⁴ The lexicons and models used in the study are: AFINN (Nielsen 2011), VADER (Hutto & Gilbert 2014), Textblob (Loria 2020), BERT (Devlin et al. 2019), T5 (Raffel et al. 2020), Scifive (Phan et al. 2021), Kennedy's lexicon (Kennedy et al. 2023), ClinicalBERT (Alsentzer et al. 2019), PubMedBERT (Gu et al. 2022), and ClinicalT5 (Lehman & Johnson 2023).

⁴<https://physionet.org/news/post/gpt-responsible-use>

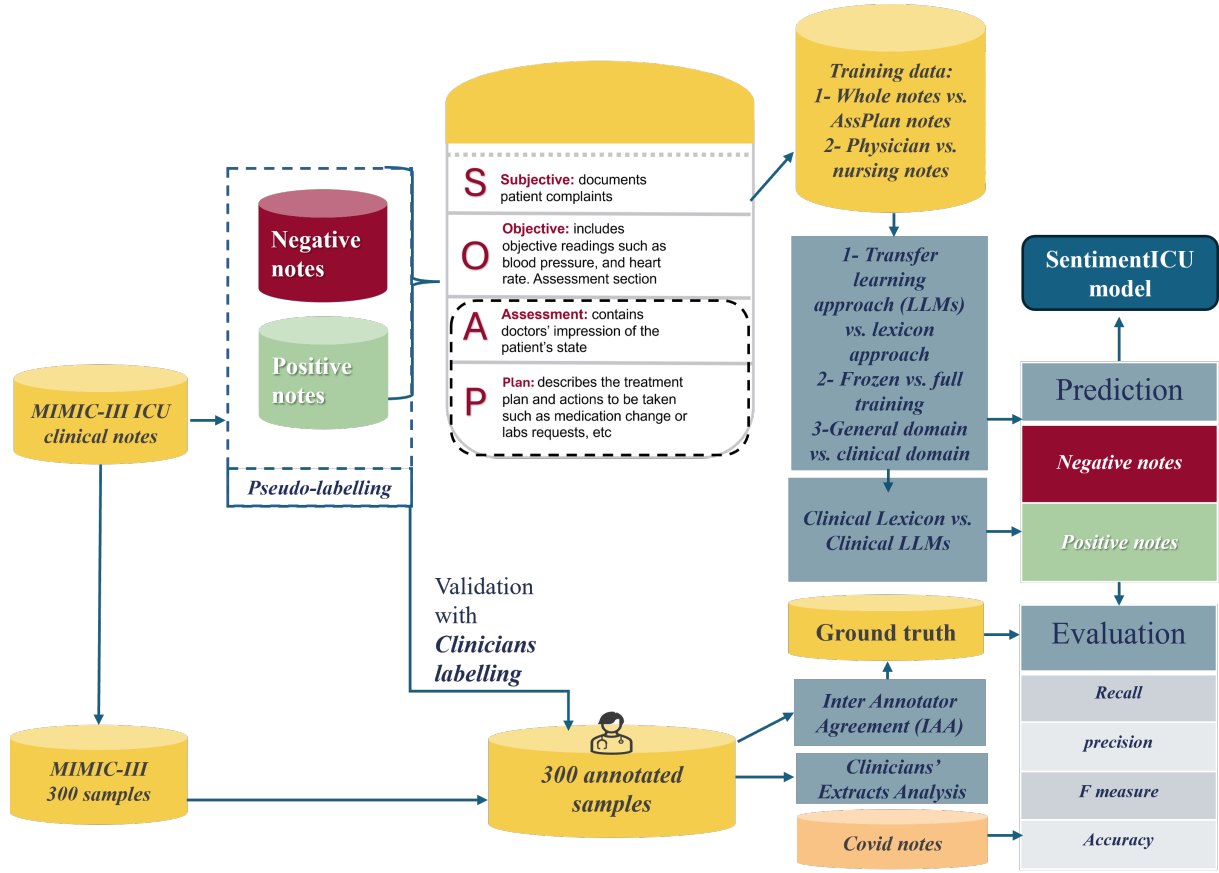


Figure 3.4: Overall method of research

Table 3.5: Definitions of TP, FP, TN and FN in sentiment analysis

	Predicted as Negative	Predicted as Positive
Actual Negative note	TN	FP
Actual Positive note	FN	TP

3.3.3 Evaluation

The metrics used for evaluation are overall accuracy, precision, recall, and F1-score, which are the commonly used performance metrics in sentiment analysis tasks (Ribeiro et al. 2016). The following provides a brief description for each metric. The formulae for these metrics are:

$$Accuracy = (TP + TN) / (TP + FP + FN + TN) \quad (3.1)$$

$$Recall(R) = TP / (TP + FN) \quad (3.2)$$

$$Precision(P) = TP / (TP + FP) \quad (3.3)$$

$$F1 - score = 2 * (R * P) / (R + P) \quad (3.4)$$

Where TP refers to true positive, TN refers to true negative, FP refers to false positive, and FN refers to false negative. Table 3.5 shows the meaning of TP, FP, TN and FN in sentiment analysis classification.

While it is important to not miss any of the classes, in the ICU context, missing negative notes which possibly include negative events could be more harmful in any tasks following up the sentiment analysis in future work i.e, risk factor prediction or decision support visual tools. This indicates the importance of recall of negative samples. On the other hand, false negatives lead to alarm fatigue, a burden in ICU environments (Lewandowska et al. 2020). Therefore, a balance point should be carefully considered for the above measures in the evaluation process when designing future prediction models. It is also recommended in machine learning to study model performance on other data, to observe variance and model learning on unseen samples (Varoquaux & Colliot 2023). Therefore, the evaluation techniques are applied on external data (the Covid dataset) as described in Chapter 6.

3.3.4 Limitations

As discussed, the research methodology provides key investigations of the pseudo-labeling, a number of lexicons and models, notes categories and segments, and learning strategies and sentiment techniques. However, there are limitations which should be considered in future work. The pseudo-labels are based on metadata in relation to patient and ICU stay attributes and not on experts' annotations. Due to the difficulty and the delay in manual annotation (explained in Chapter 5), the evaluation datasets (the MIMIC-III datasets and the Covid dataset) are relatively small. Expanding the sample size, and inclusion of other datasets in future work would enable more robust evaluation and enhance the statistical validity, and the generalizability of the results. Further limitations are discussed in Section 7.5.

3.4 Summary

This chapter has introduced the research datasets, the pseudo-labeling for labeling the datasets, and a description of the datasets. It has provided an overview of the phases of the research. The research method embraces a comparison for studying and identifying the most effective approach for clinical sentiment. Research phase RP1 has focused on reviewing the existing approaches to sentiment analysis and how much work was dedicated for clinical sentiment. Research phase RP2 investigated a various aspects using pseudo-labeling as a basis. Research phase RP3 contains the construction of ground truth and complete the final investigation on it. The chapter presents the main hypothesis, and the models and evaluation metrics. The next chapter presents the initial experimentation using the pseudo-labeling, which is Research phase RP2.

Chapter 4

Initial Investigations of Sentiment Classification Techniques

4.1 Introduction

This chapter provides an exploratory analysis of clinical sentiment classification utilizing the pseudo-labeling scheme (see Section 3.2.3) in Chapter 3. The analysis considers different segments of notes, different categories of notes (physician or nursing notes) and multiple approaches to clinical sentiment analysis including general and domain-specific approaches, and different machine learning strategies. Then, the chapter introduces the typical structure and components of a clinical note in Section 4.3. After that, it introduces the experiment hypotheses in 4.5 and study design in Section 4.6, followed by the data collection and preprocessing in Section 4.7. Models and evaluation metrics are presented in Section 4.8, Results and discussions are provided in Section 4.9. This chapter of the thesis is related to the research question RQ3 and objectives RO2-A and RO3 in Section 1.6.

4.2 Experiment Overview

As described in RO3, it is unknown which segment of notes bears more sentiment than others. This introduces uncertainty of the scope of the text to be used in detecting clinical sentiment, as different notes' segments may introduce variability in the classification. Additionally, machine learning models including large language models can only analyze specific length of text depending on the model (see Section 4.4), and therefore, segmenting notes and feeding models different or shorter segments may not result in the same outcomes. Notes are typically written in SOAP structure. Therefore, this thesis utilizes the structure in segmenting notes and studies clinical sentiment on each segment. Next section describes SOAP and this chapter compares the impact of using two segments of

notes on clinical sentiment classification.

In addition to different input, the analysis compares models of two domains: general domain and clinical domain. General domain models are built from texts from books and other texts with no specific category, whereas clinical domain models are built using clinical related texts. In approaching the research question, a collection of models from the two domains is explored.

Moreover, it has been shown in the literature that large language models may provide excellent performance in some tasks by training only a few of their layers, while keeping the rest frozen. In that case, frozen models may provide cost-effective training compared to full-layers training. To explore the effect of different training strategies on clinical sentiment, both strategies are compared in this chapter.

In summary, different text inputs, different methods, different domains, and different training strategies are used to enrich the comparison and observe their impact on the clinical sentiment identification.

The experiments of this chapter have been conducted during the second phase of investigation RP2 (The first investigation is described in Chapter 2). Therefore, all datasets used in training and testing are automatically labeled using the proposed labeling explained in Section 3.2.3.

4.3 SOAP: Clinical Notes Structure and Components

This thesis deals with clinical notes, most of which are structured into 4 sections using the SOAP (Podder et al. 2023) clinical documentation methodology. S refers to Subjective, O refers to Objective, A refers to Assessment, and P refers to Plan. Figure 4.1 shows the structure. The Subjective section documents patient complaints. The

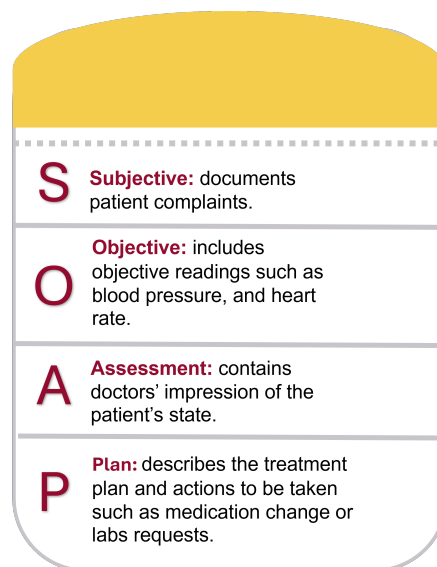


Figure 4.1: Description of SOAP structure of clinical notes

Objective section includes objective readings such as blood pressure, and heart rate. The Assessment section

contains doctors' impression of the patient's state, and the Plan section describes actions to be taken such as medication change or lab requests, etc. Figure 4.1 shows the SOAP structure visually. It is important to note that though this structure is common and recommended, clinical note writing styles vary depending on many factors including the country, the institution, policies and guidelines, and the clinician training. Therefore, notes structure may have variations in practice. For example, some clinicians use narrative style without stating SOAP headings.

Several studies have demonstrated the importance of the assessment and plan sections for clinicians (Brown et al. 2014, Coiera 2000, Rosenbloom et al. 2011). Physicians tend to review assessment and plan sections first when seeking a concise summary of patient information (Brown et al. 2014). Additionally, these segments may introduce differences in clinical sentiment if they were written by different providers since percentage of sentimental words differ by category as noted in Table 3.2 in Section 3.2.1. As mentioned in Section 3.2.1, physician and nursing notes hold great proportions of sentiment words among other types, physician notes holds 97.8% (top proportion) of sentiment keywords (positives and negatives) and nursing notes hold 81.1% of sentiment keywords. Being rich in such keywords, makes these two types of notes suitable candidates for detection of clinical sentiment, and comparing the classification using each type of interest. In experiments in the upcoming sections of this chapter, the analysis includes different SOAP segments of notes provided by physicians and nurses to study the effect on the accuracy of clinical sentiment classification. Therefore, the experiment design provided in Section 4.6 compares the different segments and different categories of notes.

4.4 Length of Clinical Notes

Providers may write notes with different length of text. They do not have restrictions on the length of notes. Unlike clinicians, automated models and even large language models are often restricted to a number of words or tokens of text. For example, BERT can only process 512 tokens, and any further text is truncated. This might have an effect on the results due to information loss missing context and details. Therefore, input text length should be considered for sentiment classification.

Since the components of the notes convey different information as described and are likely to be truncated by models, the analysis of the sentiment may differ based on the provided text and the effectiveness of the tasks may be affected (Su et al. 2021). Next section presents experiment hypothesis in relation to those aspects.

4.5 Hypothesis

The hypothesis of the experiments here are:

- **H1:** Assessment and plan sections holds important information in notes and may provide better accuracy

than depending on whole notes (see Sections 4.6, and 4.7 for further description on notes segmentation).

- **H2:** Learning-based methods may provide better accuracy than traditional lexicons, and full training may optimize the performance.
- **H3:** Domain-specific models may offer better performance over general domain models.
- **H4:** Notes types including physician notes and nursing notes may introduce variability in clinical sentiment.

The next section discusses the experiment setup and design investigating these hypotheses.

4.6 Experiment Design

Data collection and sampling using the pseudo-labeling were explained in Section 3.2.3 in Chapter 3. The experiment here includes segmentation of notes depending on notes structure: (1) Whole notes but subject to truncation by models (2) from Assessment and plan sections onwards. Both are explained in Section 4.7. Two clinical sentiment classification methods, including a set of well-known lexicons and popular machine learning large language models, are investigated and two learning strategies are explored. All model details are described in Section 4.8. The overall experiment design is shown in Figure 4.2.

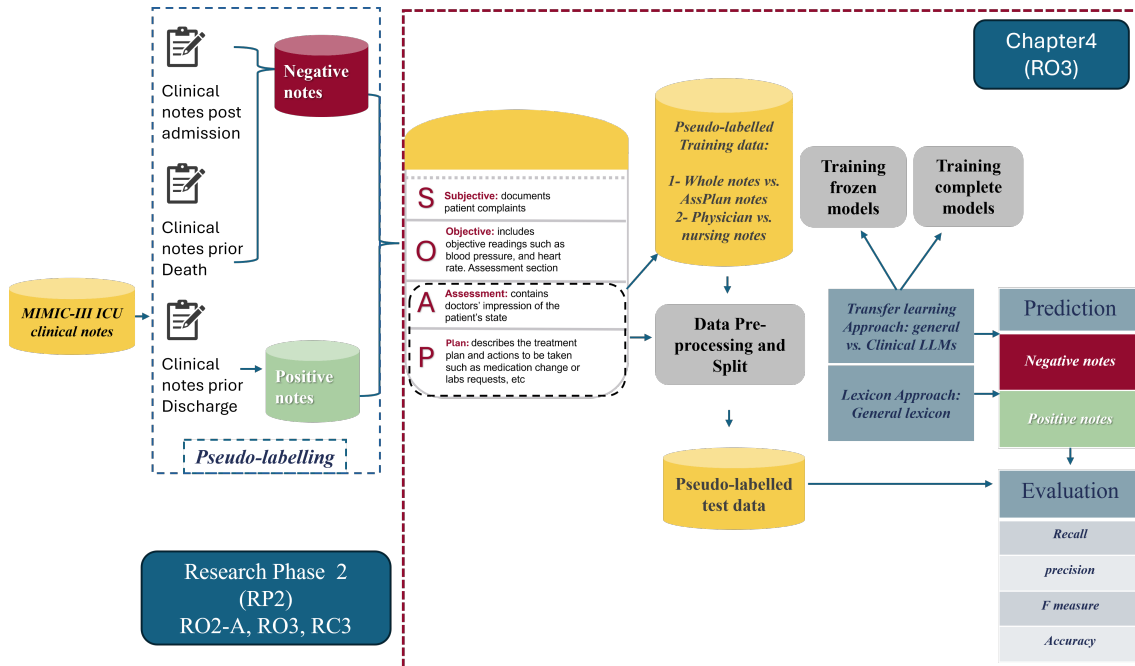


Figure 4.2: Overview of the design of experiments on the pseudo-labeled dataset. Parts related to Chapter 4 are highlighted

This experiment has four main contributions:

- An exploratory comparison between traditional lexicons, transfer learning with BERT models and prompt-based models in clinical sentiment classification for ICU clinical notes.

Table 4.1: Sizes of training, validation and test sets for each of physician and nursing notes

Category	Training	Validation	Testing	Total
physician	8612	2154	1900	12,666
Nursing	8612	2154	1900	12,666

- A comparison between clinical sentiment classification of physician notes versus nursing notes.
- A comparison between clinical sentiment classification of whole notes¹ versus using only the part of notes from SOAP Assessment header onwards (see Section 4.7) for details on segmentation.

The next section presents details about the preparation of the MIMIC ICU notes for experimentation.

4.7 Pre-processing

This section summarizes the steps taken to extract, clean, and organize the MIMIC notes dataset. NLP models typically use specific tokenizers which break text down to pieces or sub-words. All ML models used in this investigation have a maximum count of tokens that can be processed. In particular, BERT models and T5 models can only process up to 512 tokens and will truncate any text exceeding this limit. However, the truncation may differ based on the words specified by the model tokenizer. Tokens are typically words or sub-words, so a text will typically have more tokens than words, depending on the wording and the specific tokenizer. ICU notes vary in length but they can be long, especially for patients with many active health issues. Processing only 512 tokens will potentially result in information loss, perhaps even loss of key information as the more sentiment bearing elements are likely to be in the assessment and plan sections of SOAP notes which come towards the end of the note². Because of the limit enforced by the models' tokenizers, two datasets were used. A dataset of whole notes and an additional dataset with only the part of notes from Assessment header onwards is created using regular expressions (RegEx). The two datasets are used to compare the performance of the sentiment models on the whole note (for lexicon models) or truncated note (for ML models), referred to as "whole/truncated" in what follows, with performance on Assessment onwards part (referred to as "AssPlan"); the hypothesis is that the latter may perform better since many studies shows their significance to clinicians (Brown et al. 2014, Coiera 2000, Rosenbloom et al. 2011). Notes with no valid assessment and plan text are withdrawn as the comparison requires whole/truncated notes and corresponding AssPlan. Also, notes with less than 20 words from both categories (i.e., whole notes and AssPlan) are removed as they contain no meaningful content and some contain headers with no corresponding text.

The constructed dataset is divided into training, validation, and test sets. Test sets are 15% of the total data and validation sets are 20% of training data. Table 4.1 shows sizes of the three datasets. Each subset is balanced

¹As mentioned, notes might be truncated based on allowed length in pre-processing for machine learning models. Further details are explained in Section 4.7.

²Refer to Section 4.3 for information on SOAP

Table 4.2: Sizes of datasets in used in extracting training, validation and testing datasets. "pt" refers to "patients".

Category	Physician	Nursing
Size in all ICU notes	169,642	266,520
Size of notes 3 days after admission	17,021	25,026
Size of survived pt notes 3 days prior discharge	6,534	11,559
Size of died pt Notes 3 days prior death	3,412	4,706
Total of 3 days	26,967	41,291
Size of notes after extracting AssPlan	26,136	20,326
Size of negative notes after sampling	6,333	6,333
Size of positive notes after sampling	6,333	6,333
Total of 3 days after all processing	12,666	12,666

by undersampling negative samples. The undersampling is required since the pseudo-labeling strategy results in many more negative notes (those post admission and prior to death) than positive (those prior to discharge). To balance the dataset, a number equal to the positive set of randomly selected negative samples is collected, i.e., undersampling the negative class. To fairly compare different types of notes (physician notes and nursing notes), the analysis includes data sampled equally from each. In particular, the sizes of the sets of physician whole, physician AssPlan, nursing whole, and nursing AssPlan, is observed and then the sample from each contains a number equal to the smallest, the size of the positive physician notes set (6,333). Table 4.2 describes the datasets involved in pre-processing and their sizes.

The texts themselves were processed as follows. Extra spaces were removed, special characters, such as "[**" and "**]" which were used to surround de-identified elements, were removed, "w/" was expanded to "with", and "pt" or "Pt" were expanded to "patient". For T5 models, the symbol ":" is removed from all headings in the notes as this particular character is used in T5 prompting mechanism. In addition, "sst2 sentence: " is added for T5 as a prefix for all input notes and </s> for end of text and labels. The prefix prompt was pre-defined for sentiment analysis tasks by authors/developers of T5 and was named based on the dataset used for the sentiment training i.e., the Stanford Sentiment Treebank (SST-2) dataset ³. Additional pre-processing steps which are specific for each model are handled by their specific tokenizers.

4.8 Models and Metrics

A collection of different models based on different approaches was selected to investigate the ability of existing models to perform binary clinical sentiment analysis for ICU clinical notes. The analysis includes three publicly available lexicons (AFINN (Nielsen 2011), VADER (Hutto & Gilbert 2014), and TextBlob (Loria 2020)), three BERT models (BERT (Devlin et al. 2019), PubMedBERT (Gu et al. 2022), and ClinicalBERT (Alsentzer et al. 2019)) and three prompt models: T5-base (Raffel et al. 2020), Scifive (Phan et al. 2021), and ClinicalT5 (Lehman

³<https://huggingface.co/datasets/glue/viewer/sst2>.

& Johnson 2023, Goldberger et al. 2000). BERT and T5 Models are available in the HuggingFace open-source library (Wolf et al. 2020), ClinicalT5 is available at the PhysioNet website⁴. The following introduces each briefly, discusses the evaluation metrics, and describes the experiments.

AFINN and VADER are lexicon based approaches to clinical sentiment classification which involve pre-defined dictionaries. AFINN's (Nielsen 2011) dictionary contains 2,477 words, each manually assigned to scores from -5 to +5, where -5 indicates very negative and +5 indicates very positive. It is based on a collection of words and phrases from internet slang, Twitter and some publicly available dictionaries. This approach reads a text, looks up scores for individual words and computes an overall score by summing up individual words' scores. It is a very simple and lightweight approach to sentiment analysis. VADER (Hutto & Gilbert 2014) is similar to AFINN but uses a scale from -4 to +4 and is more comprehensive, incorporating words from multiple dictionaries and rules for words' intensity assessment. On the other hand, TextBlob uses PatternAnalyzer (De Smedt & Daelemans 2012) in its lexical approach. The sentiment scores are polarity scores in the range -1 to 1, where -1 is negative sentiment and +1 is positive sentiment. Lexicon approaches require no pre-training for sentiment analysis. They are built for this task. Therefore, they can be directly evaluated on the test set.

BERT models and prompt models were introduced in Section 2.5.1. Unlike lexicon models, they usually require extra steps such as architectural changes or further training for sentiment analysis if they were not trained for it or were trained for it but on unrelated context. In this study, BERT, T5-base, PubMedBERT, ClinicalBERT, ClinicalT5-base and SciFive-base-Pubmed.PMC are fine-tuned by further training for the sentiment classification task. BERT models require an additional classifier layer for this purpose, while T5 models only require changing the prompt placed as a prefix to the input text. BERT and T5-base are general domain models which were pre-trained on nonclinical text, specifically Wikipedia and Toronto Books Corpus (Zhu et al. 2015). ClinicalBERT and ClinicalT5-base were trained using clinical texts from MIMIC-III. PubMedBERT and SciFive were trained on medical abstracts and full text articles from PubMed and PubMedCentral. It is interesting to observe how different training background impact the analysis. Leaman et al. (2015) note that clinical notes contain more disorder-related vocabularies compared to biomedical texts, while the overall vocabulary size remains relatively similar.

As indicated previously in Section 2.5.2 Chapter 2, there is no individual or standard fine-tuning strategy; tuning pre-trained models can be performed in multiple ways (Sun, Qiu, Xu & Huang 2019). For example, the existing strategies for transfer learning can vary depending on the layers selected for further training. Fine-tuning all layers updates all model parameters while processing new data or the new task. This has proved effective for many tasks but is computationally expensive. On the other hand, fine-tuning the last layer involves freezing the weights of the early layers in the model and only updating the parameters of the last layer. This also has been effective for many NLP tasks especially when applied to small datasets within the same, or similar, domain. In

⁴<https://www.physionet.org/content/clinical-t5/1.0.0/>

Table 4.3: Experiment details for both frozen and non-frozen models

Model	Batch size	Learning rate	Epochs	Optimizer
All Frozen BERT	32	1e-3	3	AdamW
All Frozen prompt	64	1e-2	3	AdamW
ClinicalBERT	32	1e-4	3	AdamW
PubMedBERT	32	1e-5	3	AdamW
BERT	32	1e-5	3	AdamW
ClinicalT5	64	1e-5	5	AdamW
SciFive	64	1e-3	4	AdamW
T5-base	64	1e-3	4	AdamW

pre-training, early layers learn general language representations. Fine-tuning the last layer(s) helps the model transfer this learning and adapt to the desired specific task. Since it updates much fewer parameters, it may offer an alternative quick and efficient fine-tuning but the performance may vary depending on the complexity of task and data. The experiment compared these two strategies; resulting observations are presented in Section 4.9. In relation to experiment settings, early stopping was applied and the following hyper-parameters were explored: learning rate in the values: {1e-2, 1e-3, 1e-4, 1e-5, 3e-4, 3e-5, 5e-5}, batch size in the values: {16, 32, 64, 128}, and epochs in the values: {2, 3, 4, 5}. The best performing hyper-parameters are presented in Table 4.3.

For frozen BERT models, the models were frozen, applying fine-tuning for only the last layer. For frozen prompt models, only the last decoder and encoder block were unfrozen. The AdamW optimizer (Loshchilov & Hutter 2019) was used for all models. All models were trained using two a100 gpus. Training required 1 minute and 14 seconds for every epoch and no more than 6 minutes for training the largest models (i.e., T5-base models). In terms of parameters, BERT models contain 110 million parameters + 128x2 for classifiers, while T5-base has 220 million parameters.

The metrics used for evaluation are overall accuracy, precision, recall and F1-score, which are the commonly used performance metrics for sentiment analysis tasks (Ribeiro et al. 2016) as discussed in Section 3.3.3.

In summary, the experiments used pseudo-labels of sentiment as proposed in Section 3.2.3 and measure overall accuracy, precision, recall, and F1-score for the two classes (positive and negative), and considering the following: (1) physician notes versus using nursing notes. (2) different learning strategy where feasible: fine-tuning all layers vs. top layer for BERT models and all layers vs. top layers (last decoder and encoder block) for T5 models. (3) different section of notes: whole/truncated vs. AssPlan. The next section (Section 4.9) presents the results and discusses the observations.

4.9 Results and Discussion

This section provides the results, and reflects on different approaches and models' performance. Considering lexicon approaches first, as shown in Table 4.4, the performance of general domain lexicons is poor. Their overall

Table 4.4: Evaluation of three lexicon models: AFINN, Textblob, and VADER using overall accuracy, and recall, precision, f1-score for positive and negative classes and applied to whole notes with default truncation (whole/truncated) vs. assessment and plans sections onwards (AssPlan).

Model	Category	Notes	Acc	Neg_R	Neg_P	Neg_F1	Pos_R	Pos_P	Pos_F1
AFINN	Physician	Whole/truncated	0.5163	0.9200	0.5090	0.6554	0.1126	0.5847	0.1889
AFINN	Physician	AssPlan	0.5421	0.8242	0.5269	0.6429	0.2600	0.5966	0.3622
Textblob	Physician	Whole/truncated	0.4821	0.4874	0.4823	0.4848	0.4768	0.4819	0.4794
Textblob	Physician	AssPlan	0.5189	0.3358	0.5299	0.4111	0.7021	0.5139	0.5934
VADER	Physician	Whole/truncated	0.5500	0.7011	0.5384	0.6091	0.3989	0.5716	0.4699
VADER	Physician	AssPlan	0.5563	0.7347	0.5415	0.6235	0.3779	0.5876	0.4600
AFINN	Nursing	Whole/truncated	0.5195	0.7905	0.5126	0.6219	0.2484	0.5425	0.3408
AFINN	Nursing	AssPlan	0.5095	0.6326	0.5076	0.5633	0.3863	0.5126	0.4406
Textblob	Nursing	Whole/truncated	0.5337	0.2200	0.5904	0.3206	0.8474	0.5207	0.6450
Textblob	Nursing	AssPlan	0.5279	0.2168	0.5738	0.3147	0.8389	0.5172	0.6399
VADER	Nursing	Whole/truncated	0.5437	0.6453	0.5363	0.5858	0.4421	0.5548	0.4921
VADER	Nursing	AssPlan	0.5279	0.5242	0.5281	0.5261	0.5316	0.5277	0.5296

accuracy ranges from 0.49 to 0.55. It can be observed from Table 4.4 that negative recall of AFINN lexicon ranges from 0.63 to 0.92 and was overall greater for whole/truncated and negative samples while positive recall for TextBlob ranges from 0.47 to 0.84, and is generally greater for AssPlan (i.e., from SOAP assessment point onwards) and positive samples. The recall values of both lexicons are noticeably different for the two classes. The recall for Textblob is noticeably high for nursing notes for positive class and poor for negative class in using both sections of notes. The opposite is the case for AFINN in physician notes and both sections of notes. VADER lexicon performs similarly to AFINN's in terms of negative recall for whole/truncated samples, i.e., higher recall relative to TextBlob, but with lower range (0.52 to 0.73), and that is consistent for physician and nursing. Two factors may contribute towards the results:

- According to [Nielsen \(2011\)](#), AFINN lexicon is biased towards negative words (0.65 of its word list is negative) and therefore may assign more negative sentiment scores than positive. It is not clear if VADER has the same bias towards negative words but both VADER and AFINN's lexicons include the ANEW (Affective Norms for English Words) wordlist ([Nielsen 2011](#)).
- For all lexicon approaches, including AFINN, VADER and TextBlob, the produced sentiment scores are transformed to binary labels by considering values greater than 0 as positive and less than 0 as negative with no consideration of neutral class. This could bias the results.

The results for lexicon approaches align with [Weissman et al. \(2019\)](#) and [Holderness et al. \(2019\)](#). [Weissman et al. \(2019\)](#) indicate the failure of general domain lexicon-based sentiment approaches on clinical domain text. [Holderness et al. \(2019\)](#) show that the PatternAnalyzer lexicon ([De Smedt & Daelemans 2012](#)) used by Textblob does not capture the sentiment of many sentiment-significant words in the clinical context.

Turning next to models based on BERT, Table 4.5 shows the results for the selected BERT models. Fine-tuning all layers in BERT provides the top performance in terms of accuracy, recall, precision and F1-score. The highest

Table 4.5: Evaluation of three BERT models: BERT, clinicalBERT, and PubMedBERT using overall accuracy, and recall, precision, F1-score for positive and negative classes, and applied to whole notes with default truncation (whole/truncated) vs. assessment and plans sections onwards (AssPlan). Top half is for frozen models and lower half is for unfrozen models.

Model	Category	Notes	Acc	Neg_R	Neg_P	Neg_F1	Pos_R	Pos_P	Pos_F1
ClinicalBERT	Physician	AssPlan	0.7616	0.7968	0.7443	0.7697	0.7263	0.7814	0.7529
ClinicalBERT	Physician	Whole/truncated	0.6968	0.4947	0.8304	0.6201	0.8989	0.6402	0.7478
ClinicalBERT	Nursing	AssPlan	0.7316	0.6400	0.7835	0.7045	0.8232	0.6957	0.7541
ClinicalBERT	Nursing	Whole/truncated	0.7332	0.6695	0.7672	0.7150	0.7968	0.7068	0.7491
PubMedBERT	Physician	AssPlan	0.7189	0.7526	0.7051	0.7281	0.6853	0.7348	0.7092
PubMedBERT	Physician	Whole/truncated	0.7179	0.7053	0.7235	0.7143	0.7305	0.7125	0.7214
PubMedBERT	Nursing	AssPlan	0.6716	0.6105	0.6954	0.6502	0.7326	0.6529	0.6905
PubMedBERT	Nursing	Whole/truncated	0.6705	0.6253	0.6875	0.6549	0.7158	0.6564	0.6848
BERT	Physician	AssPlan	0.7100	0.7968	0.6789	0.7332	0.6232	0.7541	0.6824
BERT	Physician	Whole/truncated	0.7074	0.6789	0.7199	0.6988	0.7358	0.6962	0.7155
BERT	Nursing	AssPlan	0.6716	0.6316	0.6865	0.6579	0.7116	0.6589	0.6842
BERT	Nursing	Whole/truncated	0.6837	0.6779	0.6858	0.6818	0.6895	0.6816	0.6855
ClinicalBERT	Physician	AssPlan	0.9311	0.9200	0.9408	0.9303	0.9421	0.9217	0.9318
ClinicalBERT	Physician	Whole/truncated	0.9253	0.8968	0.9509	0.9231	0.9537	0.9024	0.9273
ClinicalBERT	Nursing	AssPlan	0.8079	0.7358	0.8598	0.7930	0.8800	0.7691	0.8208
ClinicalBERT	Nursing	Whole/truncated	0.8463	0.7874	0.8926	0.8367	0.9053	0.8098	0.8549
PubMedBERT	Physician	AssPlan	0.8884	0.8495	0.9212	0.8839	0.9274	0.8604	0.8926
PubMedBERT	Physician	Whole/truncated	0.8505	0.8611	0.8433	0.8521	0.8400	0.8581	0.8489
PubMedBERT	Nursing	AssPlan	0.8226	0.7821	0.8511	0.8151	0.8632	0.7984	0.8295
PubMedBERT	Nursing	Whole/truncated	0.8321	0.7937	0.8597	0.8254	0.8705	0.8084	0.8383
BERT	Physician	AssPlan	0.8868	0.8747	0.8964	0.8855	0.8989	0.8777	0.8882
BERT	Physician	Whole/truncated	0.8347	0.8232	0.8427	0.8328	0.8463	0.8272	0.8366
BERT	Nursing	AssPlan	0.8174	0.8211	0.8150	0.8180	0.8137	0.8197	0.8167
BERT	Nursing	Whole/truncated	0.8268	0.8147	0.8350	0.8247	0.8389	0.8191	0.8289

values were scored by ClinicalBERT for AssPlan segment of physician notes. In particular, accuracy is 0.93, recall is 0.92, precision is 0.94, and F1-score is 0.93 for negative class, while positive recall is 0.94, positive precision is 0.92, and positive F1-score is 0.93. All three BERT models have excellent performance (> 0.80) in most cases for both physician and nursing notes. For example, for whole/truncated nursing notes, accuracy values are 0.84, 0.83, 0.82 for BERT, ClinicalBERT, and PubMedBERT respectively. The argument here is that fine-tuning all layers updates all parameters and as such can perform well with less reliance on previous knowledge and more on current training, but the previous knowledge keeps ClinicalBERT's performance consistently the best among the three for physician AssPlan. Conversely, fine-tuning the top layer shows a variance between ClinicalBERT versus PubMedBERT, and BERT e.g. accuracy values are 0.76, 0.71, 0.71 respectively for AssPlan segment of physician notes, and 0.73, 0.67, 0.68 respectively for whole/truncated nursing notes

A similar pattern is seen for prompt models in both fine tuning strategies as observed in Table 4.6. Fine-tuning all layers in prompt models provides the top performance in terms of accuracy, recall, precision, and F1-score, but the highest values were scored by Scifive for physician AssPlan. Although the results for full fine-tuning are

Table 4.6: Evaluation for three prompt models: T5-base, ClinicalT5, and SciFive using overall accuracy, and recall, precision, F1-score for positive and negative classes, and applied to whole notes with default truncation (whole/truncated) vs. assessment and plans sections onwards (AssPlan). Top half is frozen models and lower half is unfrozen models.

Model	Category	Notes	Acc	Neg_R	Neg_P	Neg_F1	Pos_R	Pos_P	Pos_F1
ClinicalT5	Physician	AssPlan	0.8489	0.8800	0.8285	0.8535	0.8179	0.8721	0.8441
ClinicalT5	Physician	Whole/truncated	0.7637	0.8589	0.7215	0.7842	0.6684	0.8257	0.7388
ClinicalT5	Nursing	AssPlan	0.7516	0.7537	0.7505	0.7521	0.7495	0.7526	0.7511
ClinicalT5	Nursing	Whole/truncated	0.7795	0.8095	0.7637	0.7859	0.7495	0.7973	0.7727
SciFive	Physician	AssPlan	0.7858	0.7832	0.7873	0.7852	0.7884	0.7843	0.7864
SciFive	Physician	Whole/truncated	0.7711	0.8389	0.7386	0.7856	0.7032	0.8136	0.7544
SciFive	Nursing	AssPlan	0.7326	0.7147	0.7413	0.7278	0.7505	0.7246	0.7373
SciFive	Nursing	Whole/truncated	0.7637	0.7621	0.7645	0.7633	0.7653	0.7629	0.7641
T5-base	Physician	AssPlan	0.7742	0.7653	0.7792	0.7722	0.7832	0.7694	0.7762
T5-base	Physician	Whole/truncated	0.7416	0.7905	0.7200	0.7536	0.6926	0.7678	0.7283
T5-base	Nursing	AssPlan	0.7421	0.7684	0.7300	0.7487	0.7158	0.7556	0.7351
T5-base	Nursing	Whole/truncated	0.6884	0.6168	0.7199	0.6644	0.7600	0.6648	0.7092
ClinicalT5	Physician	AssPlan	0.9079	0.9368	0.8856	0.9105	0.8789	0.9330	0.9051
ClinicalT5	Physician	Whole/truncated	0.8847	0.9400	0.8464	0.8908	0.8295	0.9325	0.8780
ClinicalT5	Nursing	AssPlan	0.8484	0.8968	0.8177	0.8554	0.8000	0.8858	0.8407
ClinicalT5	Nursing	Whole/truncated	0.8700	0.9095	0.8429	0.8749	0.8305	0.9017	0.8647
SciFive	Physician	AssPlan	0.9274	0.9474	0.9109	0.9288	0.9074	0.9452	0.9259
SciFive	Physician	Whole/truncated	0.9053	0.9463	0.8745	0.9090	0.8642	0.9415	0.9012
SciFive	Nursing	AssPlan	0.8495	0.8695	0.8360	0.8524	0.8295	0.8640	0.8464
SciFive	Nursing	Whole/truncated	0.8647	0.8916	0.8462	0.8683	0.8379	0.8854	0.8610
T5-base	Physician	AssPlan	0.9084	0.9168	0.9017	0.9092	0.9000	0.9154	0.9076
T5-base	Physician	Whole/truncated	0.8947	0.8968	0.8931	0.8950	0.8926	0.8964	0.8945
T5-base	Nursing	AssPlan	0.8305	0.8705	0.8060	0.8370	0.7905	0.8593	0.8235
T5-base	Nursing	Whole/truncated	0.8584	0.9021	0.8296	0.8643	0.8147	0.8927	0.8520

close to each other, ClinicalT5 was a little slower (i.e., required an additional epoch) than Scifive and achieved a slightly lower accuracy for physician notes against what is expected.

With respect to frozen models, fine-tuning ClinicalT5 for AssPlan segment of physician notes provides the top performance in terms of accuracy, recall, precision, and F1-score as observed in Table 4.6. The variance pattern was similar to what has been observed for frozen BERT models. For instance ClinicalT5, SciFive, and T5-Base score 0.77, 0.67, 0.68 respectively for accuracy on whole/truncated nursing notes.

Overall, frozen models which were trained using clinical data (ClinicalT5 and ClinicalBERT)⁵ have better performance than models trained on medical articles (SciFive and PubMedBERT). For example, the accuracy values for ClinicalBERT range from 0.69 to 0.76 while they range from 0.67 to 0.71 for PubMedBERT. The same is observed for T5 models. As shown in Table 4.6, accuracy, and F1-score for both classes for ClinicalT5 are consistently

⁵ClinicalBERT and ClinicalT5 are trained on MIMIC dataset. Although the task introduced here is different from the tasks that these models were initially trained on, e.g. masked language modeling (MLM) or next sentence prediction (NSP), the extent of the effect of pre-training on these results is unknown.

Table 4.7: Summary average for each type of model, category, notes segment across all metrics

Model	Category	Notes	Acc	Neg_R	Neg_P	Neg_F1	Pos_R	Pos_P	Pos_F1
Frozen-T5	Physician	AssPlan	0.8030	0.8095	0.7983	0.8036	0.7965	0.8086	0.8022
	Physician	Whole/truncated	0.7588	0.8294	0.7267	0.7745	0.6881	0.8024	0.7405
	Nursing	AssPlan	0.7421	0.7456	0.7406	0.7429	0.7386	0.7443	0.7412
	Nursing	Whole/truncated	0.7439	0.7295	0.7494	0.7379	0.7583	0.7417	0.7487
Frozen-BERT	Physician	AssPlan	0.7302	0.7821	0.7094	0.7437	0.6783	0.7568	0.7148
	Physician	Whole/truncated	0.7074	0.6263	0.7579	0.6777	0.7884	0.6830	0.7282
	Nursing	AssPlan	0.6916	0.6274	0.7218	0.6709	0.7558	0.6692	0.7096
	Nursing	Whole/truncated	0.6958	0.6576	0.7135	0.6839	0.7340	0.6816	0.7065
Unfrozen-T5	Physician	AssPlan	0.9146	0.9337	0.8994	0.9162	0.8954	0.9312	0.9129
	Physician	Whole/truncated	0.8949	0.9277	0.8713	0.8983	0.8621	0.9235	0.8912
	Nursing	AssPlan	0.8428	0.8789	0.8199	0.8483	0.8067	0.8697	0.8369
	Nursing	Whole/truncated	0.8644	0.9011	0.8396	0.8692	0.8277	0.8933	0.8592
Unfrozen-BERT	Physician	AssPlan	0.9021	0.8814	0.9195	0.8999	0.9228	0.8866	0.9042
	Physician	Whole/truncated	0.8702	0.8604	0.8790	0.8693	0.8800	0.8626	0.8709
	Nursing	AssPlan	0.8160	0.7797	0.8420	0.8087	0.8523	0.7957	0.8223
	Nursing	Whole/truncated	0.8351	0.7986	0.8624	0.8289	0.8716	0.8124	0.8407
Lexicons	Physician	Whole/truncated	0.5161	0.7028	0.5099	0.5831	0.3294	0.5461	0.3794
	Physician	AssPlan	0.5391	0.6316	0.5328	0.5592	0.4467	0.5660	0.4719
	Nursing	Whole/truncated	0.5323	0.5519	0.5464	0.5094	0.5126	0.5393	0.4926
	Nursing	AssPlan	0.5218	0.4579	0.5365	0.4680	0.5856	0.5192	0.5367

greater than SciFive’s values for both segments of notes, (0.75-0.84) vs. (0.73-0.78).

Clinical sentiment classification of AssPlan segment of notes (that is from SOAP assessment point onwards), in many cases in the results (0.66), have close or comparable performance to whole/truncated notes and achieve slightly increased scores in all cases for physician notes. For instance, as observed in Table 4.6, accuracy achieved by ClinicalT5, SciFive and T5-base is higher for AssPlan segments than whole/truncated notes for physician notes for both tuning strategies (0.77 - 0.92) vs. (0.74 - 0.90). For nursing notes, AssPlan results were in most cases very close but not necessarily higher than whole/truncated. This might be due to the similarity in the sizes of both. This suggests that the AssPlan segment of the note can be used as an alternative to the whole/truncated note and provide reliable results for this two-classes sentiment task.

Table 4.7 provides a summary for all models using average accuracy, recall, precision, and F1-score while Tables 4.4, 4.5, and 4.6 include all experiments of lexicons, BERT models and prompt models.

4.10 Summary

The final conclusions of this analysis are:

1. Clinical sentiment classification of AssPlan segment of notes (that is from SOAP assessment point onwards), in many cases in the results (66%), have close or comparable performance to whole/truncated notes and

achieve slightly increased scores in all cases for physician notes. This supports hypothesis H1 (presented in Section 4.5).

2. BERT models and prompt models are capable of providing better clinical sentiment than the baseline lexicons if fine-tuned to this type of data, supporting hypothesis H2 (presented in Section 4.5).
3. Transfer learning, by only updating task specific layers on AssPlan segments of physician notes, offers a cost effective option (involving much fewer parameters) for acceptable accuracy (up to 0.76) using the smaller clinical pre-trained models like ClinicalBERT and excellent accuracy (up to 0.84) using larger clinical pre-trained models like ClinicalT5. Full fine-tuning for all models demonstrates the best capability in improving the accuracy of clinical sentiment on ICU text (up to 0.93) supporting hypothesis H2. BERT models require less computational resources for full training than prompt models while prompt models require no structural changes to models' architecture and so is more convenient in this matter than BERT models, which require those changes.
4. Existing general domain lexicon models are not reliable when applied to ICU clinical notes, and require adaption to ICU language, confirming hypothesis H3 (presented in Section 4.5).
5. Physician notes results in more accurate classification than nursing notes, supporting hypothesis H4.

As mentioned earlier, truncation creates a limitation in handling notes' texts. Notes may be truncated by models due to their pre-determined limits, and this could lead to context or information loss. Consideration should be taken to this in future work as discussed in Section 7.5.

The experiments in this chapter have been based on a dataset where sentiment labels are proxies for sentiments used in the interim for comparing lexicons and models on the experiment's settings. The ML models were able to recognize the difference in the language of notes belonging to pseudo-labeled positive and negative classes but those labels have not been compared with human labels in this phase, as the manual annotation project has been ongoing. To address this limitation, the next chapter (representing the third phase of investigation RP3), introduces a clinicians' labeled dataset, from which a gold-standard test set is created for evaluation, and investigates sentiment-related linguistic texts of notes.

Chapter 5

Manual Annotation and Construction of Ground Truth

5.1 Introduction

This chapter describes the beginning of the third and final phase of the thesis investigations. The third investigation is when the manual annotation project were completed and ground truth data labeled by clinicians was constructed. The key outcomes of this phase involve: (1) Construction of ground truth (presented in this chapter and linked to RO4). (3) Using the ground truth to validate the pseudo-labeling (presented in this chapter and linked to RO2-B). (2) Using the ground truth to conduct the final evaluation and find the best approach and model for clinical sentiment (presented in next chapter and linked to RO5).

To obtain the ground truth (RO4) and evaluate pseudo-labels against its label (RO2-B), multiple steps were taken: (1) Defining the annotation task. This includes specifying the tasks required. The task involved labeling notes with the most appropriate label for clinical sentiment, and extracting the text pieces that influence the choice of label. (2) Preparing data and corresponding documentation. (3) Recruiting clinical experts for annotation. (4) Educating annotators for the task. (5) Collecting annotation results. (6) Studying inter-annotator agreement. (7) Studying annotators extracts. (8) Forming ground truth dataset.

Section 5.2 presents a description for the manual annotation process. Section 5.3.1 presents inter-annotator agreement and the creation of ground truth dataset. Section 5.5 evaluates the suggested labeling scheme presented in Section 3.2.3 and used for experiments presented in thesis. Section 5.4 presents analysis on extracts associated with manual labeling. Figure 5.1 show the Research Phase (RP3), parts related to Chapter 5 are highlighted.

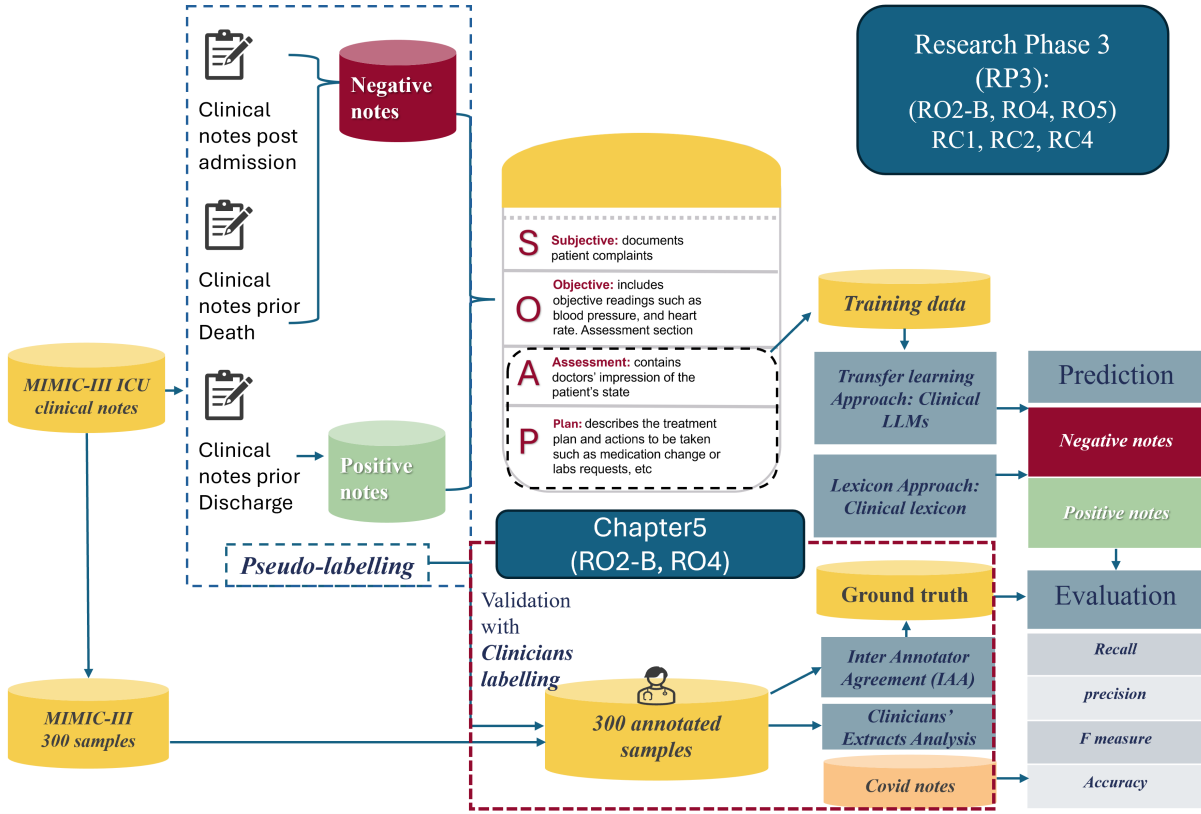


Figure 5.1: Research Phase 3, parts related to Chapter 5 are highlighted.

5.2 Manual Annotation Project

This section describes all the materials related to the manual annotation process. The subsequent sections include the motivation for the task in 5.2.1, the data collection and sampling in 5.2.2, annotators' recruitment in Section 5.2.3, annotation tool in 5.2.4, details of the task in Section 5.2.5, and details of annotation file in Section 5.2.6.

5.2.1 Motivation for manual labeling

As previously discussed, manual annotation cost of clinical text is causing a delay in the progress of NLP supervised learning. The cost of manual annotation includes recruiting annotators with the required expertise, introducing and training those experts for the task, selecting an appropriate tool for the task, preparing suitable data for the task, following up with annotators and reviewing their work iteratively and after completing the task. Based on the mentioned requirements, there has been a lag in the process for a long time. This research aims to provide a dataset for the field of NLP by which clinical sentiment approaches can be investigated and improved. The data is being published by Physionet for public use.

Experts' annotation is important for evaluating the classification of models on the experiments of the thesis. It offers evidence-based evaluation and validates the reliability and the quality of models' outcomes. Despite this

valuable role, many studies in NLP have suffered from the absence of human evaluation and the reproducibility of the evaluation [Schmidtova et al. \(2024\)](#). According to the survey in the study, 38.7% of the research papers in INLG 2023, and 36.5% of research papers in ACL 2023 have no human evaluation. Among a total of 102 publications reviewed in the survey, only 59 (56.73%) of the publications used human evaluation. Most of the publications, 98 (94.23%), used automatic evaluation metrics. The same finding was observed by [van der Lee et al. \(2021\)](#), as 95% of studies were found to have automatic metrics in both conferences, and 53 (50.96%) of the publications used both methods. The dominance of automatic metrics is due to the difficulty discussed previously in obtaining access to data and significant effort required for recruiting participants for preparing human evaluation.

As mentioned in Chapter 3, preparation of the manually annotated dataset for this research has spanned more than a year. Hence, the second research phase (RP2) completely relied on pseudo-labeling, while the manual annotation was being prepared. The third phase combines both pseudo-labeling labeling for training and human labeling for evaluation.

Using the proposed labeling and clinicians' evaluation leads to a joint benefit from training the models on large data, being exposed to great number of samples which enrich models learning and helps in recognizing patterns more ([Lu, Ehwerhemuepha & Rakovski 2022](#)), while validating on concrete human judgments.

Additionally, since the pseudo-labeling is based on patients' ICU attributes, and the manually annotated dataset is being released in Physionet, the analysis can be reproduced and progressed, adding more value to the public and open use.

The development of the dataset includes many challenges. Recruiting clinicians was difficult due to their hospital duties and workload. Each recruit had to complete data protection training before being given access to clinical notes from the MIMIC dataset. Specifically, recruitment for this task took more than a year (15 months). The average time consumed by each annotator was around 8 weeks for data access requirements and for the annotation task. More details are provided in Section 5.2.3.

5.2.2 Data Collection

The dataset presented to annotators is MIMIC-based, contained 300 notes, and consists of 52% (156/300) of physician notes and 48% (144/300) of nursing notes. The initial sampling aimed to balance the distribution but some samples were removed manually due to repetitions in contents, found mostly in some of the nursing notes. Statistics relating to the manual annotation dataset are provided in Table 5.1 The total number of samples were chosen to maximize the chance of getting a reasonable number of samples in the final ground truth considering the impact of the difficulty of the task for busy clinicians, and the potential variability of opinions. During sampling, data was examined to ensure each of the pseudo-labeled groups contains at least 20 samples.

Table 5.1: Statistics for the manually annotated dataset. ICYSTAYS refer to ICU stays. LOS refers to length of stay.

Description of groups or notes	Counts
All notes	300
Count of patients	248
Count of ICUSTAYS	251
Count of patients with $\text{LOS} \leq 6$	107
Counts of Physicians' notes	156
Counts of Nursing notes	144
Count of pseudo-labeled notes	145
Notes in days since admission ≤ 3 days	48
Notes in days prior death ≤ 3 days	23
Notes in days prior discharge ≤ 3 days	74
Notes in days in between admission and discharge	155
Count of unique discharge locations	11
The most frequent discharge locations	Home

5.2.3 Annotators Recruitment

Five annotators were recruited for the task of manual annotation for the dataset composed of ICU clinical notes. The annotation team consists of an ICU consultant, a respiratory consultant, and three ICU registrars. Individual meetings were conducted with each physician prior to the task to explain the requirements for obtaining access to the data, which included the need to complete online data protection training, and performing the manual annotation itself. Information on access requirements to data can be found on the data host's official website *PhysioNet*¹. Two individual meetings were held with each physician prior to the task. The first meeting was to explain the requirements for obtaining access to the data, which involve completing online data protection training. The second was to explain the manual annotation task. Information regarding requirements for access to the data can be found at the MIMIC dataset host official website *PhysioNet*. When an annotator completed the required training, a document including a brief background of the research and the task description was shared. Next section describes the annotation tool and a brief description for the task required in the annotation.

5.2.4 Annotation Tool

Annotation samples were shared with physicians in protected excel spreadsheets. The choice of spreadsheets was due to its simplicity and ease of its usability requiring no training for individuals with nontechnical expertise. During the search of an appropriate tool for the task, **Prodigy**² was a suitable candidate tool. Getting familiar with it and designing the interface took around two months. The design involves building the interface and adjusting it for the task by adding required components of notes' texts and class labels, adding instructions for annotators, and setting secure access through passwords. To verify the usability of the interface, two meetings

¹<https://physionet.org/content/mimiciii/1.4/>

²<https://prodi.gy/>

with two physicians were held for a consultation on how usable and easy they found the interface. Despite the suitability of Prodigy for the task, it became clear later that using it may introduce difficulty for annotators due to the following reasons:

- Data are highly sensitive and are required to be kept in a safe storage under protection of the Trinity School of Computer Science and Statistics and ADAPT³ high-performing server.
- Additional individual training will be required for annotators to familiarize them with the interface.
- Keeping the data on a Trinity College server will require approval from administrators to allow for external access by annotation team. This will introduce additional delay to the annotation due to paperwork and the process queue and time.
- Handling the annotation through a remote connection may be inconvenient due to e.g., maintenance issues causing server shutdowns or unplanned network disconnections due to e.g., weather problems.

Due to the above and to avoid potential delay or difficulty, it seems wise and convenient to choose spreadsheets for the annotation process to proceed and speed up the process as possible. This is especially wise considering that the target team is physicians, having very busy schedules and work burden and considering that the data access for MIMIC-III in Physionet require a training course which contributed in its turn to an extra difficulty and delay. The samples were saved in an excel protected by a password and shred individually with each annotator when they complete the required training for data access. Passwords were shared in the meetings conducted with the annotator prior to the annotation task.

Two files were provided to the annotation team:

1. Cover letter in Appendix A.
2. The spreadsheet of samples. Details in Section 5.2.6.

5.2.5 Task Description

Clinicians were asked to read a note and choose a label from a list, namely “Overall label”, that reflects the patient’s status. e.g., **The clinician believes that this patient is already improving or will get better**, when a note indicates that the patient is ok, has made some progress or expected to get better, **The clinician believes that this patient is deteriorating**, when a note indicates that the patient’s condition has deteriorated, **The clinician believes that this patient is neither improving nor deteriorating** when a note indicates no specific direction to improvement or deterioration, and **There is no indication of the note writer’s opinion** when no opinion can be recognized from the note.

In addition to the labels, clinicians were asked to provide extracts of texts influencing their choice of label. This requirement was not mandatory for all notes, but was asked to be provided as possible, and where there are

³Research Ireland Centre for AI Driven Digital Content Technology. See <https://www.adaptcentre.ie/>

clear indicators in the text that lead to the clinician’s decision of label. Annotators were informed that extracts should be copies from notes text (i.e., a sentence, or a phrase of a set of words but not a paragraph) reflecting a specific condition associated with improvement, deterioration, no-change or none. Clinicians were asked to separate multiple extracts and start each new extract with double dashes (-). The following describe possible examples of the extracts:

1. Extracts associated with **improvement**. This should contain copied phrases that reflect any improvement in a patient’s condition, e.g., “extubated [**4-13**], and improving on antibiotic therapy.”, “normal blood pressure, not on pressors.”
2. Extracts associated with **no change in status**. This should contain copied phrases that reflect stability in specific aspects of a patient’s condition, e.g. “Remains intubated mechanically ventilated.”
3. Extracts associated with **deterioration**. This should contain copied phrases that reflect any deteriorating aspects of a patient’s condition, e.g., “no viable therapeutic options that seem feasible for living outside of intensive care”

Annotators were also informed with the following instructions:

1. The annotation task should be completed by each physician independently.
2. The annotation is meant to target and capture the impression that is expressed by the clinicians who wrote the note (i.e., their opinion) about the patient. It is not meant to test or evaluate the clinical decision capabilities of the annotators.
3. MIMIC-III data was changed by owners in compliance with HIPPA. Therefore, Patients’ names, doctors’ names, telephone numbers, addresses, dates are all replaced by tokens surrounded by [** and **], e.g., [**Last Name**]. In addition, dates are shifted to be between the years 2100 and 2200 using random amounts. However, intervals between events are preserved as they were. Ages of some patients (over 89) is shifted to be above than 300 in compliance with HIPAA.

5.2.6 Annotation File

The provided spreadsheet includes 5 columns as follows:

- **Note:** This column contains the MIMIC-III clinical note.
- **Note ID:** This is the MIMIC-III ROW_ID from the NOTEEVENTS table. It allows researchers to link to additional information about the patient and their ICU stay.
- **Category:** the type of the note: physician or nursing, from NOTEEVENTS MIMIC table.
- **Overall label for note:** Asking if the clinician who wrote this note believes the patient’s health is (from a drop down list):

Table 5.2: Structure of Annotation File

Note_ID	Note	Category	Overall label for note	Extracts supporting the selection
---------	------	----------	------------------------	-----------------------------------

- Is already improving or will get better?
- Is deteriorating?
- Is neither improving nor deteriorating?
- OR that there is no indication of the note writer’s opinion.

- **Extracts supporting the selection of label:** where annotators are supposed to type in the text influence their choice. Annotators are asked here to start a new extract with double dashes (--).

The file appears as in Table 5.2 and, and Figure A.1 in Appendix A.

Based on the above, annotators did not have access to any information about the patient other than the note, including not knowing that the patient dies or were discharged.

5.2.7 Annotation Challenges and Takeaways

The annotation task has been very challenging for the following reasons:

- The annotators are busy unavailable clinicians who can barely do the minimum. They have limited time for communications and for performing tasks beyond their hospital duties. They have very busy schedules and a heavy work burden. They have limited technical skills requiring consideration in design and choice of annotation tool.
- Highly sensitive MIMIC data cannot be distributed in crowd sourcing platforms.
- Clinical sentiment is under-researched, meaning that there is no guidance or clear instructions on the definition of clinical sentiment or corresponding labelling.
- Limited datasets, no public data.

The following sections discuss and represent important shifting points in the project that required difficult decisions.

5.2.7.1 Issues In Relation To Sampling And Annotators

The 300 samples presented to annotators were chosen for a *pilot* annotation project, but based on the observed difficulty of the work, those samples were kept as the final annotation set.

One difficulty for annotators is that the annotation task asks for an overall opinion of notes. It requires careful reading of whole notes and comprehending the whole patient’s narrative, while analysing the opinion of notes’

writers without even physically examining the patient. Therefore, the task of providing a label for the whole note is not easy compared to for example fragments of texts as in sentence-based sentiment. Many sentiment studies (which contain larger datasets 500-1000 samples) rely on small text fragments (sentences) rather than the overall picture of the patients. We think the overall narrative or carefully extracted segments are important to get the complete picture about the patient health status.

The project involved recruiting multiple clinicians (a total of 15 physicians) over two years but some (11 physicians) quit despite their genuine interest in the research and after they engaged in the process due to the difficulty of completing the CITI training to access MIMIC-III and annotating the whole 300 samples.

To support physicians, we provided a document summarizing the steps required to access MIMIC-III in Physionet and CITI websites. For some physicians, an additional meeting was held to walk them through the access steps. Moreover, we had to make the extracts field (where they could indicate which pieces of text influenced their label) optional to ease the task for clinicians. In what was initially envisaged as the pilot study, we asked the first ICU-Registrar to familiarise themselves with the task and to start timing how many notes they managed to label within an hour. It turned out that the ratio was around 30 notes within an hour. Therefore, the estimated time for the annotation task was 10 hours. This was communicated to the annotators recruited afterwards to provide a sense of the required time for the task. Physicians found it reasonable initially but they struggled in completing the task in practice.

Some annotators did not understand the task immediately, so further discussion was required to clarify the requirements. For instance, one annotator tried to sum up the extracts into a list of criteria, as the following:

”Several criteria I used to decide wether patient is getting worse or better:

1. Take main issue as priority: for example, if treating respiratory failure, I first concern about breathing function. If breathing function is getting better, then I decide it’s improving.
2. If multiple issues are addressed, I take cardiac and respiratory function as first priority.
3. I take admission chief complaint as baseline, evaluate based on the stay in ICU or issues from 24 hours. It could be very dynamic for the same patient. But I take “the end of notes” timing as the end point and decide patient is better or not at that point.
4. If there’s only chief complaint or reason for ICU admission without other transition data (lab, vitals), I consider it as not getting better nor getting worse (not enough data to decide).
5. For psychiatric patient, I decide based on patient’s mental status. It could be comparably subjective.
6. For palliative care patient, I decide based on patient’s comfort. It could be comparably subjective.”

Though the criteria was less relevant to the required task, it provides an overview of a physician’s perspective

while doing the task and may help characterizing important aspects in future work such as choosing scores or weights for extracts' phrases while developing a lexicon.

Working with clinicians who have ongoing workload and duties was challenging. The annotators reported having exams, quitting for a few weeks due to workloads and on-calls, not remembering to come back to the task, unintentionally missing some samples while doing the task, etc. Even communication was challenging; some physicians request rescheduling the meetings multiple times and some had to quit the meeting due to sudden emergency or requirements for their round time.

In the above context, this annotation should not be regarded or treated as a standard annotation project, where tasks are distributed via crowd-sourcing platforms and participation is flexible and open to a wide range of contributors. In contrast, this was a very challenging task with very challenging collaborators who have very limited time and can barely complete the minimum. In fact, 300 samples turned out to be a lot for them. As described above, some physicians reported having difficulties, and considerable pressure was placed on them to complete it. They also had to do CITI training prior to the annotation requiring at least 6 hours, and increasing the total estimated time to be at least 16 hours.

As a result of these issues, the final annotation dataset consisted only of the 300 notes intended for the pilot.

5.2.7.2 Issues In Relation To Ground Truth

As a result of the smaller than intended annotation set, as mentioned in Section 3.3.4 in Chapter 3, the three evaluation datasets (4V, 5V, and the Covid dataset) are small, with sizes 110, 32, and 23 respectively. (See Section 5.3 for details on construction of 5V and 4V and Section 6.2 for details on the Covid dataset.) The Covid data is a secondary dataset used for an initial assessment of the generalisability to other data. Due to unavailability of other public data of clinical notes, the recruitment burden as discussed in Section 5.2.7.1, and the limited time physicians can provide, only one annotator performed the Covid annotation task, and only 23 samples were annotated. Future work should include expanded datasets and annotation.

With respect to gold ground truth (5V), and silver ground truth (4V), the sizes are dependent on inter annotator agreement as described in the next section. In such annotation with more than two annotators, the result of inter annotator agreement between all annotators is what influenced the final size for the datasets. This is not something predictable until the whole annotation project is finished. Even if the recruitment extended to a greater number of annotators, e.g. 50 annotators, the size of the final ground truth is not guaranteed to be big because it is dependent on IAA measures. Hence, in this thesis, we target 4 or more (5 or full) agreement for high reliability, which reduced the data leading to the small sample sizes. However, for future work, extending by even 1 more annotator could increase the samples in both datasets by reasonable chance because it could add more samples to both by considering the new added agreement. This could be thoughtfully done by collecting subsets of notes with 3 agreement and re-annotate them by the new annotators to see if more agreement can be found

for them. Additionally, researchers could rely on Cohen Kappa for pair-wise IAA and construct other versions of ground-truth datasets.

Although the dataset is small now, to our knowledge it is first clinical sentiment dataset and the only one publicly released. In addition, it is quite feasible to do an extension for samples in future, by recruiting extra annotators or by using unsupervised methods like label propagation from the annotated samples to other unlabeled samples in MIMIC-III notes.

The dataset offers the whole notes and assessment plan segment plus small extracts of notes. Therefore, the extracts inherit the labels assigned by clinicians for the whole notes directly because they are what influenced their choice of label. Providing the extracts enables another variant of the clinical sentiment classification, especially for larger LLMs. e.g., classifying sentiment by only using the extracts and comparing the results to whole notes, AssPlan or any other segment of notes.

With the above, the dataset provides rich alternatives for future studies, while doing the heavy lifting (a task of intense and challenging annotation and 15 months of work) in establishing this dataset.

In summary, other studies only discuss the limitation of sentiment in clinical notes without providing resources in such low-resourced area. This study opens the path for other research to expand the impact.

The key takeaways from the annotation are:

- Clinicians might benefit from research hours specified for research related tasks, and not having to interleave research with other ongoing hospital duties.
- Annotation in the clinical field may require adjustment in annotation tools and may require multiple phases of annotation. In particular, an annotation tool might have embedded add-ins to calculate a selected IAA measure in the background as annotators build up the dataset. Therefore, the utility may help locate samples with questionable majority votes (e.g., 3 agreement out of 5) and help bring them back to the attention of the 2 disagreeing annotators for a double check. Additionally, the annotation tool should be made as usable as possible to ease the task and help enhance the attention of busy clinicians with high information load. The tool should also allow for quick quality checks or flagging notes with insufficient information.
- Studying pair-wise IAA is worthwhile in such high controversial judgments, which may lead to acceptable and higher consistency as opposed to only relying on IAA for all annotators.

5.3 Preparation of Ground Truth

This section describes forming the ground truth from the manually annotated 300 notes by 5 annotators.

After the annotators are done, inter-annotator agreement (IAA) (also known as inter-rater reliability) was studied to observe the level of agreement or variability in annotators' opinions. This is the extent to which human experts' selections or opinions align with each other. The higher the consensus between annotators, the more confidence and interchangeability is allowed in their opinions. (Nguyen Hong Duc et al. 2021).

Results of inter-annotator agreement and further analysis is presented in Section 5.3.1. Overall, all 5 annotators assigned the same label to only 10% of the 300 notes and in 30% of the notes, 4 of the 5 annotators provided the same label. Further details are described in the next section.

5.3.1 Inter-Annotator Agreement

Inter-annotator agreement measures matches and mismatches in opinion-based tasks assigned to multiple participants. It allows for gaining a level of certainty in the final judgment or the overall trend towards a specific question where multiple selections are possible. Multiple measurement tools exist for investigating opinion alignments or conflicts. The selection of the measure depends on the number of annotators in the task. In this research, Fleiss inter-Annotator used to investigate the agreement of the annotators over the 300 samples.

Fleiss kappa is an extension and a generalisation of kappa coefficient, defined mathematically as in equation 5.1:

$$\text{kappa}(\kappa) = \frac{P_o - P_e}{1 - P_e} \quad (5.1)$$

where P_o = observed agreement among annotators, and P_e = expected agreement by chance, by which the level of agreement of two or more (multiple) annotators can be quantified and the possibility of agreement by chance can be eliminated. Fleiss' scale is based on the following ranges: < 0 is Poor, 0.0–0.20 is Slight; 0.21–0.40 is Fair, 0.41–0.60 is Moderate, 0.61–0.80 is Substantial, 0.81–1.0 is Almost perfect (Sylolypavan et al. 2023, Landis & Koch 1977).

The overall agreement for all 5 annotators over all 300 notes was slight using Fleiss's Kappa (0.1863). Counts matrix for annotators and classes is shown in Table 5.3. It can be seen on the count matrix that annotators have variable classes counts. For example, the first annotator (ICU-C) identifies 173/300 (58%) samples as deteriorating and improving. On the other hand, (Resp-C) identifies 215/300 (71%) for the two classes. (ICU-R1) identifies 218/300 (72%), (ICU-R2) identifies 284/300 (94%), and (ICU-R3) identifies 194/300 (64%) for the two classes.

The slight inter-annotator agreement is common even among experienced clinicians based on many previous studies (Sylolypavan et al. 2023), and could be explained by one or more of the following:

- Missing details or implicit clinical sentiment (Denecke & Reichenpfader 2023, Sylolypavan et al. 2023, Denecke & Deng 2015, Leaman et al. 2015).

Table 5.3: Counts matrix for annotators labeling on the manual annotation data. Original labels are: Is deteriorating?, Is already improving or will get better?, Is neither improving nor deteriorating?, OR that there is no indication of the note writer’s opinion., respectively. ‘ICU-R’ refers to ICU Registrar, ‘Resp-C’ refers to Respiratory consultant, and ‘ICU-C’ refers to ICU consultant.

Annotator	Deteriorating	Improving	Neither	No_opinion
ICU-C	98	75	109	18
Resp-C	80	135	38	47
ICU-R1	101	117	48	34
ICU-R2	130	154	12	4
ICU-R3	119	75	76	30

- Differing focus (Brown et al. 2014).
- Inadequate attention (Syloypavan et al. 2023, Booth & Narayanan 2024).
- Unfamiliar note style (Hultman et al. 2019, Leaman et al. 2015).
- Different backgrounds (Syloypavan et al. 2023, Joseph & Tolley 2015, AoMRC 2012).

Table 5.4 shows an example of variability of annotators’ opinions towards an individual note sample from the data. As can be seen from the example, the labeling is not consistent. In fact, even where annotators (ICU_Reg1 and Res_con) chose the same text to back up their label, they disagreed on the label. Moreover, the patient associated with the note whose labeling is presented in top section of Table 5.4 died two days later, yet two annotators (including the ICU consultant) were believing the patient is already improving or will get better. Another case is in the lower section of Table 5.4. The majority vote (3 out of 5 annotators) thought the patient is deteriorating, while the two consultants were unsure about the status. This variability reflects that the decision might not be easy for some experts and some cases.

To investigate IAA further, the correlation of annotator opinions is studied in each of the three pseudo-labeling groups (notes close to admission, discharge, and death) and notes not belonging to any of these groups. Figures 5.2, 5.3, 5.4, and 5.5 show these correlations as heatmaps, from which it can be observed that the correlation increases closer to each event (e.g., one day prior to death) and that consistency in opinions in notes prior to death is much higher than in the other two groups. Perhaps not surprisingly, it seems that negativity in clinical sentiment prior to death is more obvious and distinguishable (top correlations range from (0.48-0.66)). Clinical sentiment near admission seems less obvious to annotators. It may be that judgments are harder to make around admission time since it is early to know what is going on (i.e., patient’s status) and what is going to be happen (i.e., prognosis) at this early stage in the ICU stay (top correlation range is (0.29-0.39)). Annotators’ opinion about expected positivity or signs of improvements near discharge time is less correlated than near death time (top correlation is 0.32). As can be seen in Figure 5.5, there is low correlation in opinions for the notes that do not fall into the near admission / discharge / death groups.

Hence, not all annotated samples can be included in the evaluation dataset. Ground truth data was instead

Table 5.4: Two synthesized examples of conflicting opinions about an individual note. 'ICU-R' refers to ICU Registrar, 'Resp-C' refers to Respiratory consultant, and 'ICU-C' refers to ICU consultant.

-	ICU-R1	Resp-C	ICU-C	ICU-R2	ICU-R3
labels	is deteriorating	is neither improving nor deteriorating	is already improving or will get better	is already improving or will get better	is deteriorating
Extracts	–Pt was extubated, was improving and was transferred to 2B Ward yesterday; however, pt was transferred back to MICU after having sudden AFib.	– Pt was extubated, was improving and was transferred to 2B Ward yesterday; however, pt was transferred back to MICU after having sudden AFib.	improving - extubated	– Pt was previously intubated and had Chest X-Ray. Pt was extubated, was improving and was transferred to 2B Ward yesterday – stable BP.	– back to icu for sudden AFib.
-	ICU-R1	Resp-C	ICU-C	ICU-R2	ICU-R3
labels	is deteriorating	There is no indication of the note writer's opinion	is neither improving nor deteriorating	is deteriorating	is deteriorating
Extracts	–Pt on antibiotic at shift, agitated, HR increased to 140-160.				– reintubated due to respiratory failure not responding to NIV.

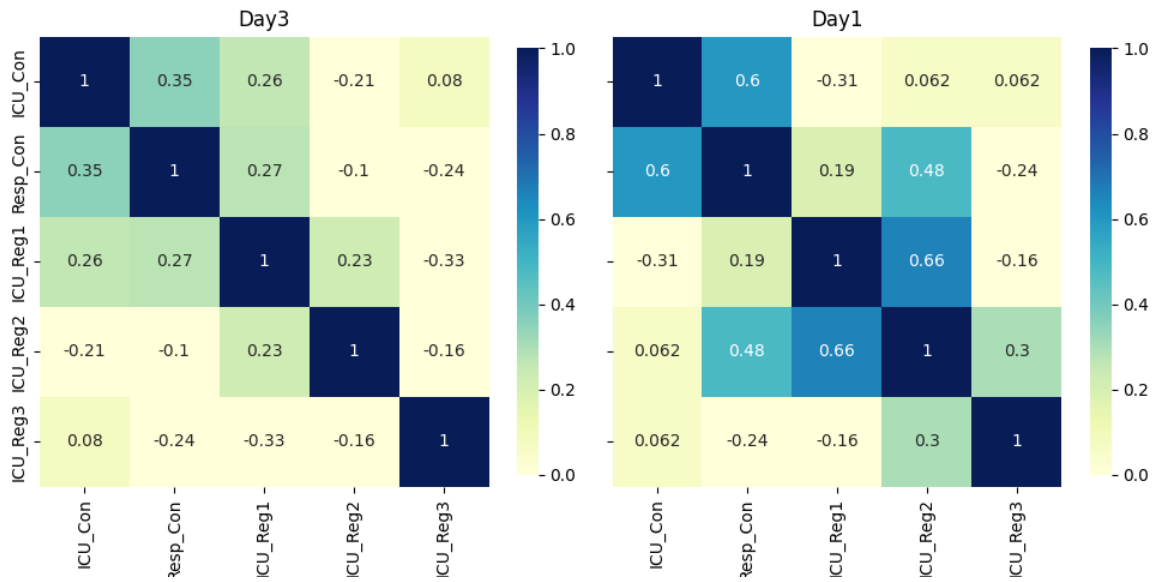


Figure 5.2: Correlation heatmap between annotators of samples of notes in 3 days, and 1 day prior death.

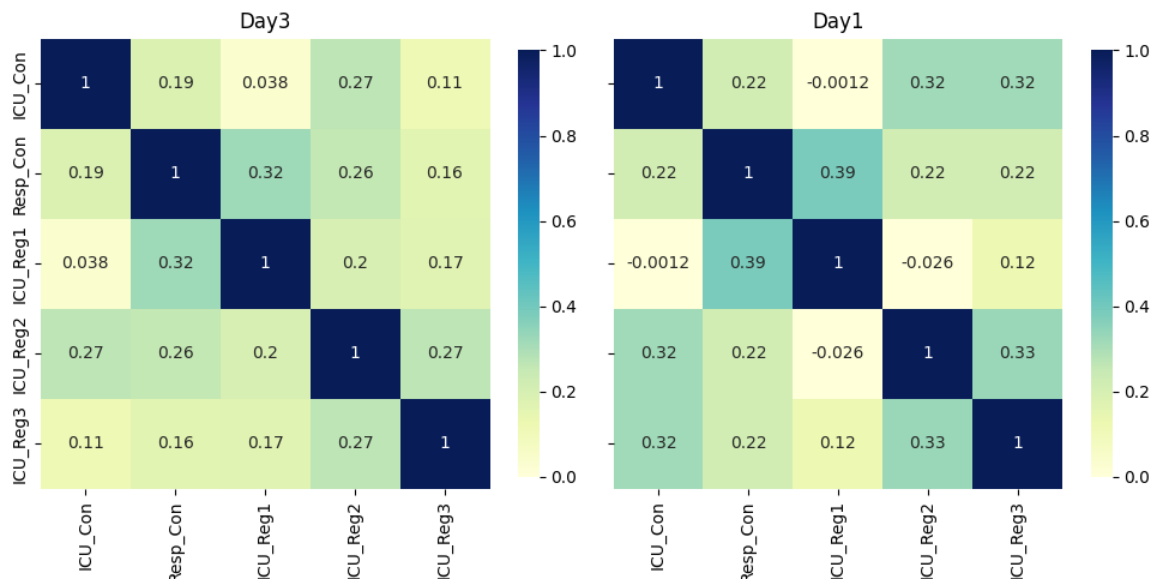


Figure 5.3: Correlation heatmap between annotators of samples of notes in 3 days, and 1 day after admission.

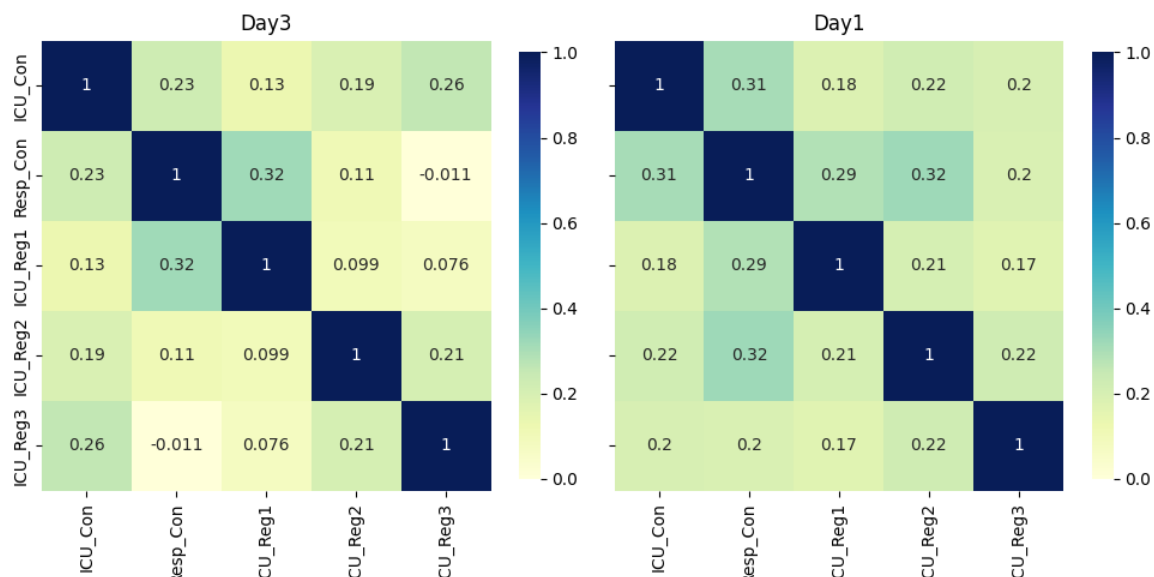


Figure 5.4: Correlation heatmap between annotators of samples of notes in 3 days, and 1 day prior discharge.

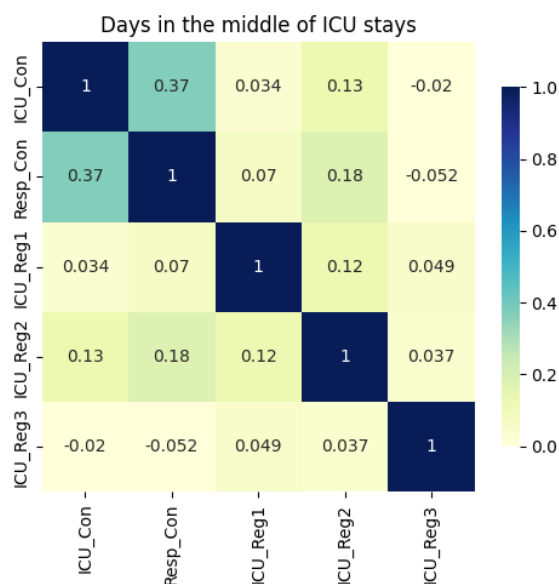


Figure 5.5: Correlation heatmap between annotators of samples of notes that don't belong to admission, death, or discharge groups.

constructed, similar to Booth & Narayanan (2024), from subsets of data with high Fleiss values (5-votes, 4-votes) or (5V, 4V), forming reliable levels of consensus and accumulated evidence supporting validity. The set of 32 notes on which all 5 annotators agree, referred to here as "5-votes" or "5V" dataset, is considered as the gold ground truth; whereas the set of 110 notes on which 4 or more annotators agree, referred to here as "4-votes" or "4V" dataset, is considered the silver ground truth. 5-votes data has perfect agreement with the Fleiss kappa (k) = 1, the Fleiss kappa for 4-votes dataset is (k) = 0.5451, which reflects moderate agreement according to Fleiss interpretation (Fleiss 1971). Both datasets have $p < 0.0001$, which indicates that kappa values are significantly different than 0, and rejecting the occurrence by chance. Figure 5.6 shows the whole path of annotation and construction of ground truth datasets.

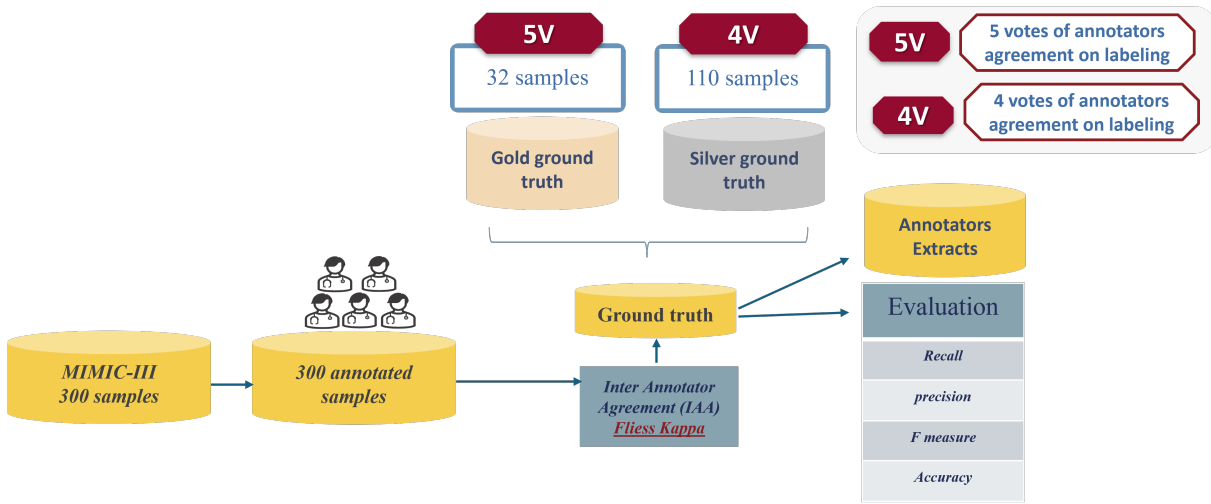


Figure 5.6: The whole path of annotation and construction of ground truth datasets.

5.4 Analysis of Extracts Associated with Manual Labeling

As mentioned in Section 5.2.5, part of the manual annotation task was for annotators to identify pieces of note text that lead to their selection of labels. We think these extracts have significant value by enabling the following:

- Investigating the most common words or phrases from texts based on a specific patient categorization e.g., patients whom annotators think are doing well.
- Investigating if annotators are paying attention to or building their decisions using the same details as each other.
- Building an ICU-specific sentiment lexicon with support from ICU clinicians.
- Using the extracts to develop AI predictive models and compare the results using the different prediction pipelines.
- Using different prompting e.g., question answering style or few shots.

To study similarity between extracts, three evaluation metrics were used: ROUGE (Recall-Oriented Understudy for Gisting Evaluation), BLEU (Bilingual Evaluation Understudy), and cosine similarity based on text embeddings. Extracts are investigated on the datasets of different votes, and the hypothesis is that high votes datasets such as 5-votes should have higher scores and lower votes datasets may have lower scores. This hypothesis is based on the fact that annotators might be looking at different aspects while examining notes as discussed in Section 5.3.1. See Table 5.5. In general, the higher the score and the closer it is to 1, the higher the match between the reference and the candidate text.

ROUGE and BLEU are evaluation metrics for automatic text evaluation rather than human evaluation. BLEU is known for machine translation tasks and ROUGE is commonly used in text summarization, but both target overlapped n-grams between reference or original text and another candidate text. [Papineni et al. \(2002\)](#) and [Lin \(2004\)](#) explain both thoroughly. Cosine similarity for two vectors A and B is calculated based on the following equation:

$$\text{cos similarity } (A, B) = \frac{A \cdot B}{\|A\| \|B\|} \quad (5.2)$$

Where: - $A \cdot B$ is the dot product of A and B . - $\|A\|$ and $\|B\|$ are the magnitudes of A and B , respectively. The magnitude of a vector A is:

$$\|A\| = \sqrt{\sum_{i=1}^n a_i^2} \quad (5.3)$$

Where a_i are the components of vector A . Text embeddings are numerical text representation obtained by a language model. We used 'paraphrase-MiniLM-L6-v2', a Sentence Transformers model ([Reimers & Gurevych 2019](#)) from the Hugging Face open-source library for this purpose.

The scores in Table 5.5 suggest that the hypothesis of greater alignment of extracts for greater agreement is indeed supported by most of the scores (i.e., ROUGE-F1=0.12, Bleu=0.037, Cosine similarity=0.27 for 5-votes dataset vs. 0.11, .038, 0.23 for 4-votes dataset respectively).

Although the metrics are widely used in NLP and NLG, and provided the anticipated increased similarity score in 5V compared to 4V, it is important to note that these metrics share a common limitation - they focus on n-gram overlap and embeddings and mainly reflect surface-level similarity. As a result, their semantic understanding is shallow and may struggle or fail to capture factual correctness and completeness ([Schmidtova et al. 2024](#), [Al Azher et al. 2025](#)).

5.5 Validation of Pseudo-Labeling

Measuring the level of agreement or disagreement between human and automatic evaluation is important in NLP tasks where both kinds of evaluation are involved ([Schmidtova et al. 2024](#)).

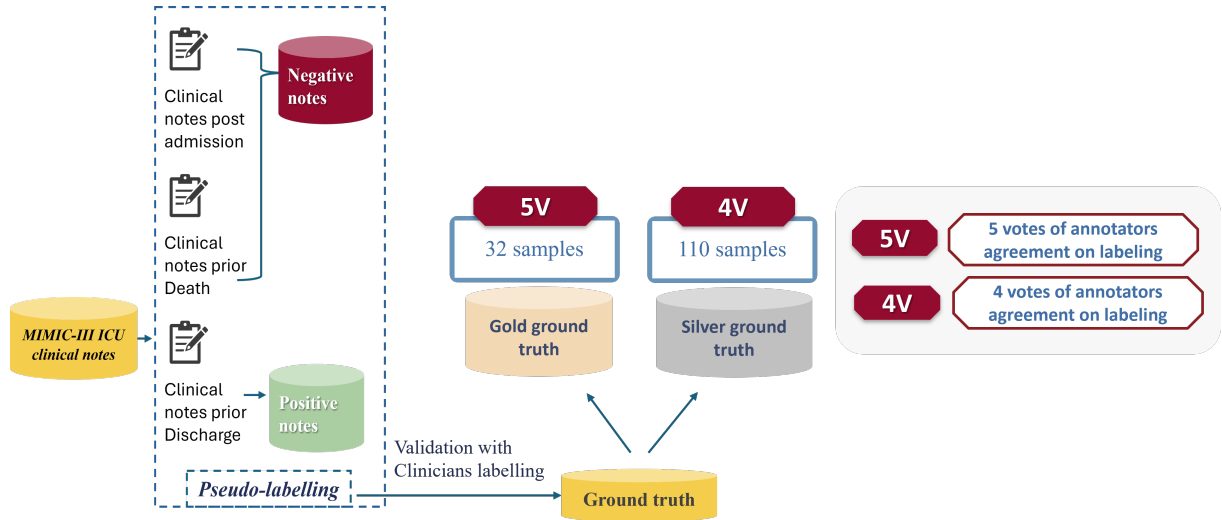


Figure 5.8: Validation with pseudo-labels.

The automated "pseudo-" labeling strategy outlined in Chapter 3 and the pseudo-labeled dataset used in Chapter 4 to evaluate clinical sentiment classification methods were partially validated by the experiments in Chapter 4 in so far as sentiment classification works well. However, now that a ground truth dataset is prepared and generated from human expert annotation, it becomes possible to assess how well the pseudo-labels align with the ground truth. This section presents that assessment. Figure 5.8 shows the overall validation.

Because the pseudo-labeling strategy only provides labels for clinical notes within a few days of admission, discharge and death, as explained in Section 3.2.3, only such notes from the manually annotated notes can be used for this analysis. There are 145 of these in the set of 300 notes annotated; 74 notes within 3 days of admission, 48 in the 3 days preceding discharge, and 23 in the 3 days preceding death. These are collected randomly during sampling but since the sampling was inspected manually afterwards for removing notes with repeated contents, a final step was taken to assure that each group has at least 20 samples. After the annotation, groups of two-classes which were annotated as 'Is deteriorating', or 'Is already improving or will get better' are extracted from admission, discharge, death groups. This is to compare the two-classes annotated by clinicians with pseudo-labels as pseudo-labels are only based on two-classes. The two-classes samples are small relative to the original death, discharge, and admission groups. Of the 23 of death samples, 2 samples were found in 5V and 10 samples were labeled using the two-classes in 4V. On the discharge group, 9 out of 74 samples were annotated with the two-classes in 5V and 33 were found in 4V. Of the 48 of the admission group, 5 samples were found in 5V and 12 in 4V. Treating manual annotation as truth, the accuracy of accuracy of pseudo-labels is computed as a percentage of correct pseudo-labels in the set of manually annotated notes. To check the validity of the pseudo-labeling, counts and percentages where pseudo labels match ground truth labels are studied. Percentage where pseudo-labels match with ground truth (5V,4V) are calculated using equation 5.4:

$$\text{Percentage of a group} = \frac{CC}{CBA} \quad (5.4)$$

Table 5.6: Statistics for matching labels between pseudo-labels and gold ground truth (5V) in middle, and silver ground truth (4V) on right. Two classes are samples labeled as 'Is deteriorating', or 'Is already improving or will get better' by annotators. '%' is percentage of correct labels compared to ground truth.

Group	Total Samples	5V			4V		
		Two-classes	Correct Labels	%	Two-classes	Correct Labels	%
Death	23	2	2	100%	10	9	90%
Discharge	74	9	9	100%	33	33	100%
Admission	48	5	2	40%	12	5	42%

where, CC is counts of correct pseudo-labels (matched with annotators labels) in the group, and CBA is counts of all samples labeled as improving or deteriorating by annotators. Using the equation, the percentage of correct labels compared to ground truth datasets (5V, 4V) are (100%, 90%) in death group, (100%, 100%) in discharge group, (40%, 42%) in admission group respectively. Table 5.6 shows counts and percentages of total samples and correctly labeled samples in the three groups. The pseudo-labeling matches ground truth in death and discharge groups. However, alignment for the admission group is much lower (around 40%). It can be observed from the table that only 12/48 were labeled as 'Is deteriorating', or 'Is already improving or will get better' by annotators. This indicates that annotators might have faced difficulty in identifying clinical sentiment in this group since the rest of notes were labeled by 2 to 3 annotators as "Is neither improving nor deteriorating", or "no indication of the note writer's opinion". The pseudo-labels have no corresponding labels to those labels.

Therefore, admission group samples are difficult to judge and validate considering that only 25% of this group can be compared with pseudo-labels. However, this group was kept in the training and evaluation datasets to maintain the diversity of samples and to observe the prediction in different settings. Within certain samples of this group in the ground truth, annotators have reported that a number of patients have exhibited signs of rapid improvement. However, identifying such cases and instances in practice can be challenging and not immediately clear. The experimentation using the pseudo-labeling in Chapter 4 was conducted on patients including those with short stays e.g., 6 days. The issue identified here by (misalignment between the pseudo-labels and annotators) reflects that this group of patients may have some exceptional quick improvements, against the pseudo-labeling, suggesting negativity of admission notes up to 3 days. The issue does not affect models trained completely on pseudo-labels in Chapter 4 because training and testing datasets have consistent labeling pattern, but it has an effect on the final evaluation as discussed in Chapter 6.

Making a final conclusion about the validity of pseudo-labels using the above details is difficult, and further investigations with increased size of annotation samples allowing for statistical significance measures are required. However, the pseudo-labels strongly align with ground truth in samples linked to two major clinical events (discharge, death), presenting validity in these two clusters.

5.6 Conclusion

This chapter describes the creation of the first public clinicians' ground truth for clinical sentiment in ICU notes. Five clinicians were recruited for annotating clinical sentiment in notes, and providing extracts of the notes that influenced their decisions. The task of annotation was challenging due to clinicians' workload and span 15 months. Inter-annotator agreement (IAA) on labels was used to decide on the ground truth dataset. Wordclouds were created from positive and negative extracts provided by annotators as representatives for their labeling choices. Pairwise comparison of extracts showed that the gold ground truth (5V) has more similar extracts than the silver ground truth (4V). Pseudo-labeling was compared against both datasets (5V) and (4V) and strong alignments were identified in samples of death and discharge notes.

Two research objectives RO2-B, RO4 are achieved in this chapter. By achieving the validation of pseudo-labeling and the construction of ground truth, the final phase of the research experiments using the pseudo-labeling and the ground truth are feasible.

The next chapter presents the final study of methods' and models' performance on the ground truth dataset, which seeks to identify the best approach for ICU clinical sentiment classification.

Chapter 6

Final Evaluation on Ground Truth Dataset

This chapter represents the final phase (RP3) of the research investigations focusing on identifying the most effective approach and model for clinical sentiment classification, addressing the thesis’s main research question. The analysis of this chapter uses findings from previous research phase (RP2) and evaluates the whole methodology, using human evaluation with the ground truth dataset generated in Chapter 5. Section 6.1 describes the comparison design. Section 6.2 describes training and evaluation datasets. Section 6.3 describes the models involved in the analysis. Section 6.4 presents the results and discussion. Section 6.5 provides the best model implementation details. Section 6.6 lists some of the technical issues encountered in the research investigation. The chapter concludes with a summary of the final findings in Section 6.7.

6.1 Experiment Method

This investigation aims to answer the main thesis research question: **Which sentiment analysis method is most effective in recognizing clinical sentiment in ICU clinical notes?** It addresses the final research objective, RO5. The previous investigation, presented in Chapter 4, concludes that methods based on clinical-specific domain are more accurate in classifying clinical sentiment than methods based on general domain. Therefore, the final experiment investigates clinical-specific methods exclusively.

Four models were studied, based on two methods: Kennedy’s keyword-based ICU-specific lexicon, and a learning-based method. The clinical lexicon is from a recent publication, while the LLMs are the ones that demonstrated the best accuracy in the second phase of investigation (RP2) which was based on the automatically labeled dataset. The aim in the previous study was to investigate the effect of domains on sentiment recognition and to

have an initial experimentation of the labeling strategy, hence all datasets were automatically (pseudo-) labeled (see Section 6.2.2). This experiment builds on the findings and evaluates those models on a human-labeled dataset, providing (1) evaluation for automatic labeling vs. human labeling; and (2) evaluation of machine learning models performance vs. human performance.

Three datasets were used in the analysis, two from MIMIC-III data: a training dataset to train LLMs, and an evaluation dataset to evaluate the lexicon and LLMs. As mentioned, the training data is the pseudo-labeled dataset used in Chapter 4 and described in Section 3.2.3. The evaluation dataset is the ground truth dataset created with clinician-labeled data as described in Section 5.3. In addition, an external dataset (the Covid dataset) by Cohen et al. (2020) is used for evaluation.

In summary, the steps involved in this phase are:

- Train clinical LLMs using training data (pseudo-labeled) and evaluate them on ground truth data.
- Evaluate all models on external data (the Covid dataset).

The experiment aims to evaluate the performance of specialized models and lexicon, seeking to identify which should be the research's contribution, SentimentICUModel. The next sections introduce the datasets involved and models adopted to test the hypothesis.

6.2 Datasets

All datasets used in the research were introduced in Section 3.2.4. The data used in the evaluations in this chapter is a subset of MIMIC-III, plus the Covid dataset (Cohen et al. 2020). Two datasets were extracted from MIMIC-III notes, one for training and the other for evaluation. The Covid dataset was also used for evaluation. This data contains clinical notes associated with *chest X-ray* and *CT images* of patients infected with or suspected of COVID-19 or other pneumonia causes. This data was mainly collected from publications, but some samples were collected independently from hospitals and physicians. The next two sub-sections describe these datasets and the labeling of each.

6.2.1 Evaluation Data

The MIMIC-III evaluation dataset and ground truth is described in Section 5.2.2, and Section 5.3 in Chapter 5. From the Covid dataset, only samples of patients who are located in ICU are extracted (158), and a respiratory consultant was asked to annotate them using the same labeling instructions used for the annotation applied to MIMIC 300 samples (see Section 5.2). Due to the physician's workload and time limit, only 23 Covid samples were annotated. The experiment method is summarised visually in the right hand side of Figure 6.1.

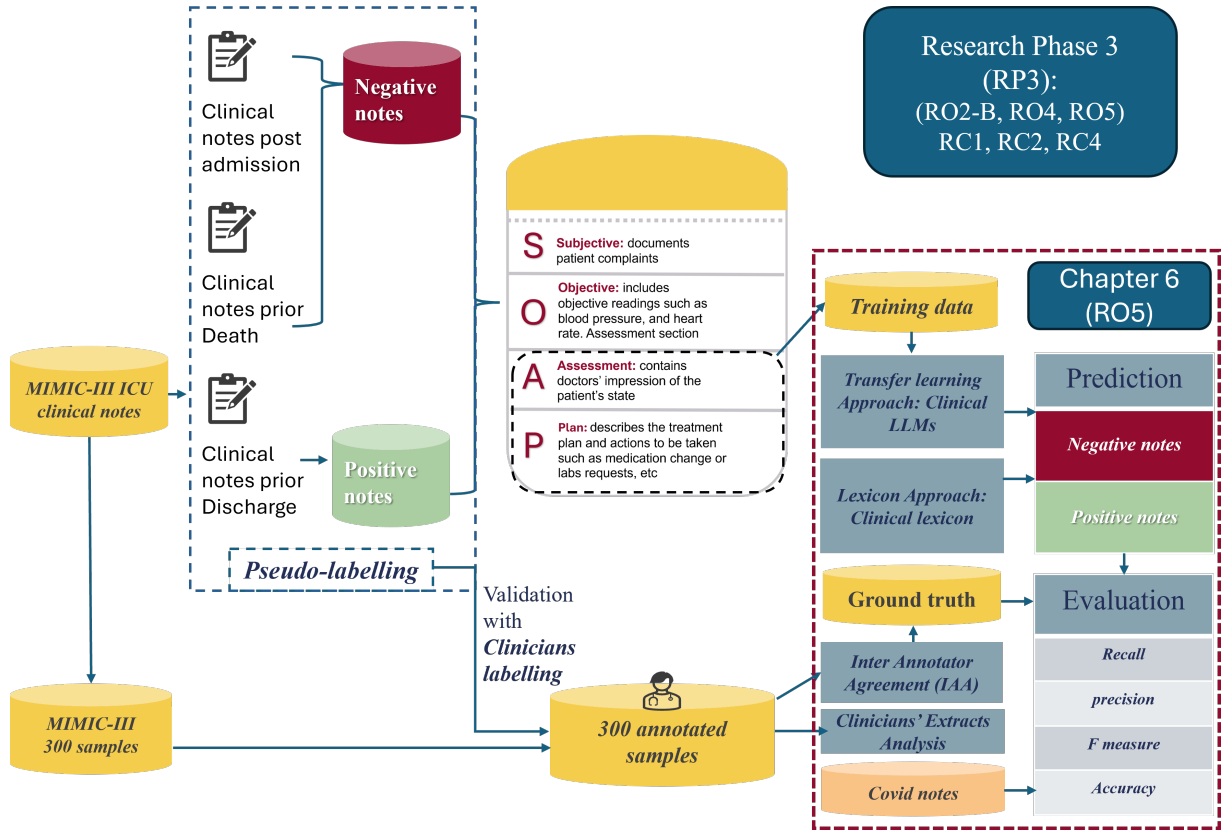


Figure 6.1: Research Phase 3, parts related to Chapter 6 are highlighted.

6.2.2 Training Data

The training data for the LLMs for the final investigation is the same data used in Chapter 4 investigations. It is a collection of MIMIC-III notes, selected based on timing attributes of patients ICU stays and labeled using the pseudo-labeling scheme explained in Section 3.2.3 and inspired by Weissman et al. (2019). In the previous phase of investigation (RP2)¹, it was found that classification of clinical sentiment of notes, specifically the part of a note that is from SOAP² assessment point onward, achieves slightly improved accuracy scores especially for physicians' notes, probably because assessment and plan contains physicians' impressions as noted in Section 4.3. Therefore, the evaluation experiment utilises this section of notes in the analysis. It was also observed that physician notes provided more accurate prediction than nursing notes. However, in this study, models were trained on each category separately to compare and contrast their effect on training and evaluation using experts' annotation. Therefore, the training data is the same that was used previously in the second phase documented in Chapter 4, except for the exclusion of Whole/truncated notes. Table 6.1 shows the size of the training dataset.

¹The study associated with this investigation is published as (Nagoor et al. 2024).

²Refer to Section 4.3 in Chapter 4 for description of SOAP.

Table 6.1: Sizes of of physician and nursing notes training dataset.

Category	Training
Physician	10766
Nursing	10766

Table 6.2: Training details for learning models

Model	Batch	Learning rate	Epochs	Optimizer	Tokens size
ClinicalT5	64	1e-3	8	AdamW	512
SciFive	64	1e-3	8	AdamW	512
ClinicalBERT	64	1e-4	4	AdamW	512

6.3 Models

As described in Section 6.1, three different LLMs were selected, plus an existing keyword-based clinical lexicon (non-learning algorithm) to investigate their capabilities in capturing clinical sentiment polarity for ICU clinical notes.

The keyword lexicon is the one from Kennedy et al. (2023) which will be referred to as Kennedy’s lexicon hereafter. This list of 103 negative and 72 positive keywords and phrases was extracted and reviewed iteratively by a group of clinicians. The sentiment in their analysis was measured by a score calculated based on equation (6.1).

$$\text{Sentiment score} = \frac{\text{NK}}{\text{NK} + \text{PK}} \quad (6.1)$$

where NK is the number of negative keywords, and PK the number of positive keywords.

The three LLMs evaluated in this phase are ClinicalBERT, Scifive and ClinicalT5. Details of these LLMs are in Section 4.8.

In experimentation with LLMs, the following hyper-parameters were explored: learning rate in the ranges $\{1e-2 \text{ to } 1e-5\}$, $\{3e-4, 3e-5\}$, and $\{5e-4, 5e-5\}$, batch size in the values $\{4, 8, 16, 32, 64, 128\}$, and epochs in the range $\{1-10\}$. Similar to the previous experiment, early stopping was applied and the best hyper-parameters are shown in Table 6.2.

6.4 Results and Discussion

This section describes the results obtained from the comparison of the clinical LLMs and the clinical lexicon with insights on models’ performance on clinical sentiment classification. Section 6.4.1 shows the results with key outcomes and highlights, Section 6.4.2 presents statistical significance of the results, Section 6.4.3 discusses cases of misclassified samples.

Table 6.3: Evaluation on physicians’ notes of ClinicalT5, SciFive, clinicalBERT, and Kennedy’s-lexicon using overall accuracy, and recall, precision, F1-score for positive and negative classes, and applied to assessment and plans sections onwards. Best result for each dataset is shown in bold.

Model	Data	Size	Acc	Neg_R	Neg_P	Neg_F1	Pos_R	Pos_P	Pos_F1
ClinicalT5	5V	17	0.8235	0.8000	0.8889	0.8421	0.8571	0.7500	0.8000
ClinicalT5	4V	45	0.8000	0.8462	0.8148	0.8302	0.7368	0.7778	0.7568
ClinicalT5	Covid	23	0.8261	0.8571	0.6667	0.7500	0.8125	0.9286	0.8667
Scifive	5V	17	0.8235	1.0000	0.7692	0.8696	0.5714	1.0000	0.7273
Scifive	4V	45	0.7333	0.9231	0.7059	0.8000	0.4737	0.8182	0.6000
Scifive	Covid	23	0.8696	0.7143	0.8333	0.7692	0.9375	0.8824	0.9091
ClinicalBERT	5V	17	0.8235	1.0000	0.7692	0.8696	0.5714	1.0000	0.7273
ClinicalBERT	4V	45	0.7556	0.9231	0.7273	0.8136	0.5263	0.8333	0.6452
ClinicalBERT	Covid	23	0.9565	1.0000	0.9412	0.9697	0.8571	1.0000	0.9231
Kennedy’s lex	5V	17	0.3529	0.5000	0.8333	0.6250	0.1429	0.5000	0.2222
Kennedy’s lex	4V	45	0.2826	0.4231	0.6471	0.5116	0.1053	0.6667	0.1818
Kennedy’s lex	Covid	23	0.1200	0.0625	1.0000	0.1176	0.0000	0.0000	0.0000

6.4.1 Results and insights

This section presents the results of predicting clinical sentiment, and provides key insights. As mentioned in Section 6.3, Clinical LLMs including ClinicalT5, Scifive, ClinicalBERT, and a clinical lexicon by Kennedy et al. (2023) are trained to predict clinical sentiment. The results of the models and Kennedy’s lexicon are reported in Table 6.3 for physicians’ notes, and Table 6.4 for nursing notes. They show overall accuracy, and recall, precision, F1-score for positive and negative classes, and applied to assessment and plan sections onward for physicians’ and nursing notes. Note that samples with no assessments or plans are excluded from ground truth 5V and 4V in tables.

Table 6.5 is a summary table averaging LLM accuracies against Kennedy’s lexicon for prediction of clinical sentiment using physician notes (left section) and nursing notes (right section).

Overall, LLMs outperform Kennedy’s lexicon on ground truth datasets and the Covid dataset (82% vs 35%) on 5V, (76% vs 28%) on 4V, and (80% vs 9%) on the Covid dataset. With respect to physician notes, all specialized LLMs achieved high accuracy (82%) for gold ground truth, i.e., samples which all annotators agree on their labeling (5V). Clinical LLMs have slightly lower accuracy (76%) on the silver ground truth, those with 4 matching opinions (4V). Predicting sentiment of nursing notes was less accurate (71% on 5V data and 68% on 4V) with similar pattern of better accuracy on the gold ground truth (5V) than on silver ground truth (4V).

On the Covid dataset (Cohen et al. 2020), clinical LLMs achieve high accuracy (88% on physician notes and 80% on nursing notes), whereas Kennedy’s lexicon’s accuracy is poor (12%, 9% respectively).

Table 6.4: Evaluation on nursing notes of ClinicalT5, SciFive, clinicalBERT, and Kennedy’s lexicon using overall accuracy, and recall, precision, F1-score for positive and negative classes, and applied to assessment and plans sections onwards. Best result for each dataset is shown in bold.

Model	Data	Size	Acc	Neg_R	Neg_P	Neg_F1	Pos_R	Pos_P	Pos_F1
ClinicalT5	5V	14	0.7143	0.6000	0.6000	0.6000	0.7778	0.7778	0.7778
ClinicalT5	4V	58	0.7241	0.6818	0.6250	0.6522	0.7500	0.7941	0.7714
ClinicalT5	Covid	23	0.8261	0.8571	0.6667	0.7500	0.8125	0.9286	0.8667
Scifive	5V	14	0.7857	0.8889	0.8000	0.8421	0.6000	0.7500	0.6667
Scifive	4V	58	0.6897	0.7222	0.7647	0.7429	0.6364	0.5833	0.6087
Scifive	Covid	23	0.7826	0.2857	1.0000	0.4444	1.0000	0.7619	0.8649
ClinicalBERT	5V	14	0.6429	0.6000	0.5000	0.5455	0.6667	0.7500	0.7059
ClinicalBERT	4V	58	0.6207	0.6818	0.5000	0.5769	0.5833	0.7500	0.6562
ClinicalBERT	Covid	23	0.7826	0.9375	0.7895	0.8571	0.4286	0.7500	0.5455
Kennedy’s lex	5V	14	0.3571	0.4000	1.0000	0.5714	0.3333	1.0000	0.5000
Kennedy’s lex	4V	58	0.1897	0.2727	0.6000	0.3750	0.1389	0.6250	0.2273
Kennedy’s lex	Covid	23	0.0435	0.0625	1.0000	0.1176	0.0000	0.0000	0.0000

On physician notes (see Table 6.3), ClinicalT5, Scifive, and ClinicalBERT achieved 82% accuracy on gold ground truth (5V). On silver ground truth (4V), ClinicalT5 achieves 80% while Scifive and ClinicalBERT achieved 73%, and 75% respectively. On the Covid dataset, ClinicalBERT achieves 95%, Scifive achieves 86%, and ClinicalT5 achieves 82%.

The three models achieve comparable and excellent results on ground-truth datasets and the Covid dataset. However, ClinicalT5 has demonstrated consistently above 80% for the two ground truth datasets and the Covid dataset. In contrast, Kennedy’s lexicon performs consistently poorly with highest accuracy of 35% on 5V.

In terms of negative and positive F1-scores, ClinicalT5 provided the top scores on 5V and 4V, while ClinicalBERT and Scifive provided more excellent scores in the Covid dataset. Kennedy’s lexicon remain poor, especially on positive F1-score.

A similar pattern is observed for nursing notes (see Table 6.4), the three clinical LLMs outperform Kennedy’s lexicon. However, accuracy scores have lower ranges than physician notes ranging from 62% to 82%. ClinicalT5 achieves consistently above 70%, (71%, 72%, 82%) for the two ground truth datasets, and the Covid dataset, respectively. Scifive achieves (78%, 68%) for 5V and 4V, and ClinicalBERT scores are (64% 62%). Both have 78% accuracy on the Covid dataset. Kennedy’s lexicon accuracy scores remain poor with top accuracy achieved for 5V of 40%.

F1-scores for negative class also have lower ranges for nursing notes than physician notes. Except for Scifive achieving excellent F1-score (84%) for negative class, the other two LLMs have higher F1-scores for positive class than negative class on 5V and 4V. This might align with findings by Kennedy et al. (2023), that nursing notes

Table 6.5: A summary table averaging clinical LLMs accuracies against keyword method. 5V indicates the gold ground truth with 5 annotators’ votes, and 4V indicates the silver ground truth with 4 annotators’ votes.

Method	Physician notes			Nursing notes		
	5V	4V	Covid	5V	4V	Covid
Clinical LLMs	0.82	0.76	0.88	0.71	0.68	0.80
Kennedy’s lex	0.35	0.28	0.12	0.40	0.25	0.09

bear more positive sentiment. However, the keyword-lexicon provided by their study does not capture much of the positivity in the samples. This might be due to the short list of positive keywords in the lexicon. During the analysis, it was noticed that accuracy values drop dramatically for samples belonging to patients with short ICU stays ($LOS \leq 6$), ranging from 24% for Kennedy’s lexicon to 46% for clinical LLMs. While examining the ground truth samples, it was found that patients with short length of stays ($LOS \leq 6$) form 60% of the samples within the admission group in the ground truth (3 out of 5 in 5V admission group, and 7 out of 12 in 4V admission group), have exhibited signs of rapid improvement, causing inconsistency in the labeling in the training and evaluation. Identifying such cases and instances in practice can be challenging and not immediately clear. However, observing the variation based on different patients groups and ICU journeys is important for understanding factors affecting clinical sentiment. Therefore, With these short LOS samples included, a slight decrease of accuracy was observed on ground truth (from top score 93% as reported in Table 4.7 in Chapter 4 to 82% on averaged LLMs performance reported here). This is because the ground truth contains 16 positive samples (7 physician notes and 9 nursing notes). Of those, 9/16 are $LOS \leq 6$ (between 1-4 days). Upon removing these samples from training and evaluation data, the accuracy was increased to 94% and 88% for 5V and 4V respectively.

From the evaluation on ground truth, it can be observed that, except for the group of short stays in ICU, the pseudo-labeling has shown effectiveness in training clinical LLMs for clinical sentiment classification. This validation of our pseudo-labeling strategy is valuable because it enables the models to be exposed to a large number of training samples in a cost-effective manner, without the cost of manual annotation. To enrich the analysis and observations, it is important in subsequent research to increase the sample size to allow for more observations on sub-groups and to do further statistical validity on both the pseudo-labeling and the prediction of clinical sentiment.

In summary, all LLMs performance aligns with annotators’ agreement, i.e, higher accuracies is observed for increased agreement on labeling. ClinicalT5 shows higher accuracy scores compared to Scifive and ClinicalBERT and all LLMs significantly outperform Kennedy’s lexicon. Models’ performance using physician notes (82%, 76%) were higher on 5V and 4V compared to nursing notes (71%, 68%), respectively. Clinical LLMs’ performance remains good on the external Covid dataset, suggesting reliable generalisability.

Overall, using the ground truth 5V and 4V, ClinicalT5 provides consistently the best accuracy on capturing clinical

sentiment on assessment and plan section of physician notes, and using full training for all layers. Finalizing the outcome of best model requires assessing the statistical significance of the accuracy of LLMs and the lexicon to identify if the results is statistically different, which is presented in next section.

6.4.2 Statistical Significance

In relation to statistical significance, the distribution of accuracy scores for each model was tested using Shapiro-Wilk test and it was found that all models' accuracies are normally distributed. Hence, ANOVA test was used to assess the significance of models' performance on data. The p-value from ANOVA test is $(8.60e-12) < 0.05$, which indicates statistically significant difference in the means of the accuracy scores ($p < 0.05$, [95% CI, 0.52, 0.63]).

When investigating model performance on the dataset, clinical LLMs were found comparable. Therefore, while ClinicalT5 shows higher accuracy scores in almost all cases compared to Scifive and clinicalBERT, there is no statistically significant difference in the mean accuracies. However, there is a statistically significant difference between all LLMs and Kennedy's lexicon ($p < 0.05$, [95% CI, -0.47, -0.28]).

The conclusion of the thesis is therefore that the most effective model in identifying clinical sentiment from ICU notes is ClinicalT5, and by considering the statistical significance, the most effective sentiment method is deep learning method using clinical large language model. Further investigations is required to replicate the study, support and confirm the conclusion.

To understand the results more, a few samples were inspected for error analysis, examining some ground truth instances that ClinicalT5 misclassified. This analysis is presented in Section 6.4.3.

6.4.3 Error Analysis

The prediction of the top performing model (ClinicalT5) is observed against ground truth labels, and samples which mismatched are collected for observation. Table 6.6 shows the instances with potential reasons for misclassifications. Overall, the misclassified samples are either short, missing essential information for judgment as in note '402461', and '142' or have some inductive words for incorrect label such as the phrase "now clinically stabilized" in note '718636', misleading the model. The complexity of patient's condition especially in cases with comorbidity may introduce some conflicting facts leading to failure in capturing the overall progress. This could be the case in notes '608843' and '197'.

In general, exposing the model to more cases which are based on different health conditions, institutions, and diverse trajectories in ICU will enrich the learning ability and support distinguishing classes more. Sparsity of data is in fact among the major challenges encountered in clinical NLP as discussed in Section 2.4. Therefore, implementation and reproducibility of clinicalT5 as SentimentICUModel is introduced next in Section 6.5 to

enable researchers to gain access to the model and extend its capabilities in future research.

6.5 Implementation and Reproducibility

ClinicalT5 demonstrates the best accuracy in clinical sentiment analysis, and so is the top performing SentimentICUModel. ClinicalT5 is provided by Physionet under protection. This is due to the potential leak of the MIMIC data on which it was trained, which can happen with encoder-decoder transformers (Lehman & Johnson 2023, Lehman et al. 2021, Carlini et al. 2020). Therefore, **SentimentICUModel** is released within Physionet for credentialed users following the required precautions. This section describes how the model was implemented and how the results can be reproduced.

SentimentICUModel was trained for 8 epochs using two A100 GPUs, and batch size of 64. The learning rate was 1e-3, optimizer is AdamW and token size was maximized to 512.

By using the model uploaded in Physionet, researchers can directly run the model on their data with no further training. It is recommended in practice to have a virtual environment for developing Python codes and experimentation. For this purpose, miniconda is a good option. After creating such environment, ClinicalT5 can be downloaded locally following the instructions on the Physionet website³. Once the model is available locally, it can be imported and used in Python script as the following:

```
from transformers import T5ForConditionalGeneration, AutoTokenizer
tokenizer = AutoTokenizer.from_pretrained("Model_path")
model = T5ForConditionalGeneration.from_pretrained("Model_path")
```

The environment used to develop the model contains the following: Operating system: Debian 5.10, Python version: 3.8.20, PyTorch version: 2.4.1, Transformers version: 4.46.3, CUDA version: 12.4, NumPy version: 1.24.4, Pandas version: 1.5.0, and Scikit-learn version: 1.3.2.

The dataset annotated by clinicians has been submitted for publication within Physionet. Because it is derived from MIMIC-III it needs to be published using the same license and restrictions as that data. Although the ground truth has been specified and used throughout the final evaluation, the complete (300 note) dataset is being published to allow researchers to have access for all samples. Researchers are encouraged to build on this resource and extend the annotation process through the following suggestions:

- Extend the annotation via recruitment of more annotators.
- Allow annotators either individually or in pairs to review the annotations and provide feedback or any potential improvements on notes quality, notes segmentation, or on the labeling options. This can be

³<https://www.physionet.org/content/clinical-t5/1.0.0/>

merged into the annotation project by e.g., adding a label reflecting a quality issue on the note or a label indicating that the note needs peer-consultation.

Clinicians extracts are also included in the submitted dataset.

By using the two resources i.e., the model and the dataset, the analysis presented above can be reproduced. Section 6.6 lists some of the technical issues encountered in implementation.

6.6 Technical Challenges

Large language models processing large data typically require multi-GPUs, which requires shifting from regular training to parallel training, and aggregation of the results. Another technical issues which was encountered while training T5 model is the adjustment of the float precision for the model's processing calculation. At training epochs, the model's calculations including the training loss suddenly zeroed out for all samples with no obvious reason. Despite many iterations of inspections for models' layers, models' hyperparameters, and batches of input texts, the problem has sustained. Later, using personal experimentation, the precision of floats was adjusted to float64 from the default float32, resolving the problem. Since no online guidance was available on this matter, documenting this finding is valuable for researchers.

6.7 Summary

The aim of the study is to identify the best and the most effective approach for classifying clinical sentiment in ICU clinical notes. The findings of the Research Phase 2 (RP2) revealed that specialised methods trained in clinical texts are better than general methods, and learning-based methods are better than dictionary-based methods. Therefore, the Research Phase (RP3) investigated specialised methods. Among the models included in the research, clinical LLMs show the best accuracy, and particularly ClinicalT5 has demonstrated the top performance. The keyword-based lexicon (Kennedy's lexicon) performance was generally poor, leaving room for improvement in this approach.

In summary, this chapter provided a comparison for clinical methods and identifies the best approach in classifying clinical sentiment. Clinical LLMs provide better accuracy than clinical lexicon. Although no statistical significance was found between clinical LLMs, ClinicalT5 demonstrates the best accuracy in the analysis, and so is the top performing **SentimentICUmodel**.

Note ID	Type	Ground truth label	ClinicalT5 prediction	Potential Reason
402461	MIMIC-III, physician	negative	positive	Only information about 'Nutren Pulmonary', 'PICC Line', 'Arterial Line' was provided in assessment and plan section, while clinicians have full access to whole note. Examples of Extracts by clinicians: '-Left lower lobe consolidation.' '-Antibiotics continued (today was due to be last day of course) given'
718636	MIMIC-III, physician	negative	positive	The assessment and plan appears mostly negative but contains some phrases which might confused the model. Examples are: 'now clinically stabilized"anemia:hctstabilized'
608843	MIMIC-III, physician	positive	negative	May be due conflicting facts. For example, an extract of annotator is: - Acute respiratory failure - Much improved -Hemoptysis - some bloody sputum but no frank hemoptysis since yesterday - Septic shock - due to pneumonia, improving wbc, afebrile, c/w ceftriax/azithro. Wean pressors.
142	The Covid dataset	negative	positive	No obvious reason but may failed to make a final conclusion since the note is very short. 'ETT and NG tube placed. New air-space shadowing with air bronchogram in left lower zone . New pneumomediastinum demonstrated.'
197	The Covid dataset	negative	positive	No obvious reason, might be due to 'mild lymphopenia($0.7 \times 10^3/\text{mm}^3$, normal range $1.0-4.0 \times 10^3/\text{mm}^3$)'

Table 6.6: Examples of samples misclassified by ClinicalT5

Chapter 7

Conclusion

7.1 Introduction

This chapter provides conclusions derived from the thesis results. Section 7.2 shows how the research objectives were fulfilled. Section 7.3 describes the contributions of the research. Section 7.4 presents the final findings. Section 7.6 provides some future directions for the research. Finally, Section 7.8 concludes the dissertation with final remarks.

Recalling the research gaps identified in Section 2.7 in Chapter 2, the state of the art has shown (1) no effective method for clinical sentiment classification, (2) no public labeled clinical sentiment dataset for the development and evaluation of sentiment methods and (3) no guidance on which method has the potential to demonstrate effective clinical sentiment classification. The research objectives and contributions, set out in response to these gaps are provided next.

7.2 Fulfillment of Research Objectives

The research question in this thesis was: **Which sentiment analysis method is most effective in recognizing clinical sentiment in ICU clinical notes?**.

The research question was addressed through the 5 Research Objectives presented in Section 1.6. The following subsections describe how the objectives have been fulfilled, addressing the research question.

7.2.1 Fulfillment of RO1

RO1: To review the literature comprehensively and reflect on the existing sentiment analysis methods that are applicable to the clinical domain. This includes four sub-tasks as follows: (a) To

review lexicon-based approaches to sentiment analysis. (b) To review machine learning, especially deep learning, approaches to sentiment analysis. (c) To review hybrid approaches (combined machine learning and lexicon-based) to sentiment analysis. (d) To review recent approaches to sentiment analysis using recent large language models.

The thesis explored the existing methods and techniques of sentiment analysis from the related work. This included classic and modern methods, and are all addressed in Chapter 2 as follows: Section 2.6.3 explored dictionary methods, fulfilling RO1-A. Section 2.6.4 presented machine learning approaches including deep learning models and recent LLMs, fulfilling RO1-B, and RO1-D. Section 2.6.5 explored hybrid approaches combining dictionary based with learning based models, fulfilling RO1-C. Section 2.7 provided an overview of sentiment analysis methods used in the clinical domain.

The survey resulted in two conclusions: (1) The previous studies of sentiment approaches pay very little attention to clinical sentiment, and mostly target non clinical domains. (2) Even in clinical studies, researchers were relying on general domain methods which have been proven to be ineffective in the clinical domain. By exploring dictionary methods, machine learning methods, hybrid approaches, and understanding the limitations of each in the clinical domain, RO1 is fulfilled and research sub-question RQ1 is addressed.

7.2.2 Fulfillment of RO2

RO2: To develop a pseudo-labeling strategy for automatic labeling of patients' ICU notes from patient outcomes and attributes of ICU stays. This objective has two aspects. RO1-A, which targets the development of the labeling, and RO1-B, which targets the validation of the labeling. Section 3.2.3 in Chapter 3 and experimentation in Chapter 4, fullfils RO1-A. The validation of pseudo-labeling against ground truth labels prepared by clinicians are in Section 5.5 in Chapter 5, fullfils RO1-B. In Chapter 4, the pseudo-labeled dataset is used for training in an initial experimentation while human labeled (ground truth) data is being constructed. The limited work of clinical sentiment in the literature is attributed to the absence of a large clinical sentiment labeled training dataset, which is necessary for training supervised machine learning models, and evaluation for all sentiment methods but is resource-intensive. The pseudo-labeling strategy has enabled exposing the model to a large number of samples in an automatic, cost effective and reproducible manner. This was achieved by utilizing some of the attributes of patients' ICU stays which are intuitively related to positivity and negativity of three major events including admission, discharge and death.

7.2.3 Fulfillment of RO3

RO3: To investigate the effect of different domains, notes' category, notes segments, and transfer learning strategies on classification of clinical sentiment of ICU clinical notes.

This objective was fulfilled in Chapter 4 and associated publication (Nagoor et al. 2024) with pseudo-labeled data

while a manually labeled ground truth dataset was being constructed. Section 3.2.1 described categories of notes and the suitability of physician and nursing notes for the study. Notes' structures and segments are presented in Section 4.3. Section 4.8 introduced training techniques including fully (all-layers) or (few-layers) and models including general domain and specific domain. Using these aspects, and by studying the performance of some of the most popular general and clinical-specific methods, RO3 was fulfilled and research sub-question RQ2 was addressed. Comparing different domains, learning techniques, notes categories, and notes segments provides a guiding analysis on which optimized capture of clinical sentiment can be further extended.

7.2.4 Fulfillment of RO4

RO4: To develop a ground truth dataset by specialized clinicians for human evaluation on models performance. This objective was fulfilled in Chapter 5 and our recent submission (see 0.1 in Associated Publications and Appendix B). Section 5.2 presented a description of the whole manual annotation process design and documentation. Section 5.3.1 computed inter-annotator agreement and showed the creation of ground truth datasets using Fleiss Kappa measure. Section 5.4 introduced an analysis on extracts associated with ground truth. By achieving RO4, achieving RO5 became feasible. Construction of ground truth allows for more robust evaluation using human labels.

7.2.5 Fulfillment of RO5

RO5: To investigate the performance of the specialized models in clinical sentiment of ICU clinical notes on the ground truth dataset. RO5 was fulfilled in Chapter 6 and our recent submission (see 0.1 in Associated Publications and Appendix B) where the final evaluation of specialized methods was carried out and evaluation was based on ground truth prepared in Chapter 5. Specialized methods were compared, some were selected based on outcomes during fulfillment of RO3. Section 6.4 identified the most effective method for clinical sentiment and **SentimentICUModel**, the most effective model for classifying clinical sentiment in ICU narratives using ground truth obtained from fulfillment of RO4. Section 6.5 provides the SentimentICUModel's implementation details for reproducibility. The final evaluation provided the most effective method for clinical sentiment and a resource that can be used for effective classification of clinical sentiment. By fulfilling RO5, research question RQ3 is addressed.

The outcomes of RO1, RO2, RO3, RO4, and RO5 collectively address the research question.

7.3 Research Contributions

In response to the research gaps mentioned in Section 7.1, this thesis has resulted in the following two major and two minor contributions to the discipline of clinical NLP.

7.3.1 Major Contributions

RC1: SentimentICUModel, A model demonstrating the best performance in identifying clinical sentiment in ICU notes as fulfillment of RO3 and RO5.

RC2: MIMIC-III-Ext-Notes-Sentiment¹ the first public ICU-specific clinical Sentiment ground-truth dataset labeled by clinicians, as fulfillment of RO4. This dataset can be used in developing methods either with complete notes or their associated fragments and validation by comparing machine output against clinicians' sentiment labels.

7.3.2 Minor Contributions

RC3: An Automatic Clinical Sentiment Labeling Strategy based on attributes of patients' ICU stays that enables training on large dataset as fulfillment of RO2. This training strategy enables training with large data while being cost-effective and reproducible.

RC4: A Comparative Analysis for evaluating the effectiveness of several popular models in clinical sentiment classification of ICU notes as fulfillment of RO3 and RO5. This comparative serve as a guiding material, and a foundation for further future investigation of clinical sentiment.

Figure 7.1 shows the contributions and Figure 7.2 shows the gaps addressed by contributions. The impact of these contributions is significant. Due to the research gaps, valuable impressions reported solely in clinical notes are unavailable for prediction and decision support in ICU. The contributions of this thesis - the SentimentICUModel, the ICU specific dataset, the labeling strategy, and the comparative analysis - collectively provide a foundation for effective clinical sentiment classification, supporting future research and integration in ICU decision support systems.

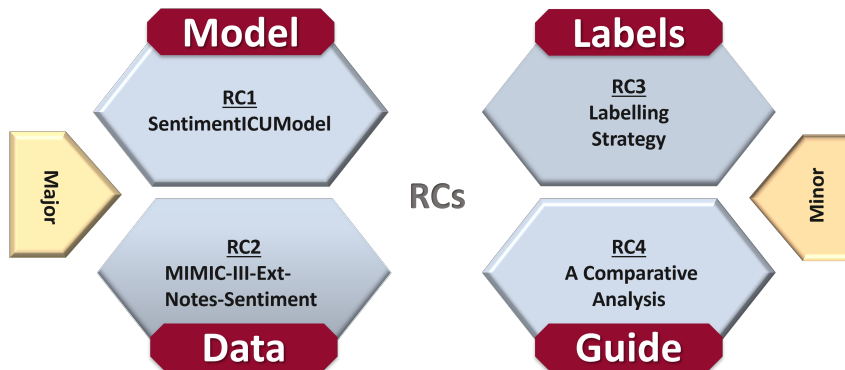


Figure 7.1: Research Contributions

¹Currently under-review in Physionet, titled: 'MIMIC-III-Ext-Notes-Sentiment: Labelled ICU Clinical Notes Dataset for Clinical Sentiment Polarity Identification'.

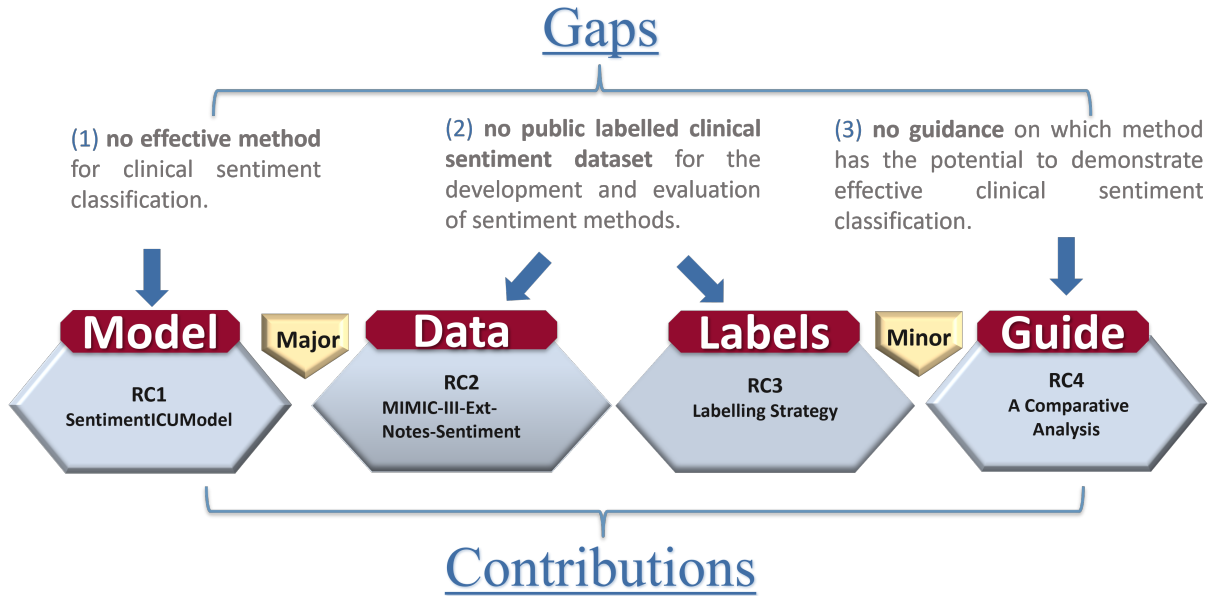


Figure 7.2: Research contributions addressing the identified research gaps.

7.4 Final Findings

The research question was: **Which sentiment analysis method is most effective in recognizing clinical sentiment in ICU clinical notes?**

The investigations conducted in the thesis have revealed that, at present, **Clinical LLMs** is the best method, and **ClinicalT5** is the best **SentimentICUModel** which demonstrates the best classification for clinical sentiment.

The key outcomes for the thesis are

- A clinical-specific sentiment models from the AI clinical LLMs models can provide a more accurate detection of clinical sentiment from ICU clinical notes than clinical-specific lexicons, the classic non-specific lexicons, and non-specific LLMs. This support the hypothesis of the research in Section 3.3.1 in Chapter 3.
- The keyword-based lexicons performance were generally poor, leaving room for future improvement in this approach.
- In terms of learning strategies, full fine-tuning optimizes the learning as opposed to light training using frozen models.
- Notes' sections of assessments and plans achieved higher performance compared to using whole notes, and detection of clinical sentiment is more accurate for physician notes than nursing notes.
- The proposed automatic labeling providing pseudo-labels for notes has been effective for training models, and shown validity when compared to clinicians' labeling. The use of both automatic and human labeling

has offered a hybrid approach by which using large training data and evaluation using human input are achieved. The labeling strategy relies on attributes of ICU stays supporting reproducibility.

- The clinician’s annotated data, and ground truth datasets are being published in MIMIC-III host website, Physionet.

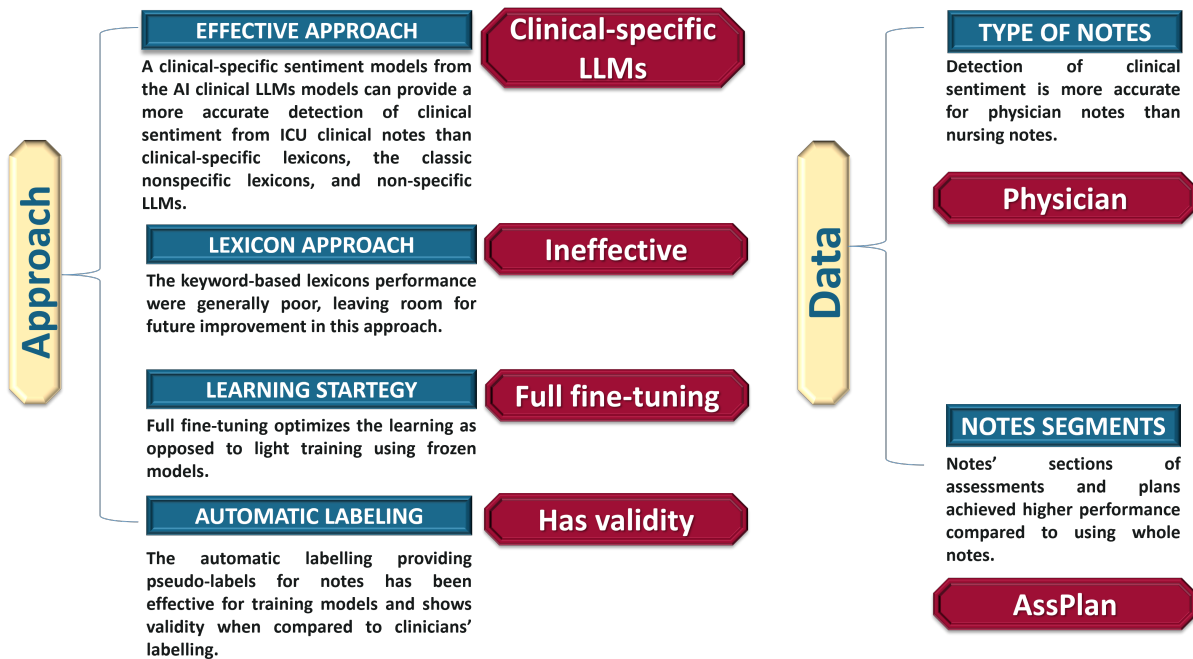


Figure 7.3: Summary of the key findings

Figure 7.3 summarises the findings. With the open resources of this thesis the analysis can be extended as suggested in future work.

7.5 Limitations

The following summarizes some acknowledged limitations of the research:

- Notes may be truncated by models due to their pre-determined limits and this could lead to information loss. Consideration should be taken to either investigate aggregation methods for long text or base the analysis on more informative sections of text.
- Training in experiments is based on a dataset whose sentiment labels are proxies for sentiments. They are neither derived by linguistic features of texts nor by experts’ annotations. The assumptions may not always be true, e.g, some patient may have fluctuations in their ICU stays or patients may experience terminal lucidity and have sudden shift in health status in the pattern ”bad, good, bad” before death.
- The conducted sentiment classification is binary (two-classes) with no neutral class included. This is unrealistic in practice. Lexicon models typically have neutral class for samples that are neither positive

nor negative samples. No reasonable strategy were found to identify neutral notes other than manual labeling.

- Researchers are advised to be aware that annotator agreement may vary depending on the samples selected for a desired investigation. Although the dataset provides a valuable resource for establishing analysis on sentiment analysis approaches, it is advised to extend the number of annotators and the number of samples, and include quality-checks by experts prior or within the annotation pipeline in future work.
- Documentation style in health care practices and conventions of writing of clinical notes may vary depending on institutions and sometimes countries. Given that MIMIC-III is based on one institution and one country, USA, it is advised to consider that this limited exposure to models may have an impact in generalizing the learning to a subsequent language-based analysis.
- Sentiment analysis relies on evaluating entire documents of notes without identification of aspect within notes' text. While analyzing complete notes is adequate and fits for the thesis purposes, employing aspect-based sentiment analysis is advisable for fine-grained scope when possible for annotators.
- Seeking dataset of patients health records (EHR) containing clinical notes within Covid-19 duration from hospitals was severely difficult due to access constraints and burden placed on hospitals. This has led to limiting the research experimentation to individual data of patients records (MIMIC-III). The secondary dataset (the Covid dataset) used for evaluation is not based on patients records. It is notes written in association with chest x-rays for patients infected by pneumonia or Covid-19. Therefore, they have very different structure from MIMIC notes, with no headings and some being quite brief, consisting of just one or two sentences that outline specific details about the patient's X-ray findings. This makes assessing the generalisability using this data inadequate. Further data and samples are required for this purpose, ideally sourced from patients physician and nursing progress notes.
- Recruiting physicians for the annotation task was challenging. It has taken 15 months with multiple unsuccessful trials, due to physicians workload and tight schedules. The primary dataset, MIMIC-III requires specific training for access. This has additionally contributed to delaying the annotation task, and delaying the final investigation using ground-truth obtained from the annotation.
- The manually annotated data is small, composed of 300 samples. Due to the slight inter-annotator agreements, the ground truth datasets were constructed on notes where all five annotators agreed (gold ground truth (5V) with 32 notes) or four agreed (silver ground truth (4V) with 110 notes) were used for evaluation. Although the ground truth datasets provided experts' evaluation on model performance, they are small and further data is required to strengthen the analysis, statistical validity of pseudo-labeling and model performance, conclusions and findings.
- The conclusions made in the thesis can be updated if further future investigations involves different settings, input data, more advanced models. Choosing the most effective method and model could be re-assessed in future work.

The next section presents the future directions for the research.

7.6 Future Work

In relation to potential future work, researchers are encouraged to build on the contributions and extend the research project through the following suggestions:

- Extending the annotation:
 1. Extend the annotated dataset via recruitment of more annotators.
 2. Allow annotators either individually or in pairs to review the annotations and provide feedback or any potential improvements on notes quality, notes segmentation, or on the labeling options. This can be merged into the annotation project by e.g., adding a label reflecting a quality issue on the note or a label indicating that the note needs peer-consultation.
- Developing predictive models:
 - Using different segments of notes to train and test different AI predictive models and report a comparative analysis.
 - Using the extracts to develop AI predictive models or keyword-based lexicons and compare the results using the different prediction pipelines.
 - Using different learning techniques or prompting e.g., question answering style, zero-shot, or few-shot.
 - Using different patients groups through clustering patients e.g., (adults vs. neonates or patients with different diagnosis).
 - Using evolving models in the state of the art of NLP.
- Integrating with decision support systems: using the provided analysis and resources to support the predictability of ICU risk factors which have a big impact on hospital management and costs. For example, scores or labels can be used independently in creating prediction models for mortality or readmission or can be fused with other features such as patient's vitals and lab results for this purpose.

7.7 Recent Large Language Models

The use of more recent large language models (LLMs) in the state of the art is recommended for the classification of clinical sentiment in ICU clinical notes. However, this may require prompt engineering involving different styles of prompting techniques. For example, asking a model a direct question using the following prompt: "Given the patient note, classify the sentiment as positive, negative, or neutral:". Another version would employ Chain-of-thought (CoT) prompting for more specific sentiment describing the severity level. For instance, asking a model to go beyond the three classes into how bad is the status of the patient from their notes. This could be formulated

as the following: 1- "Given the patient note, classify the sentiment as positive, negative, or neutral:". 2- "Using the patient's description and the generated classification, how bad or severe is the patient status?" 3- "Can you extract the pieces of text that led to that decision?"

Other advanced generation techniques such as Tree of Thoughts (ToT) by Yao et al. (2023) and Long (2023) or verbal reinforcement, known as Reflexion in Shinn et al. (2023), are also possible.

Based on findings on this research, clinical-friendly models are promising for clinical sentiment. Therefore, open-source LLMs such as MedGemma, PMC_LLaMA_13B, BioMistral-7B, Me-LLAMA or MMed-LLaMA-3 are all in the scope of future experimentation. Since they are clinical-based models, zero-shot with no training, or few shots with few samples training might provide acceptable accuracy. Samples for few shots can be collected after clustering patients into sub-groups using diagnosis or division of ICU such as Neonatal Intensive Care Unit (NICU). This can help models capture the patterns represented in their clinical notes, as ICU patients might have diverse, and complicated patterns and are not easily distinguishable by few samples.

The data privacy constraints of MIMIC-III and MIMIC-IV data enforce no usage of LLMs with external (non-local) processing, where the data is sent to be processed in external servers. Therefore, researchers should only use open-source LLMs and follow the instructions in Data Use Agreement (DUA) in compliance to ethical and legal implications by data owners. The access policy for MIMIC is restricted to researchers who gain credentialed access and sign the DUA. Obtaining credentialed access is dependent on completing training called 'CITI Data or Specimens Only Research' provided by CITI programs² as described in Physionet access details³.

The next section provides a final conclusion for the thesis.

7.8 Final Remarks

This thesis is centered on exploring various clinical sentiment methods to identify the most effective approach within the ICU narratives context. Integrating AI-based systems into ICU decision support is inevitable. Given that, and considering the crucial link to patients' survival and mortality, AI models must be created with the utmost quality, addressing all needs and specifications of critical care settings. Clinical notes have demonstrated more precise predictions compared to solely using structured data in health records, bearing clinicians' impressions and insights and data not found elsewhere in the patient's health records. Assessing the positivity or negativity of these insights using clinical sentiment classification methods is valuable for the early detection of deterioration by monitoring patient progress and making informed, valid decisions.

This work provided **SentimentICUmodel**, a model providing the most effective clinical sentiment, a labeling scheme for automatic labeling of ICU notes, a guiding comparison for models evaluation on clinical sentiment

²<https://about.citiprogram.org/>

³<https://physionet.org/about/citi-course/>

and provided **MIMIC-III-Ext-Notes-Sentiment**, the first open clinicians' annotated ground truth dataset for NLP researchers to progress clinical sentiment and other applicable NLP tasks in future work. Researchers are encouraged to shed further light on this poorly explored area, and broaden future analyses.

Finally, this work provided significant foundation to the discipline of NLP, filling a research gap that has been highly under-researched and enabling progress for future analysis. Contributions based on the thesis, support clinical linguistics, NLP applications and ultimately clinicians' decisions and healthcare practice.

Bibliography

Al Azher, I., Mokarrama, M. J., Guo, Z., Choudhury, S. R. & Alhoori, H. (2025), BAGELS: Benchmarking the Automated Generation and Extraction of Limitations from Scholarly Text, *in* C. Christodoulopoulos, T. Chakraborty, C. Rose & V. Peng, eds, ‘Findings of the Association for Computational Linguistics: EMNLP 2025’, Association for Computational Linguistics, Suzhou, China, pp. 19279–19294.

URL: <https://aclanthology.org/2025.findings-emnlp.1050/>

Ali, T., Schramm, D., Sokolova, M. & Inkpen, D. (2013), Can I Hear You? Sentiment Analysis on Medical Forums, *in* ‘Proceedings of the Sixth International Joint Conference on Natural Language Processing’, Asian Federation of Natural Language Processing, Nagoya, Japan, pp. 667–673.

URL: <https://aclanthology.org/I13-1077>

Alsentzer, E., Murphy, J., Boag, W., Weng, W.-H., Jindi, D., Naumann, T. & McDermott, M. (2019), Publicly Available Clinical BERT Embeddings, *in* ‘Proceedings of the 2nd Clinical Natural Language Processing Workshop’, ACL, Minneapolis, Minnesota, USA, pp. 72–78.

URL: <https://aclanthology.org/W19-1909>

Alshuwaier, F., Areshey, A. & Poon, J. (2022), ‘Applications and Enhancement of Document-Based Sentiment Analysis in Deep learning Methods: Systematic Literature Review’, *Intelligent Systems with Applications* **15**, 200090.

URL: <https://www.sciencedirect.com/science/article/pii/S2667305322000308>

AoMRC (2012), The benefits of consultant-delivered care, Technical report, Academy of Medical Royal Colleges (AoMRC), London.

URL: <https://rcem.ac.uk/wp-content/uploads/2021/11/TheBenefitsOfConsultantDeliveredCare.pdf>

Asghar, M. Z., Ahmad, S., Qasim, M., Zahra, S. R. & Kundi, F. M. (2016), ‘SentiHealth: creating health-related sentiment lexicon using hybrid approach’, *SpringerPlus* **5**(1), 1139.

URL: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4954801/>

Ben Abacha, A., Yim, W.-w., Adams, G., Snider, N. & Yetisgen, M. (2023), Overview of the MEDIQA-Chat

- 2023 Shared Tasks on the Summarization & Generation of Doctor-Patient Conversations, in T. Naumann, A. Ben Abacha, S. Bethard, K. Roberts & A. Rumshisky, eds, ‘Proceedings of the 5th Clinical Natural Language Processing Workshop’, Association for Computational Linguistics, Toronto, Canada, pp. 503–513.
URL: <https://aclanthology.org/2023.clinicalnlp-1.52/>
- Booth, B. M. & Narayanan, S. S. (2024), ‘People make mistakes: Obtaining accurate ground truth from continuous annotations of subjective constructs’, *Behavior Research Methods* **56**(8), 8784–8800.
URL: <https://doi.org/10.3758/s13428-024-02503-3>
- Brown, P., Marquard, J., Amster, B., Romoser, M., Friderici, J., Goff, S. & Fisher, D. (2014), ‘What do Physicians Read (and Ignore) in Electronic Progress Notes?’, *Applied Clinical Informatics* **5**(2), 430–444.
URL: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4081746/>
- Cambria, E., Poria, S., Gelbukh, A. & Thelwall, M. (2017), ‘Sentiment Analysis Is a Big Suitcase’, *IEEE Intelligent Systems* **32**(6), 74–80.
URL: <http://ieeexplore.ieee.org/document/8267597/>
- Carlini, N., Tramèr, F., Wallace, E., Jagielski, M., Herbert-Voss, A., Lee, K., Roberts, A., Brown, T. B., Song, D., Erlingsson, V., Oprea, A. & Raffel, C. (2020), Extracting Training Data from Large Language Models.
URL: <https://api.semanticscholar.org/CorpusID:229156229>
- Carrillo-de Albornoz, J., Vidal, J. R. & Plaza, L. (2018), ‘Feature engineering for sentiment analysis in e-health forums’, *PLOS ONE* **13**(11), e0207996. Publisher: Public Library of Science.
URL: <https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0207996>
- Chary, M., Parikh, S., Manini, A. F., Boyer, E. W. & Radeos, M. (2019), ‘A Review of Natural Language Processing in Medical Education’, *Western Journal of Emergency Medicine* **20**(1), 78–86.
URL: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6324711/>
- Chee, B. W., Berlin, R. & Schatz, B. (2011), ‘Predicting Adverse Drug Events from Personal Health Messages’, *AMIA Annual Symposium Proceedings* **2011**, 217–226.
URL: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3243174/>
- Chintalapudi, N., Battineni, G., Canio, M. D., Sagaro, G. G. & Amenta, F. (2021), ‘Text mining with sentiment analysis on seafarers’ medical documents’, *International Journal of Information Management Data Insights* **1**(1), 100005.
URL: <https://www.sciencedirect.com/science/article/pii/S2667096820300057>
- Choi, G., Oh, S. & Kim, H. (2020), ‘Improving Document-Level Sentiment Classification Using Importance of Sentences’, *Entropy* **22**(12), 1336.
URL: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7761344/>

- Chowdhury, M., Lim, E., Higham, A., McKinnon, R., Ventoura, N., He, Y. & De Pennington, N. (2023), 'Can Large Language Models Safely Address Patient Questions Following Cataract Surgery?', in T. Naumann, A. Ben Abacha, S. Bethard, K. Roberts & A. Rumshisky, eds, 'Proceedings of the 5th Clinical Natural Language Processing Workshop', Association for Computational Linguistics, Toronto, Canada, pp. 131–137.
URL: <https://aclanthology.org/2023.clinicalnlp-1.17/>
- Churpek, M. M., Yuen, T. C., Winslow, C., Meltzer, D. O., Kattan, M. W. & Edelson, D. P. (2016), 'Multicenter Comparison of Machine Learning Methods and Conventional Regression for Predicting Clinical Deterioration on the Wards', *Critical Care Medicine* **44**(2), 368.
URL: https://journals.lww.com/ccmjournal/abstract/2016/02000/multicenter_comparison_of_machine_learning_methods.16.aspx
- Cioffi, J. (2000), 'Recognition of patients who require emergency assistance: A descriptive study', *Heart & Lung* **29**(4), 262–268.
URL: <https://www.sciencedirect.com/science/article/pii/S0147956300306148>
- Cohen, J. P., Morrison, P., Dao, L., Roth, K., Duong, T. & Ghassem, M. (2020), 'COVID-19 Image Data Collection: Prospective Predictions are the Future', *Machine Learning for Biomedical Imaging* **1**(December 2020 issue), 1–38.
URL: <https://www.melba-journal.org/papers/2020:002.html>
- Coiera, E. (2000), 'When Conversation Is Better Than Computation', *Journal of the American Medical Informatics Association* **7**(3), 277–286.
URL: <https://doi.org/10.1136/jamia.2000.0070277>
- Collins, S. A., Cato, K., Albers, D., Scott, K., Stetson, P. D., Bakken, S. & Vawdrey, D. K. (2013), 'Relationship Between Nursing Documentation and Patients' Mortality', *American Journal of Critical Care* **22**(4), 306–313.
URL: <https://doi.org/10.4037/ajcc2013426>
- Collins, S. A. & Vawdrey, D. K. (2012), '“Reading between the lines” of flow sheet data: nurses' optional documentation associated with cardiac arrest outcomes', *Applied Nursing Research* **25**(4), 251–257. Publisher: WB Saunders.
- Cuenca-Zaldívar, J. N., Torrente-Regidor, M., Martín-Losada, L., Fernández-De-Las-Peñas, C., Florencio, L. L., Sousa, P. A. & Palacios-Ceña, D. (2022), 'Exploring Sentiment and Care Management of Hospitalized Patients During the First Wave of the COVID-19 Pandemic Using Electronic Nursing Health Records: Descriptive Study', *JMIR Medical Informatics* **10**(5), e38308.
URL: <https://medinform.jmir.org/2022/5/e38308>

- Dai, A. M. & Le, Q. V. (2015), Semi-supervised Sequence Learning, in ‘Advances in Neural Information Processing Systems’, Vol. 28, Curran Associates, Inc.
URL: <https://proceedings.neurips.cc/paper/2015/hash/7137debd45ae4d0ab9aa953017286b20-Abstract.html>
- Dang, N. C., Moreno-García, M. N. & De la Prieta, F. (2020), ‘Sentiment Analysis Based on Deep Learning: A Comparative Study’, *Electronics* **9**(3), 483. Number: 3 Publisher: Multidisciplinary Digital Publishing Institute.
URL: <https://www.mdpi.com/2079-9292/9/3/483>
- Darwich, M., Mohd Noah, S. A., Omar, N. & Osman, N. A. (2019), ‘Corpus-Based Techniques for Sentiment Lexicon Generation: A Review’, *Journal of Digital Information Management* **17**(5), 296.
URL: <http://dline.info/fpaper/jdim/v17i5/jdimv17i54.pdf>
- De Brito, B. A. O., De Moraes Barroca Filho, I., De Lima, J. M. M., Gurgel, A. M. & Malaquias, R. (2024), Natural Language Processing in the Context of Cancer: A Systematic Review, in ‘2024 IEEE Symposium on Computers and Communications (ISCC)’, pp. 1–6. ISSN: 2642-7389.
URL: <https://ieeexplore.ieee.org/document/10733573>
- De Smedt, T. & Daelemans, W. (2012), ‘Pattern for python’, *The Journal of Machine Learning Research* **13**(null), 2063–2067.
- Denecke, K. & Deng, Y. (2015), ‘Sentiment analysis in medical settings: New opportunities and challenges’, *Artificial Intelligence in Medicine* **64**(1), 17–27.
URL: <https://www.sciencedirect.com/science/article/pii/S09333365715000299>
- Denecke, K. & Reichenpfader, D. (2023), ‘Sentiment analysis of clinical narratives: A scoping review’, *Journal of Biomedical Informatics* **140**, 104336.
URL: <https://www.sciencedirect.com/science/article/pii/S1532046423000576>
- Deng, L. & Wiebe, J. (2015), MPQA 3.0: An Entity/Event-Level Sentiment Corpus, in ‘NAACL-HLT 2015’, ACL, Denver, Colorado, pp. 1323–1328.
URL: <https://aclanthology.org/N15-1146>
- Deng, Y., Declerck, T., Lendvai, P. & Denecke, K. (2016), The Generation of a Corpus for Clinical Sentiment Analysis, in H. Sack, G. Rizzo, N. Steinmetz, D. Mladeníć, S. Auer & C. Lange, eds, ‘The Semantic Web’, Springer International Publishing, Cham, pp. 311–324.
- Deng, Y., Stoehr, M. & Denecke, K. (2014), Retrieving Attitudes: Sentiment Analysis from Clinical Narratives.
URL: <http://ceur-ws.org/Vol-1276/MedIR-SIGIR2014-03.pdf>

- Devlin, J., Chang, M.-W., Lee, K. & Toutanova, K. (2019), BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, in 'Proceedings of the 2019 Conference', ACL, Minneapolis, Minnesota, pp. 4171–4186.
URL: <http://aclweb.org/anthology/N19-1423>
- Elbers, D. C., La, J., Minot, J. R., Gramling, R., Brophy, M. T., Do, N. V., Fillmore, N. R., Dodds, P. S. & Danforth, C. M. (2023), 'Sentiment analysis of medical record notes for lung cancer patients at the Department of Veterans Affairs', *PloS One* **18**(1), e0280931.
- Esterle, L. (2022), Chapter 17 - Deep learning in multiagent systems, in A. Iosifidis & A. Tefas, eds, 'Deep Learning for Robot Perception and Cognition', Academic Press, pp. 435–460.
URL: <https://www.sciencedirect.com/science/article/pii/B9780323857871000221>
- Fleiss, J. L. (1971), 'Measuring nominal scale agreement among many raters', *Psychological Bulletin* **76**(5), 378–382. Place: US Publisher: American Psychological Association.
- Gandhi, A., Adhvaryu, K., Poria, S., Cambria, E. & Hussain, A. (2023), 'Multimodal sentiment analysis: A systematic review of history, datasets, multimodal fusion methods, applications, challenges and future directions', *Information Fusion* **91**, 424–444.
URL: <https://www.sciencedirect.com/science/article/pii/S1566253522001634>
- Gao, Q., Wang, D., Sun, P., Luan, X. & Wang, W. (2021), 'Sentiment Analysis Based on the Nursing Notes on In-Hospital 28-Day Mortality of Sepsis Patients Utilizing the MIMIC-III Database', *Computational and mathematical methods in medicine* .
- Goh, K. H., Wang, L., Yeow, A. Y. K., Poh, H., Li, K., Yeow, J. J. L. & Tan, G. Y. H. (2021), 'Artificial intelligence in sepsis early prediction and diagnosis using unstructured data in healthcare', *Nature Communications* **12**, 711.
URL: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7846756/>
- Goldberger, A. L., Amaral, L. A., Glass, L., Hausdorff, J. M., Ivanov, P. C., Mark, R. G., Mietus, J. E., Moody, G. B., Peng, C. K. & Stanley, H. E. (2000), 'PhysioBank, PhysioToolkit, and PhysioNet: components of a new research resource for complex physiologic signals', *Circulation* **101**(23), E215–220.
- Grissette, H. & Nfaoui, E. H. (2021), 'Deep associative learning approach for bio-medical sentiment analysis utilizing unsupervised representation from large-scale patients' narratives', *Personal and Ubiquitous Computing* .
URL: <https://doi.org/10.1007/s00779-021-01595-4>
- Gu, Y., Tinn, R., Cheng, H., Lucas, M., Usuyama, N., Liu, X., Naumann, T., Gao, J. & Poon, H. (2022), 'Domain-Specific Language Model Pretraining for Biomedical Natural Language Processing', *ACM Transactions on*

- Computing for Healthcare* **3**(1), 1–23. arXiv:2007.15779 [cs].
URL: <http://arxiv.org/abs/2007.15779>
- Guo, X., Yu, W. & Wang, X. (2021), ‘An Overview on Fine-grained Text Sentiment Analysis: Survey and Challenges’, *Journal of Physics: Conference Series* **1757**(1), 012038. Publisher: IOP Publishing.
URL: <https://dx.doi.org/10.1088/1742-6596/1757/1/012038>
- Gupta, I., Chatterjee, I. & Gupta, N. (2022), Sentiment Analysis of COVID-19 Tweets, in ‘2022 1st International Conference on Informatics (ICI)’, pp. 229–231.
URL: <https://ieeexplore.ieee.org/document/9786887>
- Hao, B., Zhu, H. & Paschalidis, I. (2020), Enhancing Clinical BERT Embedding using a Biomedical Knowledge Base, in ‘Proceedings of the 28th International Conference on Computational Linguistics’, International Committee on Computational Linguistics, Barcelona, Spain (Online), pp. 657–661.
URL: <https://www.aclweb.org/anthology/2020.coling-main.57>
- Hoang, M., Bihorac, O. A. & Rouces, J. (2019), Aspect-Based Sentiment Analysis using BERT, in ‘Proceedings of the 22nd Nordic Conference on Computational Linguistics’, Linköping University Electronic Press, Turku, Finland, pp. 187–196.
URL: <https://aclanthology.org/W19-6120>
- Holderness, E., Cawkwell, P., Bolton, K., Pustejovsky, J. & Hall, M.-H. (2019), Distinguishing Clinical Sentiment: The Importance of Domain Adaptation in Psychiatric Patient Health Records, in ‘Proceedings of the 2nd Clinical Natural Language Processing Workshop’, ACL, Minneapolis, Minnesota, USA, pp. 117–123.
URL: <http://aclweb.org/anthology/W19-1915>
- Hu, M. & Liu, B. (2004), Mining and summarizing customer reviews, in ‘Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining’, KDD ’04, Association for Computing Machinery, New York, NY, USA, pp. 168–177.
URL: <https://doi.org/10.1145/1014052.1014073>
- Huang, K., Gray, T. F., Romero-Brufau, S., Tulskey, J. A. & Lindvall, C. (2021), ‘Using nursing notes to improve clinical outcome prediction in intensive care patients: A retrospective cohort study’, *Journal of the American Medical Informatics Association: JAMIA* **28**(8), 1660–1666.
- Huang, K., Singh, A., Chen, S., Moseley, E., Deng, C.-Y., George, N. & Lindvall, C. (2020), Clinical XLNet: Modeling Sequential Clinical Notes and Predicting Prolonged Mechanical Ventilation, in ‘Proceedings of the 3rd Clinical Natural Language Processing Workshop’, Association for Computational Linguistics, Online, pp. 94–100.
URL: <https://aclanthology.org/2020.clinicalnlp-1.11>

- Huang, K., Singh, A., Chen, S., Moseley, E. T., Deng, C.-y., George, N. & Lindvall, C. (2019), ‘Clinical XLNet: Modeling Sequential Clinical Notes and Predicting Prolonged Mechanical Ventilation’, *arXiv:1912.11975 [cs]*.
arXiv: 1912.11975.
URL: <http://arxiv.org/abs/1912.11975>
- Hultman, G. M., Marquard, J. L., Lindemann, E., Arsoniadis, E., Pakhomov, S. & Melton, G. B. (2019), ‘Challenges and Opportunities to Improve the Clinician Experience Reviewing Electronic Progress Notes’, *Applied Clinical Informatics* **10**(03), 446–453.
URL: <http://www.thieme-connect.de/DOI/DOI?10.1055/s-0039-1692164>
- Hutto, C. & Gilbert, E. (2014), ‘VADER: A Parsimonious Rule-Based Model for Sentiment Analysis of Social Media Text’, *Proceedings of the International AAAI Conference on Web and Social Media* **8**(1), 216–225.
URL: <https://ojs.aaai.org/index.php/ICWSM/article/view/14550>
- Jensen, K., Soguero-Ruiz, C., Oyvind Mikalsen, K., Lindsetmo, R.-O., Kouskoumvekaki, I., Girolami, M., Olav Skrovseth, S. & Augestad, K. M. (2017), ‘Analysis of free text in electronic health records for identification of cancer patient trajectories’, *Scientific Reports* **7**(1), 46226. Publisher: Nature Publishing Group.
URL: <https://www.nature.com/articles/srep46226>
- Jin, D., Pan, E., Oufattole, N., Weng, W.-H., Fang, H. & Szolovits, P. (2021), ‘What Disease Does This Patient Have? A Large-Scale Open Domain Question Answering Dataset from Medical Exams’, *Applied Sciences* **11**(14), 6421. Publisher: Multidisciplinary Digital Publishing Institute.
URL: <https://www.mdpi.com/2076-3417/11/14/6421>
- Johnson, A. E. W., Pollard, T. J., Shen, L., Lehman, L.-w. H., Feng, M., Ghassemi, M., Moody, B., Szolovits, P., Anthony Celi, L. & Mark, R. G. (2016), ‘MIMIC-III, a freely accessible critical care database’, *Scientific Data* **3**(1), 160035. Publisher: Nature Publishing Group.
URL: <https://www.nature.com/articles/sdata201635>
- Johnson, A., Pollard, T. & Mark, R. (2015), ‘MIMIC-III Clinical Database’.
URL: <https://physionet.org/content/mimiciii/1.4/>
- Johnson, W. G., Brennan, T. A., Newhouse, J. P., Leape, L. L., Lawthers, A. G., Hiatt, H. H. & Weiler, P. C. (1992), ‘The economic consequences of medical injuries. Implications for a no-fault insurance plan’, *JAMA* **267**(18), 2487–2492.
- Joseph, J. & Tolley, N. (2015), ‘Are consultants more likely to decide to operate than junior doctors?’, *The Bulletin of the Royal College of Surgeons of England* **97**(4), 174–177. Publisher: The Royal College of Surgeons of England.
URL: <https://publishing.rcseng.ac.uk/doi/full/10.1308/147363515X14134529302263>

- Kanayama, H. & Nasukawa, T. (2006), Fully Automatic Lexicon Expansion for Domain-oriented Sentiment Analysis, in D. Jurafsky & E. Gaussier, eds, 'Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing', Association for Computational Linguistics, Sydney, Australia, pp. 355–363.
URL: <https://aclanthology.org/W06-1642/>
- Kasthurirathne, S. N., Dixon, B. E., Gichoya, J., Xu, H., Xia, Y., Mamlin, B. & Grannis, S. J. (2016), 'Toward better public health reporting using existing off the shelf approaches: A comparison of alternative cancer detection approaches using plaintext medical data and non-dictionary based feature selection', *Journal of Biomedical Informatics* **60**, 145–152.
URL: <https://www.sciencedirect.com/science/article/pii/S1532046416000095>
- Ke, P., Ji, H., Liu, S., Zhu, X. & Huang, M. (2020), SentiLARE: Sentiment-Aware Language Representation Learning with Linguistic Knowledge, in 'Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)', Association for Computational Linguistics, Online, pp. 6975–6988.
- Kennedy, C. J., Chiu, C., Chapman, A. C., Gologorskaya, O., Farhan, H., Han, M., Hodgson, M., Lazzareschi, D., Ashana, D., Lee, S., Smith, A. K., Espejo, E., Boscardin, J., Pirracchio, R. & Cobert, J. (2023), 'Negativity and Positivity in the ICU: Exploratory Development of Automated Sentiment Capture in the Electronic Health Record', *Critical Care Explorations* **5**(10), e0960.
- Khurana, D., Koli, A., Khatter, K. & Singh, S. (2023), 'Natural language processing: state of the art, current trends and challenges', *Multimedia Tools and Applications* **82**(3), 3713–3744.
URL: <https://doi.org/10.1007/s11042-022-13428-4>
- Kong, H.-J. (2019), 'Managing Unstructured Big Data in Healthcare System', *Healthcare Informatics Research* **25**(1), 1–2. Publisher: Korean Society of Medical Informatics.
URL: <http://e-hir.org/journal/view.php?id=10.4258/hir.2019.25.1.1>
- Kramer, A. A., Dasta, J. F. & Kane-Gill, S. L. (2017), 'The Impact of Mortality on Total Costs Within the ICU', *Critical Care Medicine* **45**(9), 1457–1463.
- Kumar, V., Bajpai, R. & Roy, R. B. (2022), 'Clinical Notes Mining for Post Discharge Mortality Prediction', *IETE Technical Review* **39**(4), 953–959. Publisher: Taylor & Francis .eprint: <https://doi.org/10.1080/02564602.2021.1936224>.
URL: <https://doi.org/10.1080/02564602.2021.1936224>
- Kung, T. H., Cheatham, M., Medenilla, A., Sillos, C., Leon, L. D., Elepaño, C., Madriaga, M., Aggabao, R., Diaz-Candido, G., Maningo, J. & Tseng, V. (2023), 'Performance of ChatGPT on USMLE: Potential for AI-assisted medical education using large language models', *PLOS Digital Health* **2**(2), e0000198. Publisher: Public Library of Science.
URL: <https://journals.plos.org/digitalhealth/article?id=10.1371/journal.pdig.0000198>

- Landis, J. R. & Koch, G. G. (1977), 'The Measurement of Observer Agreement for Categorical Data', *Biometrics* **33**(1), 159–174. Publisher: International Biometric Society.
URL: <https://www.jstor.org/stable/2529310>
- Leaman, R., Khare, R. & Lu, Z. (2015), 'Challenges in clinical natural language processing for automated disorder normalization', *Journal of Biomedical Informatics* **57**, 28–37.
URL: <https://www.sciencedirect.com/science/article/pii/S1532046415001501>
- Leape, L. L., Brennan, T. A., Laird, N., Lawthers, A. G., Localio, A. R., Barnes, B. A., Hebert, L., Newhouse, J. P., Weiler, P. C. & Hiatt, H. (1991), 'The nature of adverse events in hospitalized patients. Results of the Harvard Medical Practice Study II', *The New England Journal of Medicine* **324**(6), 377–384.
- Lehman, E., Jain, S., Pichotta, K., Goldberg, Y. & Wallace, B. (2021), Does BERT Pretrained on Clinical Notes Reveal Sensitive Data?, in K. Toutanova, A. Rumshisky, L. Zettlemoyer, D. Hakkani-Tur, I. Beltagy, S. Bethard, R. Cotterell, T. Chakraborty & Y. Zhou, eds, 'Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies', Association for Computational Linguistics, Online, pp. 946–959.
URL: <https://aclanthology.org/2021.naacl-main.73/>
- Lehman, E. & Johnson, A. (2023), 'Clinical-T5: Large Language Models Built Using MIMIC Clinical Text'.
URL: <https://physionet.org/content/clinical-t5/>
- Lewandowska, K., Weisbrot, M., Cieloszyk, A., Medrzycka-Dabrowska, W., Krupa, S. & Ozga, D. (2020), 'Impact of Alarm Fatigue on the Work of Nurses in an Intensive Care Environment—A Systematic Review', *International Journal of Environmental Research and Public Health* **17**(22), 8409.
URL: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7697990/>
- Li, Q., Li, X., Du, Y., Fan, Y. & Chen, X. (2022), 'A New Sentiment-Enhanced Word Embedding Method for Sentiment Analysis', *undefined* .
URL: <https://www.semanticscholar.org/paper/A-New-Sentiment-Enhanced-Word-Embedding-Method-for-Li-Li/b2be2fbf07949a307238658f5784f10915336750>
- Lin, C.-T., McKenzie, M., Pell, J. & Caplan, L. (2013), 'Health Care Provider Satisfaction With a New Electronic Progress Note Format: SOAP vs APSO Format', *JAMA Internal Medicine* **173**(2), 160–162.
URL: <https://doi.org/10.1001/2013.jamainternmed.474>
- Lin, C.-Y. (2004), ROUGE: A Package for Automatic Evaluation of Summaries, in 'Text Summarization Branches Out', Association for Computational Linguistics, Barcelona, Spain, pp. 74–81.
URL: <https://aclanthology.org/W04-1013/>

- Liu, B. (2017), Many Facets of Sentiment Analysis, in E. Cambria, D. Das, S. Bandyopadhyay & A. Feraco, eds, 'A Practical Guide to Sentiment Analysis', Springer International Publishing, Cham, pp. 11–39.
URL: https://doi.org/10.1007/978-3-319-55394-8_2
- Liu, H. & Hosseini, A. (2025), 'SASLS: Semantic analysis of sentiment in social networks using Lexicon-Based methodology and Semi-Supervised sentiment annotation', *Ain Shams Engineering Journal* **16**(6), 103378.
URL: <https://www.sciencedirect.com/science/article/pii/S2090447925001194>
- Liu, M., Beare, R., Collyer, T., Andrew, N. & Srikanth, V. (2023), Leveraging Natural Language Processing and Clinical Notes for Dementia Detection, in T. Naumann, A. Ben Abacha, S. Bethard, K. Roberts & A. Rumshisky, eds, 'Proceedings of the 5th Clinical Natural Language Processing Workshop', Association for Computational Linguistics, Toronto, Canada, pp. 150–155.
URL: <https://aclanthology.org/2023.clinicalnlp-1.20/>
- Liu, P., Yuan, W., Fu, J., Jiang, Z., Hayashi, H. & Neubig, G. (2021), 'Pre-train, Prompt, and Predict: A Systematic Survey of Prompting Methods in Natural Language Processing'. arXiv:2107.13586 [cs].
URL: <http://arxiv.org/abs/2107.13586>
- Liu, R., Greenstein, J. L., Sarma, S. V. & Winslow, R. L. (2019), 'Natural Language Processing of Clinical Notes for Improved Early Prediction of Septic Shock in the ICU', *Annual International Conference of the IEEE Engineering in Medicine and Biology Society. IEEE Engineering in Medicine and Biology Society. Annual International Conference* **2019**, 6103–6108.
- Liu, Z., Yang, Y., Song, H. & Luo, J. (2022), 'A prediction model with measured sentiment scores for the risk of in-hospital mortality in acute pancreatitis: a retrospective cohort study', *Annals of Translational Medicine* **10**(12), 676–676. Number: 12 Publisher: AME Publishing Company.
URL: <https://atm.amegroups.org/article/view/96668>
- Locke, S., Bashall, A., Al-Adely, S., Moore, J., Wilson, A. & Kitchen, G. B. (2021), 'Natural language processing in medicine: A review', *Trends in Anaesthesia and Critical Care* **38**, 4–9.
URL: <https://www.sciencedirect.com/science/article/pii/S2210844021000411>
- Long, J. (2023), 'Large Language Model Guided Tree-of-Thought'.
URL: <https://arxiv.org/abs/2305.08291v1>
- Loria, S. (2020), 'TextBlob: Simplified Text Processing'.
URL: <https://github.com/slوريا/TextBlob>
- Loshchilov, I. & Hutter, F. (2019), Decoupled Weight Decay Regularization, in '7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019', OpenReview.net.
URL: <https://openreview.net/forum?id=Bkg6RiCqY7>

- Lu, H., Ehwerhemuepha, L. & Rakovski, C. (2022), ‘A comparative study on deep learning models for text classification of unstructured medical notes with various levels of class imbalance’, *BMC Medical Research Methodology* **22**(1), 181.
URL: <https://doi.org/10.1186/s12874-022-01665-y>
- Lu, Q., Dou, D. & Nguyen, T. (2022), ClinicalT5: A Generative Language Model for Clinical Text, in ‘Findings of the Association for Computational Linguistics: EMNLP 2022’, Association for Computational Linguistics, Abu Dhabi, United Arab Emirates, pp. 5436–5443.
URL: <https://aclanthology.org/2022.findings-emnlp.398>
- Lynch, I. P., Roberts, P. E., Keebler, J. R., Guttman, O. & Greulich, P. E. (2017), ‘Error Detection and Reporting in the Intensive Care Unit: Progress, Barriers, and Future Direction’, *Current Anesthesiology Reports* **7**(3), 310–319.
URL: <https://doi.org/10.1007/s40140-017-0228-3>
- Manning, C., Surdeanu, M., Bauer, J., Finkel, J., Bethard, S. & McClosky, D. (2014), The Stanford CoreNLP Natural Language Processing Toolkit, in ‘Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations’, Association for Computational Linguistics, Baltimore, Maryland, pp. 55–60.
URL: <https://aclanthology.org/P14-5010>
- Mao, Y., Liu, Q. & Zhang, Y. (2024), ‘Sentiment analysis methods, applications, and challenges: A systematic literature review’, *Journal of King Saud University - Computer and Information Sciences* **36**(4), 102048.
URL: <https://www.sciencedirect.com/science/article/pii/S131915782400137X>
- McCoy, T., Castro, V., Cagan, A., Roberson, A., Kohane, I. & Perlis, R. (2015), ‘Sentiment Measured in Hospital Discharge Notes Is Associated with Readmission and Mortality Risk: An Electronic Health Record Study’, *PLoS ONE* **10**(8).
- Md Suhaimin, M. S., Ahmad Hijazi, M. H., Mounq, E. G., Nohuddin, P. N. E., Chua, S. & Coenen, F. (2023), ‘Social media sentiment analysis and opinion mining in public security: Taxonomy, trend analysis, issues and future directions’, *Journal of King Saud University - Computer and Information Sciences* **35**(9), 101776.
URL: <https://www.sciencedirect.com/science/article/pii/S1319157823003300>
- Mikolov, T., Chen, K., Corrado, G. & Dean, J. (2013), Efficient Estimation of Word Representations in Vector Space.
URL: <https://api.semanticscholar.org/CorpusID:5959482>
- Mishra, P. & Desetty, R. T. (2023), NewAgeHealthWarriors at MEDIQA-Chat 2023 Task A: Summarizing Short Medical Conversation with Transformers, in T. Naumann, A. Ben Abacha, S. Bethard, K. Roberts

& A. Rumshisky, eds, ‘Proceedings of the 5th Clinical Natural Language Processing Workshop’, Association for Computational Linguistics, Toronto, Canada, pp. 414–421.

URL: <https://aclanthology.org/2023.clinicalnlp-1.44/>

Mohammad, S. M. & Turney, P. D. (2013), ‘CROWDSOURCING A WORD–EMOTION ASSOCIATION LEXICON’, *Computational Intelligence* **29**(3), 436–465. Publisher: John Wiley & Sons, Ltd.

URL: <https://onlinelibrary.wiley.com/doi/10.1111/j.1467-8640.2012.00460.x>

Moradi, M., Blagec, K., Haberl, F. & Samwald, M. (2021), ‘GPT-3 Models are Poor Few-Shot Learners in the Biomedical Domain’, *ArXiv*.

URL: <https://www.semanticscholar.org/paper/GPT-3-Models-are-Poor-Few-Shot-Learners-in-the-Moradi-Blagec/2122ed4a82bc7d8affc5f7ae5026d174ea34ea52>

Murdoch, T. B. & Detsky, A. S. (2013), ‘The Inevitable Application of Big Data to Health Care’, *JAMA* **309**(13), 1351–1352.

URL: <https://doi.org/10.1001/jama.2013.393>

Nagoor, S., Hederman, L., Koidl, K. & Martin-Loeches, I. (2024), Distinguishing Clinical Sentiment in Intensive Care Unit Clinical Notes, in ‘2024 IEEE 37th International Symposium on Computer-Based Medical Systems (CBMS)’, pp. 249–259. ISSN: 2372-9198.

URL: <https://ieeexplore.ieee.org/document/10600847>

Naseem, U., Razzak, I., Khushi, M., Eklund, P. W. & Kim, J. (2021), ‘COVIDSenti: A Large-Scale Benchmark Twitter Data Set for COVID-19 Sentiment Analysis’, *IEEE Transactions on Computational Social Systems* **8**(4), 1003–1015.

URL: <https://ieeexplore.ieee.org/document/9340540>

Neshan, S. A. S. & Akbari, R. (2020), A Combination of Machine Learning and Lexicon Based Techniques for Sentiment Analysis, in ‘2020 6th International Conference on Web Research (ICWR)’, pp. 8–14.

Nguyen Hong Duc, P., Torterotot, M., Samaran, F., White, P. R., Gérard, O., Adam, O. & Cazau, D. (2021), ‘Assessing inter-annotator agreement from collaborative annotation campaign in marine bioacoustics’, *Ecological Informatics* **61**, 101185.

URL: <https://www.sciencedirect.com/science/article/pii/S1574954120301357>

Nguyen, T. D., Diep-Phuong Nguyen, L. & Cao, T. (2017), Sentiment analysis on medical text using combination of machine learning and SO-CAL scoring, in ‘2017 21st Asia Pacific Symposium on Intelligent and Evolutionary Systems (IES)’, pp. 49–54.

Nielsen, F. (2011), ‘A new ANEW: Evaluation of a word list for sentiment analysis in microblogs: 1st Workshop on Making Sense of Microposts’, *Proceedings of the ESWC2011 Workshop on ‘Making Sense of Microposts’*:

Big things come in small packages pp. 93–98.

URL: <https://www.scopus.com/pages/publications/84874354132>

Ntiamoah, M., Xavier, T. & Lambert, J. (2024), ‘Sentiment Analysis of Patient- and Family-Related Sepsis Events: Exploratory Study’, *JMIR Nursing* **7**(1), e51720. Company: JMIR Nursing Distributor: JMIR Nursing Institution: JMIR Nursing Label: JMIR Nursing Publisher: JMIR Publications Inc., Toronto, Canada.

URL: <https://nursing.jmir.org/2024/1/e51720>

Nuthakki, S., Neela, S., Gichoya, J. W. & Purkayastha, S. (2019), ‘Natural language processing of MIMIC-III clinical notes for identifying diagnosis and procedures with neural networks’, *arXiv:1912.12397 [cs]*. arXiv: 1912.12397.

URL: <http://arxiv.org/abs/1912.12397>

Ohsfeldt, R. L., Choong, C. K.-C., Mc Collam, P. L., Abedtash, H., Kelton, K. A. & Burge, R. (2021), ‘Inpatient Hospital Costs for COVID-19 Patients in the United States’, *Advances in Therapy* **38**(11), 5557–5595.

Ohtsuka, T., Kajiwar, T., Tanikawa, C., Shimizu, Y., Nagahara, H. & Ninomiya, T. (2023), Automated Orthodontic Diagnosis from a Summary of Medical Findings, in T. Naumann, A. Ben Abacha, S. Bethard, K. Roberts & A. Rumshisky, eds, ‘Proceedings of the 5th Clinical Natural Language Processing Workshop’, Association for Computational Linguistics, Toronto, Canada, pp. 156–160.

URL: <https://aclanthology.org/2023.clinicalnlp-1.21/>

Ojeda-Hernández, M., López-Rodríguez, D. & Mora, (2023), ‘Lexicon-based sentiment analysis in texts using Formal Concept Analysis’, *International Journal of Approximate Reasoning* **155**, 104–112.

URL: <https://www.sciencedirect.com/science/article/pii/S0888613X23000130>

Papineni, K., Roukos, S., Ward, T. & Zhu, W.-J. (2002), Bleu: a Method for Automatic Evaluation of Machine Translation, in P. Isabelle, E. Charniak & D. Lin, eds, ‘Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics’, Association for Computational Linguistics, Philadelphia, Pennsylvania, USA, pp. 311–318.

URL: <https://aclanthology.org/P02-1040/>

Peng, C., Yang, X., Chen, A., Smith, K. E., PourNejatian, N., Costa, A. B., Martin, C., Flores, M. G., Zhang, Y., Magoc, T., Lipori, G., Mitchell, D. A., Ospina, N. S., Ahmed, M. M., Hogan, W. R., Shenkman, E. A., Guo, Y., Bian, J. & Wu, Y. (2023), ‘A study of generative large language model for medical research and healthcare’, *npj Digital Medicine* **6**(1), 1–10. Number: 1 Publisher: Nature Publishing Group.

URL: <https://www.nature.com/articles/s41746-023-00958-w>

Peng, Y., Yan, S. & Lu, Z. (2019), Transfer Learning in Biomedical Natural Language Processing: An Evaluation of BERT and ELMo on Ten Benchmarking Datasets, in D. Demner-Fushman, K. B. Cohen, S. Ananiadou &

- J. Tsujii, eds, 'Proceedings of the 18th BioNLP Workshop and Shared Task', Association for Computational Linguistics, Florence, Italy, pp. 58–65.
URL: <https://aclanthology.org/W19-5006/>
- Pennington, J., Socher, R. & Manning, C. (2014), GloVe: Global Vectors for Word Representation, in 'EMNLP 2014', Association for Computational Linguistics, Doha, Qatar, pp. 1532–1543.
URL: <https://aclanthology.org/D14-1162>
- Perera, S. & Sheth, A. (2015), Implicit Information Extraction from Clinical Notes, in '2015 International Conference on Healthcare Informatics', pp. 498–498.
URL: <https://ieeexplore.ieee.org/document/7349750>
- Peters, M. E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K. & Zettlemoyer, L. (2018), 'Deep contextualized word representations'. arXiv:1802.05365 [cs].
URL: <http://arxiv.org/abs/1802.05365>
- Phan, L. N., Anibal, J. T., Tran, H., Chanana, S., Bahadroglu, E., Peltekian, A. & Altan-Bonnet, G. (2021), 'SciFive: a text-to-text transformer model for biomedical literature'. arXiv:2106.03598 [cs].
URL: <http://arxiv.org/abs/2106.03598>
- Pinsky, M. R., Bedoya, A., Bihorac, A., Celi, L., Churpek, M., Economou-Zavlanos, N. J., Elbers, P., Saria, S., Liu, V., Lyons, P. G., Shickel, B., Toral, P., Tscholl, D. & Clermont, G. (2024), 'Use of artificial intelligence in critical care: opportunities and obstacles', *Critical Care* **28**(1), 113.
URL: <https://doi.org/10.1186/s13054-024-04860-z>
- Podder, V., Lew, V. & Ghassemzadeh, S. (2023), SOAP Notes, in 'StatPearls', StatPearls Publishing, Treasure Island (FL).
URL: <http://www.ncbi.nlm.nih.gov/books/NBK482263/>
- Pool, J., Indulska, M. & Sadiq, S. (2024), 'Large language models and generative AI in telehealth: a responsible use lens', *Journal of the American Medical Informatics Association: JAMIA* **31**(9), 2125–2136.
- Popescu, M., Ștefan, O. M., Ștefan, M., Văleanu, L. & Tomescu, D. (2022), 'ICU-Associated Costs during the Fourth Wave of the COVID-19 Pandemic in a Tertiary Hospital in a Low-Vaccinated Eastern European Country', *International Journal of Environmental Research and Public Health* **19**(3), 1781.
- Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W. & Liu, P. J. (2020), 'Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer'. arXiv:1910.10683 [cs, stat].
URL: <http://arxiv.org/abs/1910.10683>

- Reimers, N. & Gurevych, I. (2019), Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks, *in* ‘Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)’, Association for Computational Linguistics, Hong Kong, China, pp. 3980–3990.
URL: <https://www.aclweb.org/anthology/D19-1410>
- Ribeiro, F. N., Araújo, M., Gonçalves, P., André Gonçalves, M. & Benevenuto, F. (2016), ‘SentiBench - a benchmark comparison of state-of-the-practice sentiment analysis methods’, *EPJ Data Science* **5**(1), 1–29. Number: 1 Publisher: SpringerOpen.
URL: <https://epjdatascience.springeropen.com/articles/10.1140/epjds/s13688-016-0085-1>
- Rinker, T. (2021), ‘sentimentr: Calculate Text Polarity Sentiment’. original-date: 2015-08-16T00:55:55Z.
URL: <https://github.com/trinker/sentimentr>
- Rosenbloom, S. T., Denny, J. C., Xu, H., Lorenzi, N., Stead, W. W. & Johnson, K. B. (2011), ‘Data from clinical notes: a perspective on the tension between structure and flexible documentation’, *Journal of the American Medical Informatics Association* **18**(2), 181–186.
URL: <https://doi.org/10.1136/jamia.2010.007237>
- Salas-Zárate, M. d. P., Medina-Moreira, J., Lagos-Ortiz, K., Luna-Aveiga, H., Rodríguez-García, M. & Valencia-García, R. (2017), ‘Sentiment Analysis on Tweets about Diabetes: An Aspect-Level Approach’, *Computational and Mathematical Methods in Medicine* **2017**, e5140631. Publisher: Hindawi.
URL: <https://www.hindawi.com/journals/cmmm/2017/5140631/>
- Sarker, A., Molla, D. & Paris, C. (2011), Outcome Polarity Identification of Medical Papers, *in* ‘Proceedings of the Australasian Language Technology Association Workshop 2011’, Canberra, Australia, pp. 105–114.
URL: <https://aclanthology.org/U11-1014>
- Schmidtova, P., Mahamood, S., Balloccu, S., Dusek, O., Gatt, A., Gkatzia, D., Howcroft, D. M., Platek, O. & Sivaprasad, A. (2024), Automatic Metrics in Natural Language Generation: A survey of Current Evaluation Practices, *in* S. Mahamood, N. L. Minh & D. Ippolito, eds, ‘Proceedings of the 17th International Natural Language Generation Conference’, Association for Computational Linguistics, Tokyo, Japan, pp. 557–583.
URL: <https://aclanthology.org/2024.inlg-main.44/>
- Seinen, T. M., Kors, J. A., Mulligen, E. M. v. & Rijnbeek, P. R. (2025), ‘Using Structured Codes and Free-Text Notes to Measure Information Complementarity in Electronic Health Records: Feasibility and Validation Study’, *Journal of Medical Internet Research* **27**(1), e66910. Company: Journal of Medical Internet Research Distributor: Journal of Medical Internet Research Institution: Journal of Medical Internet Research Label: Journal of Medical Internet Research Publisher: JMIR Publications Inc., Toronto, Canada.
URL: <https://www.jmir.org/2025/1/e66910>

- Shapiro, M. & Melendez, M. (2010), Billing for Bedside Procedures, *in* 'Bedside Procedures for the Intensivist', pp. 323–331. Journal Abbreviation: Bedside Procedures for the Intensivist.
- Shin, B., Chokshi, F. H., Lee, T. & Choi, J. D. (2017), Classification of radiology reports using neural attention models, *in* '2017 IJCNN', pp. 4363–4370. ISSN: 2161-4407.
URL: <https://ieeexplore.ieee.org/document/7966408>
- Shinn, N., Cassano, F., Berman, E., Gopinath, A., Narasimhan, K. & Yao, S. (2023), 'Reflexion: Language Agents with Verbal Reinforcement Learning'. arXiv:2303.11366 [cs].
URL: <http://arxiv.org/abs/2303.11366>
- Simpson, K. R. & Lyndon, A. (2019), 'False Alarms and Overmonitoring: Major Factors in Alarm Fatigue Among Labor Nurses', *Journal of Nursing Care Quality* **34**(1), 66.
URL: https://journals.lww.com/jncqjournal/abstract/2019/01000/false_alarms_and_overmonitoring_major_factors_in.12.aspx
- Singh, G., Pan, Y., Andres-Ferrer, J., Del-Agua, M., Diehl, F., Pinto, J. & Vozila, P. (2023), Large Scale Sequence-to-Sequence Models for Clinical Note Generation from Patient-Doctor Conversations, *in* T. Naumann, A. Ben Abacha, S. Bethard, K. Roberts & A. Rumshisky, eds, 'Proceedings of the 5th Clinical Natural Language Processing Workshop', Association for Computational Linguistics, Toronto, Canada, pp. 138–143.
URL: <https://aclanthology.org/2023.clinicalnlp-1.18/>
- Singhal, K., Tu, T., Gottweis, J., Sayres, R., Wulczyn, E., Hou, L., Clark, K., Pfohl, S., Cole-Lewis, H., Neal, D., Schaekermann, M., Wang, A., Amin, M., Lachgar, S., Mansfield, P., Prakash, S., Green, B., Dominowska, E., Arcas, B. A. y., Tomasev, N., Liu, Y., Wong, R., Semturs, C., Mahdavi, S. S., Barral, J., Webster, D., Corrado, G. S., Matias, Y., Azizi, S., Karthikesalingam, A. & Natarajan, V. (2023), 'Towards Expert-Level Medical Question Answering with Large Language Models'. arXiv:2305.09617 [cs].
URL: <http://arxiv.org/abs/2305.09617>
- Sivarajkumar, S. & Wang, Y. (2023), 'HealthPrompt: A Zero-shot Learning Paradigm for Clinical Natural Language Processing', *AMIA Annual Symposium Proceedings* **2022**, 972–981.
URL: <https://pmc.ncbi.nlm.nih.gov/articles/PMC10148337/>
- Sodhi, K. & Chanchalani, G. (2022), 'Critical Care Medicine – Not Just a Profession, but a Passion, a Commitment!!', *International Journal of Applied and Basic Medical Research* **12**(3), 155–156.
URL: <https://pmc.ncbi.nlm.nih.gov/articles/PMC9484507/>
- Spinczyk, D., Nabrdalik, K. & Rojewska, K. (2018), 'Computer aided sentiment analysis of anorexia nervosa patients' vocabulary', *BioMedical Engineering OnLine* **17**(1), 19.
URL: <https://doi.org/10.1186/s12938-018-0451-2>

- Su, P. & Vijay-Shanker, K. (2022), ‘Investigation of improving the pre-training and fine-tuning of BERT model for biomedical relation extraction’, *BMC Bioinformatics* **23**, 120.
URL: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8978438/>
- Su, X., Miller, T., Ding, X., Afshar, M. & Dligach, D. (2021), ‘Classifying Long Clinical Documents with Pre-trained Transformers’, *ArXiv* .
URL: <https://www.semanticscholar.org/paper/Classifying-Long-Clinical-Documents-with-Su-Miller/161fbedea6237c7aadabcc2d373bc440fcec60b6>
- Sun, C., Huang, L. & Qiu, X. (2019), Utilizing BERT for Aspect-Based Sentiment Analysis via Constructing Auxiliary Sentence, in J. Burstein, C. Doran & T. Solorio, eds, ‘Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)’, Association for Computational Linguistics, Minneapolis, Minnesota, pp. 380–385.
URL: <https://aclanthology.org/N19-1035/>
- Sun, C., Qiu, X., Xu, Y. & Huang, X. (2019), How to Fine-Tune BERT for Text Classification?, in M. Sun, X. Huang, H. Ji, Z. Liu & Y. Liu, eds, ‘Chinese Computational Linguistics’, Springer International Publishing, Cham, pp. 194–206.
- Syloypavan, A., Sleeman, D., Wu, H. & Sim, M. (2023), ‘The impact of inconsistent human annotations on AI driven clinical decision making’, *npj Digital Medicine* **6**(1), 26. Publisher: Nature Publishing Group.
URL: <https://www.nature.com/articles/s41746-023-00773-3>
- Szep, M., Rueckert, D., Eisenhart-Rothe, R. v. & Hinterwimmer, F. (2024), ‘A Practical Guide to Fine-tuning Language Models with Limited Data’. arXiv:2411.09539 [cs].
URL: <http://arxiv.org/abs/2411.09539>
- Taboada, M., Brooke, J., Tofiloski, M., Voll, K. & Stede, M. (2011), ‘Lexicon-Based Methods for Sentiment Analysis’, *Computational Linguistics* **37**(2), 267–307.
URL: <https://doi.org/10.1162/COLI-a-00049>
- Taylor, N., Zhang, Y., Joyce, D. W., Gao, Z., Kormilitzin, A. & Nevado-Holgado, A. (2024), ‘Clinical Prompt Learning With Frozen Language Models’, *IEEE Transactions on Neural Networks and Learning Systems* **35**(11), 16453–16463.
URL: <https://ieeexplore.ieee.org/document/10215061>
- Thapa, N. B., Seifollahi, S. & Taheri, S. (2022), Hospital Readmission Prediction Using Clinical Admission Notes, in ‘Australasian Computer Science Week 2022’, ACSW 2022, Association for Computing Machinery, New York, NY, USA, pp. 193–199.
URL: <https://doi.org/10.1145/3511616.3513115>

- Thomas, E. J., Studdert, D. M., Newhouse, J. P., Zbar, B. I., Howard, K. M., Williams, E. J. & Brennan, T. A. (1999), 'Costs of medical injuries in Utah and Colorado', *Inquiry: A Journal of Medical Care Organization, Provision and Financing* **36**(3), 255–264.
- Tian, H., Gao, C., Xiao, X., Liu, H., He, B., Wu, H., Wang, H. & Wu, F. (2020), SKEP: Sentiment Knowledge Enhanced Pre-training for Sentiment Analysis, in 'Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics'.
- URL:** https://virtual.acl2020.org/paper_main.374.html
- Tran, N. & Lee, J. (2018), Using multiple sentiment dimensions of nursing notes to predict mortality in the intensive care unit, in '2018 IEEE EMBS International Conference on Biomedical & Health Informatics (BHI)', IEEE, pp. 283–286.
- URL:** <https://ieeexplore.ieee.org/abstract/document/8333424/>
- Van Aken, B., Papaioannou, J.-M., Mayrdorfer, M., Budde, K., Gers, F. & Loeser, A. (2021), Clinical Outcome Prediction from Admission Notes using Self-Supervised Knowledge Integration, in 'Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume', Association for Computational Linguistics, Online, pp. 881–893.
- URL:** <https://aclanthology.org/2021.eacl-main.75>
- van der Lee, C., Gatt, A., van Miltenburg, E. & Krahmer, E. (2021), 'Human evaluation of automatically generated text: Current trends and best practice guidelines', *Computer Speech & Language* **67**, 101151.
- URL:** <https://www.sciencedirect.com/science/article/pii/S088523082030084X>
- Varoquaux, G. & Colliot, O. (2023), Evaluating Machine Learning Models and Their Diagnostic Value, in O. Colliot, ed., 'Machine Learning for Brain Disorders', Humana, New York, NY.
- URL:** <http://www.ncbi.nlm.nih.gov/books/NBK597473/>
- Velikovich, L., Blair-Goldensohn, S., Hannan, K. & McDonald, R. (2010), The viability of web-derived polarity lexicons, in 'Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics', HLT '10, Association for Computational Linguistics, USA, pp. 777–785.
- Vij, A. & Pruthi, J. (2018), 'An automated Psychometric Analyzer based on Sentiment Analysis and Emotion Recognition for healthcare', *Procedia Computer Science* **132**, 1184–1191.
- URL:** <https://www.sciencedirect.com/science/article/pii/S1877050918307658>
- Waheeb, S. A., Ahmed Khan, N., Chen, B. & Shang, X. (2020), 'Machine Learning Based Sentiment Text Classification for Evaluating Treatment Quality of Discharge Summary', *Information* **11**(5), 281. Number: 5 Publisher: Multidisciplinary Digital Publishing Institute.
- URL:** <https://www.mdpi.com/2078-2489/11/5/281>

- Wang, H., Li, J., Wu, H., Hovy, E. & Sun, Y. (2023), ‘Pre-Trained Language Models and Their Applications’, *Engineering* **25**, 51–65.
URL: <https://www.sciencedirect.com/science/article/pii/S2095809922006324>
- Wang, Z., Chu, Z., Doan, T. V., Ni, S., Yang, M. & Zhang, W. (2025), ‘History, development, and principles of large language models: an introductory survey’, *AI and Ethics* **5**(3), 1955–1971.
URL: <https://doi.org/10.1007/s43681-024-00583-7>
- Wankhade, M., Rao, A. C. S. & Kulkarni, C. (2022), ‘A survey on sentiment analysis methods, applications, and challenges’, *Artificial Intelligence Review* **55**(7), 5731–5780.
URL: <https://doi.org/10.1007/s10462-022-10144-1>
- Waudby-Smith, I. E. R., Tran, N., Dubin, J. A. & Lee, J. (2018), ‘Sentiment in nursing notes as an indicator of out-of-hospital mortality in intensive care patients’, *PLoS ONE* **13**(6), e0198687.
URL: <https://pmc.ncbi.nlm.nih.gov/articles/PMC5991661/>
- Weissman, G. E., Ungar, L. H., Harhay, M. O., Courtright, K. R. & Halpern, S. D. (2019), ‘Construct validity of six sentiment analysis methods in the text of encounter notes of patients with critical illness’, *Journal of Biomedical Informatics* **89**, 114–121.
URL: <https://www.sciencedirect.com/science/article/pii/S1532046418302284>
- Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., Davison, J., Shleifer, S., von Platen, P., Ma, C., Jernite, Y., Plu, J., Xu, C., Le Scao, T., Gugger, S., Drame, M., Lhoest, Q. & Rush, A. (2020), Transformers: State-of-the-Art Natural Language Processing, in Q. Liu & D. Schlangen, eds, ‘Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations’, Association for Computational Linguistics, Online, pp. 38–45.
URL: <https://aclanthology.org/2020.emnlp-demos.6/>
- Workum, J. D., Volkers, B. W. S., van de Sande, D., Arora, S., Goeijenbier, M., Gommers, D. & van Genderen, M. E. (2025), ‘Comparative evaluation and performance of large language models on expert level critical care questions: a benchmark study’, *Critical Care* **29**(1), 72.
URL: <https://doi.org/10.1186/s13054-025-05302-0>
- Xu, H., Liu, B., Shu, L. & Yu, P. (2019), BERT Post-Training for Review Reading Comprehension and Aspect-based Sentiment Analysis, in J. Burstein, C. Doran & T. Solorio, eds, ‘Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)’, Association for Computational Linguistics, Minneapolis, Minnesota, pp. 2324–2335.
URL: <https://aclanthology.org/N19-1242/>

- Xu, H., Liu, B., Shu, L. & Yu, P. S. (2020), ‘DomBERT: Domain-oriented Language Model for Aspect-based Sentiment Analysis’. arXiv:2004.13816 [cs].
URL: <http://arxiv.org/abs/2004.13816>
- Yao, S., Yu, D., Zhao, J., Shafran, I., Griffiths, T. L., Cao, Y. & Narasimhan, K. (2023), Tree of thoughts: deliberate problem solving with large language models, in ‘Proceedings of the 37th International Conference on Neural Information Processing Systems’, NIPS ’23, Curran Associates Inc., Red Hook, NY, USA, pp. 11809–11822.
- Yin, D., Meng, T. & Chang, K.-W. (2020), SentiBERT: A Transferable Transformer-Based Architecture for Compositional Sentiment Semantics, in ‘Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics’.
URL: https://virtual.acl2020.org/paper_main.341.html
- Zeng, Z., Deng, Y., Li, X., Naumann, T. & Luo, Y. (2019), ‘Natural Language Processing for EHR-Based Computational Phenotyping’, *IEEE/ACM transactions on computational biology and bioinformatics* **16**(1), 139–153.
URL: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6388621/>
- Zhang, L., Wang, S. & Liu, B. (2018), ‘Deep learning for sentiment analysis: A survey’, *WIREs Data Mining and Knowledge Discovery* **8**(4), e1253.
URL: <https://wires.onlinelibrary.wiley.com/doi/10.1002/widm.1253>
- Zhou, Y., Ringeval, F. & Portet, F. (2023), A Survey of Evaluation Methods of Generated Medical Textual Reports, in T. Naumann, A. Ben Abacha, S. Bethard, K. Roberts & A. Rumshisky, eds, ‘Proceedings of the 5th Clinical Natural Language Processing Workshop’, Association for Computational Linguistics, Toronto, Canada, pp. 447–459.
URL: <https://aclanthology.org/2023.clinicalnlp-1.48/>
- Zhou, Y. & Srikumar, V. (2022), A Closer Look at How Fine-tuning Changes BERT, in ‘Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)’, Association for Computational Linguistics, Dublin, Ireland, pp. 1046–1061.
URL: <https://aclanthology.org/2022.acl-long.75>
- Zhu, Y., Kiros, R., Zemel, R., Salakhutdinov, R., Urtasun, R., Torralba, A. & Fidler, S. (2015), Aligning Books and Movies: Towards Story-Like Visual Explanations by Watching Movies and Reading Books, in ‘2015 IEEE (ICCV)’, pp. 19–27. ISSN: 2380-7504.
URL: <https://ieeexplore.ieee.org/document/7410368>
- Zou, Y., Wang, J., Lei, Z., Zhang, Y. & Wang, W. (2021), ‘Sentiment Analysis for Necessary Preview of 30-Day Mortality in Sepsis Patients and the Control Strategies’, *Journal of Healthcare Engineering* **2021**, 1713363.

Zubillaga, A., Laccourreye, P., Kerexeta, J., Larburu, N., Alonso, E., Gómez, D. J., Martínez, F. & Alonso-Arce, M. (2022), ‘Hospital Readmission Prediction via Keyword Extraction and Sentiment Analysis on Clinical Notes’, *Studies in Health Technology and Informatics* **295**, 339–342.

URL: <https://pubmed.ncbi.nlm.nih.gov/35773878/>

Appendix A

Annotation Files

The aim of the research in few sentences

The problem being addressed in this research is the lack of sentiment analysis computational models that can effectively function on ICU clinical texts. The research aims to investigate possible approaches to sentiment analysis that are applicable to the clinical domain to improve their functionality in this domain-specific frame.

Value

We anticipate that if natural language processing sentiment classification techniques succeeded in recognizing valuable signals from clinical notes, improving prediction in ICU factors will be feasible using these signals. Moreover, clinical sentiment analysis could be utilized in future work to provide a timeline visualization of a patient journey in ICU.

General definition of sentiment analysis

Identify, extract, or quantify of one or more of the following (Vij and Pruthi 2018):



Fig.1. Classification and criterion for sentiment analysis

This research will focus on opinion and polarity analysis as explained next.

Definition of Sentiment analysis in ICU notes

We define clinical sentiment as notes writers' opinion about the patient's condition and possibly prognosis, whether it's good, bad, neutral, or not mentioned. In other words, the question we aim to answer is:

Do you think that the person who wrote this note believes the patient's health:

- Is already improving or will get better?
- Is deteriorating?
- Is neither improving nor deteriorating?
- OR that there is no indication of the note writer's opinion.

Dataset

The dataset used in this study is the MIMIC-III database, which contains de-identified information about patients of the critical care units at Beth Israel Deaconess Medical Center in Boston, Massachusetts, USA in the period 2001-2012 (Johnson et al. 2016), (Johnson, Alistair, Pollard, Tom, and Mark, Roger 2015), (Goldberger et al. 2000).

Motivation for manual labeling

Manually labeled data is an important component for evaluating our sentiment models. Specifically, we need a set of notes whose “sentiment” has been assigned manually by expert annotators to test our sentiment models. No such labelled datasets of ICU notes exist. Because clinical language is so specialized, this manual labelling can only be done by clinical experts like you. We are initially aiming for 300 labelled notes.

Task Description

We are asking you to read a note and choose a label in the list of column D, namely “Overall label”, that reflects the patient’s status. e.g., **The clinician believes that this patient is already improving or will get better**, when a note indicates that the patient is ok, has made some progress or expected to get better, **The clinician believes that this patient is deteriorating**, when a note indicates that the patient’s condition has deteriorated, **The clinician believes that this patient is neither improving nor deteriorating** when a note indicates no specific direction to improvement or deterioration, and **There is no indication of the note writer’s opinion** when no opinion can be recognized from the note.

If possible, and where there are clear indicators in the text that lead to your decision of label, we would appreciate if you could copy those words or sentences into column E, namely “Extracts supporting the selection of label”. These extracts might allow us to enhance our sentiment models.

Instructions for annotation team

- 1- The annotation file is encrypted to keep it secure for transmission and storage. The password will be shared with you prior to the task.
- 2- The annotation task should be completed by each physician independently.
- 3- The annotation is meant to target and capture the impression that is expressed by the clinicians who wrote the note (i.e., their opinion) about the patient. It is not meant to test or evaluate the clinical decision capabilities of the annotators.
- 4- Data was extracted automatically by a programming script. If an annotator finds anything that appears to be an error, please let me know before completing the task.
- 5- There are some repetitive headers in the notes potentially due to copy and paste by note writers.
- 6- When you complete the task, please add your first name to the file name i.e., file name should look like: Annotation file – your first name.

The required field, and a desirable extract field

- 1- **Overall label** in column D is the main field in this annotation task. This should reflect the status of the patient as expressed in the note.
- 2- **Extracts supporting the selection of label** in column E. These should be copies from notes text (i.e., a sentence, or a phrase of a set of words but not a paragraph) reflecting a specific condition associated with improvement, deterioration, no-change or none. Please separate multiple extracts and start each new extract with double dashes (--). The following describe possible examples of this field:

A: Cover Note For Annotation Team

- a. Extracts associated with **improvement**. This should contain copied phrases that reflect any improvement in a patient's condition, e.g., "extubated [**4-13**], and improving on antibiotic therapy," "normal blood pressure, not on pressors."
- b. Extracts associated with **no change in status**. This should contain copied phrases that reflect stability in specific aspects of a patient's condition, e.g. "Remains intubated mechanically ventilated."
- c. Extracts associated with **deterioration**. This should contain copied phrases that reflect any deteriorating aspects of a patient's condition, e.g., "no viable therapeutic options that seem feasible for living outside of intensive care"

Deidentification by owners of MIMIC-III data

MIMIC-III data was changed by owners in compliance with HIPAA.

- 1- Patients' names, doctors' names, telephone numbers, addresses, dates are all replaced by tokens surrounded by [** and **], e.g., [**Last Name**].
- 2- "Dates were shifted into the future by a random offset for each individual patient in a consistent manner to preserve intervals, resulting in stays which occur sometime between the years 2100 and 2200." ¹
- 3- Ages are also modified. "Dates of birth for patients aged over 89 were shifted to obscure their true age and comply with HIPAA regulations: these patients appear in the database with ages of over 300 years." ²

References

- Goldberger, A. L., L. A. Amaral, L. Glass, J. M. Hausdorff, P. C. Ivanov, R. G. Mark, J. E. Mietus, G. B. Moody, C. K. Peng, and H. E. Stanley. 2000. "PhysioBank, PhysioToolkit, and PhysioNet: Components of a New Research Resource for Complex Physiologic Signals." *Circulation* 101 (23): E215-220. <https://doi.org/10.1161/01.cir.101.23.e215>.
- Johnson, Alistair E. W., Tom J. Pollard, Lu Shen, Li-wei H. Lehman, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G. Mark. 2016. "MIMIC-III, a Freely Accessible Critical Care Database." *Scientific Data* 3 (1): 160035. <https://doi.org/10.1038/sdata.2016.35>.
- Johnson, Alistair, Pollard, Tom, and Mark, Roger. 2015. "MIMIC-III Clinical Database." PhysioNet. <https://doi.org/10.13026/C2XW26>.
- Vij, Anneketh, and Jyotika Pruthi. 2018. "An Automated Psychometric Analyzer Based on Sentiment Analysis and Emotion Recognition for Healthcare." *Procedia Computer Science*, International Conference on Computational Intelligence and Data Science, 132 (January): 1184-91. <https://doi.org/10.1016/j.procs.2018.05.033>.

¹ <https://physionet.org/content/mimiciii/1.4/>

² <https://physionet.org/content/mimiciii/1.4/>

B: Annotation File

Note_ID	Note	Category	Overall label for note - The clinician who wrote this note believes the patient's health...	Extracts supporting the selection of label - Start a new extract by double dashes (--)
				Shahad Nagoor: (E.g., -- extracted ["**4.13**"] and improving on antibiotic therapy.

Appendix B

Submitted Manuscript

Clinical Sentiment Analysis of ICU Notes: A Dataset and a Comparison of Lexicon and Clinical Large Language Model Methods

Shahad Nagoor^{1,2*}, Lucy Hederman¹, Kevin Koidl¹,
Foad Bukhari³, Sandra Chiaka Amasike^{4†}, Antonie Bianchi^{5†},
Chris Yen-Chen Lo^{4†}, Hiral Patel^{6†}, Ignacio Martin-Loeches^{4,7}

¹School of Computer Science and Statistics, Trinity College Dublin.

²King AbdulAziz University, Rabigh, Saudi Arabia.

³Department of Medicine, College of Medicine, Taibah University,
Madinah, Saudi Arabia.

⁴Intensive Care Medicine, St. James's Hospital.

⁵Department of Anesthesiology and Intensive Care Unit North Hospital,
Aix Marseille University, Assistance Publique Hôpitaux Universitaires
de Marseille.

⁶Intensive Care Units, Beaumont Hospital.

⁷School of Medicine, Trinity College Dublin.

*Corresponding author(s). E-mail(s): nagoors@tcd.ie;

Contributing authors: hederman@tcd.ie; koidlk@tcd.ie;

fambukhari@taibahu.edu.sa; sansyke@yahoo.com;

antoine.bianchi@ap-hm.fr; chrisyoyo0910@gmail.com;

hirbhav@yahoo.co.in; imartinl@tcd.ie;

[†]These authors contributed equally to this work.

Abstract

Background: Clinicians in ICU spend valuable time documenting their clinical impressions in clinical notes to provide an overall understanding of patients' status and progress. Capturing positivity or negativity of clinicians' impressions about patients enables comprehending patients' progress in ICU stays and helps clinicians in making decisions about next steps, plans or interventions. Automation of such analysis requires applying natural language processing

techniques, specifically sentiment analysis, on clinicians' texts. This field, especially generative artificial intelligence models, has recently experienced significant advancement and achieved excellent unprecedented results in many language-related tasks, but their performance varies depending on the domain and the approach. General-domain sentiment models have been found not effective in distinguishing positivity and negativity in clinical notes and domain-specific sentiment models are recommended. Moreover, there are multiple common approaches to sentiment analysis including classical non-learning methods such as lexicons and modern learning-based methods such as generative AI models. However, there has been little work comparing and evaluating the effectiveness of specialized models, largely due to the difficulty of recruiting clinical annotators and of accessing clinical sentiment data. This study has three main contributions:

(1) **MIMIC-III-Ext-Notes-Sentiment**: the first public ICU-specific ground-truth dataset labeled by clinicians and created for investigating clinical sentiment polarity in ICU clinical notes. (2) **SentimentICUModel**: an effective model for classifying clinical sentiment in ICU narratives on the ground truth. (3) A guiding comparison of the effectiveness of a range of approaches to clinical sentiment classification on the dataset.

Method: We recruited five clinicians to annotate notes for a ground truth clinical sentiment dataset. Annotators indicated which pieces of note text influenced their labeling. Four clinical-specific models were compared on the expert annotated data: an existing clinical lexicon, clinicalBERT, clinicalT5, Scifive.

Results: The task of annotation was challenging due to clinicians' workload and spanned 15 months. The ground truth data was formed based on inter-annotator agreement analysis. Clinicians' extracts similarity aligned with their agreement level. Clinical LLMs provide comparable accuracy (up to 82%), with top score achieved by ClinicalT5 which is being released as SentimentICUModel. They outperform keyword-based lexicon ($p < 0.05$, [95% CI, -0.47, -0.28]), leaving room for future study and improvement of this approach.

Conclusion: The study provides researchers in the field with the first ICU-specific clinical sentiment ground-truth dataset labeled by clinicians, an effective clinical sentiment classifier, and insights about each sentiment approach on ICU clinical-specific narratives. Clinical LLMs have demonstrated effectiveness in identifying clinical sentiment within clinical notes, enabling early detection of sudden changes and exploring different paths and unpredictable patterns in patients' ICU stays.

Keywords: Critical Care, Clinical notes, Sentiment analysis, Clinical sentiment, Natural language processing, Generative artificial intelligence, Large language models

1 Introduction

Natural language processing (NLP) has recently experienced significant advancement. Specifically, generative artificial intelligence models have achieved excellent and unprecedented results in many

language-related tasks such as summarization, classification, inference, and generation. NLP progress and applications in medical and clinical languages is slow due to resource limitations: clinical data is not available due to patient privacy constraints and clinical researchers have

very limited time. Among clinical specialties, critical care, or intensive care unit (ICU), is a very dynamic, busy and time-sensitive environment. Therefore, decision support for clinicians in the ICU supporting patients progress, resources managements or interventions' timing is valuable.

A main concern for clinicians in ICU is monitoring the health status for each patient during their stay in ICU. The best case for a patient admitted to ICU is that they keep improving day by day until they become stable and get discharged to home or to another less critical hospital ward. However, many cases experience different paths and sometimes unpredictable patterns in their ICU stays. Early detection of sudden changes is important and can help in early interventions and management of ICU patients' status and ICU resources. For example, early detection of deterioration can help in sepsis cases where early intervention is essential to prevent death. A key element to this detection is clinicians' impressions documented in their notes. Kim et. al [1] was able to detect sepsis up to 48 hours prior to its onset when they add clinicians' notes to patients' structured data in the prediction algorithm. Clinicians in ICU spend valuable time documenting their clinical impressions and prognosis about patients in clinical notes. Impressions are essential in providing an overall understanding of patients' status and progress. Capturing positivity or negativity about patients' clinical status is an important factor in comprehending their progress in ICU stays, but can be exhausting and time-consuming especially for complicated cases. Therefore, an automatic negative/positive polarity identification for ICU notes will provide significant assistance in exploring their

progress including onset of deterioration or sudden shifts in patient status and consequently support ICU clinicians' decisions. Identifying such polarity has also been described as clinical sentiment¹.

Related work reported ranges of approaches to sentiment analysis including classical methods such as dictionary-based lexicons and modern methods such as generative large language models (LLMs). Many studies use general domain sentiment approaches and report their poor effectiveness when applied to clinical texts. Due to the particularities of clinical language, it has been suggested to not rely on general domain models but to use specialized models. However, there has been little work investigating, comparing and evaluating the effectiveness of specialized models in the literature.

To address this gap, and based on the anticipated benefits, the aim of this research is to investigate **which among the specialized methods of sentiment analysis approaches can offer the best capture of clinical polarity in ICU notes?** On that basis, this research has three main contributions:

1. **MIMIC-III-Ext-Notes-Sentiment²**: A clinical sentiment dataset of ICU clinical notes annotated manually by specialist clinicians.
2. **SentimentICUModel**: the most effective model for classifying clinical sentiment in ICU narratives on the ground truth.
3. An exploratory comparison of the effectiveness of different approaches to clinical sentiment analysis including

¹Throughout this research, we use these two terms interchangeably and we define clinical sentiment as note writers' opinion about the patient's condition and possibly prognosis, whether it is good or bad. See section 4.1.2 for more details.

²The dataset is currently under review to be published by Physionet.

an existing recent ICU-specific lexicon, and clinical LLMs in clinical sentiment classification for ICU clinical notes.

2 Motivation

Clinical sentiment has been under-researched due to difficulty of accessing appropriate data and limited time that specialised clinicians can provide for preparing such data. Preparing manually annotated data by experts in the field is essential for AI supervised learning and automated analysis comparing machine vs. human performance. Specifically, a set of notes whose “sentiment” has been assigned manually by expert annotators is required. No such labeled datasets of ICU notes exist. Because clinical language is so specialized, the existing general domain models have not succeeded in recognizing clinical polarity [2, 3]. As a consequence, very little work and attention has been paid to sentiment in this specific domain, and the development of clinical-specific sentiment approaches, including lexicons and AI models, has not been progressed. Therefore, their capabilities and performance have not been explored.

3 Related Work

The related literature includes a range of classic and more recent approaches to sentiment analysis. Multiple categorizations can be applied to the existing approaches based on (1) learnability: classic dictionary-based lexicons [4, 5], sentimentr [6], AFINN [7], EmoLex [8], and Opinion [9] vs. ML models [10, 11] (2) domain: general domain [12–15] vs. health-specific [16]. (3) scope of text: short texts e.g., sentence-level [3, 17, 18], or aspect-based [19–21], vs document-based [4, 10, 11, 22]. (4) Training: classic

ML models [10, 11] vs. recent pre-trained LLMs SentiBERT [23], DomBERT [24], SKEP [25], TransBERT [26], SentiLARE [27]. In addition to the above, there have been also hybrid approaches [17, 28], which integrate two approaches together e.g., using dictionary scores or linguistic features to train an ML approach. Our previous paper [29] includes a description of related literature with a focus on clinical-specific lexicons and recent LLMs. Overall, despite the poor effectiveness of the general domain models, most of the related work used them in capturing clinical sentiment [2, 30], and used the results in predicting important outcomes. e.g., risks of re-admission or outcomes like mortality [31–34] provides a comprehensive description of AI in ICU and concludes that AI based decision support systems are inevitably going to be integrated in ICU systems. With that, predicting risk factors and outcomes must be built on accurate and high quality predictive models including clinical sentiment prediction models. However, very little research was specifically designed or aimed to improve capturing clinical sentiment [30].

A recent study [35], concurrent to our work, is the only study which targets this gap and developed an ICU specific keyword list that can be utilised for creating an ICU-specific lexicon, and investigated its predictive accuracy against an ML model. The study offers insights, insisting on the need for ICU-specific sentiment tools, but the analysis is based on notes’ fragments and only a single BERT model (DeBERTa-v3) was tested. In our previous study [29], we adapted three lexicons, three BERT models and three T5 models from both general and clinical-specific domain and use a larger scope of notes

texts with segments starting from beginning of notes (typically by chief complaints of a patient), and from assessment and plan section onward as explained in section 4. In this study, we use the keywords provided by [35] for a specialized lexicon, and compare it against specialized clinical LLMs to investigate their ability in capturing positivity and negativity of the notes.

4 Method

As mentioned in section 1, to answer the research question: **which among the specialized methods of sentiment analysis approaches can offer the best capture of clinical polarity in ICU notes?** we studied four selected models from three families: a keyword-based ICU specific lexicon, a specialised BERT model (ClinicalBERT), and two generative prompt models (ClinicalT5, SciFive which are specialised T5-based models).

We created three datasets, two from MIMIC-III data: a training dataset to train LLMs, and an evaluation dataset to evaluate the lexicon and LLMs. Training data is automatically labeled using a labeling strategy suggested in [29], see section 4.1.1. The evaluation dataset is labeled by clinicians. In addition, we used an external Covid dataset [36] for evaluation.

In summary, the steps involved in this research study are:

- Create automatically pseudo-labeled training data.
- Recruit annotators for manual annotation of evaluation data and collect their annotations.
- Study inter-annotator agreement and construct ground truth.

- Evaluate pseudo-labeled dataset based on ground truth.
- Train clinical LLMs using training data (pseudo-labeled) and evaluate them on ground truth data.
- Evaluate models on external data (Covid data)
- Analyse text extracts influencing annotators' decisions using text similarity measures.

4.1 Datasets

The data used in this study is a subset of MIMIC-III, plus a Covid dataset [36]. MIMIC-III notes are written in English and contain de-identified information about patients in the intensive care units at Beth Israel Deaconess Medical Center in Boston, Massachusetts, USA in the period 2001-2012 [37–39]. We target physician and nursing progress notes, which belong to patients who stayed in ICU, due to their value and influence on prediction in critical care [40–42]. Additionally, based on [35] findings, these notes hold great proportions of sentiment words among other types, physician notes holds 97.8% (top proportion) of sentiment keywords (positives and negatives) and nursing notes hold 81.1% of sentiment keywords. Being rich in such keywords, makes these two types of notes suitable for detection of clinical sentiment, and comparing the classification using each type of interest. A synthesised example of a note is shown here:

65-year-old female patient presented to the ED with severe sub-sternal chest pain, known case of CAD with PCI to LCx in 2019. This pain started 30 minutes ago, patient then shocked for suddenly collapsed with VT/VE, then, ECG showed STEMI with significant trop leak. PCI was done and showed 90% lesion

in LAD with patent LCx. DES to LCx. Patient recovers well with Trop trending down. TTE: LV with HK, EF is 45-40%. HD is stable during CCU. Ongoing trop trending down.

Assessment: HR: 70 bpm. NSR. Chest discomfort post CPR improving with Tylenol.

Action: Cardiac meds started.

Response: Meds well tolerated.

Plan: Cont. STEMI protocol For discharge today.

We extracted two datasets from MIMIC-III notes, one for training and the other for evaluation. The Covid dataset [36] was also used for evaluation. This data contains clinical notes associated with *chest X-ray* and *CT images* of patients infected with or suspected of COVID-19 or other pneumonia causes. This data was mainly collected from publications, but some samples were collected independently from hospitals and physicians. We used it in a preliminary study of our work [43]. The next two sub-sections describe these datasets and the labeling of each.

4.1.1 Training Data

Training data is a collection of MIMIC-III notes, selected based on timing attributes of patients' ICU stays and labeled using a scheme introduced in [29] and inspired by findings in [2]. It is an automatic labeling strategy that is based on notes' timing relative to the following:

- Patients' admission.
- Patients' discharge to home.
- Patients' death.

In particular, notes are extracted from three subgroups of notes including:

- Group1: admission to ICU note and first 3 days³ of progress notes for patients regardless of their discharge outcomes.
- Group2: notes for last 3 days for discharged patients.
- Group3: notes for the last 3 days for deceased patients.

Based on the assumptions that patients are likely to be critically ill when admitted to ICU and near to mortality time, and are likely to be improving when discharged to home, Group1 and Group2 notes are labeled as negative; and Group3 notes are labeled as positive. In a previous study [29] we found that clinical sentiment of notes, specifically the part of a note that is from SOAP⁴ assessment point onward, achieves slightly improved accuracy scores especially for physicians' notes. In addition, assessment and plan contains physicians' impressions and is preferred by physicians to be read first for patients' information highlights and important summary [45–48]. It was also observed that physician notes provided more accurate prediction than nursing notes. However, in this study we train models on each category separately to compare and contrast their effect on training and evaluation using experts' annotation. Table 1 shows size of training dataset.

³3 days is used empirically, recommended by an ICU consultant.

⁴SOAP [44], is a clinical documentation methodology in which S refers to Subjective, O refers to Objective, A refers to Assessment, and P refers to Plan. Subjective section documents patient complaints. Objective section includes objective readings such as blood pressure, and heart rate. Assessment section contains doctors' impression of the patient's state, and Plan section describes actions to be taken such as medication change or labs requests, etc.

Table 1 Sizes of of physician and nursing notes training dataset.

Category	Training
Physician	10766
Nursing	10766

4.1.2 Manual Annotation

Experts' annotation is important for evaluating the sentiment classification of models. It offers evidence-based evaluation and validates the reliability and the quality of models' outcomes. Despite this valuable role, many studies in NLP have suffered from the absence of human evaluation and the reproducibility of the evaluation [49]. According to the survey in [49], 38.7% of the research papers in INLG 2023, and 36.5% of research papers in ACL 2023 have no human evaluation. Among a total of 102 publications reviewed in the survey, only 59 (56.73%) of the publications used human evaluation. As discussed in sections 1 and 2, manual annotation cost is causing a delay in the progress of NLP supervised learning. This research aims to provide a benchmark clinical sentiment dataset by which clinical sentiment polarity approaches can be investigated and improved. The development of the MIMIC-based dataset includes many challenges. Recruiting clinicians was difficult due to their hospital duties and workload. Each recruit had to complete data protection training before being given access to clinical notes from the MIMIC dataset. Specifically, recruitment for this task took more than a year. The average time consumed by each annotator was around 8 weeks for data access requirements and for the annotation task. The whole process took around 15 months.

Five⁵ annotators were recruited for the task of manual annotation for the dataset composed of ICU clinical notes. The annotators were asked to examine 300 samples from the clinical notes and to provide a sentiment label from a list. The annotation team consists of an ICU consultant, a respiratory consultant, and three ICU registrars. Two individual meetings were held with each physician prior to the task. The first meeting was to explain the requirements for obtaining access to the data, which involved completing online data protection training. The second was to explain the manual annotation task. Information regarding requirements for access to the data can be found at the MIMIC dataset host official website *PhysioNet*⁶. When an annotator completed the required training, a document including a brief background of the research and the task description was shared (see supplementary file 1). Our definition of clinical sentiment is the note writers' opinion about the patient's condition and possibly prognosis, whether it is good, bad, neutral, or not mentioned. On that basis, we asked annotators to label the notes by answering the following question from the provided list: **Do you think that the person who wrote this note believes the patient's health:**

- Is already improving or will get better?
- Is deteriorating?
- Is neither improving nor deteriorating?
- OR that there is no indication of the note writer's opinion.

Once annotations were collected, we studied inter-annotator agreement (also known as inter-rater reliability) to

⁵The number of annotators is greater than the number of classes to allow for majority votes in case of opinion conflicts.

⁶<https://physionet.org/content/mimiciii/1.4/>

Table 2 Statistics for the manually annotated dataset.

Group	Counts
Patients	248
ICU stays	251
Patients with LOS ≤ 6	107
Physicians' notes	156
Nursing notes	144

observe the level of agreement or variability in annotators' opinions. This is the extent to which human experts' selections or opinions align with each other. The higher the consensus between annotators, the more confidence and interchangeability between their opinions is provided in subsequent analysis [50].

Results of inter-annotator agreement is used to construct ground truth datasets. Further description and analysis is presented in section 5. Next is a description of the sampling process for the two manual annotation datasets (MIMIC and Covid).

4.1.3 Sampling

The MIMIC-based manual annotation dataset consists of 52% (156/300) physician notes and 48% (144/300) nursing notes. The initial sampling aimed to balance the distribution but some samples were removed manually due to repetitions in contents, found mostly in some of the nursing notes. We ensured there were at least 20 samples from each of the three groups of pseudo-labeling (i.e., notes after admission, prior to discharge home, and prior to death) so that we can compare and investigate the labeling validity with labeling of human annotators (see section 5.3). We also included samples that don't belong to the three groups to allow us to study annotators' agreements on notes out of the scope of the pseudo-labeling.

Statistics relating to the manual annotation dataset are provided in Table 2. The 300 notes relate to 248 patients and 251 ICU stays.

From the Covid dataset, we extracted only samples of patients who are located in ICU (158), and asked a respiratory consultant to do the annotation using the same labeling instructions used for the annotation applied to MIMIC 300 samples (see section 4.1.2). Due to the physician's workload and time limit, only 23 samples were annotated. Research method is summarised visually in Figure 1.

4.2 Models

As described early in this section, we select three different LLMs and an existing keyword-based lexicon (non-learning algorithm) to investigate their capabilities in capturing clinical sentiment polarity for ICU clinical notes.

The keyword lexicon is the keyword list from [35] which we will refer to as Kennedy's lexicon. This list of 103 negative and 72 positive keywords and phrases was extracted and reviewed iteratively by a group of clinicians. Sentiment in their analysis was measured by a score calculated based on equation (1).

$$\text{Sentiment score} = \frac{\text{NK}}{\text{NK} + \text{PK}} \quad (1)$$

where NK is the number of negative keywords, and PK the number of positive keywords.

The three LLMs we use are ClinicalBERT [51] from BERT models family, and Scifive [52], and ClinicalT5 [39, 53], both from the T5 family. ClinicalBERT and Scifive are open-source and publicly available in the Hugging Face library [54]; ClinicalT5 is available under

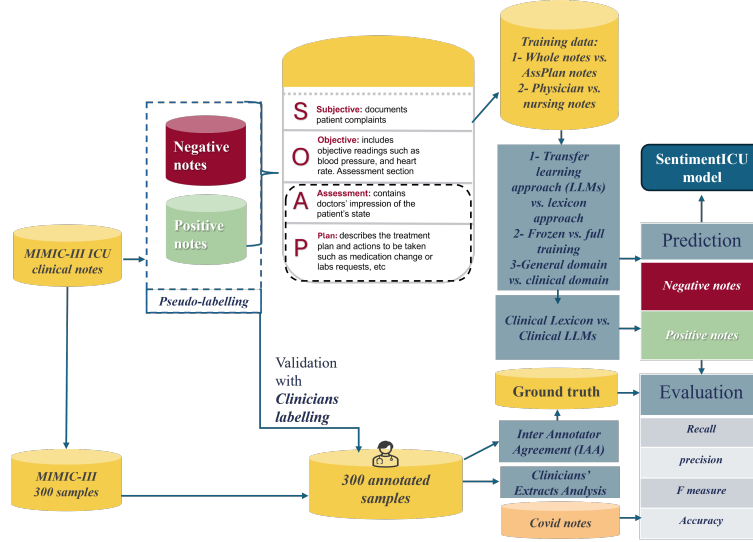


Fig. 1 Research Method

protected access through PhysioNet website⁷. We described each LLM's pre-processing requirements in [29]. ClinicalBERT and ClinicalT5 were trained on MIMIC-III text data⁸. SciFive was trained on abstracts and articles from PubMed and PubMedCentral. These three models have demonstrated the top accuracy scores in our previous work [29] against general domain lexicons and models. The aim in the previous study was to investigate the effect of domains on sentiment recognition and to have an initial assessment of the labeling strategy, hence

all datasets were automatically (pseudo-) labeled (see section 4.1.1). In this work, we build on the findings and evaluate those models on a human-labeled dataset, providing (1) evaluation of automatic labeling vs. human labeling; and (2) evaluation of machine learning models performance vs. human performance.

In experimentation with LLMs, we explored the following hyper-parameters: learning rate in the ranges: {1e-2 to 1e-5}, {3e-4, 3e-5}, and {5e-4, 5e-5}, batch size in the values: {4, 8, 16, 32, 64, 128}, and epochs in the range {1-10}. We applied early stopping and selected the best hyper-parameters, shown in Table 3.

⁷<https://www.physionet.org/content/clinical-t5/1.0.0/>

⁸They have been trained on initial tasks, e.g. masked language modeling (MLM) or next sentence prediction (NSP), which are not related to the tasks here, but we do not know the extent of the effect of the previous training on the results of this research.

Table 3 Experiment details for learning models

Model	Batch	Learning rate	Epochs	Optimizer	Tokens size
ClinicalT5	64	1e-3	8	AdamW	512
SciFive	64	1e-3	8	AdamW	512
clinicalBERT	64	1e-4	4	AdamW	512

5 Results

This section presents the results and discusses the manual annotation, construction of ground truth, validation of pseudo-labeling, evaluation of sentiment models on the annotated dataset, and annotators’ extracts of texts.

5.1 Inter-Annotator Agreement

We study Inter-Annotator Agreement (IAA) using Fleiss kappa [55]. Fleiss kappa is an extension and a generalisation of kappa coefficient, defined mathematically as

$$\text{kappa}(\kappa) = \frac{P_o - P_e}{1 - P_e} \quad (2)$$

where P_o = observed agreement among annotators, and P_e = expected agreement by chance, by which the level of agreement of two or more (multiple) annotators can be quantified and the possibility of agreement by chance can be eliminated. Fleiss’ scale is based on the following ranges: < 0 is Poor, 0.0–0.20 is Slight; 0.21–0.40 is Fair, 0.41–0.60 is Moderate, 0.61–0.80 is Substantial, 0.81–1.0 is Almost perfect [56, 57].

The overall agreement for all annotators over all 300 notes was slight using Fleiss’s Kappa (0.1863) based on the definition and interpretation by this method [57]. Counts matrix for classes distribution by annotators is shown in supplementary file 3.

The slight inter-annotator agreement is common even among very professional clinicians based on many previous studies [56], and could be explained by one or more of the following:

- Missing indicative details or implicit clinical sentiment, [12, 56, 58, 59].
- Differing focus [45].
- Inadequate attention [56, 59].
- Unfamiliar note style [48, 58].
- Different backgrounds [56, 60, 61].

To investigate IAA further, we study correlation of annotator opinions in each of the three groups (admission, discharge, death) and notes not belonging to any groups (notes in the middle). Figure 2, and figures in supplementary file 2 show the correlation between annotators for notes prior to death, near admission, prior to discharge, and the rest of the notes, respectively. From the heatmaps, it can be observed that the correlation increases closer to each event (e.g., one day prior to death) and that consistency in opinions in notes prior to death is much higher than in the other two groups. Perhaps not surprisingly, it seems that negativity in clinical sentiment prior to death is more obvious and distinguishable (top correlations range from (0.48–0.66)). Clinical sentiment near admission seems less obvious to annotators. It may be that judgments are harder to make around admission time since it is early to know what is going on (i.e., patient’s status) and what is going to be happen (i.e., prognosis) at this early stage in the

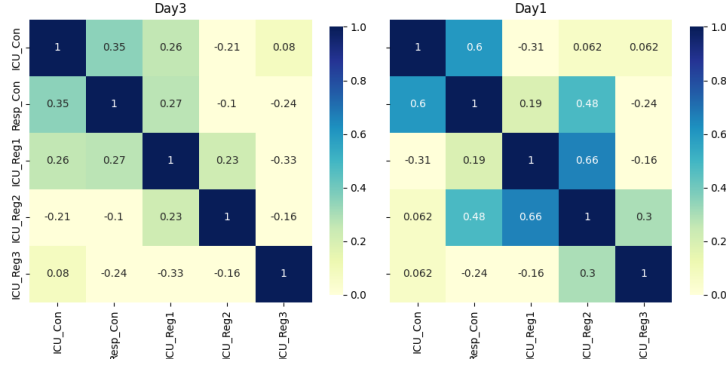


Fig. 2 Correlation heatmap between annotators of samples of notes in 3 days, and 1 day prior death.

ICU stay (top correlation range is (0.29-0.39). Also, annotators' opinions about expected positivity or signs of improvement near discharge time seems less correlated compared to their opinions of expected negativity near death time (top correlation is 0.32).

As can be seen in supplementary file 2, there is low correlation in opinions for the notes that do not fall into the near admission / discharge / death groups. The highest correlation between ICU consultant and respiratory consultant is 0.37, and no other correlation is noticeable. This might indicate that this group is the most difficult to distinguish for annotators.

5.2 Construction of Ground Truth

Based on the above context, annotators labeling cannot be used interchangeably in all samples. Ground truth data should have an accumulated evidence supporting validity.

Therefore, to construct the ground truth, we inspected the dataset and

looked for subsets forming reliable levels of consensus by annotators, similar to the method used by [59]. Although that study investigated constructing ground truth from continuous annotation, the same logic is used here to generate candidates for ground truth and select the appropriate one.

We consider subsets of data with high Fleiss values as ground truth (5-votes, 4-votes datasets or 5V, 4V). 5-votes data has perfect agreement with the Fleiss kappa (k) = 1, the Fleiss kappa for 4-votes dataset is (k) = 0.5451, which reflects moderate agreement according to Fleiss interpretation [55]. Both datasets have $p < 0.0001$, which indicates that kappa values are significantly different than 0, and rejecting the occurrence by chance.

While we consider the dataset of 5-votes or 5V (32 notes) as the perfect candidate for ground truth (gold ground truth), we see that 4-votes or 4V (110 notes) data as reliable too (silver ground truth). Therefore, the evaluation will be based on these two datasets.

Table 4 Research datasets sizes.

Category	Physician	Nursing
Pseudo-labeled training data	10766	10766
Clinicians' annotated data	156	144
MIMIC Ground Truth (5V)	17	15
MIMIC Ground Truth (4V)	46	64
Covid data	23	-

Ground truth datasets and other research datasets are described in Table 4.

5.3 Evaluation of Pseudo-Labeling

There are 145 pseudo-labeled notes in the set of 300 notes annotated; 74 notes within 3 days of admission, 48 in the 3 days preceding discharge, and 23 in the 3 days preceding death. These are collected randomly during sampling but since the sampling was inspected manually afterwards for removing notes with repeated contents, a final step was taken to assure that each group has at least 20 samples. After the annotation, notes which were annotated as 'Is deteriorating', and notes annotated 'Is already improving or will get better' are extracted from admission, discharge and death groups. This allows us to compare the two classes annotated by clinicians with pseudo-labels which only have two classes. The two-classes samples are small relative to the original death, discharge, and admission groups. Of the 23 of death samples, 2 samples were found in 5V and 10 samples were labeled using the two-classes in 4V. On the discharge group, 9 out of 74 samples were annotated with the two-classes in 5V and 33 were found in 4V. Of the 48 of the admission group, 5 samples were found in 5V and 12 in 4V.

Treating manual annotation as truth, the accuracy of pseudo-labels is computed as a percentage of correct pseudo-labels in the set of manually annotated notes. To check the validity of the pseudo-labeling, counts and percentages where pseudo labels match ground truth labels are studied. Percentage where pseudo-labels match with ground truth (5V,4V) are calculated using equation 3:

$$\text{Percentage of a group} = \frac{CC}{CBA} \quad (3)$$

where CC is counts of correct pseudo-labels (matched with annotators' labels) in the group, and CBA is counts of all samples labeled as improving or deteriorating by annotators. Using the equation, the percentage of correct labels compared to ground truth datasets (5V, 4V) are (100%, 90%) in death group, (100%, 100%) in discharge group, (40%, 42%) in admission group respectively. Table 5 shows counts and percentages of total samples and correctly labeled samples in the three groups. The pseudo-labeling matches ground truth in death and discharge groups. However, alignment for the admission group is much lower (around 40%). It can be observed from the table that only 12/48 were labeled as 'Is deteriorating', or 'Is already improving or will get better' by annotators. This indicates that annotators faced difficulty in identifying clinical sentiment in the post-admission group; the rest of the post-admission notes were labeled by 2 to 3 annotators as "Is neither improving nor deteriorating", or "no indication of the note writer's opinion". The pseudo-labels have no corresponding labels to those labels.

Table 5 Statistics for matching labels between pseudo-labels and gold ground truth (5V) in middle, and silver ground truth (4V) on right.

Group	Total Samples	5V			4V		
		Two-classes	Correct Labels	%	Two-classes	Correct Labels	%
Death	23	2	2	100%	10	9	90%
Discharge	74	9	9	100%	33	33	100%
Admission	48	5	2	40%	12	5	42%

Two classes are samples labeled as 'Is deteriorating', or 'Is already improving or will get better' by annotators. '%' is percentage of correct labels compared to ground truth.

Therefore, admission group samples are difficult to judge and validate considering that only 25% of this group can be compared with pseudo-labels. However, this group was kept in the training and evaluation datasets to maintain the diversity of samples and to observe the prediction in different settings. Within certain samples of this group in the ground truth, annotators have reported that a number of patients have exhibited signs of rapid improvement. However, identifying such cases and instances in practice can be challenging and not immediately clear. The experimentation using the pseudo-labeling in our previous work [29] was conducted on patients including those with short stays e.g., 6 days. The issue identified here by (misalignment between the pseudo-labels and annotators) reflects that this group of patients may have some exceptional quick improvements, against the pseudo-labeling, suggesting negativity of admission notes up to 3 days. The issue does not affect models trained completely on pseudo-labels in [29], because training and testing datasets have consistent labeling pattern, but it has an effect on the final evaluation as discussed in section 5.4.

Making a final conclusion about the validity of pseudo-labels using the above details is difficult, and further investigations with increased size of annotation

samples allowing for statistical significance measures are required. However, the pseudo-labels strongly align with ground truth in samples linked to two major clinical events (discharge, death), presenting validity in these two clusters.

5.4 Model Evaluation

Table 6 shows a summary table averaging LLMs accuracies against Kennedy's lexicon for prediction of clinical sentiment using physician notes (left) and nursing notes (right). Overall, LLMs outperform Kennedy's lexicon on ground truth datasets and Covid data (82% vs 35%) on 5V, (76% vs 28%) on 4V, and (80% vs 9%) on Covid data. With respect to physician notes, all specialized LLMs achieved high accuracy (82%) for gold ground truth, i.e., samples which all annotators agree on their labeling (5V). Clinical LLMs have slightly lower accuracy (76%) on the silver ground truth, those with 4 matching opinions (4V). Predicting sentiment of nursing notes was less accurate (71% on 5V data and 68% on 4V) with similar pattern of better accuracy on the gold ground truth (5V) than on silver ground truth (4V).

On the Covid dataset [36], clinical LLMs achieve high accuracy (0.88

Table 6 A comparison of models averaged accuracy across ground truth datasets 5V, and 4V.

Method	Physician notes			Nursing notes		
	5V	4V	Covid	5V	4V	Covid
Clinical LLMs	0.82	0.76	0.88	0.71	0.68	0.80
Kennedy’s lex	0.35	0.28	0.12	0.40	0.25	0.09

on physician notes and 0.80 on nursing notes), whereas Kennedy’s lexicon’s accuracy is poor (0.12, 0.09 respectively).

Supplementary file 3 shows the evaluation for ClinicalT5, SciFive, clinicalBERT, and Kennedy’s lexicon using overall accuracy, and recall, precision, F1-score for positive and negative classes, and applied to assessment and plan sections onward for physicians’ and nursing notes. Note that samples with no assessments or plans are excluded from ground truth 5V and 4V in tables.

During the analysis, it was noticed that accuracy values drop dramatically for samples belonging to patients with short ICU stays ($LOS \leq 6$), ranging from 24% for Kennedy’s lexicon to 46% for clinical LLMs. While examining the ground truth samples, it was found that patients with short length of stays ($LOS \leq 6$) form 60% of the samples within the admission group in the ground truth (3 out of 5 in 5V admission group, and 7 out of 12 in 4V admission group), have exhibited signs of rapid improvement, causing inconsistency in the labeling in the training and evaluation. As discussed in section 5.3, identifying such cases and instances in practice can be challenging and not immediately clear. However, observing the variation based on different patients groups and ICU journeys is important for understanding factors affecting clinical sentiment. Therefore, with these short LOS samples included, a slight decrease of accuracy was observed on ground truth compared

to on pseudo-labeled data (from top score 93% as reported in [29] to 82% on averaged LLMs performance reported here on ground truth). This is because the ground truth contains 16 positive samples (7 physician notes and 9 nursing notes). Of those, 9/16 are $LOS \leq 6$ (between 1-4 days). Upon removing these samples from training and evaluation data, the accuracy was increased to 94% and 88% for 5V and 4V respectively. From the evaluation on ground truth, it can be observed that, except for the group of short stays in ICU, the pseudo-labeling has shown effectiveness in training clinical LLMs for clinical sentiment classification. This validation of our pseudo-labeling strategy is valuable because it enables the models to be exposed to a large number of training samples in a cost-effective manner, without the cost of manual annotation. To enrich the analysis and observations, it is important in subsequent research to increase the sample size to allow for more observations on sub-groups and to do further statistical validity on both the pseudo-labeling and the prediction of clinical sentiment.

In summary, all LLMs performance align with annotators’ agreement, i.e., higher accuracies is observed for increased agreement on labeling. ClinicalT5 shows higher accuracy scores compared to Scifive and clinicalBERT and all LLMs significantly outperform Kennedy’s lexicon. Models’ performance using physician notes (0.82, 0.76) were higher on 5V and 4V compared to

nursing notes (0.71, 0.68), respectively. Clinical LLMs' performance remain good on external Covid dataset, suggesting reliable generalisability.

In relation to statistical significance, we tested the distribution of accuracy scores for each model using Shapiro-Wilk test, and found that all models accuracies are normally distributed. Hence, we use ANOVA test and found statistically significant difference in the means of the accuracy scores ($p < 0.05$, [95% CI, 0.52, 0.63]). When investigating model performance on all data we found that clinical LLMs are comparable. Therefore, while ClinicalT5 shows higher accuracy scores in almost all cases compared to Scifive and clinicalBERT, there is no statistically significant difference in the mean accuracies. In contrast, there is a statistically significant difference between all LLMs and Kennedy's lexicon ($p < 0.05$, [95% CI, -0.47, -0.28]), confirming that their performance differ significantly.

5.5 Analysis of Annotators' Extracts

As mentioned in section 4.1.2, part of the manual annotation task was to extract pieces of texts leading to annotators' selections of labels. We aimed to study how each annotator digests information from the notes and what piece of text helped them make their final decision about a patient's status or prognosis. We think these extracts have significant value by enabling the following:

- Investigating the most common words or phrases from texts based on a specific patient categorization e.g., patients whom annotators think are doing well.

- Investigating if annotators are paying attention to or building their decisions using the same details.
- Building an ICU-specific sentiment lexicon with support from ICU clinicians.
- Using the extracts to develop AI predictive models and compare the results using the different prediction pipelines.
- Using different prompting e.g., question answering style or few shots.

In this work, we address the first two (i.e., investigating most common words or phrases and studying extracts similarity between annotators), and we keep the rest for future work.

To explore the most repeated words in the extracted text, we use word clouds to visually show the repeated texts in extracts relating to (1) patients whom annotators agree (by majority vote) that their status is improving. (2) patients whom annotators agree (by majority vote) that their status is deteriorating.

Figure 3 shows the word clouds for each patient group. In the clouds, the size of each word represents its frequency so the largest word indicates that it is the most frequent. These words can be considered as more indicative of the group more than other less frequent words. e.g., it seems that patients who were deteriorating in the selected samples are likely to have respiratory issue due to the repetition of the words 'respiratory', and 'vent' (ventilation). On the other hand, 'extubated', 'stable', and 'comfortable' are noticeable on the other group. However, these words may differ depending on the context they were used in. Therefore, we recommend that building any further or subsequent analysis on them e.g., creating a lexicon requires involving a specialized clinician for review and supervision.

Table 7 Overall ROUGE-F1, BLEU, cosine similarity scores comparing texts from annotators' extracts by pairwise analysis.

Pair	5V_R	5V_B	5V_C	4V_R	4V_B	4V_C
ICU_Con.Resp_Con	0.1475	0.0240	0.3023	0.1364	0.0429	0.2901
ICU_Con.ICU_Reg1	0.0817	0.0486	0.1479	0.0430	0.0137	0.0955
ICU_Con.ICU_Reg2	0.1172	0.0335	0.3129	0.1063	0.0443	0.2503
ICU_Con.ICU_Reg3	0.0627	0.0000	0.2535	0.0302	0.0000	0.1788
Resp_Con.ICU_Reg1	0.1315	0.0649	0.2212	0.0875	0.0500	0.1448
Resp_Con.ICU_Reg2	0.3404	0.1734	0.5184	0.3185	0.1650	0.4764
Resp_Con.ICU_Reg3	0.1343	0.0000	0.3199	0.1404	0.0000	0.2842
ICU_Reg1.ICU_Reg2	0.0745	0.0283	0.1931	0.1107	0.0645	0.2066
ICU_Reg1.ICU_Reg3	0.0482	0.0000	0.1435	0.0248	0.0000	0.1131
ICU_Reg2.ICU_Reg3	0.0865	0.0000	0.3543	0.1164	0.0000	0.3306
Mean	0.1224	0.0373	0.2767	0.1114	0.0380	0.2370

5V_R indicates ROUGE-F1 for 5-votes dataset, 5V_B indicates Bleu score for 5V (5-votes) dataset. 5V_C indicates cosine similarity for 5V dataset. The remaining columns indicates the same scores for 4V (4-votes) dataset.

To study similarity between extracts, we used three evaluation metrics: ROUGE (Recall-Oriented Understudy for Gisting Evaluation), BLEU (Bilingual Evaluation Understudy), and cosine similarity based on text embeddings.

ROUGE and BLEU are evaluation metrics for automatic text evaluation rather than human evaluation. BLEU is known for machine translation tasks and ROUGE is commonly used in text summarization, but both target overlapped n-grams between reference or original text and another candidate text. The studies [62] and [63] explain both thoroughly. In general, the higher the score and the closer it is to 1, the higher matches of n-grams between the reference and the candidate text. We utilize these metrics to evaluate the similarity between extracts captured by clinicians. We investigated those extracts on the datasets of different votes, and we hypothesize that high votes datasets such as 5-votes should have higher scores and lower votes datasets may have lower scores. This hypothesis is based on the fact that annotators

might be looking at different aspects while examining notes. Cosine similarity for two vectors A and B is calculated based on the following equation:

$$\cos \text{ similarity } (A, B) = \frac{A \cdot B}{\|A\| \|B\|} \quad (4)$$

Where: - $A \cdot B$ is the dot product of A and B . - $\|A\|$ and $\|B\|$ are the magnitudes of A and B , respectively. The magnitude of a vector A is:

$$\|A\| = \sqrt{\sum_{i=1}^n a_i^2} \quad (5)$$

Where a_i are the components of vector A . Text embeddings are numerical text representation obtained by a language model. We used 'paraphrase-MiniLM-L6-v2', a Sentence Transformers model [64] from the Hugging Face open-source library for this purpose. Table 7 provides ROUGE (Recall-Oriented Understudy for Gisting Evaluation),

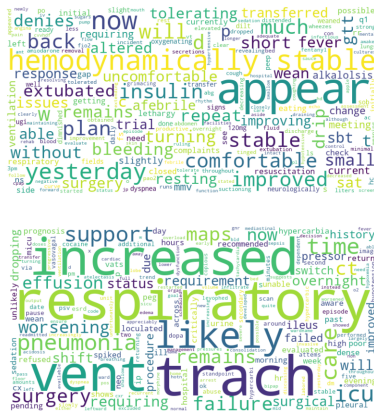


Fig. 3 WordClouds for positive notes' extracts belonging to patients who are believed by majority of physicians to be improving in ICU (above) and negative notes' extracts belonging to patients who are believed by majority of physicians to be deteriorating in ICU (below)

BLEU (Bilingual Evaluation Understudy), and cosine similarity scores comparing extracts from annotators. The score obtained from Table 7 suggest that our hypothesis is indeed supported by most of the scores (i.e., ROUGE-F1=0.12, Bleu=0.037, Cosine similarity=0.27 for 5V dataset vs. 0.11, .038, 0.23 for 4V dataset respectively).

6 Conclusion

This study provided three contributions. First, the first public ICU-specific clinical Sentiment ground truth dataset labeled by clinicians. Clinical sentiment has been highly under-researched due to the specialty of the clinical language, unavailability of public patients data, and the difficulty of recruiting clinicians to create labeled dataset, which is required

in human vs. machine predictions. Five clinicians were recruited for distinguishing clinical sentiment in notes, and providing insightful extracts of notes influencing their decisions. Afterwards, Inter-annotator agreement (IAA) was investigated for providing ground truth dataset, and extracts were investigated in pairwise comparison. The task of annotation was challenging due to clinicians' workload and span through 15 months. This research also investigated clinical sentiment using a selection of clinical LLMs and an ICU-specific keyword lexicon, applied to MIMIC-III ICU clinical notes. The objective of the analysis was to observe which among the clinical specialized models can offer the best accuracy in this specific sentiment polarity classification within the ICU-specific framework. Specialized LLMs including clinicalT5 (provided top accuracy scores, and so is the top performing **SentimentICUmodel**), Scifive, clinicalBERT outperformed the specialized lexicon. The keyword-based lexicon (Kennedy's lexicon) performance was generally poor, but positive keywords were especially less effective in capturing sentiment from notes, leaving room for improvement in this approach.

In summary, this work provided SentimentICUmodel, a model providing the most effective clinical sentiment, a guiding comparison for evaluating the effectiveness of several popular models in clinical sentiment classification of ICU notes, and provided the first clinicians' annotated sentiment dataset for NLP researchers to progress clinical sentiment and other applicable NLP tasks in future work.

Declarations

Ethics

Patients information and notes included

in this resource is extracted from MIMIC-III which was originally reviewed and exempt from informed consents based on a decision made by Beth Israel Deaconess Medical Center review board.

Acknowledgments

This work was completed with resources provided by ADAPT SFI Research Centre at Trinity College Dublin.

Data and Model Availability

Dataset license is available at data host website Physionet⁹. The annotation dataset is currently under-review and will be available for credentialed users of Physionet. Models were trained privately due to restrictions of MIMIC-III data. SentimentICUModel will be available for credentialed users of Physionet with details on implementation and reproducibility.

Competing interests

None

Funding

This research was conducted with the financial support of King AbdulAziz University.

Authors' Contributions

S.N. is the first author and investigator. L.H. is the primary supervisor and K.K. is co-supervisor. F.B. provided a synthesised sample of a note and annotated the Covid dataset. S.A., A.B., F.B., C.L. and H.P. carried out the annotation of the MIMIC dataset, provided subsequent advice on annotation, and reviewed the final paper. F.B. and I.M. provided clinical insights on the research.

Abbreviations

AI: Artificial intelligence

ICU: Intensive care unit

LLMs: Large language models

NLP: Natural Language Processing

References

- [1] Goh, K.H., Wang, L., Yeow, A.Y.K., Poh, H., Li, K., Yeow, J.J.L., Tan, G.Y.H.: Artificial intelligence in sepsis early prediction and diagnosis using unstructured data in healthcare. *Nature Communications* **12**, 711 (2021) <https://doi.org/10.1038/s41467-021-20910-4> . Accessed 2025-06-22
- [2] Weissman, G.E., Ungar, L.H., Harhay, M.O., Courtright, K.R., Halpern, S.D.: Construct validity of six sentiment analysis methods in the text of encounter notes of patients with critical illness. *Journal of Biomedical Informatics* **89**, 114–121 (2019) <https://doi.org/10.1016/j.jbi.2018.12.001>
- [3] Holderness, E., Cawkwell, P., Bolton, K., Pustejovsky, J., Hall, M.-H.: Distinguishing Clinical Sentiment: The Importance of Domain Adaptation in Psychiatric Patient Health Records. In: *Proceedings of the 2nd Clinical Natural Language Processing Workshop*, pp. 117–123. ACL, Minneapolis, Minnesota, USA (2019). <https://doi.org/10.18653/v1/W19-1915> . <http://aclweb.org/anthology/W19-1915> . Accessed 2022-09-02
- [4] Spinczyk, D., Nabrdalik, K., Rojewska, K.: Computer aided sentiment analysis of anorexia nervosa

⁹<https://physicianionet.org/content/mimiciii/view-license/1.4/>

- patients' vocabulary. *BioMedical Engineering OnLine* **17**(1), 19 (2018). <https://doi.org/10.1186/s12938-018-0451-2>. Accessed 2022-10-18
- [5] De Smedt, T., Daelemans, W.: Pattern for python. *The Journal of Machine Learning Research* **13**(null), 2063–2067 (2012)
- [6] Rinker, T.: sentimentr. original-date: 2015-08-16T00:55:55Z (2022). <https://github.com/trinker/sentimentr>. Accessed 2022-10-21
- [7] Nielsen, F.: A new ANEW: Evaluation of a word list for sentiment analysis in microblogs. *arXiv*. *arXiv:1103.2903 [cs]* (2011). <https://doi.org/10.48550/arXiv.1103.2903>
- [8] Mohammad, S.M., Turney, P.D.: Crowdsourcing a Word-Emotion Association Lexicon. *arXiv*. *arXiv:1308.6297 [cs]* (2013). <https://doi.org/10.48550/arXiv.1308.6297>
- [9] Hu, M., Liu, B.: Mining and summarizing customer reviews. In: *Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. KDD '04, pp. 168–177. Association for Computing Machinery, New York, NY, USA (2004). <https://doi.org/10.1145/1014052.1014073>. <https://doi.org/10.1145/1014052.1014073>. Accessed 2022-09-30
- [10] Sarker, A., Molla, D., Paris, C.: Outcome Polarity Identification of Medical Papers. In: *Proceedings of the Australasian Language Technology Association Workshop 2011*, Canberra, Australia, pp. 105–114 (2011). <https://aclanthology.org/U11-1014>. Accessed 2022-10-17
- [11] Choi, G., Oh, S., Kim, H.: Improving Document-Level Sentiment Classification Using Importance of Sentences. *Entropy* **22**(12), 1336 (2020). <https://doi.org/10.3390/e22121336>. Accessed 2022-09-28
- [12] Denecke, K., Deng, Y.: Sentiment analysis in medical settings: New opportunities and challenges. *Artificial Intelligence in Medicine* **64**(1), 17–27 (2015). <https://doi.org/10.1016/j.artmed.2015.03.006>. Accessed 2022-08-31
- [13] Cuenca-Zaldívar, J.N., Torrente-Regidor, M., Martín-Losada, L., Fernández-De-Las-Peñas, C., Florencio, L.L., Sousa, P.A., Palacios-Ceña, D.: Exploring Sentiment and Care Management of Hospitalized Patients During the First Wave of the COVID-19 Pandemic Using Electronic Nursing Health Records: Descriptive Study. *JMIR Medical Informatics* **10**(5), 38308 (2022). <https://doi.org/10.2196/38308>. Accessed 2024-05-16
- [14] Zubillaga, A., Laccourreye, P., Kerexeta, J., Larburu, N., Alonso, E., Gómez, Mez, F. D. Jesús and Martínez, Alonso-Arce, M.: Hospital Readmission Prediction via Keyword Extraction and Sentiment Analysis on Clinical Notes. In: *Advances in Informatics, Management and Technology in Healthcare*, pp. 339–342. IOS Press, Amsterdam, Netherlands (2022). https://doi.org/10.3233/9781608919711_17

[//doi.org/10.3233/SHTI220732](https://doi.org/10.3233/SHTI220732)

Accessed 2022-09-05

- [15] Deng, Y., Stoeck, M., Denecke, K.: Retrieving Attitudes: Sentiment Analysis from Clinical Narratives. (2014). <http://ceur-ws.org/Vol-1276/MedIR-SIGIR2014-03.pdf> Accessed 2022-12-04
- [16] Asghar, M.Z., Ahmad, S., Qasim, M., Zahra, S.R., Kundi, F.M.: SentiHealth: creating health-related sentiment lexicon using hybrid approach. SpringerPlus **5**(1), 1139 (2016) <https://doi.org/10.1186/s40064-016-2809-x> . Accessed 2022-12-04
- [17] Nguyen, T.D., Diep-Phuong Nguyen, L., Cao, T.: Sentiment analysis on medical text using combination of machine learning and SO-CAL scoring. In: 2017 21st Asia Pacific Symposium on Intelligent and Evolutionary Systems (IES), pp. 49–54 (2017). <https://doi.org/10.1109/IESYS.2017.8233560>
- [18] Carrillo-de-Albornoz, J., Vidal, J.R., Plaza, L.: Feature engineering for sentiment analysis in e-health forums. PLOS ONE **13**(11), 0207996 (2018) <https://doi.org/10.1371/journal.pone.0207996> . Publisher: Public Library of Science. Accessed 2022-11-26
- [19] Hoang, M., Bihorac, O.A., Rouces, J.: Aspect-Based Sentiment Analysis using BERT. In: Proceedings of the 22nd Nordic Conference on Computational Linguistics, pp. 187–196. Linköping University Electronic Press, Turku, Finland (2019). <https://aclanthology.org/W19-6120>
- [20] Salas-Zárate, M.d.P., Medina-Moreira, J., Lagos-Ortiz, K., Luna-Aveiga, H., Rodríguez-García, M.A., Valencia-García, R.: Sentiment Analysis on Tweets about Diabetes: An Aspect-Level Approach. Computational and Mathematical Methods in Medicine **2017**, 5140631 (2017) <https://doi.org/10.1155/2017/5140631> . Publisher: Hindawi. Accessed 2022-09-04
- [21] Sun, C., Huang, L., Qiu, X.: Utilizing BERT for Aspect-Based Sentiment Analysis via Constructing Auxiliary Sentence. arXiv. arXiv:1903.09588 [cs] version: 1 (2019). <https://doi.org/10.48550/arXiv.1903.09588> . <http://arxiv.org/abs/1903.09588> Accessed 2022-09-05
- [22] Taboada, M., Brooke, J., Tofiloski, M., Voll, K., Stede, M.: Lexicon-Based Methods for Sentiment Analysis. Computational Linguistics **37**(2), 267–307 (2011) <https://doi.org/10.1162/COLL.a.00049> . Accessed 2022-10-17
- [23] Yin, D., Meng, T., Chang, K.-W.: SentiBERT: A Transferable Transformer-Based Architecture for Compositional Sentiment Semantics. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (2020). https://virtual.acl2020.org/paper_main.341.html Accessed 2022-12-12
- [24] Xu, H., Liu, B., Shu, L., Yu, P.S.: DomBERT: Domain-oriented Language Model for

- Aspect-based Sentiment Analysis. arXiv. arXiv:2004.13816 [cs] (2020). <https://doi.org/10.48550/arXiv.2004.13816>. <http://arxiv.org/abs/2004.13816>. Accessed 2022-12-12
- [25] Tian, H., Gao, C., Xiao, X., Liu, H., He, B., Wu, H., Wang, H., Wu, F.: SKEP: Sentiment Knowledge Enhanced Pre-training for Sentiment Analysis. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (2020). https://virtual.acl2020.org/paper_main.374.html. Accessed 2022-12-12
- [26] Li, Q., Li, X., Du, Y., Fan, Y., Chen, X.: A New Sentiment-Enhanced Word Embedding Method for Sentiment Analysis. undefined (2022). Accessed 2022-12-08
- [27] Ke, P., Ji, H., Liu, S., Zhu, X., Huang, M.: SentiLARE: Sentiment-Aware Language Representation Learning with Linguistic Knowledge. In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 6975–6988. Association for Computational Linguistics, Online (2020). <https://doi.org/10.18653/v1/2020.emnlp-main.567>
- [28] Neshan, S.A.S., Akbari, R.: A Combination of Machine Learning and Lexicon Based Techniques for Sentiment Analysis. In: 2020 6th International Conference on Web Research (ICWR), pp. 8–14 (2020). <https://doi.org/10.1109/ICWR49608.2020.9122298>
- [29] Nagoor, S., Hederman, L., Koidl, K., Martin-Loeches, I.: Distinguishing Clinical Sentiment in Intensive Care Unit Clinical Notes, pp. 249–259. IEEE Computer Society, Guadalajara, Mexico (2024). <https://doi.org/10.1109/CBMS61543.2024.00049>
- [30] Denecke, K., Reichenpfader, D.: Sentiment analysis of clinical narratives: A scoping review. Journal of Biomedical Informatics **140**, 104336 (2023) <https://doi.org/10.1016/j.jbi.2023.104336>. Accessed 2024-05-27
- [31] Gao, Q., Wang, D., Sun, P., Luan, X., Wang, W.: Sentiment Analysis Based on the Nursing Notes on In-Hospital 28-Day Mortality of Sepsis Patients Utilizing the MIMIC-III Database. Computational and mathematical methods in medicine (2021) <https://doi.org/10.1155/2021/3440778>
- [32] Waudby-Smith, I.E.R., Tran, N., Dubin, J.A., Lee, J.: Sentiment in nursing notes as an indicator of out-of-hospital mortality in intensive care patients. PloS One **13**(6), 0198687 (2018) <https://doi.org/10.1371/journal.pone.0198687>
- [33] Zou, Y., Wang, J., Lei, Z., Zhang, Y., Wang, W.: Sentiment Analysis for Necessary Preview of 30-Day Mortality in Sepsis Patients and the Control Strategies. Journal of Healthcare Engineering **2021**, 1713363 (2021) <https://doi.org/10.1155/2021/1713363>
- [34] Pinsky, M.R., Bedoya, A., Biho-rac, A., Celi, L., Churpek, M., Economou-Zavlanos, N.J., Elbers, P., Saria, S., Liu, V., Lyons, P.G.,

- Shickel, B., Toral, P., Tscholl, D., Clermont, G.: Use of artificial intelligence in critical care: opportunities and obstacles. *Critical Care* **28**(1), 113 (2024) <https://doi.org/10.1186/s13054-024-04860-z> . Accessed 2025-06-29
- [35] Kennedy, C.J., Chiu, C., Chapman, A.C., Gologorskaya, O., Farhan, H., Han, M., Hodgson, M., Lazzareschi, D., Ashana, D., Lee, S., Smith, A.K., Espejo, E., Boscardin, J., Pirracchio, R., Cobert, J.: Negativity and Positivity in the ICU: Exploratory Development of Automated Sentiment Capture in the Electronic Health Record. *Critical Care Explorations* **5**(10), 0960 (2023) <https://doi.org/10.1097/CCE.0000000000000960> . Accessed 2024-05-27
- [36] Cohen, J.P., Morrison, P., Dao, L.: COVID-19 Image Data Collection. arXiv. arXiv:2003.11597 [cs, eess, q-bio] (2020). <https://doi.org/10.48550/arXiv.2003.11597> . <http://arxiv.org/abs/2003.11597> . Accessed 2022-11-23
- [37] Johnson, A.E.W., Pollard, T.J., Shen, L., Lehman, L.-w.H., Feng, M., Ghassemi, M., Moody, B., Szolovits, P., Anthony Celi, L., Mark, R.G.: MIMIC-III, a freely accessible critical care database. *Scientific Data* **3**(1), 160035 (2016) <https://doi.org/10.1038/sdata.2016.35> . Number: 1 Publisher: Nature Publishing Group. Accessed 2022-11-24
- [38] Johnson, A., Pollard, T., Mark, R.: MIMIC-III Clinical Database. *PhysioNet* (2015). <https://doi.org/10.13026/C2XW26> . <https://physionet.org/content/mimiciii/1.4/> . Accessed 2022-11-24
- [39] Goldberger, A.L., Amaral, L.A., Glass, L., Hausdorff, J.M., Ivanov, P.C., Mark, R.G., Mietus, J.E., Moody, G.B., Peng, C.K., Stanley, H.E.: PhysioBank, PhysioToolkit, and PhysioNet: components of a new research resource for complex physiologic signals. *Circulation* **101**(23), 215–220 (2000) <https://doi.org/10.1161/01.cir.101.23.e215>
- [40] Cioffi, J.: Recognition of patients who require emergency assistance: A descriptive study. *Heart & Lung* **29**(4), 262–268 (2000) <https://doi.org/10.1067/mhl.2000.108327> . Accessed 2022-11-14
- [41] Collins, S.A., Vawdrey, D.K.: “Reading between the lines” of flow sheet data: nurses’ optional documentation associated with cardiac arrest outcomes. *Applied Nursing Research* **25**(4), 251–257 (2012)
- [42] Collins, S.A., Cato, K., Albers, D., Scott, K., Stetson, P.D., Bakken, S., Vawdrey, D.K.: Relationship Between Nursing Documentation and Patients’ Mortality. *American Journal of Critical Care* **22**(4), 306–313 (2013) <https://doi.org/10.4037/ajcc2013426> . Accessed 2022-11-14
- [43] Nagoor, S., Hederman, L., Koidl, K., Caputo, A.: COVID_icu_bert: A Fine-tuned Language Model for COVID-19 Intensive Care Unit Clinical Notes. *International Journal of Cognitive and Language Sciences* **17**(8), 445–452 (2023). Accessed 2024-11-19

- [44] Podder, V., Lew, V., Ghassemzadeh, S.: SOAP Notes. In: StatPearls. StatPearls Publishing, Treasure Island (FL) (2023). <http://www.ncbi.nlm.nih.gov/books/NBK482263/> Accessed 2023-09-26
- [45] Brown, P.J., Marquard, J.L., Amster, B., Romoser, M., Friderici, J., Goff, S., Fisher, D.: What do Physicians Read (and Ignore) in Electronic Progress Notes? *Applied Clinical Informatics* **5**(2), 430–444 (2014) <https://doi.org/10.4338/ACI-2014-01-RA-0003> . Accessed 2025-07-11
- [46] Coiera, E.: When Conversation Is Better Than Computation. *Journal of the American Medical Informatics Association* **7**(3), 277–286 (2000) <https://doi.org/10.1136/jamia.2000.0070277> . Accessed 2025-07-11
- [47] Rosenbloom, S.T., Denny, J.C., Xu, H., Lorenzi, N., Stead, W.W., Johnson, K.B.: Data from clinical notes: a perspective on the tension between structure and flexible documentation. *Journal of the American Medical Informatics Association* **18**(2), 181–186 (2011) <https://doi.org/10.1136/jamia.2010.007237> . Accessed 2025-07-11
- [48] Hultman, G.M., Marquard, J.L., Lindemann, E., Arsoniadis, E., Pakhomov, S., Melton, G.B.: Challenges and Opportunities to Improve the Clinician Experience Reviewing Electronic Progress Notes. *Applied Clinical Informatics* **10**(3), 446–453 (2019) <https://doi.org/10.1055/s-0039-1692164> . Accessed 2025-07-11
- [49] Schmidtova, P., Mahamood, S., Balloccu, S., Dusek, O., Gatt, A., Gkatzia, D., Howcroft, D.M., Platek, O., Sivaprasad, A.: Automatic Metrics in Natural Language Generation: A survey of Current Evaluation Practices. In: Mahamood, S., Minh, N.L., Ippolito, D. (eds.) *Proceedings of the 17th International Natural Language Generation Conference*, pp. 557–583. Association for Computational Linguistics, Tokyo, Japan (2024). <https://aclanthology.org/2024.inlg-main.44/> Accessed 2025-01-06
- [50] Nguyen Hong Duc, P., Torterotot, M., Samaran, F., White, P.R., Gérard, O., Adam, O., Cazau, D.: Assessing inter-annotator agreement from collaborative annotation campaign in marine bioacoustics. *Ecological Informatics* **61**, 101185 (2021) <https://doi.org/10.1016/j.ecoinf.2020.101185> . Accessed 2025-06-16
- [51] Alsentzer, E., Murphy, J., Boag, W., Weng, W.-H., Jindi, D., Naumann, T., McDermott, M.: Publicly Available Clinical BERT Embeddings. In: *Proceedings of the 2nd Clinical Natural Language Processing Workshop*, pp. 72–78. ACL, Minneapolis, Minnesota, USA (2019). <https://doi.org/10.18653/v1/W19-1909> . <https://aclanthology.org/W19-1909> Accessed 2022-08-31
- [52] Phan, L.N., Anibal, J.T., Tran, H., Chanana, S., Bahadroglu, E., Peltekian, A., Altan-Bonnet, G.: SciFive: a text-to-text transformer model for biomedical literature. *arXiv. arXiv:2106.03598 [cs]* (2021). <https://doi.org/>

- [10.48550/arXiv.2106.03598](https://arxiv.org/abs/2106.03598) .
<http://arxiv.org/abs/2106.03598>
 Accessed 2023-08-17
- [53] Lehman, E., Johnson, A.: Clinical-T5: Large Language Models Built Using MIMIC Clinical Text. PhysioNet. <https://doi.org/10.13026/RJ8X-V335> . <https://physionet.org/content/clinical-t5/1.0.0/> Accessed 2024-01-28
- [54] Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., Davison, J., Shleifer, S., Platen, P., Ma, C., Jernite, Y., Plu, J., Xu, C., Scao, T.L., Gugger, S., Drame, M., Lhoest, Q., Rush, A.M.: HuggingFace's Transformers: State-of-the-art Natural Language Processing. arXiv. arXiv:1910.03771 [cs] (2020). <https://doi.org/10.48550/arXiv.1910.03771>
- [55] Fleiss, J.L.: Measuring nominal scale agreement among many raters. Psychological Bulletin **76**(5), 378–382 (1971) <https://doi.org/10.1037/h0031619> . Place: US Publisher: American Psychological Association
- [56] Sylolypavan, A., Sleeman, D., Wu, H., Sim, M.: The impact of inconsistent human annotations on AI driven clinical decision making. npj Digital Medicine **6**(1), 26 (2023) <https://doi.org/10.1038/s41746-023-00773-3> . Publisher: Nature Publishing Group. Accessed 2025-07-12
- [57] Landis, J.R., Koch, G.G.: The Measurement of Observer Agreement for Categorical Data. Biometrics **33**(1), 159–174 (1977) <https://doi.org/10.2307/2529310> . Publisher: International Biometric Society. Accessed 2025-06-27
- [58] Leaman, R., Khare, R., Lu, Z.: Challenges in clinical natural language processing for automated disorder normalization. Journal of Biomedical Informatics **57**, 28–37 (2015) <https://doi.org/10.1016/j.jbi.2015.07.010> . Accessed 2023-11-04
- [59] Booth, B.M., Narayanan, S.S.: People make mistakes: Obtaining accurate ground truth from continuous annotations of subjective constructs. Behavior Research Methods **56**(8), 8784–8800 (2024) <https://doi.org/10.3758/s13428-024-02503-3> . Accessed 2025-06-27
- [60] Joseph, J., Tolley, N.: Are consultants more likely to decide to operate than junior doctors? The Bulletin of the Royal College of Surgeons of England **97**(4), 174–177 (2015) <https://doi.org/10.1308/147363515X14134529302263> . Publisher: The Royal College of Surgeons of England. Accessed 2025-07-11
- [61] AoMRC: The benefits of consultant-delivered care. Technical report, Academy of Medical Royal Colleges (AoMRC), London (2012). https://rcem.ac.uk/wp-content/uploads/2021/11/The_Benefits_of_Consultant_Delivered_Care.pdf
- [62] Papineni, K., Roukos, S., Ward, T., Zhu, W.-J.: Bleu: a Method for Automatic Evaluation of Machine

- Translation. In: Isabelle, P., Char-
niak, E., Lin, D. (eds.) Proceed-
ings of the 40th Annual Meet-
ing of the Association for Com-
putational Linguistics, pp. 311–318.
Association for Computational Lin-
guistics, Philadelphia, Pennsylva-
nia, USA (2002). <https://doi.org/10.3115/1073083.1073135>
- [63] Lin, C.-Y.: ROUGE: A Package
for Automatic Evaluation of Sum-
maries. In: Text Summarization
Branches Out, pp. 74–81. Associa-
tion for Computational Linguistics,
Barcelona, Spain (2004). <https://aclanthology.org/W04-1013/>
Accessed 2025-06-17
- [64] Reimers, N., Gurevych, I.: Sentence-
BERT: Sentence Embeddings using
Siamese BERT-Networks. In: Pro-
ceedings of the 2019 Conference on
Empirical Methods in Natural Lan-
guage Processing and the 9th Inter-
national Joint Conference on Natu-
ral Language Processing (EMNLP-
IJCNLP), pp. 3980–3990. Associa-
tion for Computational Linguistics,
Hong Kong, China (2019). <https://doi.org/10.18653/v1/D19-1410>