
Do Emotions Matter?

*An Investigation of Human Emotions and Financial
Decision Making in the Digital Era*

By

SUWAN(CHENG) LONG

Supervised by

PROF BRIAN M. LUCEY



School of Business

TRINITY COLLEGE DUBLIN

THE UNIVERSITY OF DUBLIN

A DISSERTATION submitted to the University
of Dublin in accordance with the requirements
of the degree of DOCTOR OF PHILOSOPHY in
the School of Business.

MARCH 2024

DECLARATION

I declare that this thesis has not been submitted as an exercise for a degree at this or any other university and it is entirely my own work. I agree to deposit this thesis in the University's open-access institutional repository or allow the Library to do so on my behalf, subject to Irish Copyright Legislation and Trinity College Library conditions of use and acknowledgement.

Juré sur l'honneur en connaissance des conséquences juridiques.

SIGNED: Cheng Long DATE: 02/05/2024

SUMMARY

This doctoral dissertation thoroughly examines the influence of social media sentiment on financial markets, focusing particularly on the GameStop short squeeze and the cryptocurrency market behavior. The dissertation includes three main research papers that together offer new insights into how discussions on internet forums can sway market trends in both stocks and cryptocurrencies.

The first paper in this dissertation offers a detailed investigation of the GameStop short squeeze event, emphasizing the important role of social media platforms, with a particular focus on the r/WallStreetBets subreddit, in influencing the volatility and pricing of Gamestop stock price. A standout achievement of this study is the creation of a tailored Reddit dictionary based on VADER(Hutto and Gilbert (2014)), developed to examine the complex language and sentiment among investors on the forum. This innovative tool enables a more accurate analysis of how investor sentiment, particularly towards 'meme stocks,' can lead to significant price fluctuations. Over 10.8 million textual data were collected from r/WallStreetBets, through a combination of qualitative and quantitative analysis, the paper demonstrates the direct impact of collective online sentiment on the stock market, challenging traditional financial theories by illustrating the power of social media.

Followed with the Reddit-tailored VADER dictionary, the second paper progresses the discussion by developing a sophisticated sentiment analysis model specifically designed for the cryptocurrency markets. There is a research gap in the field of sentiment analysis. As general sentiment analysis tools are not able to capture the sentiments of specific terms in the special alternative finance market. By adopting a machine learning-based textual analysis approach, Logistic Regression, Random Forest, and XGBoost were chosen based on their ability to tackle multiclass classification, given the diverse sentiments expressed across platforms like Reddit threads, posts, and Twitter tweets. The chosen optimal model is refined with a lexicon enriched with cryptocurrency-specific terminology, making it a novel instrument for precise mapping and evaluating sentiment trends within these digital markets. The development of such a tool has substantial value to practitioners in the rapidly evolving world of cryptocurrency trading.

In the third paper, the initial studies are expanded to examine the broader implications of sentiment analysis across the cryptocurrency market. More than 600 million text data are collected between January 1, 2018, to June 30, 2021, through various subchannels and keywords on Reddit and Twitter, to examine the intraday interconnectedness between crypto market sentiments and cryptocurrency price volatilities. This comprehensive analysis highlights the time-varying dynamic relationship between retail investor sentiment and cryptocurrency price volatility. It details the efficacy of high-frequency data in uncovering complex market patterns, sentiment-driven trading behaviors, and the interconnectedness of different cryptocurrencies. The results show that market sentiment is the net recipient of the network shocks overtime, both at low- and high- frequency. Market volatility, especially the prices volatility from Bitcoin and Ripple, play the shocks transmitter

role in the network. By illustrating the critical role of timely and detailed data in determining market trends, the paper advances the field of sentiment analysis, proposing innovative methodologies for predicting market movements. This research underscores the importance of sentiment analysis in understanding the mechanisms of market volatility, especially in the fast-paced and increasingly growing cryptocurrency market.

In summary, this dissertation clarifies the powerful effect that online discussions can have on market movements, covering both stock and cryptocurrency markets. It introduces new tools and methods for analyzing markets, and examines how the online discussion act to the market volatility, providing valuable perspectives for investors, analysts, academics, and policymakers. This thesis offers an replicable methods to develop sentiment analysis tools for any specific fields, and a ready-to-use sentiment lexicon for cryptocurrency market. This research opens doors for further exploration in the ever-changing areas of behavioral finance and market analysis, aiming to deepen the academic comprehension of how social media sentiment influences financial markets.

DEDICATION

To whom may concern.

ACKNOWLEDGMENTS

In 2021, I wrote a letter to my 27-year-old self, outlining my expectations for doctoral studies and my post-PhD career direction. Time has flown by, and I have completed the final draft of my doctoral thesis; I am about to graduate.

I used to joke that when I graduate with my Ph.D., my acknowledgments section would probably be over three thousand words long because I received so much help during this process. It is their support, encouragement, and help that motivated me to finish my studies and this doctoral dissertation. The help I received was not only academic but also contributed to shaping my character and moral conduct. Support from family, teachers, and friends has made me a better person, which is more important than earning a doctoral degree.

Firstly, I would like to express my gratitude to my doctoral supervisor, Prof. Brian M. Lucey. Throughout my doctoral journey, and indeed in my life, one of the luckiest things that happened to me was having such an excellent mentor. He not only taught me how to be a scholar but also how to be a better person. He led by example in academic research, teaching me to study diligently, approach each research question rigorously, and take responsibility for my academic output. He would recognize my research questions and help me to explore them one by one to solve any problems I might have. Prof. Lucey is an impeccable PhD supervisor.

Whenever I bragged about my supervisor to my friends, they were envious. How many supervisors, like Prof. Lucey, promptly respond to your academic queries late at night (he has been doing this since I met him until now, as I am about to graduate), proudly recommend you to his academic collaborators to help you network and participate in more research projects, and are proud of all the successes, big and small, you achieve, never sparing praise and recognition? Meanwhile, Prof. Lucey would 'force' me to take breaks and go on vacations. He never wanted me to sacrifice my time off to work. He always told me not to burn out, research and study are not everything, but to enjoy the beauty of life. I sometimes wonder can I be as good as Prof. Lucey when I become a Ph.D. supervisor? I will continue learning from him for the rest of my life.

I would also like to thank Prof. Michael Dowling from Dublin City University for his support and help. Prof. Dowling taught me a lot of analytical and coding skills in our collaborative projects. He also gave me the opportunity to work as a lecturer and teach independently in the early stages of my PhD, equipped me with the essential skills I shall have in academia. I would also like to thank Prof. Larisa Yarovaya from the University of Southampton, Prof. Ying Xie from the University of Cranfield, Prof. Zhengyuan Zhou from New York University, Prof. Dayong Zhang from Southwest University of Finance and Economics, and Prof. Andrew Urquhart from the University of Reading for their help in my academic research. Special thanks to my colleague Alexander Neumueller at the Cambridge Centre for Alternative Finance. He helped me understand more about cryptocurrency and generate new research ideas. I am grateful to them for tolerating me as an immature junior researcher, helping me to improve my academic skills, believing in me, and encouraging me. Without their help and support, I would not have been

able to participate in so many academic projects or have so many academic outputs during my doctoral studies.

None of this would have been possible without the support and love of my parents. I want to express my gratitude here to my father, Siyi Long, and my mother, Jianxiang Chen. Neither of them attended university, so my parents had no concept of these things when I chose to study abroad and pursue a doctorate. However, they did not hinder me at all because they didn't understand; on the contrary, they wholeheartedly supported me in everything I wanted to do. My parents often say to me, "Take responsibility for your own choices." Because of this, I am very independent and responsible, carefully considering every decision before making it. They often feel indebted to me and always say to me, " Mom and Dad don't know anything and can't help you." The fact is my parents have sacrificed too much for me, and I will never be able to repay their love and kindness in my lifetime. I also want to thank my younger brother, Tao Long, for taking care of our parents during my years abroad. I want to express my gratitude to my fiancé, Xing Zhao. Thank you for your love and support. My family is my rock, giving me the courage and strength to pursue what I want to do and complete my studies.

During this journey of pursuing my doctorate, I have many friends to express gratitude towards. Thank you to Weiran Huang for tirelessly and patiently teaching me coding skills while I studied programming. Thank you to Dr. Jacqueline Rossovski for her continuous encouragement during my doctoral studies. Thank you to Xiaoqin Peng and her family for taking care of me while I lived in Ireland. Thank you to my good friend Dorothy Yang, I hope her doctoral studies go smoothly. I especially want to thank my best friend Demi Han, who has been my emotional

support, whether I was laughing or having a breakdown, she was always there for me. And thank you to the ones I cannot be able to list them all here. I will pass on your kindness to others as you did.

Lastly, I would like to thank myself. Thank you to the self who has always strived relentlessly and pursued continuous improvement. From the dilapidated primary school in a small village to gaining admission to a provincial key high school, to receiving a scholarship to study at University College Dublin in Ireland, and then receiving a full scholarship to pursue a doctorate at Trinity College Dublin, and further becoming a visiting scholar at the University of Cambridge. Only I understand the hardships endured along the way. I made it. But I didn't come this far just to come this far. I hope I never forget my original intentions and continue to strive to become the person I aspire to be.

TABLE OF CONTENTS

	Page
List of Tables	xvii
List of Figures	xviii
1 Introduction	1
1.1 Background and Motivation	4
1.1.1 Paper 1: Gamestop Short Squeeze	4
1.1.2 Paper 2: Cryptocurrency Sentiment Lexicon	6
1.1.3 Paper 3: Sentiment Analysis and Cryptocurrency Market Dynamics	8
1.2 Aims and Objectives	9
1.3 Contribution	11
1.3.1 Paper 1: GameStop Short Squeeze	11
1.3.2 Paper 2: Cryptocurrency Sentiment Lexicon	12
1.3.3 Paper 3: Sentiment Analysis and Cryptocurrency Market Dynamics	13
1.4 Research Philosophy and Structure	15
1.4.1 Research Philosophy	15
1.4.2 Thesis Structure	17
1.5 Publications from this Doctoral Thesis	18

TABLE OF CONTENTS

2	Literature Review	21
2.1	Behavioural Finance	22
2.1.1	Psychology Foundation	24
2.1.2	Power Dynamics	29
2.1.3	Social media & Investor Decisions	33
2.2	Cryptocurrency and Market Volatility	37
2.3	Textual Analysis	40
2.3.1	Approaches	41
2.3.2	Applications	47
2.4	Exploratory Research Questions	50
3	Paper 1 - Gamestop Short Squeeze	55
3.1	Data and Methods	56
3.1.1	Data	56
3.1.2	Methods	57
3.2	Results	63
3.2.1	GameStop overview	63
3.2.2	A new lexicon and sentiments in r/WallStreetBets	67
3.2.3	The impact of Reddit net sentiment on GME	74
3.3	Conclusion	82
4	Paper 2 - Cryptocurrency Sentiment Lexicon	89
4.1	Data and Methods	91
4.1.1	Data	91
4.1.2	Methods	93
4.1.2.1	Bag-of-Words and TF-IDF Representations	96
4.1.2.2	Logistic Regression for Multiclass Classification	97

4.1.2.3	Random Forest for Sentiment Analysis	99
4.1.2.4	Extreme Gradient Boosting (XGBoost) for Sentiment Analysis	101
4.1.2.5	Training and Testing Dataset Construction	103
4.1.2.6	Machine Learning Model Development	105
4.1.2.7	Further Adjustment	106
4.2	Results and Discussion	108
4.2.1	Evaluation of Machine Learning Models on Initial Datasets .	108
4.2.2	Performance Analysis with Enriched Datasets	109
4.2.3	A More Accurate Model in the Market	111
4.2.4	A New Crypto-Sentiment Lexicon	112
4.3	Conclusion	113
5	Paper 3: Sentiment Analysis and Cryptocurrency Market Dynamics	117
5.1	Data and Methods	119
5.1.1	Data	119
5.1.2	Methods	121
5.1.2.1	Sentiment Analysis	121
5.1.2.2	Granger Causality	122
5.1.2.3	Time Varying Parameter Vector Autoregression (TVP-VAR)	123
5.2	Results	129
5.3	Conclusion	142
6	Conclusion	157
6.1	Summary of Results	157
6.1.1	GameStop Short Squeeze	157

TABLE OF CONTENTS

6.1.2	Crypto Sentiment Lexicon	159
6.1.3	Sentiment Analysis and Cryptocurrency Market Dynamics .	160
6.2	Future Research and Outlook	162
	Bibliography	165

LIST OF TABLES

TABLE	Page
3.1 Number of r/WallStreetBets threads and comments	57
3.2 New words dictionary and their Valence Scores	68
3.3 Updated VADER keywords and valence scores	69
3.4 Granger causality results at 1-min, 5-min, 10-min, and 30-min frequency	81
4.1 Reddit subchannels	93
4.2 Twitter Bitcoin&BTC Tweets	93
4.3 Initial Datasets Model Performance	109
4.4 Enriched Datasets Model Performance	111
4.5 Sample Cryptocurrency Sentiment Lexicon	116
5.1 Averaged total positive (negative) dynamic connectedness table	135
5.2 Averaged long-term positive (negative) dynamic connectedness table . .	136
5.3 Averaged short-term positive (negative) dynamic connectedness table .	136

LIST OF FIGURES

FIGURE	Page
2.1 Muller-Lyer illusion	26
3.1 GME and Russell 2000 index, one-minute frequency	65
3.2 Word cloud of all posts	66
3.3 Positive and Negative Sentiments at 1 Minute Frequency	70
3.4 Positive and Negative Sentiments at 5 Minute Frequency	71
3.5 Positive and Negative Sentiments at 10 Minute Frequency	71
3.6 Positive and Negative Sentiments at 30 Minute Frequency	72
3.7 Net Sentiments at 1 Minute Frequency	72
3.8 Net Sentiments at 5 Minute Frequency	73
3.9 Net Sentiments at 10 Minute Frequency	73
3.10 Net Sentiments at 30 Minute Frequency	73
3.11 1-minute net sentiment and GME prices	75
3.12 1-minute net sentiment and GME returns	75
3.13 1-minute net sentiment and GME trading volume	75
3.14 5-minute net sentiment and GME prices	76
3.15 5-minute net sentiment and GME returns	77
3.16 5-minute net sentiment and GME trading volume	77
3.17 10-minute net sentiment and GME prices	78

3.18	10-minute net sentiment and GME returns	79
3.19	10-minute net sentiment and GME trading volume	79
3.20	30-minute net sentiment and GME prices	80
3.21	30-minute net sentiment and GME returns	80
3.22	30-minute net sentiment and GME trading volume	80
5.1	BTC Granger Causality P values - Sentiments to Volatility	130
5.2	ETH Granger Causality P values - Sentiments to Volatility	130
5.3	XRP Granger Causality P values - Sentiments to Volatility	130
5.4	DOGE Granger Causality P values - Sentiments to Volatility	131
5.5	BTC Granger Causality P values - Volatility to Sentiments	132
5.6	ETH Granger Causality P values - Volatility to Sentiments	133
5.7	XRP Granger Causality P values - Volatility to Sentiments	133
5.8	DOGE Granger Causality P values - Volatility to Sentiments	133
5.9	Total, long-term and short-term connectedness	137
5.10	BTC Net total, long-term, and short-term directional connectedness measures	139
5.11	ETH Net total, long-term, and short-term directional connectedness measures	139
5.12	XRP Net total, long-term, and short-term directional connectedness measures	139
5.13	DOGE Net total, long-term, and short-term directional connectedness measures	140
5.14	SENTIMENT Net total, long-term, and short-term directional connect- edness measures	140
5.15	Connectedness Spikes and Their Associated Whales Activities	141

LIST OF FIGURES

5.16	Total, long-term and short-term connectedness at 100 days rolling window	147
5.17	BTC Net total, long-term, and short-term directional connectedness measures at 100 days rolling window	148
5.18	ETH Net total, long-term, and short-term directional connectedness measures at 100 days rolling window	148
5.19	XRP Net total, long-term, and short-term directional connectedness measures at 100 days rolling window	149
5.20	DOGE Net total, long-term, and short-term directional connectedness measures at 100 days rolling window	149
5.21	Total, long-term and short-term connectedness at 150 days rolling window	150
5.22	BTC Net total, long-term, and short-term directional connectedness measures at 150 days rolling window	151
5.23	ETH Net total, long-term, and short-term directional connectedness measures at 150 days rolling window	151
5.24	XRP Net total, long-term, and short-term directional connectedness measures at 150 days rolling window	152
5.25	DOGE Net total, long-term, and short-term directional connectedness measures at 150 days rolling window	152
5.26	Total, long-term and short-term connectedness at 200 days rolling window	153
5.27	BTC Net total, long-term, and short-term directional connectedness measures at 200 days rolling window	153
5.28	ETH Net total, long-term, and short-term directional connectedness measures at 200 days rolling window	154

5.29	XRP Net total, long-term, and short-term directional connectedness measures at 200 days rolling window	154
5.30	DOGE Net total, long-term, and short-term directional connectedness measures at 200 days rolling window	155

INTRODUCTION

Behavioral finance is an in-depth study of the psychological factors and emotional reactions that influence investment decisions and market behavior, through which we can explore human emotions and financial decision-making in the financial markets. As we step into the digital era, the integration of social media has reshaped the way information is disseminated and consumed, greatly affecting investor behavior and market dynamics. Contrary to the traditional stock markets, cryptocurrency markets operate round the clock and lack the typical 'circuit breaker mechanism' (Lee et al., 1994; Corwin and Lipson, 2000; Christie et al., 2002; Goldstein and Kavajecz, 2004; Abad and Pascual, 2010; Chakrabarty et al., 2011; Hautsch and Horvath, 2019), make it has the characters of more volatile price fluctuations. Not only the cryptocurrency market, but the traditional stock market also faces the power of social media. Gamestop(GME) short squeeze is a very good example. This market phenomenon highlights the need for a deeper understanding of how emotions expressed through social media drive investor decisions and influence movements in all financial markets, from stock exchanges to digital currency plat-

forms.

Historically, it has been posited by the consensus in traditional finance that significant sway over market dynamics is held by institutional investors, a principle that has been substantiated by their performance across conventional equity markets. Strict examination of publicly available information, leading to advantageous investment decisions, is often credited to these institutional entities. However, retail investors have often been deemed as less informed, due to their susceptibility to psychological and cognitive biases, resulting in potentially sub-optimal investment choices (Nofsinger and Sias (1999); Shleifer and Summers (1990)).

However, a notable shift in these power dynamics has been suggested by recent research. The influence that retail investors can have on market trends has been highlighted by emerging evidence, illustrated starkly by events like the 2021 GameStop short squeeze. Seen as an alternative to traditional finance, the cryptocurrency market also presents a novel platform for further examination of this evolving relationship.

Against this background, the study set two main objectives. The first is to develop a clear, replicable method for designing sentiment analysis tools to accurately capture and measure retail investor sentiment expressed on various social media platforms. This approach aims to provide detailed guidance on how to examine investor sentiment in social media discourse. Different markets, even the different sectors within the market, have different terminology and expressions. A general sentiment analysis model will not be sufficient enough to capture the actual human emotions. The second objective is to use this methodological framework to investi-

gate the dynamics between retail investor sentiment, the influence of major traders, and price movements in specific investment areas, including cryptocurrencies.

This research is expected to make significant contributions at both theoretical and practical levels. Theoretically, it aims to gain a deeper understanding of behavioral finance and asset pricing mechanisms in the context of digital and traditional financial markets and explore the impact of emotions and cognitive biases on market behavior. On a practical level, the ready-to-use cryptocurrency sentiment lexicon will be beneficial to market participants. The methods of how to create a sentiment analysis will help other fields to develop their own sophisticated analysis tools. These tools will help investors understand and respond to sentiment-driven price changes and the potential impact of the actions of important market participants.

In conclusion, this thesis aims to explain the complex relationship between investor sentiment, primary trading activity, and market volatility by introducing an approach to developing specialized sentiment analysis tools. The results of this study are expected to enhance our understanding of financial market dynamics, providing investors and regulators with new perspectives for informed decision-making and effective market regulation.

1.1 Background and Motivation

1.1.1 Paper 1: Gamestop Short Squeeze

The significance of social media platforms in shaping investment decisions and stock market outcomes has been gaining considerable academic attention. Various platforms like Twitter, Facebook, Weibo, StockTwits, and others have been scrutinized for their potential to influence market sentiments and, by extension, stock prices (Chahine et al. (2015), Ahmad et al. (2016), An et al. (2020), Bajo and Raimondo (2017), Antweiler and Frank (2004), Behrendt and Schmidt (2018), Al Guindy (2021), Danbolt et al. (2015), Feng and Johansson (2019), Cookson and Niessner (2019)). Reddit, however, has emerged as a new and unique contributor to this phenomenon, as evidenced by the GameStop case in early 2021.

An unusual dynamic in the financial markets was demonstrated by the GameStop saga, where an unprecedented surge in the stock price of the video game retailer was driven by an army of retail investors rallying on the subreddit r/WallStreetBets channel. This phenomenon was not driven by the company's financial fundamentals but by a mass sentiment wave against institutional investors who had bet heavily against GameStop. The potential of online discussion platforms, now capable of mobilizing retail investors in ways that challenge the traditional market structure and power dynamics, was underscored by the event.

The social media-driven frenzy around GameStop's stock is a manifestation of how market sentiment can be generated and amplified, even in the absence of any significant change in a company's underlying business fundamentals. Behavioral finance has been brought to the fore, offering a fresh lens through which investor

biases and herd behavior, and their potential impacts on market outcomes, can be examined. Intriguing research questions about the role and power of social media platforms in shaping investor sentiments and stock price movements have been prompted by these phenomena.

Several key motivators drive this research. First, the research aims to explore the significant shift in power dynamics enabled by social media, where market trends can be affected collectively by retail investors, defying traditional financial structures. Second, to investigate the behavioral aspect of finance, examining how investor sentiments on social media platforms can lead to herd behavior, speculative bubbles, and potential market disruptions. Finally, to understand the limitations of existing sentiment analysis tools when applied to unconventional communication styles prevalent on platforms like Reddit.

There is a significant gap in the literature, presented by the lack of a Reddit-specific lexicon. This research designed a unique lexicon to capture the essence of investment sentiments expressed on Reddit, marked by its peculiar language forms that often defy conventional linguistic norms. It contributes to the existing sentiment analysis literature, offering a new tool for dissecting investor sentiments on this influential platform.

Additionally, this research broadens our knowledge of how social media sentiment affects stock prices by examining the high-frequency, intraday stock prices of GameStop. This analysis is expected to provide valuable insights to both academics and market professionals, especially considering the significant influence platforms like Reddit now have on investor sentiment.

1.1.2 Paper 2: Cryptocurrency Sentiment Lexicon

The increasing importance of the cryptocurrency market, known for its volatility and unique trading dynamics, is captured by the attention of both the investing community and academic researchers. The second paper of this dissertation aims to extend the scope of sentiment analysis, traditionally applied to stock markets, to the domain of cryptocurrencies. Inspired by previous work on customizing sentiment analysis tools like VADER for specific market contexts, such as the GameStop event, this research proposes the development of a machine learning-based sentiment analysis model specifically for cryptocurrency markets.

The need for such a model has been underscored by the distinct nature of the cryptocurrency market. Characterized by its own set of jargon, investor demographics, and behavioral patterns, the crypto market is unlike traditional financial markets. These unique attributes demand a specialized approach to sentiment analysis. Existing sentiment analysis tools, primarily tailored for more conventional markets, are often found short in capturing the nuances and complexities of cryptocurrency-related communications. The development of a sentiment analysis model that accurately interprets the sentiments of cryptocurrency traders expressed on various social media platforms is aimed to be accomplished by this research.

The motivation behind this research is varied. Firstly, the high volatility of the cryptocurrency market makes it a prime subject for sentiment analysis. Investor sentiments, as reflected in social media discussions, can have a profound impact

on market movements. Understanding these sentiments is therefore crucial for predicting market trends and making informed trading decisions. The key to understanding the sentiment is to have state-of-the-art machine learning techniques, tailored to the specific linguistic and behavioral patterns of the cryptocurrency market.

Another driving factor behind this research is its practical implications. By providing investors and market analysts with a tool specifically designed for cryptocurrency sentiment analysis, this study is aimed to aid in better investment decision-making. The development of such a tool is of academic interest and substantial value to practitioners in the rapidly evolving world of cryptocurrency trading.

In conclusion, the development of a novel sentiment analysis model for the cryptocurrency market is pursued by this research, addressing a significant gap in the current literature. Its findings are anticipated to provide an efficient and accurate tool to help better examine the relationship between social media sentiments and cryptocurrency market movements, benefiting both academic researchers and financial market participants.

1.1.3 Paper 3: Sentiment Analysis and Cryptocurrency

Market Dynamics

Marked by its distinctive characteristics and explosive expansion, the cryptocurrency sector stands apart from conventional financial markets, especially in terms of its operational mechanisms and market conduct. Cryptocurrencies exhibit volatility that is markedly higher than that of traditional financial instruments, such as stocks and bonds. This volatility is not only a reflection of the speculative nature of these digital assets but also a consequence of several unique factors. These include the relatively new cryptocurrency market, which leads to less liquidity and can cause price movements to be more pronounced based on smaller transactions. Additionally, regulatory news, security breaches, technological advancements, and shifts in investor sentiment can all have immediate and dramatic effects on cryptocurrency prices.

Moreover, the decentralized nature of cryptocurrencies, free from central banks or government intervention, means that price determination is largely influenced by market forces of supply and demand. However, this also leaves them particularly susceptible to sentiment-driven swings, as investor perceptions and social media hype can lead to rapid price increases or crashes.

The main motivation for this study is the need to understand the factors that contribute to the high volatility characteristics of the cryptocurrency market. The market sentiment reflects the collective mood and expectations of investors and plays a crucial role in driving market trends and triggering rapid price fluctuations. Social media platforms are increasingly becoming central hubs for information dissemination and opinion formation in the cryptocurrency space. By analyzing the

sentiment expressed on these platforms, valuable insights can be gleaned about investor behavior and its subsequent impact on market movements.

The practical applications of this research are various, extending benefits to various market stakeholders, including investors, traders, market analysts, and regulatory bodies. A deeper understanding of the correlation between social media sentiment and market volatility can significantly improve investment strategies, risk management practices, and policy-making in the cryptocurrency market.

Finally, a contribution to the broader domain of behavioral finance is made by this research. By examining how investor sentiments, communicated through social media, influence the highly reactive and speculative cryptocurrency markets, the study offers new insights into the interplay between investor behavior and market dynamics.

1.2 Aims and Objectives

This thesis aims to clear the complex relationship between online sentiment and market movements, focusing on two main areas: the stock market (as seen in the GameStop short squeeze) and the ever-changing cryptocurrency market.

First, this study aims to delve into how social media, specifically platforms like Reddit, influence stock market movements. The GameStop saga will serve as a central case study, in which an explosion of activity on the r/WallStreetBets subreddit severely impacted the company's stock price. By analyzing approximately 10.8 million comments on the Reddit subreddit from January 1st to February 28th,

2021, the research develops a Reddit-specific dictionary based on VADER(Hutto and Gilbert (2014)) to capture the unique language and slang of the Reddit subreddit community. The dictionary provides valuable insights into the emerging dynamics of social media-driven investment strategies and their impact on traditional stock markets. The GameStop phenomenon highlights the transformative potential of digital platforms, effectively converting them into a critical space for trading activity that can shape stock prices and market direction.

Expanding into the cryptocurrency space, the research investigates how sentiment expressed on social media correlates with volatility observed in cryptocurrency markets. This exploration will include collecting and analyzing data from popular social media platforms such as Reddit and Twitter, focusing on important cryptocurrencies such as Bitcoin, Ethereum, Ripple, and Dogecoin. A key part of this analysis is leveraging advanced machine-learning techniques and statistical models to create a sentiment lexicon specifically designed for the cryptocurrency market. Econometric analysis models, such as Granger causality and time-varying parameter vector autoregression (TVP-VAR), are being used to uncover the complex connections between market sentiment, external events, and cryptocurrency price fluctuations.

The study provides theoretical insights and empirical results that have practical implications for a wide range of participants in financial markets. By deepening our understanding of the impact of network sentiment on market behavior and volatility, the research aims to support more strategic decision-making and planning in the stock and cryptocurrency markets.

Furthermore, this study aims to lay the foundation for future research in these dynamically developing areas. The project aims to enrich the academic fields of behavioral finance and market analysis by developing a market-specific vocabulary, applying sophisticated analytical methods, and capturing empirical evidence. This work is expected to pave the way for further research efforts to explore the detailed interactions between network sentiment and market movements, thereby broadening our understanding of these complex relationships.

1.3 Contribution

This comprehensive research series, encompassing three interconnected studies, is provided with a multifaceted contribution to the field of financial market analysis, particularly focusing on the impact of social media sentiments in stock and cryptocurrency markets.

1.3.1 Paper 1: GameStop Short Squeeze

In the first paper, an in-depth analysis of the GameStop short squeeze is offered, highlighting the critical role of social media in influencing stock market dynamics. A major contribution of this study is the development of a unique Reddit-specific lexicon. This lexicon, akin to the seminal work of Loughran and McDonald (2011), enables a nuanced understanding of the language and sentiments prevalent in Reddit's investment discussions to be understood. By analyzing 10.8 million comments from r/WallStreetBets, it is explained by the research how sentiments on social media platforms can significantly impact stock prices, particularly in the context of

'meme stocks'.

The complexities of market dynamics driven by social media are revealed in the findings, demonstrating both the power and limitations of online communities in influencing Gamestop(GME) stock price. A stronger correlation between *NET* Investment Sentiments and GameStop's short-term returns during bullish market phases is identified by the study, offering insights into the mechanics of social media influence during different market conditions. Practical implications are provided for a diverse range of stakeholders, including investors, policymakers, and media professionals, by providing a cautionary note on the reliability of social media sentiments in stock market investments.

1.3.2 Paper 2: Cryptocurrency Sentiment Lexicon

The second paper fills a key gap in current research by crafting a sophisticated sentiment analysis model designed specifically for the cryptocurrency market. At the core of the model is an innovative sentiment lexicon that has been carefully developed to reflect the unique language and expressions used by cryptocurrency traders on social media. This specialized lexicon helps to accurately identify and assess the emotions prevalent in the cryptocurrency trading environment.

This research primarily focuses on developing the tools necessary to understand the emotions expressed by cryptocurrency market participants. Unlike traditional sentiment analysis models that may not capture the unique characteristics of the cryptocurrency community in terms of vocabulary and rapid sentiment changes, this model utilizes cryptocurrency-specific terminology to provide detailed insights

into trader sentiment.

A distinguishing feature of this research is the introduction of labeling the dataset using ChatGPT, which serves as a modern alternative to the traditional time-consuming task of manual labeling. This methodological innovation represents a leap forward in the field of sentiment analysis and has the potential to simplify the process of collecting and analyzing sentiment data. By simplifying and improving the accuracy of sentiment analysis in the cryptocurrency space, the model provides a valuable resource for investors and analysts looking to cope with the complexities of the cryptocurrency market through more informed decision-making tools.

1.3.3 Paper 3: Sentiment Analysis and Cryptocurrency

Market Dynamics

This study makes a key contribution to the understanding of the dynamics of cryptocurrency markets, particularly by revealing that market sentiment is the primary recipient of shocks within connected networks. This finding illuminates the central role that investor sentiment plays in absorbing and responding to market volatility and external events, significantly affecting the overall volatility and behavior of cryptocurrency markets.

By carefully analyzing the interconnections within cryptocurrency markets, this study shows that changes in market sentiment not only reflect immediate reactions to external shocks but also serve as an indicator of the market's sensitivity and resilience to such events. This insight is critical to stakeholders across the financial

field, including investors, analysts, and policymakers, as it allows for a deeper understanding of the mechanisms driving market movements and the sentiment to predict or mitigate the impact of market-wide shocks' potential.

The revelation that market sentiment is primarily on the receiving end of shocks within cryptocurrency networks adds a new dimension to discussions of market analysis and investor behavior. It highlights the importance of monitoring and analyzing investor sentiment as a key factor in predicting market reactions to external influences.

Furthermore, this finding provides a new perspective on the dynamics of financial market sentiment analysis and enriches the academic field. It challenges existing theories and models by highlighting the need to consider temporal and directional aspects of market connectivity and sentiment flows. This contribution lays the foundation for future research aimed at interpreting the complex factors that influence market volatility, particularly in the rapidly evolving and highly speculative cryptocurrency trading environment.

1.4 Research Philosophy and Structure

1.4.1 Research Philosophy

The philosophical foundations of the research project act as key navigational tools that profoundly shape the course of the research. These foundations guide the choice of methodology and data collection methods, as well as the interpretation of research findings. More than just providing context, research philosophy is integrated into the structure of the research design, influencing every decision and every step in the process.

This section delves into the philosophical underpinnings that set the stage for this research project. It emphasizes the ontological and epistemological positions embedded in this research - the former addresses the nature of reality about the research, while the latter focuses on the nature of knowledge and the criteria for it to be considered valid in the context of the research.

This study is primarily rooted in the positivist research philosophy. Positivism originated in the field of natural science and advocates the existence of an independent, objective reality that exists beyond human cognition. Under this lens, social phenomena are considered to be similar to physical phenomena and can be systematically observed, quantified, and analyzed in an objective and replicable manner (Collins (2018)).

Positivism holds that patterns in social reality can be discovered through the use of empirical methods and statistical analysis. Therefore, the philosophical principles of positivism attempt to empirically understand the dynamics of cryp-

to currency markets, investor sentiments, and their corresponding decisions.

The chosen philosophy of positivism underpins the entire research design of this thesis. To understand the cryptocurrency market, the study acknowledges the reality of the market as an external, quantifiable entity. Likewise, investor sentiment and decision-making are considered measurable phenomena that can be analyzed objectively.

Methodologically, this translates into an empirical analysis of cryptocurrency markets, using sentiment analysis models to quantify investor sentiment and econometric models to reveal the relationship between sentiment and investment decisions. Philosophical consistency with positivism strengthens methodological choices and ensures the robustness of research designs (Edmondson and McManus (2007)).

Despite its commitment to positivism, the study acknowledges that human behavior is inherently complex and multifaceted. While quantitative data are used to discern patterns and relationships, this study also recognizes that non-quantifiable factors may influence observed phenomena. The study does not seek to provide a comprehensive understanding of the research problem; rather, it strives to provide valuable insights within the confines of its philosophical and methodological framework.

In summary, the positivist research philosophy led to the adoption of empirical analysis methods in this study. The following section details the methodological implications of this idea, outlining the research design, data collection, and analysis

strategies deployed. By strictly adhering to the principles of positivist philosophy, the research aims to conduct a systematic and transparent exploration of the stock and cryptocurrency market and investor behavioral phenomena.

1.4.2 Thesis Structure

This chapter has introduced the research background, motivation, aims, contributions, and also the research philosophy, and structure of the thesis. Following the introductory chapter, the rest of this thesis is structured as follows:

The chapter 2 provides a thorough literature review of the relevant studies in behavioral finance, cryptocurrency, market volatility, and textual analysis approaches and applications. Moreover, it points out the research questions and research hypotheses which have been tested in this doctoral thesis.

The chapter 3, chapter 4 and chapter 5 present the outcomes of the empirical analyses, showcasing the insights gained through a broad spirit of inquiry rather than through the testing of specific hypotheses. This approach reflects the exploratory nature of the research questions, which are designed to be open-ended and capable of being interpreted in various ways by different models. Consequently, this study does not set out formal hypotheses, like hypothesizing a positive relationship between certain variables. Instead, it delves into a wide-ranging exploration of how online sentiments impact market dynamics, particularly within the volatile cryptocurrency market. The findings from these analyses not only contribute to academic discourse but also hold practical and social relevance for investors, policymakers, and the academic community, all of which are discussed in detail in these

chapters.

The doctoral thesis concludes with chapter 6, which summarises the main findings, implications, and contributions and also points out several possible future research directions.

1.5 Publications from this Doctoral Thesis

Some parts of chapter 1, chapter 2, chapter 3, chapter 4, and chapter 5 are used for following academic papers. They are the main research papers of this doctoral thesis:

Long, Suwan, Brian Lucey, Ying Xie, and Larisa Yarovaya. "I just like the stock: The role of Reddit sentiment in the GameStop share rally." *Financial Review*, 2023. ABS 3, Impact Factor: 3.2

Long, Suwan(Cheng) and Chatziantoniou, Ioannis, and Gabauer, David and Lucey, Brian M., "Do Social Media Sentiments Drive Cryptocurrency Intraday Price Volatility? Evidence from Dynamic Return and Volatility Connectedness." *European Journal of Finance*, 2024. ABS 3, Impact Factor: 2.5

Long, Suwan(Cheng) and Xie, Ying and Zhou, Zhengyuan and Lucey, Brian M. and Urquhart, Andrew, "From Whales to Waves: The Role of Social Media Sentiment in Shaping Cryptocurrency Markets". *Under Review at Information Systems Research*, 2024, ABS 4*, Impact Factor: 4.9

Some parts of chapter 2, chapter 3, chapter 5 contribute to the following papers. They are the finished side projects during the doctoral research.

Pham, Linh, Sitara Karim, Muhammad Abubakr Naeem, and **Cheng Long**. "A tale of two tails among carbon prices, green and non-green cryptocurrencies." *International Review of Financial Analysis*, 2022. ABS 3, Impact Factor: 8.23

Long, Suwan, Brian Lucey, Satish Kumar, Dayong Zhang, and Zhiwei Zhang. "Climate finance: What we know and what we should know?." *Journal of Climate Finance*, 2022

Pan, Shuiyang, **Suwan Cheng Long**, Yiming Wang, and Ying Xie. "Nonlinear asset pricing in Chinese stock market: A deep learning approach." *International Review of Financial Analysis*, 2023, ABS 3, Impact Factor: 8.23

Long, Suwan Cheng, Brian Lucey, Dayong Zhang, and Zhiwei Zhang. "Negative elements of cryptocurrencies: Exploring the drivers of Bitcoin carbon footprints." *Finance Research Letters*, 2023. ABS 2, Impact Factor: 9.85

The following unpublished papers benefit from insights in chapter 2, chapter 4, chapter 5, and the visiting research at Judge Business School, University of Cambridge.

Mahdavi Ardekani, Aref and Bertz, Julie and Dowling, Michael M. and **Long, Suwan(Cheng)**, "FinSentGPT: A universal financial sentiment engine?" *Interna-*

tional Review of Financial Analysis. 2024. ABS 3, Impact Factor: 8.23

Bryce, Cormac and Dowling, Michael M. and **Long, Suwan(Cheng)** and Wardman, Jamie, "In Search of Infodemics: US Media Amplification of Risk". *Revise and Resubmit at Risk Analysis, 2024. ABS 4, Impact Factor: 4*

Long, Suwan(Cheng), Bryan, Zhang, Mingfeng, Liu and Alexander, Neumueller, The economic geography of Bitcoin mining: a global analysis, *Working paper, Cambridge Centre for Alternative Finance, 2023*

LITERATURE REVIEW

The literature review chapter thoroughly investigates the field of behavioral finance, with a special focus on the cryptocurrency market alongside the psychological foundations of market behaviors, the influence of power dynamics within financial markets, and the significant role social media plays in shaping investor decisions. This section also discusses text analysis within financial research, examining various methodologies and their applications.

By organizing and presenting these topics, the literature review aims to bridge the identified research gaps through a set of detailed research questions. These questions pave the way for discovering fresh insights into the relationship between social media sentiment and financial market volatility, particularly within the context of the ever-evolving cryptocurrency market.

2.1 Behavioural Finance

Behavioral finance represents a paradigm shift from the traditional financial theories that have long dominated our understanding of financial markets and investment decision-making. The Efficient Market Hypothesis (EMH), which posits that asset prices fully reflect all available information, has been a cornerstone of traditional finance. This hypothesis, stated by Fama in the 1970s (Fama (1970)), suggests that price fluctuations only respond to new, rational information. However, the advent of behavioral finance in the latter part of the 20th century marked a significant departure from this view, introducing the idea that psychological factors significantly influence market behaviors and investment decisions.

The transition from models based on rational expectations and market efficiencies to those incorporating psychological insights reflects a broader recognition of the anomalies within traditional models. The work of scholars like Merton (1973) and Lucas Jr (1978) laid the groundwork for integrating economic fundamentals with rational expectations. Yet, it was the empirical and theoretical advancements in the 1990s, highlighted by the pioneering work of Campbell et al. (1998), and Shefrin (2002), that signaled a shift toward understanding the psychological dimensions of financial markets. These studies underscored the limitations of existing models in capturing the complexities of market dynamics, paving the way for behavioral finance.

Behavioral finance offers a different perspective from traditional financial theories, arguing that the weak descriptive power of some models is due to the complex and nonlinear processes behind financial asset prices. Peters (1994) pointed out that stock prices and returns are cyclical and unpredictable, showing nonlinearities

and potential chaotic behavior because of time-varying positive feedback. This idea is supported by similar behaviors observed in complex social phenomena (Eve et al. (1997)), suggesting a strong link to behavioral finance. Research on individual decision-making highlights how the process can lead to positive feedback and non-linear price movements of assets (Baumol and Benhabib (1989), Richards (1990), Sterman (1988)).

Anchoring, where decision-makers rely too much on specific information (Baumol and Benhabib (1989)), along with the short-term inelastic supply of new stocks, tends to increase positive feedback effects. This often results in sharper price drops compared to rises, reflecting investors' loss aversion. Under stress, investors focus more on negative information, overestimate the probability of bad events, and show increased loss aversion as their investments shrink, which contradicts the control assumed by neoclassical economic models (Abolafia (2001)). These models, based on stable equilibrium and efficient market hypotheses, struggle during financial market disruption, when uncertainty and positive feedback are heightened.

Behavioral finance seeks to explain why investor behavior often diverges from traditional financial theory. It emphasizes cognitive biases and emotional factors affecting investor behavior and market outcomes, providing a more detailed view of financial markets than traditional models.

The following literature review will delve into the main ideas, theories, and empirical findings in behavioral finance, including pioneering work and recent developments. It will examine cognitive biases like heuristics, availability bias, overconfidence, and herding, their effects on investor behavior, and their impact on

financial markets. Additionally, it will explore the connection between behavioral finance and broader financial phenomena, such as the dynamics between institutional and retail investors, market volatility, and asset pricing anomalies.

The aim of the following review in the psychology foundation of behavioral finance and power dynamics is not only to offer a comprehensive overview of existing research in behavioral finance but also to identify gaps in current understanding.

2.1.1 Psychology Foundation

The compelling field of behavioral finance is located at the intersection of finance, economics, and psychology. Traditional finance theories, which assume rational behavior of individuals and efficient market operations, are contrasted by behavioral finance, as they acknowledge human decision-making's imperfect nature and psychology's pervasive influence on financial behavior. The entire field of behavioral finance is built upon this recognition of psychological factors, signifying a significant departure from traditional finance.

Behavioral finance challenges the notion of market stability and rational decision-making by highlighting how cognitive biases and emotional factors influence investor behavior. These biases lead to patterns in asset prices that deviate from the predictions of models based on rational expectations, suggesting a more complex and nuanced understanding of markets (Peters (1994), Eve et al. (1997)).

Prospect Theory, as established by Kahneman and Tversky (Kahneman and Tversky (1979)), represents a foundational shift in understanding investor behav-

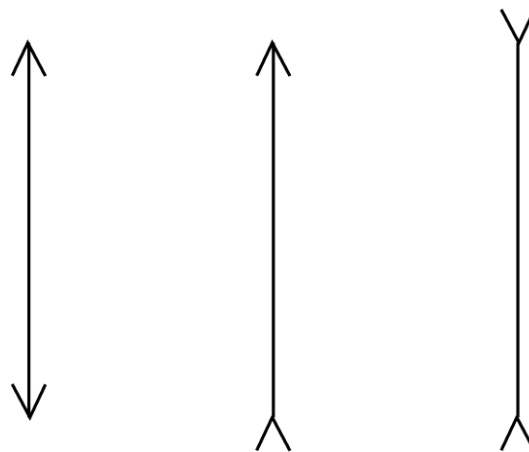
ior, particularly in the context of asset pricing and trading dynamics. This theory diverges from the Expected Utility Theory, which posits rational decision-making in financial markets (Von Neumann and Morgenstern (1947)), by introducing the concept of loss aversion and the differential weighting of gains versus losses. Cumulative Prospect Theory (Tversky and Kahneman (1992)) further refines this concept, illustrating how individuals' decisions under risk are not solely driven by outcomes but are significantly influenced by the perception of gains relative to losses.

The relevance of Prospect Theory extends beyond theoretical finance, offering insights into real-world market phenomena, such as the GameStop short squeeze event. This incident underscores the significant impact of social media on investor decisions, where digital platforms facilitated a collective movement among retail investors, challenging traditional market predictions. The event highlighted how social media can amplify behavioral biases, particularly heuristics, overconfidence, and herding, leading to market outcomes that deviate from those expected under rational market conditions.

Heuristics, or mental shortcuts, play a critical role in human decision-making, allowing individuals to make judgments swiftly and efficiently. Kahneman and Tversky's (Kahneman et al. (1982)) seminal work on heuristics highlighted their utility in navigating complex decisions, such as evaluating statistical probabilities or making choices based on incomplete data. While these shortcuts are indispensable for processing vast amounts of information quickly, they also introduce cognitive biases and errors in judgment, which can significantly impact financial decision-making (Ricciardi and Simon (2001)).

The human brain has evolved to use mental shortcuts for processing visual information, and these vision heuristics are generally effective. Yet, when these cognitive shortcuts are applied inappropriately, they can lead to heuristic biases, manifesting as optical illusions. The famous Muller-Lyer illusion, which of the three vertical lines is longer?

Figure 2.1: Muller-Lyer illusion



Heuristic bias is very difficult to overcome. As it is known that the three lines have the same length, the right sideline still appears to be longer. Just as our eyes can be deceived into making us see things that are not there (or not see things that are), our minds can be led astray by making us rely on shortcuts that don't always yield accurate judgments.

The advent of social media has amplified the impact of heuristic biases on financial markets. Platforms like Twitter and Reddit not only facilitate rapid information dissemination but also create echo chambers that reinforce heuristic-driven

perceptions. This dynamic is particularly evident in the context of market trading activities, where the *Availability Bias*, *Overconfidence*, and *Herding* significantly influence investor behavior.

The Availability Heuristic, as described by Tversky and Kahneman (Tversky and Kahneman (1992)), illustrates how easily accessible information can disproportionately influence an individual's perception of an event's likelihood. This bias becomes particularly potent in the context of social media, where information about cryptocurrencies or specific stocks can rapidly circulate, shaping retail investors' perceptions and decisions. Social media discussions can rapidly elevate specific events or sentiments, making them more "available" and perceived as more significant by investors. This phenomenon can lead to misinformed investment strategies, where decisions are based more on the visibility of information than on its veracity or relevance.

Overconfidence, a well-documented cognitive bias, has profound implications for financial markets. Studies by Fischhoff et al. (1977) and Odean (1998) illustrate how overconfidence can lead to excessive trading and misguided investment strategies. This bias encourages investors to overestimate their knowledge and underappreciate the risks involved, often resulting in suboptimal financial outcomes.

The research on overconfidence has identified its impact across various market settings. In static environments, overconfident investors are prone to trade more aggressively, while in dynamic settings, biases like self-attribution can lead to misinterpretations of market movements and flawed decision-making processes (Daniel et al. (1998), Taylor and Brown (1988), Chan et al. (2001), and Jiang et al. (2005)).

The literature points to a consistent overestimation of one's ability to predict and capitalize on market trends, leading to price overreactions and subsequent corrections. Moreover, the role of herding, where investors mimic the actions of others, is a significant factor in market anomalies like bubbles and crashes. This behavior is often driven by psychological factors such as the fear of missing out (FOMO), social influence, and the desire for conformity. When investors move together, deviating from fundamental values, they collectively contribute to price distortions. This raises concerns about market efficiency and stability.

The theoretical exploration of herding traces back to early models and discussions by Bikhchandani and Sharma (2001), who underscore the difficulties in empirically validating theoretical models of herding due to the inherent challenges in controlling for market fundamentals in real-world settings (Bikhchandani and Sharma (2001)). This gap between theory and empirical evidence prompts a reliance on experimental studies to test hypotheses under controlled conditions, as advocated by Hey and Morone (2004). Their work suggests that while markets have the potential to correct misinformed herd behavior, certain conditions may still foster it, leading to inefficient outcomes (Hey and Morone (2004)).

The interplay between heuristics and social media in financial markets underscores the need for a deeper understanding of how digital platforms influence retail investor decisions and market outcomes. The 2021 GameStop events present a unique case study in the context of herding behavior. This event, characterized by a group of individual investors on platforms like Reddit's r/WallStreetBets coordinating to buy shares of GameStop, led to a significant short squeeze. The GameStop events align with the perspectives of irrational herding and the impact of investor

psychology on market dynamics found in behavioral finance literature. It also highlights the role of digital communities and social media platforms in influencing investment decisions, a phenomenon not extensively covered in traditional herding behavior theories.

Similar to the GameStop events, the cryptocurrency market has witnessed instances where investor actions seemed to be driven more by community sentiments than traditional market indicators. Forums and social media platforms dedicated to cryptocurrencies, such as specific subreddits, Twitter handles, and other online groups, have become influential in directing investor sentiment and decision-making. This influence is often characterized by rapid swings in market sentiment, leading to sudden and substantial market movements.

The research question that emerges in this context is: How do digital communities and social media platforms impact herding behavior in the cryptocurrency market, and how does this compare to traditional financial markets? This inquiry aims to extend the understanding of herding behavior from the stock market, as evidenced by the GameStop short squeeze, to the relatively new and unregulated domain of cryptocurrencies.

2.1.2 Power Dynamics

In the sector of financial markets, the interactions and power dynamics between institutional and retail investors greatly influence market behavior, shape trends and determine outcomes in interesting ways. Behavioral finance, which has a strong interest in understanding investor behavior, provides a perspective through which

these power dynamics can be explored more effectively.

Traditionally, institutional investors - with their extensive resources, specialized knowledge, and ability to influence market trends - have dominated the landscape. They are often seen as the rational actors in financial markets, deploying sophisticated strategies that are presumed to steer markets towards efficiency.

On the contrary, retail investors, often termed 'the crowd', have typically been seen as less informed, prone to psychological biases, and hence, more likely to make suboptimal investment decisions. Their collective influence, though substantial in terms of numbers, has traditionally been discounted due to their perceived irrationality.

However, with the emergence of technology and the democratization of financial information, this traditional power dynamic is increasingly being questioned. As shown by events like the 2021 surge in GameStop stock prices, a new generation of retail investors, armed with digital tools, crowdsourced knowledge, and a desire to challenge the status quo, is being seen as a significant force.

The exploration of the power dynamics between institutional and retail investors through the lens of behavioral finance is aimed in this section. It examines how the behavior of these different groups of investors affects market outcomes, how their relative influence evolves, and what these changes mean for the future of financial markets. By gaining a deeper understanding of these power dynamics, meaningful insights into the workings of financial markets and the roles that different actors play within them can be generated.

The evolution of power dynamics in financial markets has been significantly shifted, influenced by changes in investor behavior and shifting ownership patterns. Berle and Means (1991) introduced the concept of the "rational indifference" of individual shareholders, leading to the rise of corporate "managerial control." This model is critical to understanding the dynamics of corporate governance and shareholder influence.

However, the landscape has shifted towards concentrated ownership, primarily driven by the rise of institutional investors. By 2016, institutional investors held 63% of outstanding public corporate equity, a significant increase from the 6.1% in 1950. By 2017, they owned about 78% of the Russell 3000 index and 80% of the S&P 500 index, representing substantial dollar values. This consolidation has profoundly impacted corporate governance and shareholder activism. Studies by Hartzell and Starks (2003) and Aghion et al. (2013) highlight the influence of institutional investors on corporate governance.

Recent research further expands these dynamics. Institutional investors typically outperform the market before fees, often at the expense of retail investors who tend to underperform. Kang et al. (2022)'s study in the Chinese stock market showed how institutional investors exploit the irrational behaviors of retail investors for excess gross returns, though these are often offset by mutual funds' total expenses. Additionally, institutional investors frequently take the 'wrong side' of stock return anomalies, buying overvalued stocks and selling undervalued ones. This behavior, aligned with the 'sophisticated institutions hypothesis,' suggests that institutional trading may contribute to market mispricing (Hartzell and Starks

(2003), Aghion et al. (2013)).

Concurrently, the role of media and social media in influencing market activities has become increasingly significant. Market sentiment and psychology, shaped by comments and opinions on social media, play a crucial role in stock-holding decisions. Baker et al. (2017), Jiao et al. (2020), Light (2012) and Conway (2012) note that institutional investors analyze social media content for insights into investor and consumer sentiment.

The rise of the GameStop phenomenon in early 2021, driven by retail investors coordinating through platforms like Reddit's r/WallStreetBets, challenges the traditional dominance of institutional investors and raises questions about the true nature of power in contemporary financial markets. This event underscores an emergent form of shareholder activism, not accounted for in traditional models of corporate governance and market power.

This evolving scenario reveals a research gap in understanding the interplay between institutional and retail investors in the modern financial context. Questions arise about how the assertiveness of retail investors, as demonstrated in the GameStop events, alters established power dynamics in financial markets, the implications for corporate governance and market behavior, and the extent to which these changes challenge traditional theories of market power and shareholder apathy as posited by Berle and Means (1991). Additionally, understanding the impact of market psychology and sentiment on financial performance, influenced by media and social media, is crucial to enhance market efficiency and mitigate potential risks.

2.1.3 Social media & Investor Decisions

As we delve into the complexities of behavioral finance, it's important to highlight the transformative impact social media has on investor decision-making. Today, internet-enabled social platforms have emerged as powerful mediums for disseminating financial information and shaping public opinion about market trends, thereby directly impacting investment decisions.

Social media platforms, such as Twitter, Reddit, and others, are revolutionizing the way information is shared and consumed. They have opened up new channels for investors to access real-time information, share their perspectives, and form collective investment strategies. More than ever before, investment decisions are now shaped by trending topics, popular sentiments, and viral posts, transcending the boundaries of traditional financial news and analyst reports.

An exciting dimension to behavioral finance is brought by these social media dynamics, adding to the complexity of our understanding of investor behavior. A unique platform in which cognitive biases and herd instincts manifest themselves vividly, often resulting in dramatic market moves, is provided by them.

The 2021 GameStop short squeeze, largely attributed to discussions in the r/WallStreetBets Reddit subreddit, is seen as a case study of the profound impact that social media can have on investor decision-making.

Investor behavior and market trends are profoundly influenced by a confluence

of factors, often marked in the notion of 'market sentiment.' This concept, rooted in the belief that the market possesses its own unique psychology, assists traders in anticipating market movements (Baker et al. (2017)). The evolving dynamics of this market psychology are increasingly shaped by the discourse on social media platforms, where a diverse array of comments and opinions offer valuable insights into collective investor sentiment. This trend is especially notable among institutional investors, who rely on social media to refine their understanding of market sentiments and inform their stock-holding decisions. Such is the significance of social media in the financial domain that content related to the stock market now occupies a substantial portion of these platforms (Jiao et al. (2020)). In a testament to this growing influence, anecdotal reports indicate that hedge funds and other financial institutions are harnessing services like Gnip for aggregating and analyzing social media content, aiming to glean insights into broader investor and consumer sentiments (Light (2012), Conway (2012)).

The interactive nature of social media allows not only for the observation of market trends but also for direct participation by investors in shaping market psychology. Their contributions to these platforms exert a considerable impact on market perceptions and behaviors (Baker et al. (2017)). This interaction has spurred a shift among traders, who are now keenly focused on devising methods to quantify market sentiment and determine its direct influence on financial market performance. A rich collection of research underscores the dual role of social media in this context. On one hand, it enhances market efficiency by integrating individual trading choices, social media inputs, and high-speed networking to yield more accurate price predictions. On the other hand, it brings to the fore challenges such as the propagation of misinformation on social networks and the unpredictability

of automated trading, exacerbated by phenomena like herding behavior (Hirschey et al. (2000), Chen et al. (2014), Jame et al. (2016), Bartov et al. (2018), Ma and McGroarty (2017)).

In the broader context of market dynamics, the media's role in guiding investor attention is critical. Notably, companies with limited investor bases tend to experience lower stock prices and higher expected returns. Media coverage can significantly enhance a company's visibility, expanding its investor base, thereby elevating market value and moderating expected returns. This correlation between media presence, investor attention, and market performance further underscores the complicated interplay among market sentiment, media influence, and financial outcomes. Consequently, there is a pressing need for robust methodologies to accurately measure market sentiment and comprehend the influence of market psychology on financial performance, thereby enhancing market efficiency and mitigating potential risks (Tetlock (2015)).

Recent research shows how social media affects investment decisions, pointing out that younger investors, mainly males with smaller investment portfolios and less financial knowledge, often look to social media for advice. This group tends to trade more frequently, with about 26 percent making over ten trades a month, compared to 17 percent of those not using social media. Their trading behavior is influenced by social interaction, peer pressure, social responsibility, and a desire to learn about investing. This situation raises important questions about whether social media acts more as an alternative or an addition to traditional financial advice. The significant increase in social media use, from 5 percent in 2005 to 72 percent in 2020, has established it as a key source of financial information and

advice(Hasler et al. (2021)). However, this reliance on social media brings risks, such as the meme stock phenomenon and unstable cryptocurrency investments, alongside the benefits of broader access to financial information and user satisfaction. The challenge remains to improve the consistency and reliability of financial advice on these platforms.

The extensive review of existing literature on the role of social media in influencing investor decisions uncovers a landscape rich with potential for further exploration. Key areas that merit deeper investigation include understanding how demographic factors like age, gender, and financial literacy influence reliance on social media for investment advice. There's a pressing need to develop reliable mechanisms for assessing the quality of financial advice on these platforms. Moreover, it is crucial to examine the specific behavioral influences of social media trends, such as the emergence of 'finfluencers' and the dynamics behind phenomena like meme stock frenzies.

Another significant question revolves around whether social media serves as a substitute for or complement to traditional financial advisory services, and how this interaction impacts investment strategies and outcomes. Finally, the ethical and regulatory implications of disseminating financial advice through social media platforms call for a more structured approach, ensuring investor protection and the integrity of financial markets. These gaps and questions pave the way for future research, aiming to navigate and harness the complex interplay between social media and investment behaviors in the evolving digital finance landscape.

2.2 Cryptocurrency and Market Volatility

The advent of cryptocurrencies, a revolutionary concept introduced by the pseudonymous Satoshi Nakamoto in 2008, has marked a new epoch in the domain of financial transactions (Nakamoto (2008)). As a foundational element of blockchain technology, cryptocurrencies have rapidly gained traction in both academic research and the financial industry, primarily for their innovative decentralized nature. Unlike traditional stock markets, the cryptocurrency markets operate continuously, without the conventional 'circuit breaker mechanism' that helps curb extreme volatility in traditional financial markets (Lee et al., 1994; Corwin and Lipson, 2000; Christie et al., 2002; Goldstein and Kavajecz, 2004; Abad and Pascual, 2010; Chakrabarty et al., 2011; Hautsch and Horvath, 2019). This absence of a market freeze mechanism, despite the presence of significant market volatility, distinctly differentiates cryptocurrency markets from established financial systems. The introduction and subsequent evolution of futures markets within the cryptocurrency landscape have led to mixed impacts on its operations and market dynamics (Corbet et al. (2018a), Jalan et al. (2021)).

The increasing complicity and distinct attributes of the cryptocurrency market have spurred a wealth of academic studies, examining this novel market from various economic and financial perspectives. These research aims can be broadly categorized into three principal areas: the exploration of risk volatility spillover within the cryptocurrency market as well as between it and other financial markets; the examination of connectedness and volatility spillover among different cryptocurrencies; and the investigation into the influence of retail investor sentiment on market volatility.

In the field of risk and volatility in investments, many studies have looked at how the prices and volatility of cryptocurrencies interact with other types of investments. For example, research by Bouri et al. (2017) suggests that Bitcoin might protect against changes in commodity prices and various measures of uncertainty. Corbet et al. (2018b) have studied how well-known cryptocurrencies relate to a wide range of financial assets over different times and analysis levels. On the other hand, Pele et al. (2021) argue that cryptocurrencies are a unique type of investment, different from traditional ones like stocks, commodities, bonds, and currencies. This idea is supported by Cheah et al. (2022), who say that Bitcoin's value changes are more affected by factors specific to Bitcoin and outside uncertainties, rather than by typical market trends.

Looking more closely at the cryptocurrency market, research by Liu (2019), using GARCH models, and Borri (2019), using CoVaR methods, have examined how volatility affects returns in leading cryptocurrencies and the spread of volatility within the crypto market. Other researchers, like Le et al. (2021), have looked into how different financial areas, including Fintech, green bonds, and cryptocurrencies, are connected, while Nugroho (2021) has studied the relationship between gold-backed cryptocurrencies and gold over time. In uncertain economic times, like after COVID-19, Huang et al. (2022) found that cryptocurrencies can provide diversification benefits in both stable and unstable conditions, offering extra benefits for investors who prefer less risk. These studies give us a lot of information about how cryptocurrencies relate to other investments, what influences their value, and how volatility spreads within the crypto market and to other areas. However, a common limitation in this research is its focus on daily data, which means it doesn't cover the faster-paced changes in the market.

The internal market dynamics of cryptocurrencies have also been rigorously examined, focusing on the connectedness and volatility spillover among different digital currencies (Corbet et al., 2018b; Caporale et al., 2018; Ji et al., 2019; Fang et al., 2021; Mensi et al., 2019; Trimborn and Härdle, 2018; Frijns and Margaritis, 2008; Gkillas et al., 2021). These scholarly efforts have provided valuable insights into the behavior and persistence of cryptocurrencies, the prediction of volatility using advanced machine learning methodologies, and the impact of structural breaks on the memory levels of cryptocurrency prices.

Another area of significant interest in the financial literature is the role of retail investor sentiment in influencing market volatility. Extensive research has been conducted to understand the relationship between investor sentiment and market performance across various contexts, such as near-term stock market returns (Brown and Cliff, 2004), the cross-section of stock returns (Baker and Wurgler, 2006), and household market-level sentiment (Da et al., 2015). These explorations have now extended into the cryptocurrency domain, examining the impact of retail trading and investor sentiment on the volatility of this burgeoning market. Studies like those by Karaa et al. (2021) have investigated intraday cryptocurrency data, uncovering a strong correlation between positive feedback trading and high or improving investor sentiment, as well as a notable association between trading volume, liquidity, and sentiment. Moreover, Yao et al. (2021) have observed that increased investor attention can significantly diminish cryptocurrency-specific risk by enhancing liquidity, especially for smaller-cap cryptocurrencies.

In conclusion, exploring the volatility of cryptocurrency markets and the role

of social media on retail investor decisions has opened a new avenue for research. This area combines the distinct features of cryptocurrencies with the dynamic nature of online investor communities, presenting both challenges and opportunities for researchers, market players, and policymakers. As the cryptocurrency market expands and becomes more interactive with traditional financial systems, further investigation is crucial to grasp its effects on global financial stability, investment strategies, and regulatory policies. The development of a sentiment analysis tool specifically for cryptocurrency-related communications stands out as a significant research need. This tool aims to directly link social media sentiments with fluctuations in the cryptocurrency market, addressing an important gap in current financial research. As we continue to explore this rapidly evolving field, the intersection of sentiment analysis, social media influence, and cryptocurrency market dynamics is likely to provide valuable insights, reshaping our understanding of financial markets in the digital age.

2.3 Textual Analysis

In today's digital world, the growth of online communication and content creation has made textual analysis a key method for understanding the vast amounts of text available online. This approach is particularly valuable in finance, where social media platforms and forums have changed how information is shared and consumed. Through textual analysis, researchers and analysts can interpret and quantify data from texts to examine sentiments, opinions, and trends within digital discussions among crypto users, retail investors, and the financial community at large.

The rise of cryptocurrencies and digital assets has led to increased interest from investors, with discussions on social media platforms providing important insights into market sentiment and investor behavior. These online conversations have a notable impact on market volatility, underscoring the importance of textual analysis in identifying trends, understanding investor sentiment, and forecasting market changes. This literature review explores how textual analysis has evolved in the financial sector, particularly its application in sentiment analysis of social media discourse related to cryptocurrencies and retail investment. Focusing on the latest advancements, including the use of large language models (LLMs), this review aims to highlight the role of textual analysis in gaining a clearer understanding of what drives market movements in the cryptocurrency domain and beyond.

2.3.1 Approaches

The evolution of textual analysis as a research tool reflects the broader technological and methodological advancements in data analysis and computational linguistics. Initially, textual analysis was primarily qualitative, relying on manual coding and interpretation of texts to identify themes and patterns. However, the limitations of these manual approaches, including their time-consuming nature and potential for bias, led to the development of more quantitative, automated methods.

Approaches to textual analysis can broadly be divided into two primary categories, namely dictionary-based techniques and machine-learning methodologies. Dictionary-based methods predominantly involve utilizing pre-determined dictionaries comprising words and their corresponding sentiment scores to gauge the sentiment inherent in a specific piece of text. Conversely, machine learning

approaches consist of training algorithms on pre-classified datasets intending to generate a model that can be subsequently deployed to analyze the sentiment embedded in a given piece of text.

Dictionary-Based Methods

The advent of dictionary-based methods marked a significant shift towards the quantitative analysis of text. These methods utilize predefined dictionaries or lexicons that contain words associated with specific sentiments or themes (Stone et al. (1966)).

Dictionary-based methods, rooted in the early days of computational linguistics, provide a straightforward approach to textual analysis by leveraging predefined lists of words associated with specific sentiments or themes. Researchers quantify the sentiment of a text by calculating the sum of sentiment scores based on the presence of dictionary words within the text. The simplicity and interpretability of these methods have made them particularly appealing for applications where transparency and explainability are crucial. Earlier research has commonly employed dictionaries like the General Inquirer (GI) built-in dictionary and DICTION for sentiment analysis (Tetlock (2007), Engelberg (2008), Feldman et al. (2008), Tetlock et al. (2008), Henry and Leone (2016)). However, one major drawback of these general language dictionaries is their non-specificity to particular domains, rendering them potentially insufficient for deriving accurate results within specific contexts such as finance. Recognizing this, significant advancements have been made in developing specialized dictionaries tailored to particular domains. The Loughran and McDonald (2011) dictionary for financial sentiment analysis exemplifies this trend, demonstrating improved accuracy over general-purpose dictionaries

in its domain. This specialization indicates the method's adaptability to the unique linguistic features of specific fields. These domain-specific tools have become increasingly prevalent in past studies, indicating their value and effectiveness within the financial domain (Doran et al. (2012), Huang et al. (2014), Jegadeesh and Wu (2013), Chen et al. (2013), Liu and McConnell (2013), Loughran and McDonald (2013)).

However, despite the advancements in dictionary-based sentiment analysis, there remain inherent limitations, especially in adapting to contextual details, idiomatic expressions, and the evolving nature of language. Efforts to introduce dynamic updating mechanisms and context-aware sentiment scoring have been proposed to address these challenges, aiming to improve the adaptability of these methods to the changing linguistic landscape.

In the context of cryptocurrency markets, where the communication among users is dynamic and filled with domain-specific jargon, there is a notable research gap. Existing sentiment analysis methods, including advanced domain-specific dictionaries, do not fully accommodate the unique lexicon and sentiment expressions prevalent in cryptocurrency forums and social media platforms. This gap highlights the need for developing a customized sentiment analysis package tailored to the cryptocurrency community. Such a tool would need to be highly adaptable, and capable of understanding and interpreting the rapidly evolving language and sentiments within this space, thereby providing more accurate and relevant insights for researchers and practitioners interested in the intersection of behavioral finance and social media sentiments within the cryptocurrency domain.

Machine Learning-Based Methods

The limitations of dictionary-based methods, including their lack of adaptability to new contexts and sensitivity to nuanced language use, paved the way for machine learning-based approaches. These methods leverage annotated datasets to train algorithms capable of classifying text according to sentiment or thematic categories (Pang et al. (2008)). The evolution from simple models, such as the bag-of-words, to more complex neural network-based models, like convolutional neural networks (CNNs) and long short-term memory (LSTM) networks, marked a significant advancement in the field (Le and Mikolov (2014); Tang et al. (2015); Yang et al. (2016)). These models are capable of capturing semantic relationships and contextual nuances, offering a more sophisticated understanding of textual data.

Relevantly, for the context of this study, sentence-level sentiment classification is employed, which involves discerning the sentiment expressed within a particular sentence (Wiebe et al. (1999)). Early research in this area incorporated parse trees along with original words as input to neural models. However, more recent advances have seen the growing popularity of CNNs and RNNs that use word embeddings as input (Wang et al. (2015)). An assortment of deep learning models has been proposed for this purpose, including recursive autoencoders networks (RAE), matrix-vector recursive neural networks (MV-RNN), recursive neural tensor networks (RNTN), tag-guided recursive neural networks (TG-RNN), tag-embedded recursive neural networks (TE-RNN) (Poria et al. (2016)). Furthermore, LSTM- and CNN-based deep neural network models have also been introduced for claim classification (Guggilla et al. (2016)), and tree-structured LSTM models have been put forth for encoding syntactic knowledge(Huang et al. (2017)).

Machine learning-based methods have revolutionized textual analysis by offering models that learn from data, thus capable of capturing the subtleties and complexities of language beyond the reach of dictionary-based methods.

The emergence of Large Language Models (LLMs) like GPT-4 has significantly impacted the field of textual analysis, heralding a new era where text generation, context understanding, and sentiment analysis are conducted with remarkable precision. This evolution, noted by researchers such as Bommasani et al. (2021), Yang and Menczer (2023), and Zhao et al. (2023), underscores the breadth of LLMs' capabilities, extending beyond specific tasks to general natural language processing and even domain-specific applications. Particularly in sentiment analysis, LLMs have shown to outperform traditional methods, with studies by Liang et al. (2022) and Zeng et al. (2022) demonstrating their high accuracy in detecting emotional inclinations in text.

In the finance sector, the application of LLMs for sentiment analysis has become a focal point of interest. The ability of models like ChatGPT to analyze financial news and social media sentiment accurately offers investors and analysts novel insights into market trends and investor sentiment. Furthermore, the examination of earnings call transcripts through LLMs, discussed by Loughran and McDonald (2016), provides a deeper understanding of company performance and market expectations.

The development of hybrid models, which merge the interpretability of dictionary-based methods with the predictive power of machine learning approaches, repre-

sents a significant advancement in textual analysis. The work of Liang et al. (2022) on GLM-130B and the application of ChatGPT in generating credibility ratings for news outlets by Yang and Menczer (2023) exemplify the potential of these hybrid approaches in offering nuanced insights while maintaining analytical transparency.

The critical examination of LLMs' application in sentiment analysis, especially within the finance sector, uncovers a rich area for future research. The insights derived from LLMs' analysis of financial discourse can revolutionize investment strategies and market analysis. As the field continues to evolve, the integration of LLMs with traditional textual analysis methodologies promises to deepen our understanding of financial markets, investor behavior, and the overall dynamics of the digital economy.

Despite these advancements, there remains a significant challenge in accurately analyzing social media texts, especially given their unique characteristics in the digital age. Social media's concise, informal, and rapidly changing content presents a complex challenge that current models struggle to address. This gap underscores the need for specialized sentiment analysis tools designed for the social media environment, aiming to improve our ability to analyze and predict market trends based on social media sentiment accurately. Addressing this challenge is crucial for leveraging sentiment analysis to its full potential in capturing the nuanced and dynamic sentiments expressed across social media platforms.

2.3.2 Applications

The application of textual analysis in finance, especially through the lens of social media sentiment analysis, has emerged as a critical tool in interpreting market sentiments and influencing investment strategies. This domain has witnessed a considerable expansion, driven by the increase of social media and the dynamic exchange of opinions and sentiments that these platforms facilitate. Below, we explore the impact of textual analysis on financial markets, emphasizing its role in social media sentiment analysis, backed by a series of studies that underscore its growing importance and application in finance.

Textual analysis has significantly advanced in the finance sector, offering novel insights into market trends and investor behaviors through the analysis of financial news, analyst reports, and, increasingly, social media content. Bollen et al. (2011) demonstrated the potential of Twitter sentiment to predict stock market movements, illustrating the power of social media in forecasting financial trends. This finding is echoed by Shen et al. (2019), who found that social media sentiment could serve as a more reliable indicator of market dynamics than traditional metrics like Google Trends.

The analysis of earnings call transcripts is another area where textual analysis has proved invaluable, providing deeper insights into company performance and future prospects. Loughran and McDonald (2016) highlighted how subtle linguistic cues in these transcripts could predict future financial performance, offering investors a nuanced tool for decision-making beyond raw financial data.

Recent studies have further expanded the application of textual analysis in fi-

nance, exploring the influence of social media sentiment on cryptocurrency markets. Chen et al. (2019), Gurdgiev et al. (2019), and Kraaijeveld and De Smedt (2020) have all highlighted the predictive power of social media sentiment, particularly Twitter sentiment, on cryptocurrency returns and market trends, suggesting a strong link between online sentiment and market performance.

Moreover, the integration of social media insights with trading decisions has been shown to enhance market efficiency, as discussed by Hirschey et al. (2000) and Chen et al. (2014). This integration, however, is not without challenges, including the spread of misinformation and the complexities introduced by automated trading systems, which can amplify herd behavior and market volatility (Ma and McGroarty (2017)).

Textual analysis has broadened its utility in the finance sector, moving into areas like risk management and regulatory compliance (Gentzkow et al. (2019), Engelberg et al. (2020), Carvalho and Plastino (2021), Ebadi et al. (2021), Jiménez-Zafra et al. (2019)). Through careful examination of sentiments and themes in financial communication, institutions can more effectively navigate the financial landscape. This process helps in identifying potential risks and ensuring that regulatory standards are met, demonstrating the critical role of textual analysis in maintaining financial stability and compliance.

The integration of textual analysis with Large Language Models (LLMs) represents a significant step forward in financial analysis. This combination enhances the traditional financial indicators with insights into the dynamic and often unpredictable sentiments expressed on social media. By offering a more complete

picture of market sentiment, textual analysis improves the understanding of market trends, facilitating a more informed approach to financial decision-making.

Despite advancements in sentiment analysis, from straightforward rule-based methods to complex machine learning models, several challenges remain. These challenges include the complexity of language, sensitivity to context, and the ever-changing nature of online slang and idiomatic expressions. Machine learning models, including neural networks and transformers like BERT and GPT, have improved the ability to capture nuanced expressions of sentiment. However, they require large amounts of data and significant computational power. Additionally, the need for sentiment analysis tools specifically designed for the cryptocurrency sector has not been fully addressed. The sector's unique language, rapidly evolving sentiments, and the global, multilingual nature of discussions on social media platforms highlight the necessity for specialized models and lexicons. These tools must be designed to accurately interpret and analyze sentiments in the highly volatile cryptocurrency market.

This area presents a significant research opportunity. There is an urgent need for sentiment analysis models that are tailored to the fast-paced and global nature of social media discourse, especially for the cryptocurrency market. These models should be capable of understanding the complexities of online communication, including the use of specific jargon, emojis, and hashtags, to provide precise insights into market sentiment. Developing such tools would not only improve our understanding of cryptocurrency market dynamics but also equip investors and analysts with valuable predictive tools. As the field of sentiment analysis evolves, tackling these challenges and pursuing innovative methods will be crucial for fully

exploiting the potential of textual analysis in today's digital age, particularly in the rapidly expanding cryptocurrency sector.

2.4 Exploratory Research Questions

This thesis undertakes an exploratory journey through the complex field of behavioral finance, cryptocurrency market dynamics, and the impact of social media on financial markets. By leveraging textual analysis, this work delves into how sentiments expressed on social media platforms influence investor behavior and market movements. Rather than positing specific hypotheses, this investigation adopts a spirit of inquiry, open to various outcomes that these complex relationships may reveal. The following exploratory questions guide this thesis:

Exploratory Question One:

- How does social media contribute to shaping herding behavior among investors within traditional stock and cryptocurrency markets?

This research question probes the degree to which social media influences herding behavior among investors in both traditional stock and cryptocurrency markets. It seeks to understand how online interactions on platforms like Twitter, Facebook, and Reddit may lead to the formation of collective investment patterns, where investors move en masse, often disregarding their individual analysis or the fundamental value of assets.

Exploratory Question Two:

How does emotional contagion through social media platforms impact investor sentiment and behavior, especially during market highs and lows?

The focus here is on understanding how emotional contagion, the phenomenon where emotions like fear or euphoria spread among users, affects investor sentiment and behavior through social media, particularly during periods of extreme market fluctuations. This question explores the potential of social media platforms to amplify emotional responses, thereby influencing market highs and lows.

Exploratory Question Three:

How do social media platforms, particularly Reddit, influence stock market movements, as evidenced by the GameStop short squeeze?

This question aims to unravel the specific role that social media platforms, with a particular focus on Reddit, play in shaping investment decisions and outcomes in the stock market. The GameStop short squeeze serves as a pivotal case study to illustrate how collective sentiment and discussions on such platforms can significantly impact stock prices and market dynamics.

Exploratory Question Four:

What are the unique characteristics of cryptocurrency market dynamics, and how can sentiment analysis be tailored to understand these markets?

The research seeks to identify the unique characteristics of cryptocurrency market dynamics and how sentiment analysis can be tailored to understand and predict these market behaviors. It considers the distinct volatility and trading patterns in cryptocurrency markets and how these might require specialized sentiment

analysis approaches.

Exploratory Question Five:

How does the crypto-specific sentiment lexicon compare to traditional financial analysis tools in terms of predicting market movements and understanding investor behaviors?

This question evaluates the effectiveness of a crypto-specific sentiment lexicon in predicting market movements and understanding investor behaviors, compared to traditional financial analysis tools. It aims to assess whether this tailored approach provides a more accurate and nuanced understanding of the cryptocurrency market.

Exploratory Question Six:

Can the newly developed crypto sentiment lexicon provide reliable predictions about cryptocurrency market trends based on the analysis of social media sentiments?

The research aims to determine the reliability and accuracy of the newly developed crypto sentiment lexicon in forecasting cryptocurrency market trends based on social media sentiments. This inquiry focuses on the lexicon's ability to interpret and predict market fluctuations and investor responses accurately.

Exploratory Question Seven:

How do sentiments expressed on social media platforms impact trends and volatility in the cryptocurrency market?

This question delves into how sentiments expressed on social media platforms influence the trends and volatility within the cryptocurrency market. It aims to uncover the relationship between online sentiment, as manifested through platforms like Reddit and Twitter, and the inherent volatility and price movements in the cryptocurrency market.

By crafting these questions within a broader context of inquiry, this thesis opens itself to diverse analytical models and interpretations, acknowledging the complicated and often unpredictable nature of the interplay between social media, investor sentiment, and market dynamics.

PAPER 1 - GAMESTOP SHORT SQUEEZE

The early 2021 surge in GameStop shares, driven by speculative trading by individual investors on the r/WallStreetBets subreddit, showcases the significant influence of online forums on the financial market. This event underlines the power of social media sentiment in altering financial markets, challenging the conventional views on investment analysis by revealing how retail investors can move markets through coordinated online actions. Prior studies have explored the impact of media sentiment on market dynamics across various platforms, including Twitter, Facebook, and Weibo, demonstrating social media's critical role in shaping financial market perceptions and actions (Renault (2017), Cookson and Niessner (2020), Corbet et al. (2021), Chahine et al. (2015), Ahmad et al. (2016), An et al. (2020), Bajo and Raimondo (2017), Antweiler and Frank (2004), Behrendt and Schmidt (2018), Al Guindy (2021), Danbolt et al. (2015), Feng and Johansson (2019)). These studies set the stage for analyzing the specific influence of the r/WallStreetBets subreddit on GameStop's intraday stock prices, looking into how the sentiment and tone of forum discussions affected stock movements during January and February 2021.

Addressing the unique linguistic challenges of the Reddit platform, marked by its specific slang, emojis, and meme culture, this research introduces a specialized Reddit investment lexicon, based on the work of Vader(Hutto and Gilbert (2014)). This tool aims to bridge the gap in sentiment analysis methodologies, which typically struggle to accurately interpret the complex language prevalent on Reddit forums (Das and Chen (2007)). The findings highlight the noticeable impact of Reddit-derived sentiments on GameStop’s intraday returns, showing how the volume and tone of subreddit discussions relate to stock price changes. This research adds to the understanding of social media’s effect on financial markets by detailing how high-frequency intraday stock price movements can be driven by online forum discussions. It aligns with the broader field of behavioral finance, underscoring the significant role of investor sentiment and cognitive biases in influencing capital markets, as discussed in key studies by Daniel et al. (2001), Corbet et al. (2020), Guegan and Renault (2021), and Akyildirim et al. (2020).

3.1 Data and Methods

3.1.1 Data

This paper utilizes high-frequency stock price data for GameStop (GME) from the 1st of January 2021 to the 28th of February 2021. The GME data and Russell 2000 index (of which GME is a constituent) are collected at 1-minute intervals from Bloomberg.

For sentiment analysis, 10.8 million comments from the r/WallStreetBets sub-

reddit for the same observation period (the first two calendar months of 2021) are collected.

Table 3.1 shows the number of comments collected for each group out of the 10.8 million comments.

Table 3.1: Number of r/WallStreetBets threads and comments

Total Threads	846,628
Total Comments	10,024,617
Total	10,871,245

3.1.2 Methods

The analysis began by extracting sentiments from texts scraped from the r/WallStreetBets subreddit. Text Sentiment Analysis is identified as a trending field with a substantial amount of academic research behind it. In the existing literature, two main approaches are employed: the lexical approach and the machine learning approach. Lexical approaches are aimed at mapping words to sentiments by building a lexicon or a "dictionary of sentiment". This dictionary can be used to assess the sentiment of phrases and sentences, without the need to look at anything else. In lexical approaches, the sentiment category or score of each word in the sentence is looked at, and the sentiment category or score of the whole sentence is decided. This approach has been utilized by Loughran and McDonald (2011) and Renault (2017), among others. On the other hand, machine learning approaches look at previously labeled data in order to determine the sentiment of never-before-seen sentences. In the machine learning approach, a model is trained using previously

seen text to predict/classify the sentiment of some new input text. The paper adapts a lexicon-based method.

VADER (Valence Aware Dictionary and Sentiment Reasoner) is a sentiment package that is explicitly sensitive to feelings expressed in social media text. VADER relies on a lexicon and five general rules to map lexical features to sentiment scores (Hutto and Gilbert (2014)). When compared with feature-based machine learning methods, VADER has a few advantages:

1) VADER was designed with a focus on social media data, with an emphasis on the rules that captured the essential meaning of social media texts. The lexicon and rules used by VADER are directly accessible and can be easily inspected and updated in a discipline context.

2) VADER does not require any training data, and it utilizes a human-validated sentiment lexicon and general rules that are related to grammar and syntax.

Threads and comments posted on Reddit are usually short sentences with emojis and rich use of punctuation, hence, VADER was selected to conduct the sentiment analysis. VADER sentiment analysis package has been employed widely across different disciplines, see Pano and Kashef (2020), Shelar and Huang (2018), Sivasangari et al. (2018), and Oliveira et al. (2016). However, the lexicon developed by Hutto and Gilbert (2014) for VADER is not specific to the finance field, thus some key financial words are excluded, such as "Bear" and "Bull". Key phrases and hashtags unique to financial social media platforms are also missing from the VADER lexicon, such as "Diamond Hands". Directly applying VADER to the

analysis of sentiment of the r/WallStreetBets subreddit may misclassify words when gauging tone in financial applications.

Latent Dirichlet Allocation (LDA) is a generative probabilistic model that is extensively used for topic modeling in natural language processing (Blei et al. (2003)). In the context of the research, the main topics present in the Reddit discussions related to GME are identified using LDA.

The LDA model is based on two fundamental assumptions:

1. A small number of topics are mixed in each document (post or comment in this context).
2. One of the document's topics is attributed to each word in the document.

In applying LDA, each document is represented as a 'bag of words,' with grammar and word order being disregarded, while multiplicity is kept track of.

The topics denoted as $\beta_{1:K}$ where K is the predefined number of topics, are represented. Each topic β_k is a multinomial distribution over the entire set of words, represented as $\beta_k = (\beta_{k1}, \beta_{k2}, \dots, \beta_{kV})$ where V is the size of the vocabulary.

The topic mixture of the d -th document is denoted as θ_d , which is a multinomial distribution over the topics, represented as $\theta_d = (\theta_{d1}, \theta_{d2}, \dots, \theta_{dK})$.

Finally, for each word w_{di} in the d -th document, a topic assignment z_{di} is made,

indicating from which topic the word is generated.

The generative process for each document in LDA can be described as follows:

1. Choose $\theta_d \sim \text{Dirichlet}(\alpha)$, where α is the Dirichlet prior.
2. For each word w_{di} in the document,
 - a. Choose a topic $z_{di} \sim \text{Multinomial}(\theta_d)$.
 - b. Choose a word $w_{di} \sim \text{Multinomial}(\beta_{z_{di}})$.

LDA aims to infer the hidden topic structure $\beta_{1:K}$, $\theta_{1:D}$, and $z_{1:D,1:N_d}$ given the observed documents, where D is the number of documents and N_d is the length of the d -th document. This can be achieved by computing the posterior distribution:

$$(3.1) \quad P(\beta_{1:K}, \theta_{1:D}, z_{1:D,1:N_d} | w_{1:D,1:N_d}) \propto P(\beta_{1:K}, \theta_{1:D}, z_{1:D,1:N_d}, w_{1:D,1:N_d})$$

However, the exact inference of LDA is rendered intractable due to the coupling between θ and β in the above equation. Therefore, approximation methods like Gibbs sampling or variational inference are often resorted to.

Using LDA modeling, the threads, and comments are first classified into 49 topics, and then a list of the 130 most common words used on r/WallStreetBets is created by analyzing the content and key information in the 49 topics. The number of LDA modeling topics is determined in such a way that the topic coherence score is maximized. To construct and validate the new VADER lexicon, a human-centered approach is adopted, and 10 annotators are hired to manually estimate the sentiment valence (intensity) of each keyword.

This approach is consistent with the methods used by Hutto and Gilbert (2014) in developing the original VADER lexicon. The lexicon in VADER is updated by 1) adding new r/WallStreetBets subreddit words and corresponding valence scores to the original lexicon, and 2) replacing the original valence scores with new valence scores if some words already exist in VADER. The updated valence scores are in the range of $[-4, +4]$, with $[-4]$ being Extremely Negative, $[0]$ being Neutral, and $[+4]$ being Extremely Positive.

The *SentimentIntensityAnalyser* object from the VADER package¹ is used to extract the *polarity_scores*. *polarity_scores* provide the overall sentiment metrics(compound score) for the comments. The compound score was computed by taking the sum of the valence scores of each word in the lexicon, adjusted according to the five general rules defined by Hutto and Gilbert (2014), and then normalized to be between -1 (most negative) and +1 (most positive).

When the compound score is greater than 0.05, it denotes a positive sentiment. When the compound score is less than -0.05, it denotes a negative sentiment. A compound score between 0.05 and -0.05 denotes a neutral sentiment. In addition to the compound score, VADER *SentimentIntensityAnalyser* also returns Positive, Negative, and Neutral scores for a text. These scores are calculated as the sum of Positive, Negative, and Neutral valence scores in the lexicon, respectively.

After the sentiment data of each thread and comment is obtained, the value 1 is assigned to the text labeled as positive, -1 to negative, and 0 to neutral as their sentiment value. The difference between the Positive and Negative measures is then taken to obtain the net sentiment score of a text. A new variable *NET* is

¹See for details: <https://www.nltk.org/api/nltk.sentiment.html>

defined as the sum of the net sentiment scores of the texts posted in a time period. The variable *AVERAGE*, which is the mean value of the net sentiment scores of all the texts posted in a time period, is also introduced. The net sentiment and average sentiment of all texts at 1-minute, 5-minute, 30-minute, and 1-hour frequency are then calculated.

The net sentiment is calculated as follows:

$$(3.2) \quad \text{net sentiment} = \text{sum of positive} + \text{sum of negative} + \text{sum of neutral}$$

The average sentiment is calculated as follows:

$$(3.3) \quad \text{average sentiment} = \frac{\text{net sentiment}}{\text{total number of texts}}$$

After sentiment variables are constructed, the impact of *NET* and *AVERAGE* sentiments on the GME intraday prices are analyzed using the standard Granger Causality approach (Granger (1969)). The Granger causality test is a statistical hypothesis test for determining whether one-time series is useful in forecasting another. Essentially, whether lagged values of Reddit sentiments can significantly explain the current values of GME prices and returns is examined.

The structuring of a Vector Autoregression (VAR) model with GME prices (or returns) and Reddit sentiments is required for the application of the Granger causality test. Specifically, if the GME prices are denoted as P_t and Reddit sentiment as S_t , the VAR model can be specified as follows:

$$(3.4) \quad P_t = \alpha + \sum_{i=1}^n \beta_i P_{t-i} + \sum_{i=1}^n \gamma_i S_{t-i} + \epsilon_t$$

$$(3.5) \quad S_t = \delta + \sum_{i=1}^n \phi_i S_{t-i} + \sum_{i=1}^n \theta_i P_{t-i} + \mu_t$$

where n is the number of lags, ϵ_t and μ_t are error terms, and $\alpha, \beta_i, \gamma_i, \delta, \phi_i, \theta_i$ are parameters to be estimated.

The null hypothesis for the Granger causality test from Reddit sentiments to GME prices is that the coefficients of the lagged values of S_t (γ_i) in the first equation are jointly zero. If this hypothesis is rejected, it implies that Reddit sentiments Granger-cause GME prices. Similarly, the null hypothesis for the Granger causality test from GME prices to Reddit sentiments is that the coefficients of the lagged values of P_t (θ_i) in the second equation are jointly zero. Rejection of this hypothesis indicates that GME prices Granger-cause Reddit sentiments.

The statistical software Stata will be used to estimate the VAR model and conduct the Granger causality tests. Insights into the causal relationship between Reddit sentiments and GME prices are provided by the results, presenting the influence of social media sentiments on GME stock prices and returns.

3.2 Results

3.2.1 GameStop overview

During the COVID-19 pandemic, businesses were struggling to stay afloat and the economic landscape was ripe for aggressive short-selling strategies. Often employed by sizable institutional investors and hedge funds, which possess significant

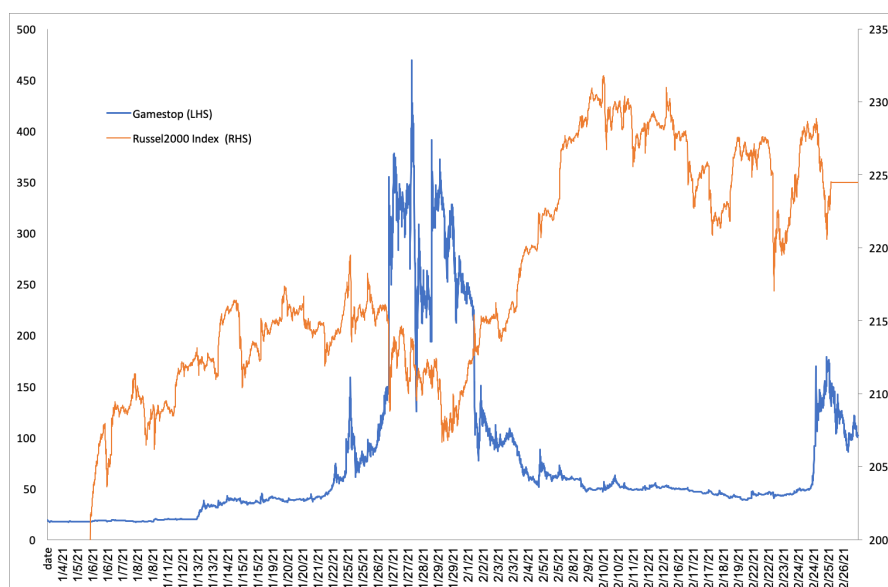
financial and human capital, these tactics can significantly impact market trends.

A prominent hedge fund, Melvin Capital, anticipating a further decrease in stock prices during the pandemic, opened short positions against GameStop shares. GameStop, a game and gaming retail business, was facing hardship amid the economic trouble induced by COVID-19. Around that period, GameStop shares were typically traded at approximately \$7 per share. However, the company soon observed a noticeable rise in its share price, instigating enough volatility for the hedge fund to initiate a short-selling strategy.

It is widely suggested that a coordinated movement of retail investors, primarily from the subreddit r/WallStreetBets, instigated a swift escalation in GameStop's share price, an occurrence that later came to be known as the GameStop short squeeze. This unexpected turn of events proved to be catastrophic for Melvin Capital, which required nearly \$3b in supplementary capital, with the cumulative loss from the short squeeze approximating \$25b². Figure 3.1 vividly illustrates the volatility of GME shares in juxtaposition with Russell 2000 over the span from January 2021 to March 2021, highlighting the wild ride GME stocks experienced in January-February 2021.

²Refer to <https://www.ai-cio.com/news/gamestops-robinhood-boosters-clobbering-hedge-funds/> for a thorough overview

Figure 3.1: GME and Russell 2000 index, one-minute frequency



The GameStop episode garnered extensive public attention and widespread media discussion. Factors such as the gamification of trading, the accessibility of financial markets to retail investors through online trading platforms like Robinhood, and the inherent volatility of this scenario resulted in extensive debates. The event culminated in a series of congressional hearings and lawsuits. It underscored the expansion of the decentralized financial system and the potential disruptive power of technology on financial markets, thereby propelling the GameStop case to prominence from a policy perspective.

Beyond GameStop, other so-called “*meme stocks*” and assets were targeted by Reddit’s retail traders, revealing Reddit as a newly discovered avenue for potential market manipulation. Compared to other platforms like Twitter, Reddit exhibits a much higher degree of disorder. Extracting sentiments from Reddit’s subreddits is, therefore, a semantically challenging task. It is noteworthy that previous research on media sentiment has mostly relied on aggregated data from traditional news or social media outlets. Few studies have focused on micro-blogging trading platforms

3.2.2 A new lexicon and sentiments in r/WallStreetBets

Developing an idiosyncratic Reddit-specific lexicon was a complicated three-step process. To initiate this process, every thread from the r/WallStreetBets subreddit was methodically categorized into 49 thematic topics via Latent Dirichlet Allocation (LDA) modeling. In the second step, a comprehensive list of the 130 most recurrent words was compiled. Subsequently, ten annotators were engaged in a manual ranking process to determine the sentiment valence of each word on this list. Table 3.2 presents a complete roster of words constituting the new lexicon, alongside their respective sentiment scores.

The implementation of this human-centric approach to valence ranking was especially beneficial in assigning sentiment scores to colloquial or jargon terms. For instance, 'to the moon' was assigned a [+3.5] score, while the terms 'yolo' and 'diamond hand' scored [+2.4] and [+3] respectively. To convey negative sentiments, users of the subreddit often resorted to standard negative words such as 'loss' [-2.5], 'wrong' [-1.8], and 'fake' [-2.3], in conjunction with a variety of curse words.

Table 3.2: New words dictionary and their Valence Scores

Word	Score	Word	Score	Word	Score	Word	Score
available	0.8	diamond_hand	3	cash	0.6	advice	1.3
awesome	3.7	dip	-0.4	concern	-1.3	alternative	0.9
baby	1.2	dumb	-1.9	crash	-3.2	amazing	3.2
bad	-2.7	earning	1.8	crazy	0.7	ass	-1.9
ball	0.4	easy	1.6	crypto	0.5	attack	-1.9
bull	2.8	end	-0.8	damn	-1.7	capital	1
bullshit	-2.4	enough	0.1	diamond	2.9	fact	0.3
buy	1.9	hype	1.2	hard	-1.1	fake	-2.3
call	0.9	idiot	-2.6	hedge	0.5	fight	-1.2
future	1.1	illegal	-3.2	hell	-2.5	fine	1.3
gain	2.2	interest	1.1	high	2.4	flair	1.4
gamma	0	issue	-1.1	hodl	2.8	fuck	-2.8
gang	-0.3	joke	-0.5	hold	1.5	fucking	-2.7
gold	2	jump	1.4	holding	1.6	fun	1.9
good	2.5	least	-0.4	hope	1.5	funny	1.9
great	3.1	legal	1.9	limit	-0.4	problem	-2.3
green	2	manipulation	-2.3	lmao	2.6	profit	2.5
hand	0.1	margin	-0.1	lol	1	proud	2.1
party	0.8	moment	0.7	long	1.8	pump	-0.5
penny	-0.2	moon	2.1	loss	-2.5	purchase	1.3
poor	-1.9	movement	0.9	love	2.3	push	0.5
possible	0.8	naked	-1.1	low	-1.7	quick	0.8
potential	1.4	nice	2	luck	2.1	retard	-2.2
power	2.2	order	0.4	revolution	2	share	0.8
pretty	2.3	panic	-3	rich	2.5	shit	-2.6
probably	0.4	straight	1	ride	1	short	-1.8
top	2.4	strong	2.1	rocket	2.8	silver	-0.2
trade	0.6	stupid	-2.1	sale	-0.7	small	-0.3
value	1.3	support	2.2	scare	-2.3	squeeze	-1.6
win	2.7	target	1.3	scared	-2.6	star	2.4
worth	1.9	tendie	1.7	sell	-1.8	stonk	1.5
wrong	-1.8	to_the_moon	3.5	seller	-1.3	stop	-0.8
yolo	2.4			selling	-1.9		

This table shows the list of words used in the new lexicon with their associated valence scores.

The third phase involved updating the existing lexicon in VADER. This entailed introducing new Reddit-specific words with their respective valence scores and revising the scores of words present in the original VADER lexicon. Table 3.3 presents a snapshot of this updating process. It's crucial to note that the updated scores not only represent variations in sentiment intensity but frequently denote shifts in the overall sentiment tone. For example, the term 'crazy' transitioned from a negative score of [-1.4] to a positive score of [0.7].

Table 3.3: Updated VADER keywords and valence scores

Existing Word	Original Score	New Scores
crazy	-1.4	0.7
crash	-1.7	-3.2
interest	2	1.1
loss	-1.3	-2.5
profit	1.9	2.5

This table shows the valence scores assigned to the *new* words and updated scores assigned to the *existing* words in the lexicon.

Figure 3.3, 3.4, 3.5 and 3.6 presented here show fluctuations in positive and negative emotion measures during the sample period, captured at different sampling frequencies. These visual representations reveal a striking symmetry between trends in positive and negative emotions, with positive emotions consistently showing higher value. The red line represents positive sentiment and the blue line represents negative sentiment.

Figure 3.3: Positive and Negative Sentiments at 1 Minute Frequency

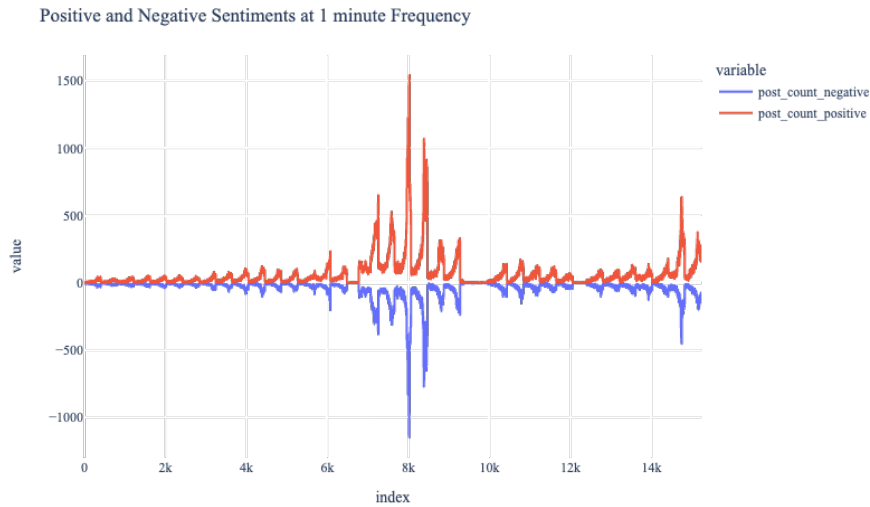


Figure 3.4: Positive and Negative Sentiments at 5 Minute Frequency

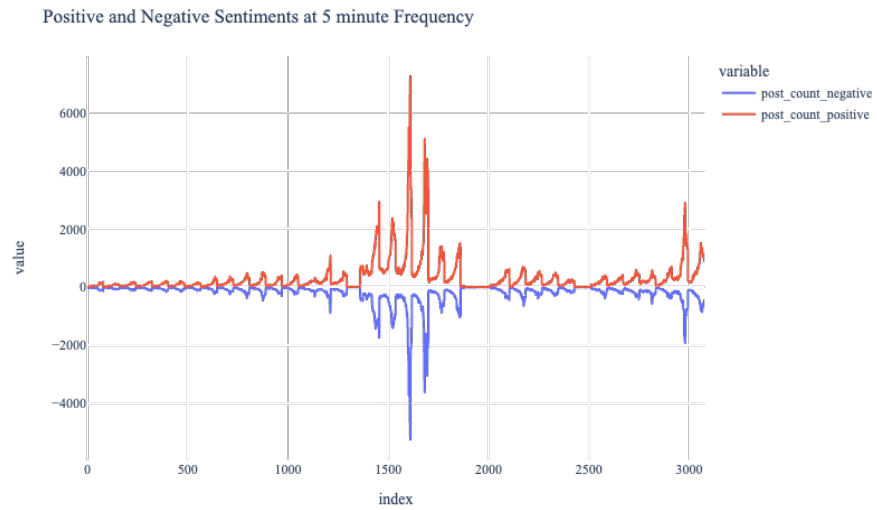


Figure 3.5: Positive and Negative Sentiments at 10 Minute Frequency

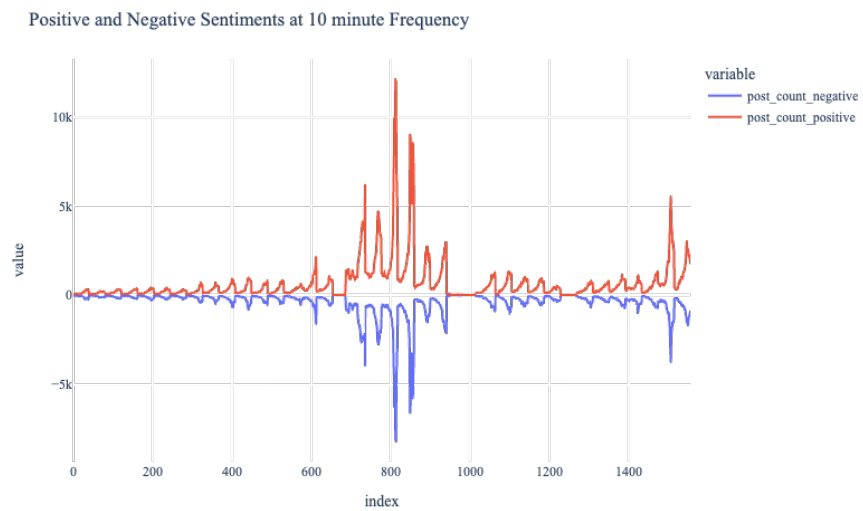


Figure 3.6: Positive and Negative Sentiments at 30 Minute Frequency

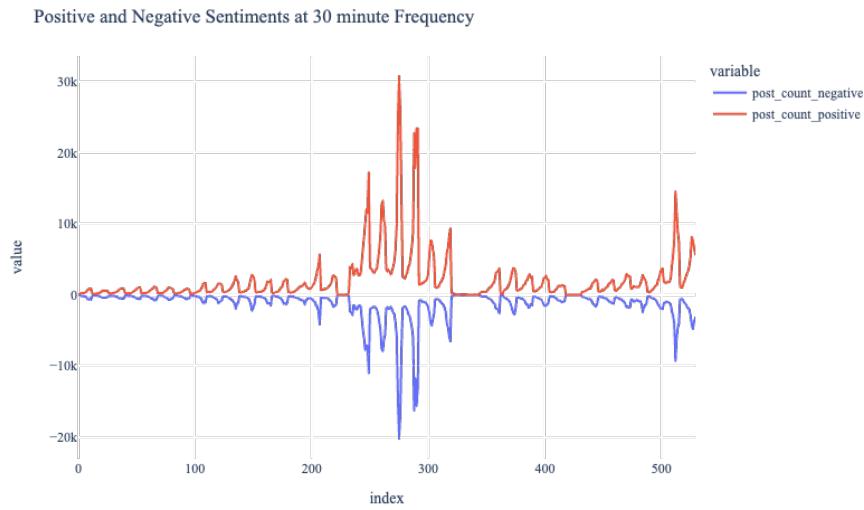


Figure 3.7, 3.8, 3.9 and 3.10 showing net sentiment at different frequencies provides a view of investor sentiment swings over a selected time frame. These charts highlight the ups and downs of net sentiment, effectively capturing changes in investor sentiment towards the cryptocurrency market. Notably, these figures show that optimism outweighed pessimism.

Figure 3.7: Net Sentiments at 1 Minute Frequency

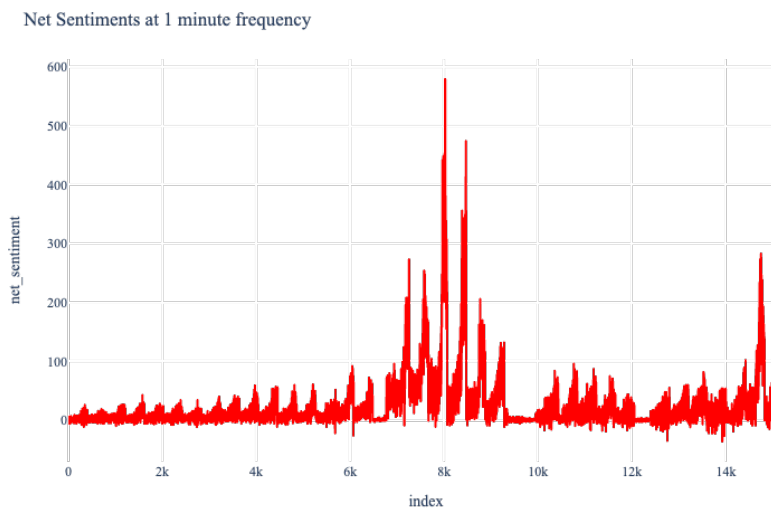


Figure 3.8: Net Sentiments at 5 Minute Frequency

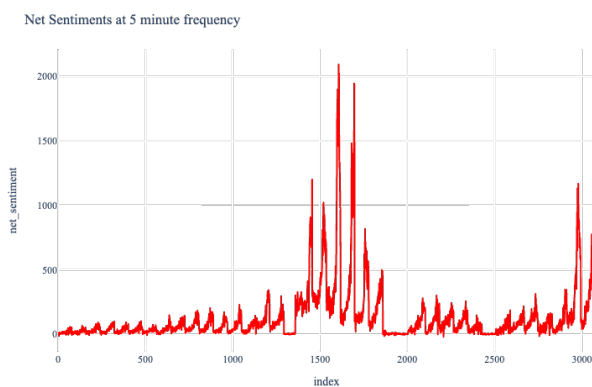


Figure 3.9: Net Sentiments at 10 Minute Frequency

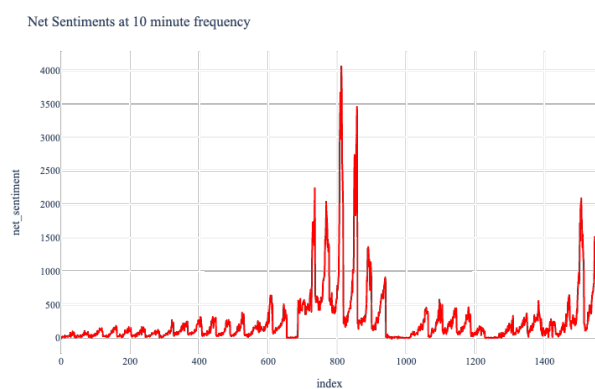
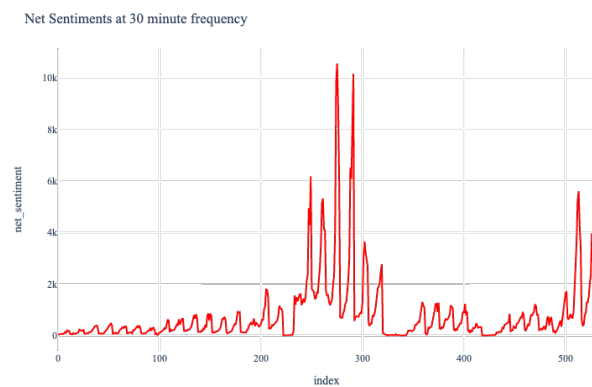


Figure 3.10: Net Sentiments at 30 Minute Frequency



3.2.3 The impact of Reddit net sentiment on GME

The impact of Reddit sentiments on GameStop is evaluated with a detailed description of the dynamics of the *NET* Sentiment (Positive - Negative), GME's opening and closing prices, Open-to-Open and Close-to-Close returns, and trading volume at 1-min, 5-min, 10-min, and 30-min frequencies. Figure 3.11, 3.12 and 3.13 display the results for the 1-min frequency.

According to Figure 3.11, the relationships between the *NET* Sentiments and GME's opening and closing prices at the 1-min frequency were considerably stronger during up-market days. This specifically refers to the period from January 20, 2021, to January 27, 2021. The active engagement of Reddit users in the discussion forum seemed to significantly influence opening and closing prices during bullish market conditions.

However, when the GME stock transitioned into bearish market, a noticeable decoupling of 1-min sentiments and opening and closing prices occurred, with each moving in contrasting directions. This phenomenon is particularly pronounced in Figure 3.11 post January 31, 2021. Despite a surge in *NET* Sentiment (Positive>Negative), the GME price continued its downward trajectory. This suggests that the positive comments and encouragement to hold GME stocks expressed on Reddit were insufficient to curb or prevent this price decline.

Figure 3.11: 1-minute net sentiment and GME prices

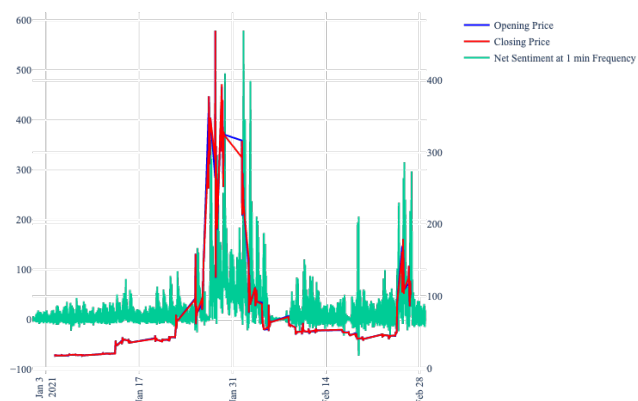


Figure 3.12: 1-minute net sentiment and GME returns



Figure 3.13: 1-minute net sentiment and GME trading volume

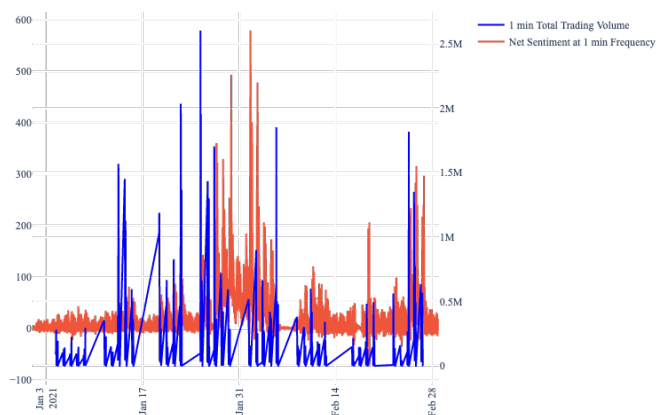


Figure 3.14, 3.15 and 3.16, which uses 5-min data, present these patterns with even greater clarity. There are observable breakpoints in the relationships between *NET* Sentiment and GME prices and returns. However, during the bearish market phase, the positive linkages become weaker. This vital data reveals the sentiment formation mechanism on r/WallStreetBets and its effect on the GME share rally, and it demonstrates that while Reddit discussions helped to fuel interest in buying GME stock, they were unable to sustain it when the price re-aligned with fair value.

Figure 3.14: 5-minute net sentiment and GME prices

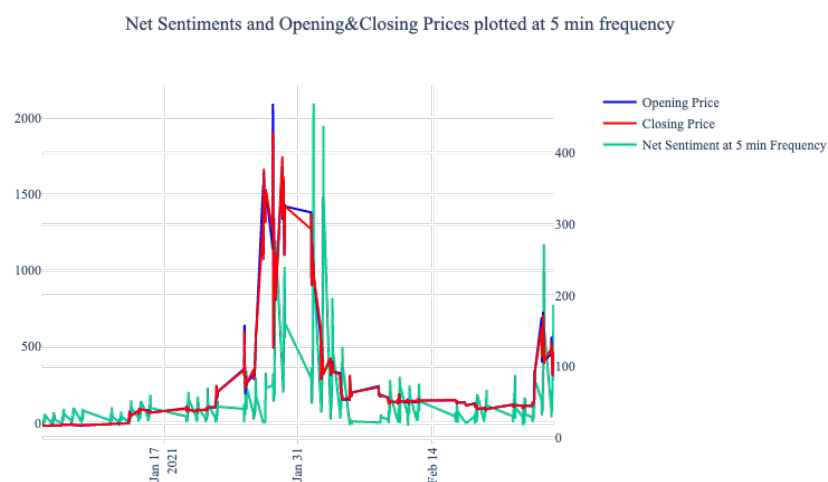


Figure 3.15: 5-minute net sentiment and GME returns

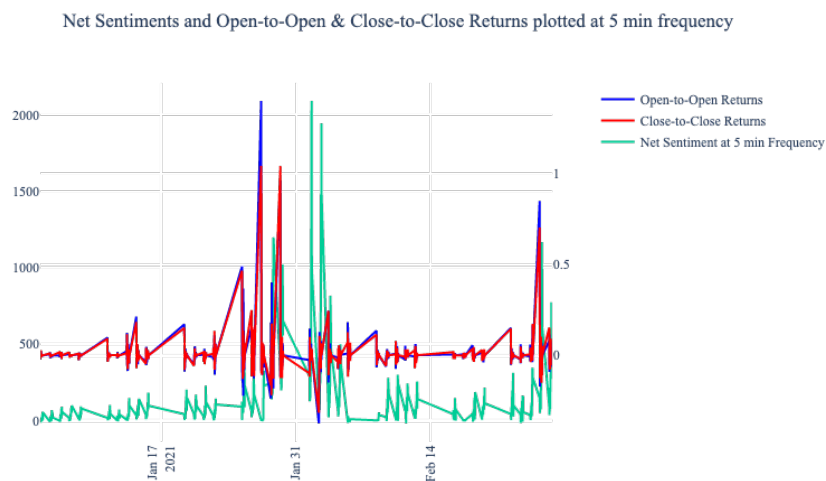
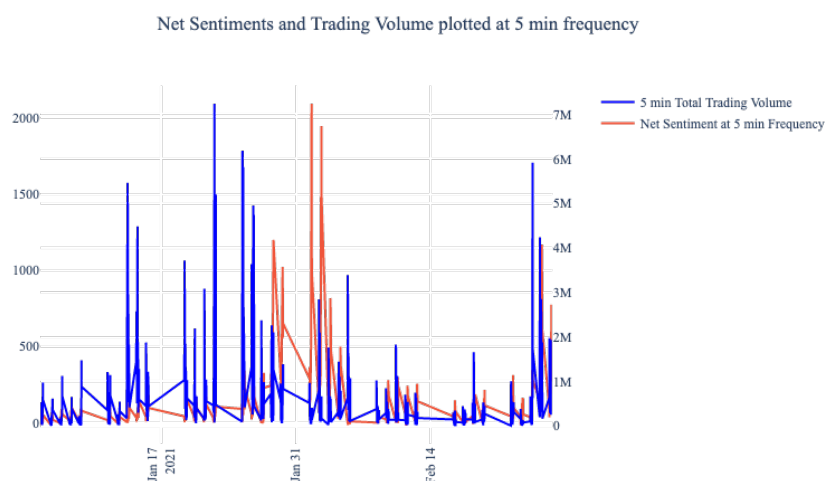


Figure 3.16: 5-minute net sentiment and GME trading volume



This pattern is consistently observed even at lower frequencies. Figures 3.17, 3.18 and 3.19 for 10-minute frequency, and figures 3.20, 3.21 and 3.22 display similar trends. At 10-min and 30-min frequencies, the intensification of relationships is observed predominantly during two brief periods of GME price surge. However, outside these two periods, when there was a growth in *NET* Sentiment (i.e., an increase in the number of positive comments over negative comments in r/WallStreetBets), the relationships are weaker.

Figure 3.17: 10-minute net sentiment and GME prices

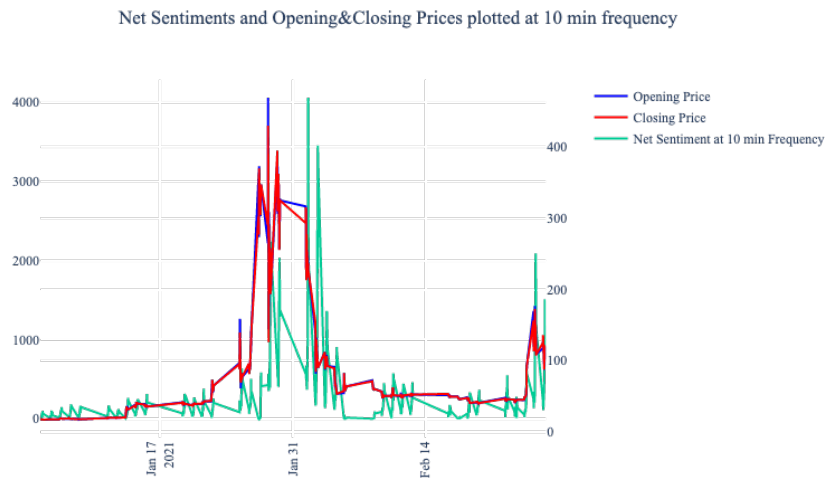


Figure 3.18: 10-minute net sentiment and GME returns

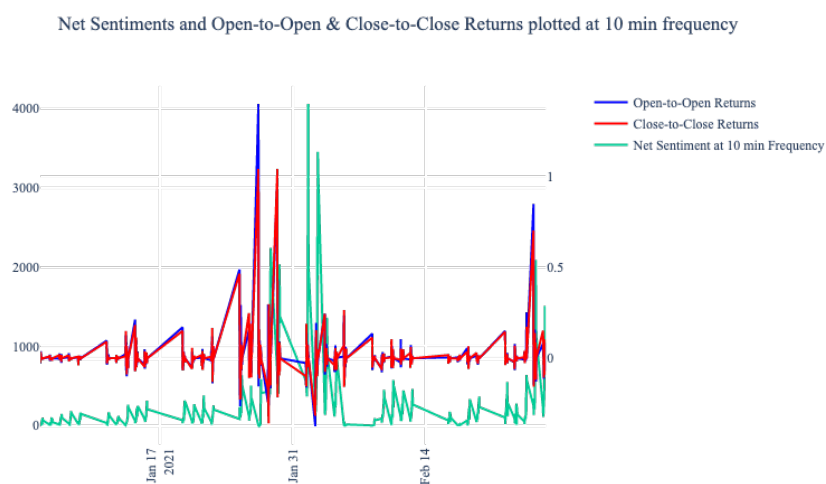


Figure 3.19: 10-minute net sentiment and GME trading volume

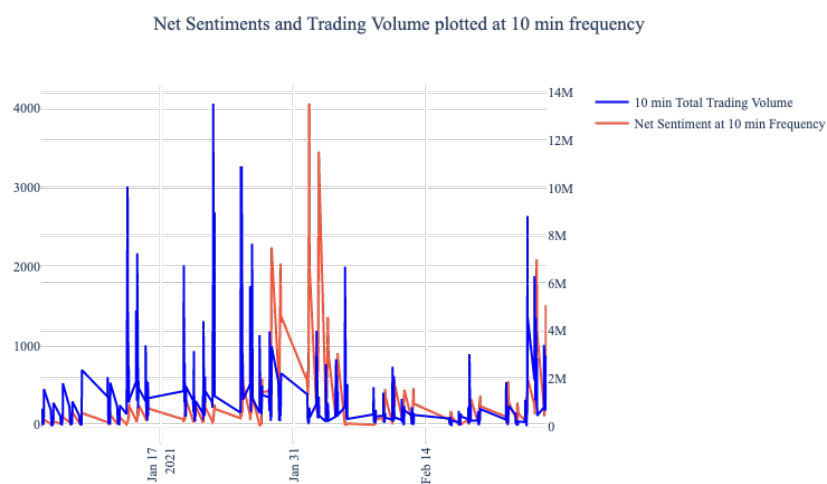


Figure 3.20: 30-minute net sentiment and GME prices

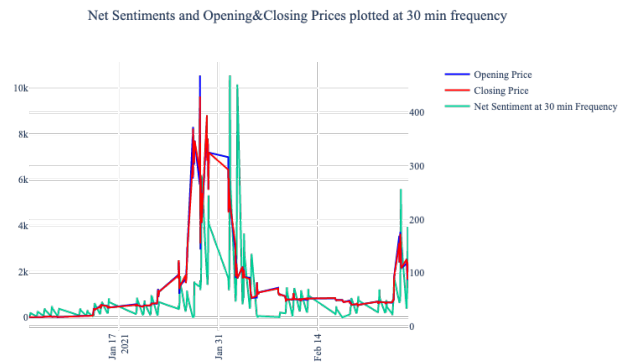


Figure 3.21: 30-minute net sentiment and GME returns

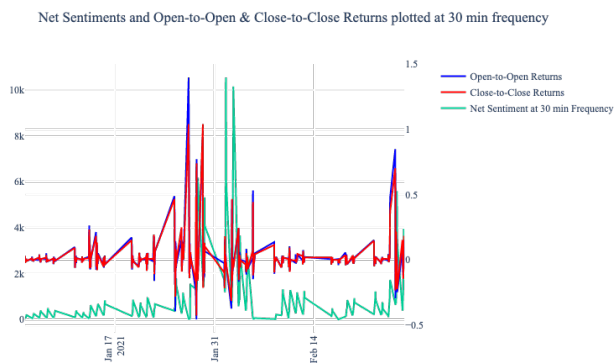
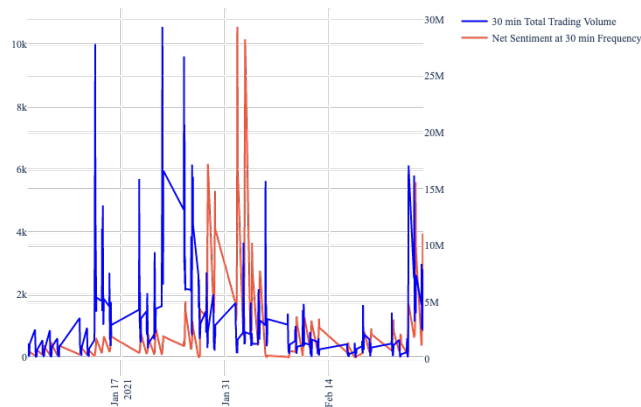


Figure 3.22: 30-minute net sentiment and GME trading volume



To further investigate the causal relationships between *NET* Sentiments and GME returns, the standard Granger causality test (Granger (1969)) is employed, and the results are outlined in Table 3.4. At a 1-min frequency, it is revealed that GME's Open-to-Open returns are influencing *NET* Sentiments and *AVERAGE* Sentiments, but the converse is not found to be true. Neither *NET* nor *AVERAGE* Sentiments are showing any noticeable effect on GME's 1-min returns.

Table 3.4: Granger causality results at 1-min, 5-min, 10-min, and 30-min frequency

1 minute				
GMEO_R does not Granger Cause NET_SENTIMENT	15227	2.39737*	0.09	
NET_SENTIMENT does not Granger Cause GMEO_R	15227	1.12244	0.33	
AVERAGE_SENTIMENT does not Granger Cause GMEO_R	15227	1.15125	0.32	
GMEO_R does not Granger Cause AVERAGE_SENTIMENT	15227	4.02723**	0.02	
5 minutes				
NET_SENTIMENT does not Granger Cause GMEO_R	3074	5.95583*	0.00	
GMEO_R does not Granger Cause NET_SENTIMENT	3074	0.43033	0.65	
AVERAGE_SENTIMENT does not Granger Cause GMEO_R	3074	0.40559	0.67	
GMEO_R does not Granger Cause AVERAGE_SENTIMENT	3074	0.95974	0.38	
10 minutes				
NET_SENTIMENT does not Granger Cause GMEO_R	1554	6.6658***	0.0013	
GMEO_R does not Granger Cause NET_SENTIMENT	1554	0.00043	1.00	
GMEO_R does not Granger Cause AVERAGE_SENTIMENT	1554	1.43995	0.24	
AVERAGE_SENTIMENT does not Granger Cause GMEO_R	1554	0.45851	0.63	
30 minutes				
NET_SENTIMENT does not Granger Cause GMEO_R	528	3.2623**	0.04	
GMEO_R does not Granger Cause NET_SENTIMENT	528	0.22412	0.8	
AVERAGE_SENTIMENT does not Granger Cause GMEO_R	528	1.8199	0.16	
GMEO_R does not Granger Cause AVERAGE_SENTIMENT	528	1.79278	0.17	

Note: *, **, and *** denote significance at the 10%, 5%, and 1% levels, respectively.

Interestingly, the influence of *NET* sentiments on GME return became substantial only at lower frequencies. Clear evidence of a positive impact is observed for 5-min, 10-min, and 30-min data. However, no causal relationships between *AVERAGE* sentiment and GME returns were identified at these frequencies.

This detailed investigation provides valuable insights into the dynamic interplay between Reddit sentiments and the infamous GameStop short squeeze, providing the potential implications of social media chatter on stock market performance.

3.3 Conclusion

The focus of this comprehensive paper is to explain the influence of sentiments derived from Reddit discussions on GameStop's intraday returns. By employing a substantial data set of approximately 10.8 million comments extracted from the r/WallStreetBets subreddit during the early months of 2021, a distinctive and customized Reddit-specific lexicon has been developed. This lexicon, fine-tuned to interpret and comprehend investment sentiments, was obtained using a human-centric approach that caters to the unique linguistic characteristics and jargon used on Reddit. It was built with the intent of aiding future research works that aim to explore and analyze investment sentiments that originate from such unique online forums.

In order to demonstrate the efficacy of this lexicon in capturing investment sentiments and translating them into tangible effects, the GameStop share rally phenomenon is delved into. This unique event in early 2021 served as the perfect case study to validate the lexicon's power. Insights are offered into how a social media platform can become a digital trading floor, facilitating dialogues that have an impact on GameStop's stock price.

In the detailed analyses, the role of *NET* Investment Sentiments on various GME open-to-open returns, specifically the 5-minute, 10-minute, and 30-minute returns, is discerned. This is seen as an observation that confirms the influence of social media sentiments on stock performance. However, it is also noted that *AVERAGE* Sentiments do not exhibit the same influence on GME performance. Interestingly, in the 1-minute data, a different causality trend is shown. GME returns are found to Granger cause both *NET* and *AVERAGE* Sentiments, suggesting a reverse relationship where stock performance influences the sentiments on the platform.

In the comprehensive examination of the dynamics of the *NET* Sentiments, in conjunction with GME prices, returns, and trading volumes, an interesting pattern is further revealed. The relationship between sentiments and the GME stock is observed to be more profound during bullish market phases compared to bearish ones. This illustrates the capability of Reddit to transmit and magnify positive sentiments during upward market trends. However, the limitations of social media influence during downturns and its inability to halt a stock's decline are also underscored. Therefore, while Reddit's role in the GameStop share rally is noticeable, this influence is neither absolute nor universal.

The broader implications of this paper are extended beyond academia and offer insights valuable to a range of stakeholders, including institutional and retail investors, policymakers, as well as media professionals. Novel empirical evidence on the impact of social media sentiments on stock prices is provided, enhancing the understanding of the 'meme stocks' phenomenon, with the GameStop saga being a prime example. Although the influence of Reddit discussions on the stock market

is confirmed, a cautionary note for individual investors is simultaneously served. The discussions and sentiments on social media platforms, despite their growing influence, may not provide sufficient protection to investments when markets take a bearish turn.

This research also reiterates the growing power of retail investors when they act collectively, as observed in forums like r/WallStreetBets. However, it also stresses that this influence is yet to be powerful enough to safeguard individual investors from losses in the volatile and unpredictable arena of 'meme' stock markets. Consequently, while the democratization of investing is a positive trend, it still needs a degree of caution and sophistication to navigate through these highly speculative investment markets.

Appendix

The Appendix outlines the methodology that is adopted to construct and validate a customized, context-specific sentiment lexicon, in line with the human-centered approach proposed by Hutto and Gilbert (2014). This sentiment lexicon is intended to gauge the sentiment valence, or intensity, of each keyword extracted from posts and comments on the r/WallStreetBets subreddit on Reddit.

In accordance with the wisdom-of-the-crowd approach described by Surowiecki (2005), the help of ten annotators is enlisted to manually determine the sentiment valence of these context-free keywords. This method of crowd-sourcing is intended to provide a reliable point of estimation for the sentiment valence score.

The methodology was implemented in two main phases, detailed as follows:

- Recruitment, Training, and Selection of Annotators:

The primary phase of the methodology was dedicated to the selection and training of the annotators. This phase was carefully executed to ensure the highest possible quality of the manually annotated data and to confirm that the scores received from the annotators were meaningful (Hutto and Gilbert (2014)). This phase encompassed the following steps:

- 1. Initial Screening: Each potential annotator underwent a pre-screening phase where they were assessed for their proficiency in the English language and comprehension of stock market terminology. To demonstrate

their understanding of finance-related terms, each candidate was required to score above 80% in an exercise.

- 2. Sentiment Rating Training: Post the screening, every annotator was provided with a training session (either online or in-person) in sentiment rating. They were instructed to assign scores ranging from [-4, +4] to each keyword. Here, a score of [-4] was attributed to Extremely Negative sentiments, [0] represented Neutral (or Neither, N/A) sentiments, and [+4] indicated Extremely Positive sentiments. The training was facilitated using a list of words and pre-validated sentiment scores extracted from the VADER dictionary.
- 3. Test Annotation: After the training session, each annotator was given a list of test lexical items that had pre-validated scores. The list included individual words, emotions, acronyms, and tweets, extracted from four distinct gold standard ground truth sentiment annotated corpora and the VADER lexicon developed by Hutto and Gilbert (2014). An annotator was deemed eligible if they achieved a score of 90% or above for matching the known revalidated sentiment score of these test items. The second and third steps of this phase were instrumental in ensuring the annotator's consistency in using the rating rubric.
- 4. Quality Check: Each batch of 25 keywords contained five "golden words" with revalidated sentiment scores and standard deviations. These "golden words" were selected from the VADER lexicon developed by Hutto and Gilbert (2014). An annotator's scores were discarded if they deviated by more than 1 standard deviation from the mean of the known distribution on three or more of the five "golden words". In such cases, the annotator was asked to re-annotate.

- Instructions for Manual Annotation:

Following the first phase, the second phase consisted of instructions for manual annotation which were implemented as follows:

- 1. Annotators had to successfully complete the screening, training, and selection process outlined in the first phase.
- 2. Sentiment ratings were obtained either through Amazon Mechanical Turk, finance academics, or students studying finance. A total of 10 annotators were appointed.
- 3. The appointed annotators were asked to rate each keyword on a scale ranging from " [-4] Extremely Negative" to "[4] Extremely Positive", with "[0] being Neutral (or Neither, N/A)".
- 4. An aggregate of the 10 independent ratings for each keyword was used to compute the mean score and the standard deviation for that word.
- 5. Keywords with a non-zero mean score and a standard deviation of 2.5 or less were retained.
- 6. The list of these retained keywords, along with their corresponding valence sentiment scores, made up the EXTRA VADER sentiment lexicon.
- 7. The VADER lexicon was updated with the EXTRA VADER sentiment lexicon. Words and their corresponding valence sentiment scores in the EXTRA VADER replaced the equivalent entries in the VADER lexicon, and new words from the EXTRA VADER lexicon were appended at the end of the VADER lexicon.

- 8. New sentiment scores were calculated using the updated VADER lexicon.

PAPER 2 - CRYPTOCURRENCY SENTIMENT LEXICON

The rapid expansion of the cryptocurrency market has catalyzed interest in the development of advanced tools for sentiment analysis. This study introduces a machine learning-based sentiment analysis model specifically designed to capture the sentiment in cryptocurrency traders' social media posts with high accuracy. By concentrating on the distinctive dynamics of cryptocurrency markets, the research aims to provide a sophisticated tool for investors and researchers, facilitating a deeper comprehension of market sentiment dynamics. Traditional sentiment analysis methodologies, including dictionary-based approaches and standard machine learning techniques, face challenges in financial contexts due to their limited capacity to grasp the complex sentiment found in social media texts related to cryptocurrencies (Loughran and McDonald (2011), Doran et al. (2012), Huang et al. (2014)).

To overcome these obstacles, the research advocates for the creation of a cryptocurrency-specific lexicon, in conjunction with advanced machine learning

models, to improve the precision of sentiment detection in the cryptocurrency market's fluctuating environment. The limitations of general sentiment dictionaries in financial applications have prompted the development of specialized lexicons that more accurately reflect the subtleties of financial discourse (Jegadeesh and Wu (2013), Chen et al. (2013), Liu and McConnell (2013)). The unique and rapidly changing terminology of the cryptocurrency sector, however, necessitates further refinement of these tools.

Neural network-based word embedding techniques, including convolutional neural networks (CNNs) and long-short-term memory (LSTM) models, mark significant advancements in sentiment analysis by overcoming the limitations of traditional bag-of-words models, which struggle with recognition word order and semantics (Bahdanau et al. (2014), Le and Mikolov (2014), and Tang et al. (2015)). Additionally, attention-based models enhance sentiment analysis by concentrating on the most informative parts of texts, thus refining the ability to discern sentiment in complex financial discussions (Yang et al. (2016)). At the core of this research is an emphasis on a feature-based sentiment analysis approach, primarily utilizing three machine learning models: Logistic Regression, Random Forest, and XGBoost. The selection of these models is driven by their robust performance in multiclass classification problems, crucial for accurately interpreting the varied sentiments found in Reddit threads, posts, and Twitter tweets. These sentiments often extend beyond simple positive, negative, or neutral categories, highlighting the importance of these models' capabilities in handling the complexity of sentiment analysis across different social media platforms.

The contribution of this paper lies in its development of a machine-learning

model that merges a cryptocurrency-specific lexicon with cutting-edge text analysis techniques. This approach seeks to fill the existing gap in sentiment analysis tools, providing a more accurate and context-aware mechanism for evaluating sentiment among cryptocurrency traders. This initiative aligns with the broader research trend towards enhancing sentiment analysis methods for financial markets, indicating a move towards the integration of machine learning and artificial intelligence in navigating the complexities of financial sentiment analysis (Pang et al. (2008), Loughran and McDonald (2013)). Through this initiative, the goal is to arm investors and analysts with a refined instrument for forecasting market movements based on social media sentiment, contributing towards more strategic decision-making and planning in the cryptocurrency domain.

4.1 Data and Methods

4.1.1 Data

In the expansive field of cryptocurrency, the central role is played by Bitcoin, which is undeniable. It is perceived as the big tree in the forest that everyone recognizes. Its significance is attributed not only to its large presence in the market but also to the sentiment it evokes among enthusiasts and skeptics alike. More often than not, the reflection of people's sentiment towards the entire cryptocurrency arena is seen in how Bitcoin is felt and talked about. That's precisely why a central place in the research was given to it.

The data in this study is gathered through the web, specifically web scraping, as the main tool. This approach allows for a deep dive into online platforms to extract

valuable information. A span from January 1, 2018, to June 30, 2021, is covered. This period, rich with events, market shifts, and evolving public opinions, is chosen as a prime time frame for the study.

Reddit is chosen as one of the primary sources, and various subreddits are delved into for diverse perspectives. Directly, r/bitcoin and r/btc are selected because they are solely focused on Bitcoin. A peek into r/eth is also taken, given Ethereum's growing prominence in the crypto conversation. But for a broader sense, r/cryptomarkets and r/Cryptocurrency are chosen too. These subreddits, like the town squares of the crypto community on Reddit, help in understanding the general feelings and attitudes of the Reddit community towards the crypto world, acting as mirrors reflecting the collective sentiment.

Twitter, on the other hand, offers a dynamic pulse of real-time reactions. Tweets that have mentions of 'bitcoin' and 'btc' are collected. This platform, with its flurry of daily discussions, provides a snapshot of how the wind is blowing on any given day in the crypto sentiment landscape.

In the collected dataset, the observation of a substantial presence of extraneous entries, including placeholder messages such as "[deleted]" and "[removed]", as well as automated responses like *"I am a bot, and this action was performed automatically. Please contact the moderators of this subreddit if you have any questions or concerns."* is made. To ensure data integrity and relevance for the analysis, the data was carefully cleaned to remove these irrelevant entries. The resulting dataset represents the refined, usable data for the study.

The cleaned data from both platforms are shown in below tables (Table 4.1 and Table 4.2):

Table 4.1: Reddit subchannels

Subchannel	Number of Posts
r/bitcoin	3,653,207
r/btc	1,496,968
r/eth	738678
r/cryptocurrency	7,998010
r/cryptomarkets	269937
Total	14,156,800

Table 4.2: Twitter Bitcoin&BTC Tweets

Glossary	Number of Tweets
Bitcoin, Btc	55,889,578

4.1.2 Methods

Sentiment analysis is a computational technique that determines and categorizes opinions found within text data. Its primary purpose is to gauge the writer’s sentiment about a particular topic, product, or service. Typically, these sentiments are categorized as positive, negative, or neutral, though the details can vary based on the specific requirements of the analysis.

Within the vast landscape of sentiment analysis, there are several methodologies. One of the earliest and most direct is the rule-based approach. In this method, specific words or phrases are assigned predetermined sentiment scores. For example, Python libraries like TextBlob¹ and VADER (Hutto and Gilbert (2014)) use

¹See, <https://textblob.readthedocs.io/en/dev/>

this approach. They have a predefined list of words with corresponding sentiment values, which, when applied to a piece of text, provide an aggregate score indicating the overall sentiment.

On the other hand, feature-based methods are more complex. These methods extract certain attributes or features from text data. These features may range from simple word frequencies to more complex attributes such as lexical tags or n-grams. Once the features have been extracted and defined, various machine-learning techniques are used to train the data model.

Among these machine-learning techniques, Logistic Regression is recognized as a popular choice. It is a statistical method that measures the probability of a data point falling into a particular category. In the field of sentiment analysis, the likelihood of a text message being positive, negative, or neutral is comprehended.

Simultaneously, ensemble models, which consist of multiple algorithms or even the repeated application of a single algorithm, present a robust solution for sentiment analysis. Random Forest operates by constructing numerous decision trees during training. An output is provided based on the mode of the classes of these individual trees. On the other hand, XGBoost, another ensemble method, works slightly differently. It is a gradient-boosting algorithm that sequentially builds decision trees, with each new tree aiming to correct the errors made by its predecessor.

In the heart of the research lies the focus on a feature-based approach to sentiment analysis, predominantly centered on three machine learning models: Logistic Regression, Random Forest, and XGBoost. The significance of choosing

these models stems from their exceptional ability to tackle multiclass classification, a key aspect considering the diverse sentiments expressed across platforms like Reddit threads, posts, and Twitter tweets. Such sentiments often range beyond the simplistic categories of positive, negative, or neutral, making the capability of these models indispensable.

Beyond the range of sentiments, the rationale behind model selection is also driven by a combination of computational efficiency and predictive accuracy. These models have consistently demonstrated a balanced performance, offering trustworthy results without imposing undue computational demands. The ambition, as delved deeper into the research, is to utilize the strengths of these models, aiming to decode and understand the prevailing sentiment trends from the massive social media messages.

An essential step before delving into the analysis is the transformation of textual data from these platforms into a numerical format suitable for machine learning models. For this crucial step, two methods are primarily employed: the Bag-of-Words (BoW) and the Term Frequency-Inverse Document Frequency (TF-IDF) techniques. These methodologies aid in converting unstructured text into structured numerical data, laying the foundation for the sentiment analysis algorithms to operate effectively.

The text representations and machine learning models are explained in detail in the following subsections.

4.1.2.1 Bag-of-Words and TF-IDF Representations

Bag-of-Words (BoW)

The Bag-of-Words model is a simplistic method of text representation that involves treating a document as a 'bag' or 'multiset' of words. It disregards the order of the words but maintains their frequency. In the BoW representation, each document is represented as a vector in a multidimensional space, where each dimension corresponds to a unique word from the entire corpus. The frequency of a word in a document corresponds to the value in the vector at the dimension corresponding to that word.

Mathematically, the Bag-of-Words (BoW) model represents a document d_i as a vector $v_i \in \mathbb{R}^M$ in a multidimensional space, where each dimension corresponds to a unique word in the corpus. Given a corpus D with N documents $\{d_1, d_2, \dots, d_N\}$ and a dictionary W containing M unique words $\{w_1, w_2, \dots, w_M\}$, the frequency of word w_j in document d_i is denoted as v_{ij} . Thus, the BoW vector for document d_i can be expressed as:

$$(4.1) \quad v_i = [v_{i1}, v_{i2}, \dots, v_{iM}]$$

where v_{ij} is the frequency of the word w_j in the document d_i , for $j = 1, 2, \dots, M$.

Term Frequency-Inverse Document Frequency (TF-IDF)

The Term Frequency-Inverse Document Frequency (TF-IDF) method is a more advanced model for text representation. Unlike the BoW model, which only considers the frequency of a word in a document, the TF-IDF model also takes into

account the importance of the word in the entire corpus.

The TF-IDF representation of a word is the product of two statistics: Term Frequency (TF) and Inverse Document Frequency (IDF). TF is the frequency of a word in a document, and IDF is the log inverse of the fraction of documents in the corpus that contain the word.

Mathematically, given a document d_i with words $w_1, w_2, \dots, w_{n_i}$ and a corpus D with N documents, the TF-IDF representation of a word w_j in document d_i is given by:

$$(4.2) \quad \text{TF-IDF}(w_j, d_i, D) = \text{TF}(w_j, d_i) \times \text{IDF}(w_j, D)$$

where

$$(4.3) \quad \text{TF}(w_j, d_i) = \frac{\text{Frequency of } w_j \text{ in } d_i}{\text{Total number of words in } d_i}$$

and

$$(4.4) \quad \text{IDF}(w_j, D) = \log \left(\frac{N}{\text{Number of documents in } D \text{ with } w_j} \right)$$

4.1.2.2 Logistic Regression for Multiclass Classification

Logistic Regression is a statistical model primarily used for binary classification problems. However, for multiclass classification problems such as the one where sentiment is classified into three categories (positive, negative, and neutral), the algorithm is generalized to the multiclass case using techniques such as one-vs-rest

(OvR) or softmax regression.

In the research, logistic regression is applied to the task of sentiment analysis on Reddit threads and posts and Twitter tweets. The textual data is first preprocessed and transformed into a numerical representation using either Bag-of-Words (BoW) or Term Frequency-Inverse Document Frequency (TF-IDF) approaches. Subsequently, the logistic regression model is trained on these numerical vectors using multiple training datasets.

Softmax regression (or multinomial logistic regression) extends logistic regression to multiclass problems. It computes the probability that a given input sample belongs to each class, and predicts the class with the highest estimated probability.

Given an input vector $x \in \mathbb{R}^n$, the softmax regression model first computes a score $s_k(x)$ for each class k , then estimates the probability of each class by applying the softmax function (also called the normalized exponential) to the scores.

The score for each class is computed as:

$$(4.5) \quad s_k(x) = x^T w_k + b_k$$

where w_k is the weight vector associated with class k , b_k is the bias term for class k , x^T is the transpose of the input vector x , and k ranges over the number of classes.

Once the scores are computed, the probability $\hat{p}_k$ that the input sample belongs to class k is estimated using the softmax function:

$$(4.6) \quad \hat{p}_k = \sigma(s(x))_k = \frac{\exp(s_k(x))}{\sum_{j=1}^K \exp(s_j(x))}$$

where $\sigma(s(x))_k$ is the estimated probability that the input sample x belongs to class k given the scores of each class for that sample, K is the number of classes, and $\exp$ is the exponential function.

The prediction $\hat{y}$ of the softmax regression model for an input sample x is then the class with the highest estimated probability:

$$(4.7) \quad \hat{y} = \arg\max_k \sigma(s(x))_k = \arg\max_k s_k(x) = \arg\max_k (x^T w_k + b_k)$$

where $\arg\max_k$ is the function that returns the value of k that maximizes the expression to its right.

In conclusion, logistic regression, and in particular its extension to the multiclass case via softmax regression, provides a simple and efficient method for sentiment analysis. Despite its simplicity, when used with proper preprocessing methods, it can achieve good performance on tasks like this study.

4.1.2.3 Random Forest for Sentiment Analysis

Random Forest is an ensemble learning method that is characterized by constructing a multitude of decision trees during training. It operates by outputting the class that is the mode of the classes for classification, or the mean prediction of the individual trees for regression. The habit of decision trees to overfit to their training set is corrected by Random Forest.

In the context of research, Random Forest is applied for sentiment analysis on Reddit threads, posts, and Twitter tweets. The textual data is first preprocessed and transformed into a numerical form using either Bag-of-Words (BoW) or Term Frequency-Inverse Document Frequency (TF-IDF) approaches. Following this, the Random Forest model is trained on these numerical vectors using multiple training datasets.

The algorithm for Random Forest can be summarized as follows:

- Draw n bootstrap samples from the original data.
- For each of the bootstrap samples, grow an unpruned classification (or regression) tree, with the following modification: at each node, rather than choosing the best split among all predictors, randomly sample m of the predictors and choose the best split from among those variables. (Common choices of m are $m = \sqrt{p}$ for classification and $m = p/3$ for regression, where p is the number of predictors.)
- Predict new data by aggregating the predictions of the n trees (i.e., majority votes for classification, average for regression).

The training of each decision tree can be formalized as follows:

Given a training set $(x_i, y_i)_{i=1}^N$, where x_i is the i -th instance and y_i is the corresponding label, a decision tree recursively splits the space such that at each node, it chooses the split that maximizes the reduction in impurity. The impurity at node m is calculated as:

$$(4.8) \quad G_m = \sum_{k=1}^K \hat{p}_{mk}(1 - \hat{p}_{mk})$$

where K is the number of classes, and $\hat{p}_{mk}$ is the estimated probability that an instance in node m belongs to class k .

The prediction of the Random Forest model for an input sample x is the majority vote of the predictions of the n trees:

$$(4.9) \quad \hat{y} = \arg\max_k \frac{1}{n} \sum_{i=1}^n I(y_i = k)$$

where $I(\cdot)$ is the indicator function, which is equal to 1 if $y_i = k$ (the prediction of the i -th tree is k), and 0 otherwise.

In conclusion, Random Forest, by leveraging the power of ensemble learning, provides an effective method for sentiment analysis. Its ability to avoid overfitting while maintaining high predictive power makes it suitable for multi-class classification tasks.

4.1.2.4 Extreme Gradient Boosting (XGBoost) for Sentiment Analysis

Extreme Gradient Boosting, or XGBoost, is a sophisticated, high-performing machine learning algorithm based on gradient boosting frameworks. It is designed to be highly efficient, flexible, and portable. In the context of the study, XGBoost is employed for the sentiment analysis of Reddit threads, posts, and Twitter tweets. The textual data are transformed into numerical vectors using either Bag-of-Words (BoW) or Term Frequency-Inverse Document Frequency (TF-IDF) approaches.

Then, XGBoost is applied on these vectors with multi-source training datasets for the 3-class classification (positive, negative, and neutral sentiments).

The XGBoost algorithm iteratively trains weak learners (in this case, decision trees) to complement the shortcomings of the preceding models. At each iteration, a new function is learned by the model that predicts the residuals (i.e., differences) between the current predictions and the true values. This is achieved by minimizing a regularized objective function.

$$(4.10) \quad \mathcal{L}(\phi) = \sum_{i=1}^n l(y_i, \hat{y}_i) + \sum_{k=1}^K \Omega(f_k)$$

where ϕ denotes the set of all decision tree functions, $l(y_i, \hat{y}_i)$ is the loss function that measures the difference between the true label y_i and the predicted label $\hat{y}_i$ for the i -th instance, $\Omega(f_k)$ is the regularization term that penalizes the complexity of the k -th decision tree, and K is the number of trees.

Typically, the loss function used for classification tasks in XGBoost is the log-loss function, defined as:

$$(4.11) \quad l(y_i, \hat{y}_i) = -y_i \log(\hat{y}_i) - (1 - y_i) \log(1 - \hat{y}_i)$$

and the regularization term is defined as:

$$(4.12) \quad \Omega(f_k) = \gamma T + \frac{1}{2} \lambda |w|^2$$

where T is the number of leaves in the tree, w represents the leaf weights, γ and λ are regularization hyperparameters.

In conclusion, XGBoost is recognized as a powerful algorithm for classification tasks, including sentiment analysis. The reduction of bias and variance through gradient boosting, coupled with an integrated regularization term to prevent overfitting, makes it an ideal choice for the task.

4.1.2.5 Training and Testing Dataset Construction

To facilitate an effective evaluation of the selected machine learning models, it's imperative to build robust training and test datasets tailored for sentiment analysis. In this research, sentiments have been categorized into three distinct labels: positive, neutral, and negative. The process of dataset creation encompassed a series of methodological steps:

a. Noisy Labeling using Emojis: To begin with, a massive volume of 100,000 social media messages was processed using the 'noisy labeling' method. Noisy labeling, in this context, refers to the practice of determining the sentiment of a message based on the emojis it contains. Emojis, due to their expressive nature, can often provide a heuristic to gauge the sentiment of a text. For instance, messages containing the "😊" emoji were earmarked as 'positive', whereas those with the "😞" emoji were designated as 'negative'. In cases where a single message incorporated both positive and negative emojis, the sentiment was determined based on the majority presence. That is, the sentiment label would align with the emoji that appeared more frequently within that message.

b. Manual Annotation by Human Annotators: Subsequently, a subset of 10,000 social media messages was carefully annotated by a team of eight human annotators. These annotators, specifically hired for this task, underwent a comprehensive training schedule to ensure consistency and accuracy in their annotations. Their primary role was to read and categorize each message into one of the three sentiment labels.

c. Annotation by ChatGPT: Parallely, the identical set of 10,000 messages, previously annotated manually, was also processed by ChatGPT for sentiment labeling. This step aimed to juxtapose human intuition with the capabilities of an advanced language model in sentiment classification.

d. Segregation of Human-Annotated Data: From the corpus of 10,000 messages that were manually annotated, a subset of 3,000 was curated and reserved as a test set. This test set would serve as a benchmark to evaluate the performance of machine learning models. The remaining 7,000 messages from this corpus are designated as the training set, forming the foundation upon which the models would be trained.

e. Incorporation of ChatGPT-Annotated Data: From the 10,000 messages labeled by ChatGPT, the same 3,000 messages earmarked in the previous step were harnessed to augment the training set. This strategy aimed to enrich the training dataset by introducing varied labeling perspectives. The residual 7,000 messages, processed by ChatGPT, are integrated into the primary training set.

This study combines human judgments and algorithmic accuracy to build a

detailed dataset, laying the groundwork for thorough machine learning analysis.

4.1.2.6 Machine Learning Model Development

The objective of this phase is to transform raw message data into actionable insights using a structured machine-learning pipeline. This involved a series of detailed steps designed to enhance the quality of the data and subsequently predict sentiments.

a. Pre-processing of Messages: The initial step was the refinement of the raw messages. This entailed several processes designed to enhance the clarity and relevance of the data. Symbols, numbers, and common words (often referred to as 'stop words') that lack significant semantic value were systematically eliminated. Subsequent to this cleansing process, the messages are tokenization. Tokenization, in this context, means breaking down the text into smaller units, such as individual words or combinations of words. These units were categorized into unigrams (single words), bigrams (two-word combinations), and trigrams (three-word combinations).

b. Text Representation and Feature Extraction: Once the messages were pre-processed, the next task was to represent the text in a manner conducive to machine learning. Two predominant techniques were employed for this: the 'Bag of Words' (BoW) model and the 'Term Frequency-Inverse Document Frequency' (TF-IDF). While BoW represents text data based on the frequency of words without considering their order, TF-IDF assigns importance to words based on how frequently they appear in a document relative to their occurrence across other documents. Both techniques transform textual data into numerical vectors, making them suitable for machine learning algorithms.

c. Deployment of Machine Learning Models: With the vectorized text data at hand, the next step is to feed this data into three distinct machine-learning models: Logistic Regression, Random Forest, and XGBoost. Each model was tasked with predicting the sentiments of the text messages. To optimize their performance, hyperparameter tuning was conducted using 5-fold cross-validation on the training dataset. This means the training data was divided into five subsets; each subset was used as a validation set while the others formed the training set. This process was repeated five times, ensuring comprehensive model evaluation. All three models were equipped to process both BoW and TF-IDF representations of unigrams, bigrams, and trigrams. From a vast array of over 160,000 potential units (unigrams, bigrams, and trigrams), a cap was set, and only the 30,000 most prevalent features (vectorized social messages) were chosen for model training.

d. Weighted Importance of Datasets: An innovative step was taken to assign differential importance to datasets. The datasets labeled by both ChatGPT and human annotators have explained a weight ten times greater than their original value. This strategic amplification was implemented to underscore their importance, especially when the size of the datasets labeled through noisy labeling methods is much greater than the human label. This approach was based on the premise that manually annotated and ChatGPT-labeled datasets exhibit higher reliability than those labeled using heuristic methods like noisy labeling.

4.1.2.7 Further Adjustment

The result of the model is the average of the results of 10 random draws, strictly ensuring the mutual exclusivity between the data sets drawn each time. This

approach was implemented to enhance the robustness of the findings, minimize potential bias, and ensure the generalizability of the results. After selecting the optimal model, an additional post-processing step is added to refine the assignment of emotion labels. This step involves calculating ratios to identify the dominance of a particular emotion. Specifically, if the ratio of the probability of being positive to the sum of the probabilities of being neutral and negative exceeds 1, the message is explicitly categorized as "positive". Conversely, if the probability of negativity divided by the probability of the combination of positivity and neutrality exceeds 1, the message is explicitly categorized as "negative". This approach ensures that sentiment labels are assigned only when there is a clear dominance of evidence in favor of the sentiment, thus improving the accuracy of the sentiment classification process.

$$(4.13) \quad \text{Categorize as "positive"} \iff \frac{P(\text{positive})}{P(\text{neutral}) + P(\text{negative})} > 1.$$

$$(4.14) \quad \text{Categorize as "negative"} \iff \frac{P(\text{negative})}{P(\text{positive}) + P(\text{neutral})} > 1.$$

In this approach, after the best performing model is selected, SHAP (SHapley Additive exPlanations) analysis is implemented for complex feature extraction. SHAP, originating from a game theory background, provides a sophisticated way to interpret the output of any machine learning model (Lundberg and Lee (2017)). This is done by assigning a specific predicted importance value to each feature, based on the concept of the Shapley value from cooperative game theory.

Using SHAP analysis, the state-of-the-art approach drills down into the model's output, providing detailed insights into how each feature contributes to the overall

decision-making process. Characteristics are carefully divided into three different emotion groups: positive, negative, and neutral. This classification helps reveal the complex interplay between various types of emotions and their impact on cryptocurrency market communication. It allows the subtle ways in which different emotions influence market dynamics and investor behavior to be understood.

To further refine the analysis, manual adjustments are performed. This critical step improves the accuracy of the approach and allows the sentiment dictionary to be tailored to work seamlessly with the complexity and unique characteristics of cryptocurrency market communications. By customizing this lexicon, market sentiment is ensured to be more accurately expressed and understood, capturing the volatile and often unpredictable nature of the cryptocurrency market.

4.2 Results and Discussion

4.2.1 Evaluation of Machine Learning Models on Initial Datasets

The first set of experiments begins with a training dataset comprising 50,000 entries marked with a noisy labeling process supplemented by 7,000 entries from a blend of human annotations. The model is evaluated by its F1 score. The F1 score is a measure used to evaluate the accuracy of a classification model, which combines precision and recall into a single metric. High F1 scores are generally desirable, indicating that the model has good precision and recall balance. As shown in table 4.3, for the Bag-of-Words (BoW) text representation, the models exhibit the following F1 scores: Random Forest at 0.249, XGBoost at 0.408, and Logistic

Table 4.3: Initial Datasets Model Performance

		50k noisy labelled + 7k Human Labelled+0K ChatGPT	50k noisy labelled + 3k Human Labelled + 4k ChatGPT Labelled	50k noisy labelled + 0k Human Labelled + 7k ChatGPT Labelled
BoW text representative	Random Forest	0.248573334	0.275745163	0.370369709
	XGBoost	0.408288196	0.469971777	0.522243358
	Logistic Regression	0.478677283	0.501658579	0.526203851
TF-IDF text representative	Random Forest	0.274349678	0.29182069	0.368280668
	XGBoost	0.535755378	0.535780409	0.537719266
	Logistic Regression	0.605989225	0.584845726	0.574363693

Regression at 0.479. Switching to the TF-IDF representation, the F1 scores improve across the board, with Random Forest achieving 0.274, XGBoost reaching 0.536, and Logistic Regression leading the pack with 0.606.

Introducing a dataset with 3,000 human-labeled and 4,000 ChatGPT-labeled entries into the noisy 50k set, the BoW representation sees increases in F1 scores to 0.276 for Random Forest, 0.470 for XGBoost, and 0.502 for Logistic Regression. The TF-IDF representation scores for this mix are 0.292 for Random Forest, 0.536 for XGBoost, and 0.585 for Logistic Regression.

Replacing the human labels entirely with 7,000 ChatGPT-labeled entries in the noisy 50k set, the BoW representation delivers a Random Forest F1 score of 0.370, XGBoost at 0.522, and Logistic Regression at 0.526. For the TF-IDF representation, the results show a Random Forest F1 score of 0.368, XGBoost at 0.538, and Logistic Regression at 0.574.

4.2.2 Performance Analysis with Enriched Datasets

Building on the first experiment, each dataset configuration was enriched by incorporating additional labeled data. In the second set of experiments, the training datasets were augmented by incorporating an additional dataset consisting of 3,000

unseen messages and their corresponding sentiment labels predicted by ChatGPT. This enrichment aimed to explore if the inclusion of ChatGPT-labeled data could enhance the predictive performance of the machine learning models.

In the enriched BoW representation, as shown in Table 4.4, the F1 score of the Random Forest model improved from 0.249 (initial dataset) to 0.276 with the partial replacement of human-labeled datasets by ChatGPT-labeled datasets, and further to 0.370 with the complete replacement of human-labeled datasets by ChatGPT-labeled datasets. XGBoost's F1 score increased from 0.408 (initial dataset) to 0.470 with the partial replacement, and then to 0.522 with the complete replacement. The Logistic Regression model's F1 scores progressed from 0.479 (initial dataset) to 0.502 with the partial replacement, and then to 0.526 with the complete replacement.

In the enriched TF-IDF representation, the F1 score of the Random Forest model slightly increased from 0.274 (initial dataset) to 0.292 with the partial replacement, and then further increased to 0.368 with the complete replacement. XGBoost's scores remained relatively consistent, ranging from 0.536 (initial dataset) to 0.538 with the complete replacement. However, the Logistic Regression model's performance decreased slightly from 0.606 (initial dataset) to 0.585 with the partial replacement, and then further decreased to 0.574 with the complete replacement.

Table 4.4: Enriched Datasets Model Performance

		50k noisy labelled + 7k Human Labelled + 0k ChatGPT Labelled + ChatGPT predicted X_rtest	50k noisy labelled + 3k Human Labelled + 4k ChatGPT Labelled + ChatGPT predicted X_rtest	50k noisy labelled + 0k Human Labelled + 7k ChatGPT Labelled + ChatGPT predicted X_rtest
BoW text representative	Random Forest	0.245004708	0.297104653	0.471910568
	XGBoost	0.527693503	0.556743656	0.600037199
	Logistic Regression	0.661206728	0.658178137	0.665137706
TF-IDF text representative	Random Forest	0.231520678	0.286180517	0.452761071
	XGBoost	0.580730669	0.57889932	0.582477354
	Logistic Regression	0.670106578	0.67400368	0.675956276

The optimal model performance is observed with the Logistic Regression model using the TF-IDF text representation. By leveraging a training set that excludes manually labeled data and utilizes 7,000 ChatGPT-labeled entries alongside 3,000 ChatGPT-predicted unseen labels, Logistic Regression yields the highest F1 score of 0.676. This superior result demonstrates the strength of Logistic Regression in sentiment analysis, particularly when the model is trained on a dataset rich with ChatGPT’s algorithmic annotations and predictions.

4.2.3 A More Accurate Model in the Market

In the provided text, a comparative analysis of sentiment models is delved into, juxtaposing them with two well-established sentiment analysis tools: Vader (Hutto and Gilbert (2014)) and TextBlob². For this evaluation, a dataset of 2,000 messages from the data is selected. Not only are these messages randomly selected to ensure an unbiased assessment, but they are also unseen by any machine learning model before, putting its predictive power to a strict test.

The messages are manually annotated before being analyzed through the model. This step is crucial to establish concrete benchmarks, allowing the accuracy of the

²<https://textblob.readthedocs.io/en/dev/>

model to be evaluated in a controlled environment. The results of this analysis are enlightening and show that significant progress in sentiment classification is made by the model. An accuracy of 88.7% is achieved by the model, far exceeding Vader (42.85%) and TextBlob (43.15%).

This significant difference in accuracy not only demonstrates the effectiveness of the model in sentiment classification but is also a key indicator of its potential to set a new benchmark in the field. The model not only demonstrates its superiority over existing tools but also lays the foundation for future innovations in sentiment analysis, especially in the field of financial market communications.

4.2.4 A New Crypto-Sentiment Lexicon

The development and utility of a specialized sentiment analysis dictionary tailored to the cryptocurrency market are delved into. Inspired by Loughran and McDonald (2011)'s approach, the lexicon is a comprehensive amalgamation of more than 10,000 terms, carefully selected to cover a wide range of cryptocurrency-specific terms. A wide range of terms is included, from basic cryptocurrency terms to more complex lexical and grammatical variants. This includes not only nouns and verbs but also adjectives and adverbs, as well as their different tenses and forms, providing an exhaustive understanding of the language used in cryptocurrency-related texts.

The dictionary has been developed to capture the subtle emotional details that are usually embedded in the language dedicated to cryptocurrency communications. Its inclusivity is a key factor in ensuring a detailed and accurate understanding of

the emotions conveyed in these texts. From capturing subtle differences in sentiment expressed through various adjective forms to understanding the impact of tense changes on emotional tone, the lexicon provides a solid foundation for deep sentiment analysis.

The validity of the lexicon is not merely theoretical. To empirically demonstrate its usefulness, an example of the lexicon is included in the appendix of the paper (see Table 4.5). The example provides a glimpse into the depth and breadth of the lexicon, illustrating how it can be utilized to improve the accuracy and reliability of sentiment analysis in the cryptocurrency space.

4.3 Conclusion

In summary, this study takes an in-depth look into sentiment analysis within the cryptocurrency market, with a specific focus on Bitcoin due to its significant standing. Much like the tallest tree in a forest stands out, Bitcoin's influence is prominent in the cryptocurrency world. It's observed that the general sentiment towards Bitcoin frequently mirrors the wider sentiment toward the entire cryptocurrency market.

This method uses web scrabbing technology to accumulate a rich data set from January 1, 2018, to June 30, 2021. This period was marked by major events and changing market dynamics, which provides the basis for the study. Different viewpoints from various parts of Reddit, from Bitcoin-specific forums to broader cryptocurrency discussions, are used to capture the collective sentiment of the community. Meanwhile, real-time comments on Twitter add a dynamic dimension to

the data, providing insights into the daily up and down of cryptocurrency sentiment.

One of the most important findings is the development of a specialized sentiment analysis lexicon that was carefully designed to understand the unique language of the cryptocurrency market. Inspired by Loughran and McDonald (2011)'s methodology, the lexicon contains more than 10,000 terms essential for accurately capturing the nuances of sentiment in cryptocurrency-related texts. Its comprehensiveness and consideration of various grammatical forms help ensure a nuanced interpretation of sentiment, thus greatly increasing the depth and reliability of the sentiment analysis.

The model develops a detailed evaluation process, compared against established sentiment analysis tools like Vader and TextBlob, utilizing a dataset of 2,000 carefully selected messages. This evaluation served as a benchmark and highlighted the model's enhanced ability to accurately classify sentiment, achieving an impressive 88.7% accuracy rate, a notable improvement over existing models and a potential new standard in cryptocurrency sentiment analysis.

This study highlights the model's effectiveness, offering a more capable alternative to current sentiment analysis tools, especially in the context of financial market communication. It also introduces a customized cryptocurrency sentiment dictionary, reflecting a dedication to ongoing innovation and refinement to meet the unique demands of cryptocurrency market analysis.

To wrap up, the research represents a blend of cutting-edge methods, technology, and sector-specific knowledge, resulting in a model that sets a new precedent for

sentiment analysis within the cryptocurrency market. This foundation not only helps develop further scholarly inquiry but also holds significant implications for a broad scope of market participants, deepening the collective comprehension of the complicated world of cryptocurrency sentiment.

Appendix

Table 4.5: Sample Cryptocurrency Sentiment Lexicon

Positive	Neutral	Negative
Ability: able abilities ability Absolute: absolute absolutely absoluteness Accelerate: accelerate accelerates accelerated accelerating acceleration Accept: accept accepts accepted accepting acceptance Accept Crypto: accepts crypto accepting crypto accepted crypto accept crypto Access: access accesses accessed accessing accessibility accessible Accomplish: accomplish accomplishes accomplished accomplishing accomplishment accomplishments	Abroad: abroad abroads Academy: academy academies academic Accord: accord accords accorded according accordance accordances Account: account accounts accounted accounting accountable accountability accountant accountancy accountants Act: act acts acted acting actor actors action action actions Activity: activity activities Actually: actually	Afraid: Afraid Angry: Angry Angrier Angriest Angrily Annoying: Annoy Annoys Annoyed Annoying Annoyingly Annoyer Anxiety: Anxious Anxieties Anxiously Anxiousness Argue: Argue Argues Argued Arguing Argument Arguments Arrest: Arrest Arrests Arrested Arresting Arrestable Arrestor Attack: Attack Attacks Attacked Attacking Attacker

PAPER 3: SENTIMENT ANALYSIS AND CRYPTOCURRENCY MARKET DYNAMICS

The introduction of cryptocurrencies, indicated by Satoshi Nakamoto's 2008 paper, has fundamentally altered the landscape of financial transactions through the advent of blockchain technology (Nakamoto (2008)). These digital currencies differ markedly from traditional financial markets, operating continuously without the typical 'circuit breaker mechanisms' found in traditional stock markets (Lee et al. (1994), Corwin and Lipson (2000)). Research in the cryptocurrency domain has broadly focused on risk volatility spillover, the interconnectedness within the cryptocurrency market, and the impact of retail investor sentiment on market volatility. Studies have highlighted the unique aspects of cryptocurrency markets, such as their potential to act as a hedge against traditional market fluctuations (Bouri et al. (2017) and their distinct asset class characteristics, with volatility and return dynamics that diverge significantly from other financial assets (Pele et al. (2021), Cheah et al. (2022)).

The dynamics within the cryptocurrency market, particularly concerning the volatility spillover among various cryptocurrencies, have been extensively explored. Researchers have employed a range of methodologies to uncover the complex relationships between market volatility, returns, and the broader financial ecosystem, with findings indicating significant exposure to tail-risk within the crypto market itself (Liu (2019), Borri (2019)). Moreover, the influence of retail investor sentiment, especially as mediated through social media platforms, has emerged as a crucial factor in shaping market volatility. Studies have demonstrated the predictive power of online sentiment, such as Twitter sentiment, on cryptocurrency returns, underscoring the evolving nature of market analysis in the age of digital communication (Chen et al. (2019), Gurdgiev et al. (2019), Kraaijeveld and De Smedt (2020)).

This research seeks to bridge the existing gap by utilizing high-frequency data to examine the interconnectedness and volatility spillover within the cryptocurrency market, alongside assessing the impact of retail investor sentiment on these dynamics. The findings suggest that the connectedness within the cryptocurrency network is inherently time-varying, significantly influenced by exogenous shocks, and demonstrates a notable shift in dynamics over time. Such insights underscore the importance of incorporating high-frequency data and retail investor sentiment in understanding the complexities of the cryptocurrency market, providing valuable perspectives for investors, policymakers, and regulators in navigating this evolving financial landscape.

5.1 Data and Methods

5.1.1 Data

The sentiment data utilized in this study is derived from the comprehensive analysis conducted in the previous study. This dataset encompasses a range of sentiments extracted from diverse textual sources, offering insights into public perception and mood variations over time. The dataset is synchronized with the time frame of the cryptocurrency market data to maintain coherence in temporal analysis.

Data on 'whale' activities within the cryptocurrency market is included as a key element in the analysis. Sourced from Whale Alert (whale-alert.io), this dataset provides an in-depth view of large-scale transactions and movements executed by major market participants. Whale Alert's platform monitors millions of blockchain transactions in real time, combining these with off-chain data from numerous sources. The database, arguably the most comprehensive of its kind, encompasses billions of transactions and hundreds of millions of addresses, covering the entire span of the study period. This data is instrumental in examining the potential impact of whale activities on market volatility.

This study introduces a new element by incorporating cryptocurrency price data at 30-minute intervals. The focus is on Bitcoin (BTC), Ethereum (ETH), Ripple (XRP), and Dogecoin (DOGE) based on their roles played in the cryptocurrency market from 01/01/2018 to 30/06/2021. Thirty minutes of trading data is obtained from firsttradedata.com, the leading provider of the high-resolution intraday stock market, cryptocurrency, futures, and FX rate data. Bitcoin (BTC), Ethereum (ETH), and Ripple (XRP) have huge market capitalizations and represent a wide range

of technologies, adoption levels, and investor interest. Despite its smaller market capitalization, Dogcoin (DOGE) is chosen as a representative of the meme coins, which have gained significant cultural significance and influence in recent years. This selection provides a comprehensive view of the cryptocurrency market, including both major, established cryptocurrencies and novel, culturally influential cryptocurrencies.

For analytical purposes, the frequency of price data is determined to be 30 minutes. This timeframe is carefully chosen to be consistent in studying the interaction between market sentiment and cryptocurrency market volatility. The 30-minute interval is short enough to capture rapid changes in market sentiment, which is critical in the highly volatile and sentiment-driven cryptocurrency market. Compared to shorter time intervals such as 1, 5, or 10 minutes, 30-minute data smoothes out minute-by-minute price fluctuations, providing a clearer picture of market trends and sentiment changes. It also balances the need for a detailed view of intraday price movements with the practicality of managing a comprehensive data set for extensive statistical analysis. Suitable for both short-term traders and long-term investors, the timeframe provides insight into intraday trends and reflects broader market movements of interest to a wide range of market participants.

To examine the asymmetric volatility spillovers between cryptocurrencies and the crypto sentiment, positive (good) and negative (bad) cryptocurrency volatilities are created. Due to the non-stationary nature of cryptocurrency prices - according to the ERS unit-root test (Elliott et al. (1996)) - 30-minute returns are computed by, $y_t = \frac{z_t - z_{t-1}}{z_{t-1}}$. In the next step, daily positive and negative absolute returns (Baruník et al. (2016))¹ are computed as follows:

¹Baruník et al. (2016) uses positive and negative semivariances for the asymmetric connected-

$$(5.1) \quad x_t = \frac{1}{T} \sum_{t=1}^T |y_t|$$

$$(5.2) \quad I_t = \begin{cases} 0, & \text{if } y_t < 0 \\ 1, & \text{if } y_t \geq 0 \end{cases}$$

$$(5.3) \quad x_t = \left(\frac{1}{T} \sum_{t=1}^T (1 - I_t) |y_t| \right) + \left(\frac{1}{T} \sum_{t=1}^T I_t |y_t| \right)$$

$$(5.4) \quad x_t = x_t^+ + x_t^-$$

x_t^+ and x_t^- represent the daily average positive and negative absolute returns which can also be seen as good and bad volatilities, respectively.

5.1.2 Methods

5.1.2.1 Sentiment Analysis

Based on the comprehensive analysis presented in Paper 4, this study uses the most effective model for sentiment analysis. The sentiment of each forum post and comment is processed using a value assignment system: text identified as positive is labeled with a value of 1, negative sentiment is labeled with -1, and neutral statements are labelled with 0. This numerical sentiment coding is essential for alignment.

Average sentiments are calculated as follows. Sentiment is averaged over 30-minute segments to correspond to specific time data points in the cryptocurrency market for further examination.

ness approach, however, as squaring returns frequently generate outliers, the focus is on absolute returns.

The average sentiment is calculated as follows:

$$(5.5) \quad \text{average sentiment} = \frac{\text{net sentiment}}{\text{total number of texts}}$$

5.1.2.2 Granger Causality

After the sentiment variable is constructed, the impact of sentiment on cryptocurrency price volatility is analyzed using the standard Granger causality approach (Granger (1969)). Granger causality testing is a statistical hypothesis test used to determine whether one time series helps to predict another. Essentially, the lagged value of social media sentiment is tested to see if it significantly explains the current value of cryptocurrency price volatility.

Specifically, if cryptocurrency is denoted as P_t and social media sentiment as S_t , then the VAR model can be specified as follows:

$$(5.6) \quad P_t = \alpha + \sum_{i=1}^n \beta_i P_{t-i} + \sum_{i=1}^n \gamma_i S_{t-i} + \epsilon_t$$

$$(5.7) \quad S_t = \delta + \sum_{i=1}^n \phi_i S_{t-i} + \sum_{i=1}^n \theta_i P_{t-i} + \mu_t$$

where n is the number of lags, ϵ_t and μ_t are error terms, and $\alpha, \beta_i, \gamma_i, \delta, \phi_i, \theta_i$ are parameters to be estimated.

The null hypothesis for the Granger causality test from social media sentiments to cryptocurrencies' price volatility is that the coefficients of the lagged values of S_t (γ_i) in the first equation are jointly zero. If this hypothesis is rejected, it implies that social media sentiments Granger-cause cryptocurrencies' prices. Granger causality

from cryptocurrencies' price volatility to social media sentiments is also tested.

5.1.2.3 Time Varying Parameter Vector Autoregression (TVP-VAR)

In the forthcoming phase of this study, the Time-Varying Parameter Vector Autoregression (TVP-VAR) frequency connectedness method is employed to estimate the measures separately for both positive and negative components. The TVP-VAR approach is demonstrated to outperform and mitigate several notable limitations of the rolling-window Vector Autoregression (VAR) methodology. Among the addressed limitations are: (i) the necessity to subjectively select the size of the rolling window, (ii) the loss of observations, and (iii) the susceptibility of the parameters to outliers.

In particular, a TVP-VAR(2) is estimated as suggested by the Bayesian information criterion (BIC) which can be outlined as:

$$(5.8) \quad x_t = \Phi_t x_{t-1} + \epsilon_t \quad \epsilon_t \sim N(0, \Sigma_t)$$

$$(5.9) \quad vec(\Phi_t) = vec(\Phi_{t-1}) + v_t \quad v_t \sim N(0, R_t)$$

where x_t , x_{t-1} and ϵ_t are $N \times 1$ dimensional vectors in t , $t - 1$, and the corresponding error term, respectively. Φ_t and Σ_t are $N \times N$ dimensional matrices demonstrating the time-varying VAR coefficients and the time-varying variance-covariances whereas $vec(\Phi_t)$ and v_t are $N^2 \times 1$ dimensional vectors and R_t is a $N^2 \times N^2$ dimensional matrix.

Since the concept of the generalized forecast error variance decomposition (GFEVD) introduced by Koop et al. (1996) and Pesaran and Shin (1998) is built upon the Wold representation theorem, the estimated TVP-VAR model has to be trans-

formed into its TVP-VMA process by the following equality: $x_t = \sum_{i=1}^p \Phi_{it}x_{t-i} + \epsilon_t = \sum_{j=0}^{\infty} \Psi_{jt}\epsilon_{t-j}$. The GFEVD is preferred over its orthogonal counterpart as the retrieved results are completely invariant with the variable ordering. Additionally, Wiesen et al. (2018) stress, that the GFEVD should be employed if no theoretical framework - which would allow identifying the error structure - is available.

Subsequently, the GFEVD is computed which can be interpreted as the effect a shock in variable j has on variable i in terms of its forecast error variance:

$$(5.10) \quad \theta_{ijt}(H) = \frac{(\Sigma_t)_{jj}^{-1} \sum_{h=0}^H ((\Psi_{ht}\Sigma_t)_{ijt})^2}{\sum_{h=0}^H (\Psi_{ht}\Sigma_t\Psi'_{ht})_{ii}}$$

$$(5.11) \quad \tilde{\theta}_{ijt}(H) = \frac{\theta_{ijt}(H)}{\sum_{i=1}^N \theta_{ijt}(H)}$$

where $\tilde{\theta}_{ijt}(H)$ represents the contribution of the j th variable to the variance of the forecast error of the i th variable at horizon H , the rows of $\tilde{\theta}_{ijt}(H)$ are not summed up to unity. Therefore, each variance contribution is normalized by dividing by its row sum. By applying this normalization technique, the following equations are obtained: $\sum_{i=1}^N \tilde{\theta}_{ijt}(H) = 1$ and $\sum_{j=1}^N \sum_{i=1}^N \tilde{\theta}_{ijt}(H) = N$.

After the GFEVD is calculated, all connectedness measures are computed. On the bilateral level, the net pairwise connectedness measures are calculated.

$$(5.12) \quad NPDC_{ijt}(H) = \tilde{\theta}_{ijt}(H) - \tilde{\theta}_{jit}(H).$$

If $NPDC_{ijt}(H) > 0$ ($NPDC_{ijt}(H) < 0$), variable j influences variable i more (less) than vice versa which means that variable j dominates (is dominated) by

variable i .

The *total directional connectedness TO others* measures the influence a shock in variable i has on all other variables j :

$$(5.13) \quad TO_{it}(H) = \sum_{j=1, j \neq i}^N \tilde{\theta}_{jit}(H)$$

while the *total directional connectedness FROM others* measures how much variable i is influenced by shocks in all other variables j :

$$(5.14) \quad FROM_{it}(H) = \sum_{j=1, j \neq i}^N \tilde{\theta}_{ijt}(H).$$

The *net total directional connectedness* is the difference between the total directional connectedness TO others and the total directional connectedness FROM others and is interpreted as the net influence variable i has on the analyzed network.

$$(5.15) \quad NET_{it}(H) = TO_{it}(H) - FROM_{it}(H)$$

If $NET_{it} > 0$ ($NET_{it} < 0$), variable i influences all others j more (less) than being influenced by them. Thus, variable i is considered as a net transmitter (receiver) of shocks.

Instead of the originally proposed total connectedness index (TCI), the corrected TCI of Chatziantoniou et al. (2021) and Gabauer (2021) is applied to measure the degree of network interconnectedness:

(5.16)

$$TCI_t(H) = \frac{N}{N-1} \sum_{i=1}^N TO_{it}(H) = \frac{N}{N-1} \sum_{i=1}^N FROM_{it}(H) \quad 0 \leq TCI_t(H) \leq 1 \text{ if } H \rightarrow \infty$$

The TCI measures the network interconnectedness, which is illustrated by the average effect a shock in one variable has on all others. The higher the TCI, the higher the market risk.

Up to this point, attention has been focused solely on the connectedness measures in the time domain. Analogously, the exploration continues with the connectedness measures in the frequency domain. Following the spectral decomposition method of Stiasny (1996), the connectedness relationship in the frequency domain is examined. First, the frequency response function, $\Psi(e^{-i\omega}) = \sum_{h=0}^{\infty} e^{-i\omega h} \Psi_h$, where $i = \sqrt{-1}$ and ω denotes the frequency, and the spectral density of x_t at frequency ω are defined as a Fourier transformation of the TVP-VMA(∞):

$$(5.17) \quad S_{xt}(\omega) = \sum_{h=-\infty}^{\infty} E(x_t x'_{t-h}) e^{-i\omega h} = \Psi_t(e^{-i\omega h}) \Sigma_t \Psi'_t(e^{+i\omega h}).$$

The frequency GFEVD is seen as the combination of the spectral density and the GFEVD. In the frequency domain, the frequency GFEVD needs to be normalized as well.

$$(5.18) \quad \theta_{ijt}(\omega) = \frac{(\Sigma_t)_{jj}^{-1} |\sum_{h=0}^{\infty} (\Psi_t(e^{-i\omega h}) \Sigma_t)_{ijt}|^2}{\sum_{h=0}^{\infty} (\Psi_t(e^{-i\omega h}) \Sigma_t \Psi'_t(e^{i\omega h}))_{ii}}$$

$$(5.19) \quad \tilde{\theta}_{ijt}(\omega) = \frac{\theta_{ijt}(\omega)}{\sum_{k=1}^N \theta_{ikt}(\omega)}$$

where $\tilde{\theta}_{ijt}(\omega)$ illustrates the portion of the spectrum of variable i at a given frequency ω that is attributed to a shock in the variable j .

To examine short-term and long-term connectedness measures rather than connectedness measures at a single frequency, all frequencies within a specific range, $d = (a, b)$: $a, b \in (-\pi, \pi), a < b$, are aggregated:

$$(5.20) \quad \tilde{\theta}_{ijt}(d) = \int_a^b \tilde{\theta}_{ijt}(\omega) d\omega$$

Now, the same connectedness measures as outlined in Diebold and Yılmaz (2012, 2014) are computed. These measures have the same interpretation, however, in this case, they refer to frequency connectedness measures that provide information about spillovers in a certain frequency range d :

$$(5.21) \quad NPDC_{ijt}(d) = \tilde{\theta}_{ijt}(d) - \tilde{\theta}_{jit}(d)$$

$$(5.22) \quad TO_{it}(d) = \sum_{j=1, j \neq i}^N \tilde{\theta}_{jit}(d)$$

$$(5.23) \quad FROM_{it}(d) = \sum_{j=1, j \neq i}^N \tilde{\theta}_{ijt}(d)$$

$$(5.24) \quad NET_{it}(d) = TO_{it}(d) - FROM_{it}(d)$$

$$(5.25) \quad TCI_t(d) = \frac{N}{N-1} \sum_{i=1}^N TO_{it}(d) = \frac{N}{N-1} \sum_{i=1}^N FROM_{it}(d)$$

As all measures provide information about a specific frequency range, however, not of the overall impact, Baruník and Křehlík (2018) suggest weighting all contributions with respect to the overall system using $\Gamma(d) = \sum_{i,j=1}^N \tilde{\theta}_{ijt}(d)/N$:

$$(5.26) \quad NPD\tilde{C}_{ijt}(d) = \Gamma(d) \cdot NPDC_{ijt}(d)$$

$$(5.27) \quad TO_{it}^{\sim}(d) = \Gamma(d) \cdot TO_{it}(d)$$

$$(5.28) \quad FROM_{it}^{\sim}(d) = \Gamma(d) \cdot FROM_{it}(d)$$

$$(5.29) \quad NET_{it}^{\sim}(d) = \Gamma(d) \cdot NET_{it}(d)$$

$$(5.30) \quad TCI_t^{\sim}(d) = \Gamma(d) \cdot TCI_t(d)$$

Finally, the relationship between time-domain and frequency-domain measures can be represented as follows.

$$(5.31) \quad NPDC_{ijt}(H) = \sum_d NPDC_{ijt}(d)$$

$$(5.32) \quad TO_{it}(H) = \sum_d TO_{it}(d)$$

$$(5.33) \quad FROM_{it}(H) = \sum_d FROM_{it}(d)$$

$$(5.34) \quad NET_{it}(H) = \sum_d NET_{it}(d)$$

$$(5.35) \quad TCI_t(H) = \sum_d TCI_t(d).$$

A Connectedness Table is provided as the final output, offering a snapshot of the interdependence between variables at each point in time. This table clearly shows both the influence of each variable on the others and its susceptibility to shocks from them. The measures and methodology described are not only helpful in understanding the asymmetric connectedness structure among the variables under study but also in providing insights into how they change over time.

5.2 Results

The relationship between market sentiment and cryptocurrency volatility is analyzed, revealing interesting dynamics as evidenced by the Granger causality test. The test plays a key role in understanding the predictive power of past sentiment values on current market volatility, which varies across cryptocurrencies. In some cases, the p-value is below the 0.01 critical value, indicating a statistically significant relationship between past sentiment values and current market volatility. However, in other cases, the p-value is found close to 1, indicating that past sentiment is a very small predictor of current market volatility. This pattern is consistent across cryptocurrencies.

The Granger causality test results are shown in Figures 5.1, 5.2, 5.3, and 5.4. The figures below display the p-values obtained from the Granger causality test conducted on rolling windows of 250 days. The test aims to determine the causality between sentiments and the volatility of various cryptocurrency coins. Each point on the graph represents the p-value for a specific 250-day window, with the window moving forward in time across the dataset. Lines indicating significance levels at 0.05(red) and 0.01(green) are also shown.

Figure 5.1: BTC Granger Causality P values - Sentiments to Volatility

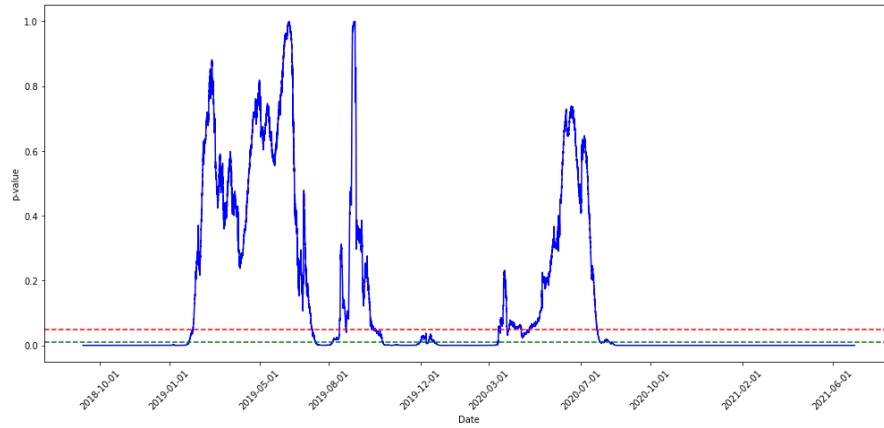


Figure 5.2: ETH Granger Causality P values - Sentiments to Volatility

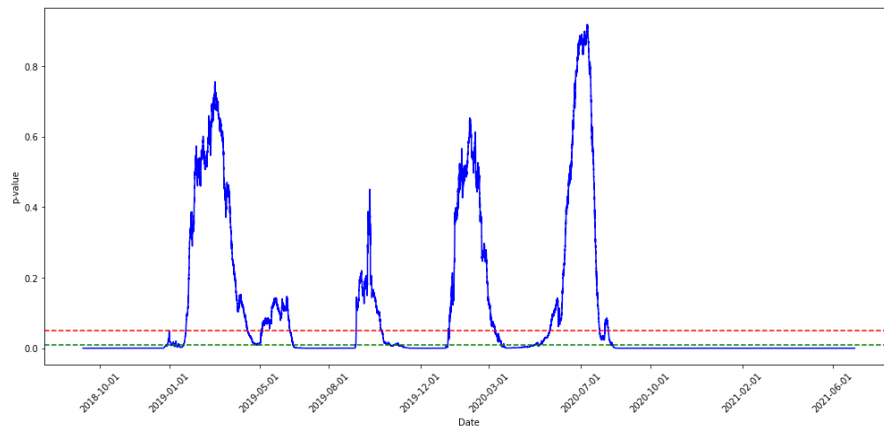


Figure 5.3: XRP Granger Causality P values - Sentiments to Volatility

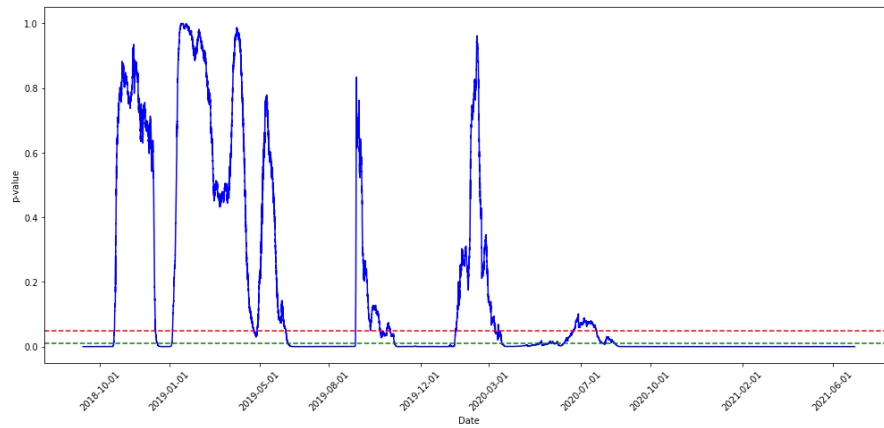
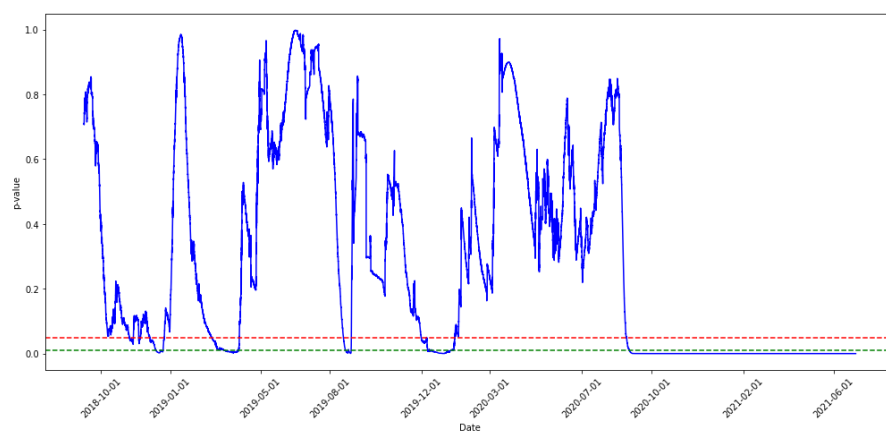


Figure 5.4: DOGE Granger Causality P values - Sentiments to Volatility



The reversed Granger causality test (volatility to sentiments) results are shown in Figures 5.5, 5.6, 5.7 and 5.8.

Figure 5.5: BTC Granger Causality P values - Volatility to Sentiments

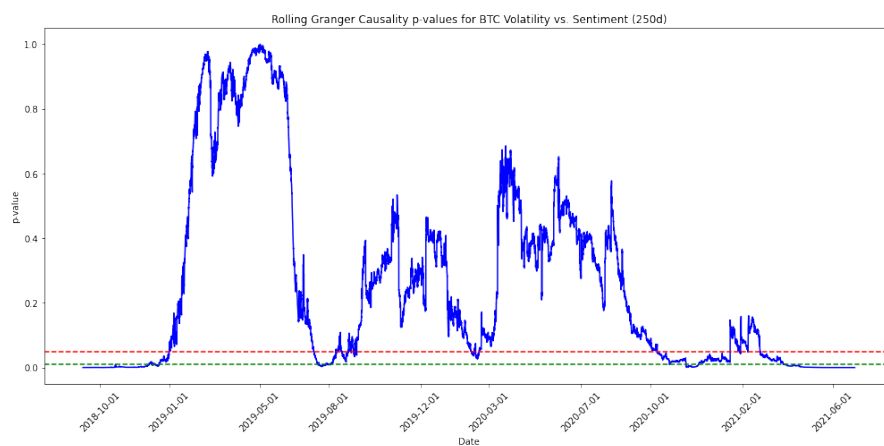


Figure 5.6: ETH Granger Causality P values - Volatility to Sentiments

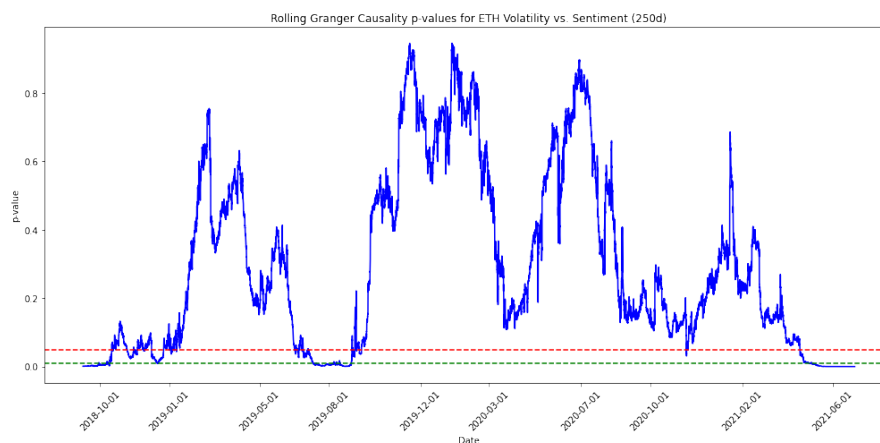


Figure 5.7: XRP Granger Causality P values - Volatility to Sentiments

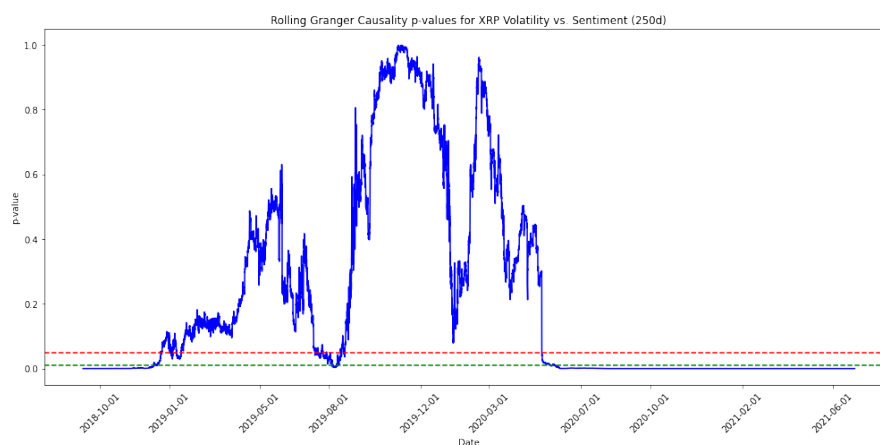
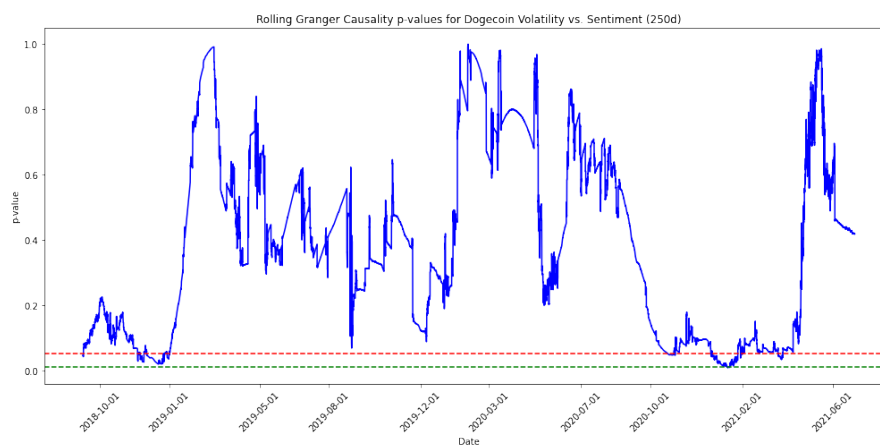


Figure 5.8: DOGE Granger Causality P values - Volatility to Sentiments



Several factors are attributed to the volatile nature of these p-values. External shocks such as regulatory changes, technological advances, macroeconomic events, or geopolitical crises are found to have a considerable impact on the link between sentiment and market volatility. Another characteristic identified in the cryptocurrency market is the variation of volatility over time, swinging between peaks and valleys. The impact of sentiment on market volatility is seen to vary across time, necessitating the introduction of time-varying parameter vector autoregression (TVP-VAR) in the analysis.

The results of the reverse Granger causality test are consistent with the results of the original test, suggesting the existence of a feedback loop, i.e., market volatility affects sentiment, which feeds back into market volatility. This dynamic interaction evolves over time, creating a periodic effect whereby positive sentiment may drive an increase in trading activity, thereby increasing market volatility. Increased volatility in turn affects market sentiment, either reinforcing the pre-existing sentiment or generating the opposite sentiment. The observed patterns suggest that market participants not only react to price volatility but also influence future price movements through collective sentiment.

The average connectedness tables (Table 5.1) provide a comprehensive overview of the dynamics over different time periods, rather than focusing on individual events. The Total Connectedness Index (TCI) shows moderate connectivity among the network variables, but there is a clear asymmetry, i.e., positive shocks are more strongly connected than negative shocks. Regardless of the nature of the shocks, the division of variables into net receivers and net transmitters remains consistent. Sentiment, Ethereum (ETH), and dogecoin (DOGE) are the main net receivers,

Table 5.1: Averaged total positive (negative) dynamic connectedness table

	BTC	ETH	XRP	DOGE	Sentiments	FROM
BTC	48.37 (51.03)	27.35 (26.11)	17 (15.78)	4.52 (5.42)	2.76 (1.66)	51.63 (48.97)
ETH	29.95 (24.44)	44.09 (48.65)	19.27 (20.23)	4.04 (4.95)	2.66 (1.73)	55.91 (51.35)
XRP	15.91 (10.74)	14.91 (13.11)	61.7 (67.05)	4.42 (5.58)	3.06 (3.52)	38.3 (32.95)
DOGE	5.48 (5.95)	3.95 (3.59)	8.04 (8.11)	81.15 (81.45)	1.38 (0.9)	18.85 (18.55)
Sentiments	5.05 (2.51)	4.8 (3.02)	5.66 (3.46)	2.27 (1.73)	82.22 (89.29)	17.78 (10.71)
TO	56.39 (43.64)	51.01 (45.83)	49.96 (47.57)	15.25 (17.68)	9.86 (7.81)	182.47 (162.53)
Inc.Own	104.76 (94.67)	95.1 (94.48)	111.66 (114.62)	96.41 (99.13)	92.08 (97.09)	TCI
NET	4.76 (-5.33)	-4.9 (-5.52)	11.66 (14.62)	-3.59 (-0.87)	-7.92 (-2.91)	36.49(32.51)
NPT	3 (2)	1 (2)	4 (3)	2 (2)	0 (1)	

Notes: Results are based on a TVP-VAR model with a lag length of order two (BIC) and a 100-step ahead generalized forecast error variance decomposition. Values represent positive and negative - in parentheses - connectedness measures, respectively.

while Ripple (XRP) and Bitcoin (BTC) are mainly net transmitters of positive and negative shocks.

The long-term connectedness, indicated in Table 5.2, shows a minimal degree of TCI asymmetry, whereas the short-term connectedness, presented in Table 5.3, reveals more pronounced asymmetry. These results highlight the varying impact of shocks on the network in the long and short term. It's noteworthy that the classification into net recipients and net transmitters in the long-term and short-term TCI deviates slightly from the results in Table 5.1, with Ethereum (ETH) assuming a net transmitter role in both cases. However, sentiment remains a significant net recipient across all frequencies.

Table 5.2: Averaged long-term positive (negative) dynamic connectedness table

	BTC	ETH	XRP	DOGE	Sentiments	FROM
BTC	45.02 (43.65)	30.44 (30.16)	16.82 (18.08)	4.04 (5.7)	3.67 (2.41)	54.98 (56.35)
ETH	30.63 (25.44)	41.48 (43.48)	17.56 (19.22)	5.08 (7.36)	5.25 (4.5)	58.52 (56.52)
XRP	20.28 (15.52)	20.13 (19.7)	50.04 (54.56)	5.57 (5.99)	3.98 (4.23)	49.96 (45.44)
DOGE	6.22 (8.69)	7.05 (8.79)	9.56 (8.06)	70.23 (66.18)	6.95 (8.28)	29.77 (33.82)
Sentiments	5.25 (4.54)	7.1 (4.73)	4.58 (3.9)	11.53 (16.34)	71.53 (70.49)	28.47 (29.51)
TO	62.38 (54.19)	64.72 (63.38)	48.53 (49.25)	26.23 (35.39)	19.85 (19.43)	221.7 (221.65)
Inc.Own	107.4 (97.84)	106.2 (106.86)	98.57 (103.81)	96.45 (101.57)	91.38 (89.91)	cTCI/TCI
NET	7.4 (-2.16)	6.2 (6.86)	-1.43 (3.81)	-3.55 (1.57)	-8.62 (-10.09)	55.42(55.41)/44.34(44.33)
NPT	4 (2)	3 (4)	2 (1)	1 (1)	0 (1)	

Notes: Results are based on a TVP-VAR model with a lag length of order two (BIC) and a 100-step ahead generalized forecast error variance decomposition. Values represent positive and negative - in parentheses - connectedness measures, respectively.

Table 5.3: Averaged short-term positive (negative) dynamic connectedness table

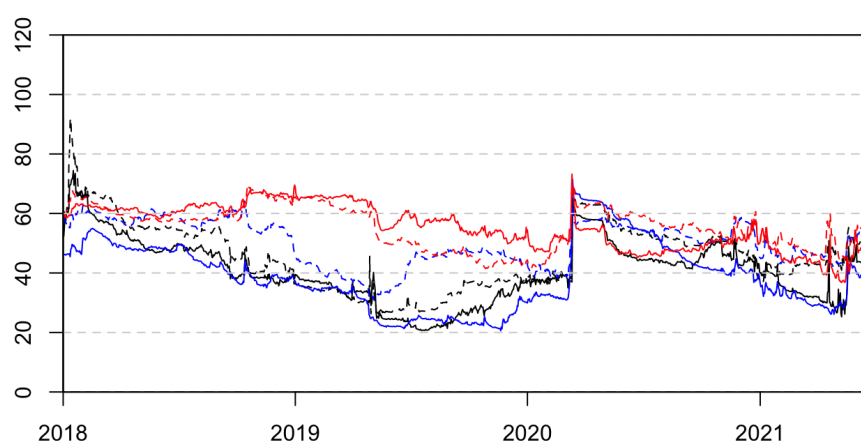
	BTC	ETH	XRP	DOGE	Sentiments	FROM
BTC	45.32 (58.3)	29.29 (25.07)	19.42 (10.3)	2.76 (4.07)	3.22 (2.26)	54.68 (41.7)
ETH	26.61 (24.02)	48.41 (54.3)	20.23 (15.05)	2.7 (2.94)	2.05 (3.69)	51.59 (45.7)
XRP	16.89 (10.79)	18.82 (16.45)	57.49 (66.83)	3.58 (2.38)	3.21 (3.55)	42.51 (33.17)
DOGE	5.68 (6.02)	5.58 (4.41)	8.23 (4.93)	77.22 (83.15)	3.28 (1.49)	22.78 (16.85)
Sentiments	7.22 (3.28)	7.43 (4.97)	10.6 (6.68)	3.67 (3.29)	71.08 (81.77)	28.92 (18.23)
TO	56.41 (44.12)	61.12 (50.9)	58.48 (36.96)	12.71 (12.68)	11.77 (10.99)	200.5 (155.64)
Inc.Own	101.73 (102.42)	109.53 (105.2)	115.97 (103.8)	89.92 (95.83)	82.85 (92.76)	cTCI/TCI
NET	1.73 (2.42)	9.53 (5.2)	15.97 (3.8)	-10.08 (-4.17)	-17.15 (-7.24)	50.12 (38.91) / 40.10 (31.13)
NPT	2 (3)	3 (4)	4 (2)	1 (1)	0 (0)	

Notes: Results are based on a TVP-VAR model with a lag length of order two (BIC) and a 100-step ahead generalized forecast error variance decomposition. Values represent positive and negative - in parentheses - connectedness measures, respectively.

The dynamic connectedness results are instrumental in understanding the evolution of connectedness over time, particularly in response to disruptive events. Fluctuations in connectedness strength often indicate strong co-movement among network variables, hinting at potential contagion dynamics. The analysis of TCI, depicted in Figure 5.9, reveals considerable variations over time, with total connectedness being sensitive to external shocks and specific events. Notable peaks in

total connectedness were observed at the end of 2018 and mid-2019, and in early 2020, both positive and negative connectedness values remained high, indicating strong co-movement and contagion dynamics.

Figure 5.9: Total, long-term and short-term connectedness



Note: Results are based on a TVP-VAR model with a lag length of order two (BIC) and a 100-step ahead generalized forecast error variance decomposition. Black, red, and blue lines stand for the total, long-term, and short-term connectedness measures, respectively. Solid lines represent negative connectedness measures while dashed lines represent positive connectedness measures.

The asymmetry in time-domain connectedness is evident, with long-term negative connectedness values dominating until early 2020. Positive connectedness briefly exceeds negative values during the COVID-19 outbreak, but the difference between positive and negative connectedness remains small throughout 2020. Toward the end of the sample, positive correlations regain their dominance, reflecting the typical risk-averse nature of investors. When positive perceptions of market risk outweigh negative perceptions, positive correlation metrics outweigh negative metrics, and vice versa.

The frequency-domain assessment further clarifies the dominance of low-frequency

(long-term) dynamics in the initial period, with the exception of early- to mid-2020. By late 2020 and early 2021, low-frequency connectedness again dominates, suggesting that the effects of structural changes in the market are long-lasting. The pattern of asymmetry in the frequency domain is consistent with the pattern of asymmetry in total connectedness, as shown in Figure 5.9.

The directional connectedness, represented in Figures 5.10, 5.11, 5.12, 5.13 and 5.14, reveal shifts between net-transmitting and net-receiving roles among various cryptocurrencies and sentiment. Sentiment is consistently subject to shocks, suggesting that it is vulnerable to cryptocurrency market dynamics. XRP, primarily a net transmitter, shifts to a net-receiving role around the end of 2018 to early 2019, suggesting that fundamental changes influenced this transition. Bitcoin (BTC) exhibits fluctuations between roles until the end of 2020, with high-frequency dynamics becoming more prevalent after 2021. Ethereum assumes a net transmitting role until mid-2020, dominates by long-term dynamics, and then transitions to a net recipient role, indicative of quick shock absorption. Dogecoin fluctuates between roles until the end of 2020 and becomes a net transmitter after that, with low-frequency dynamics becoming more prevalent after 2021.

Figure 5.10: BTC Net total, long-term, and short-term directional connectedness measures

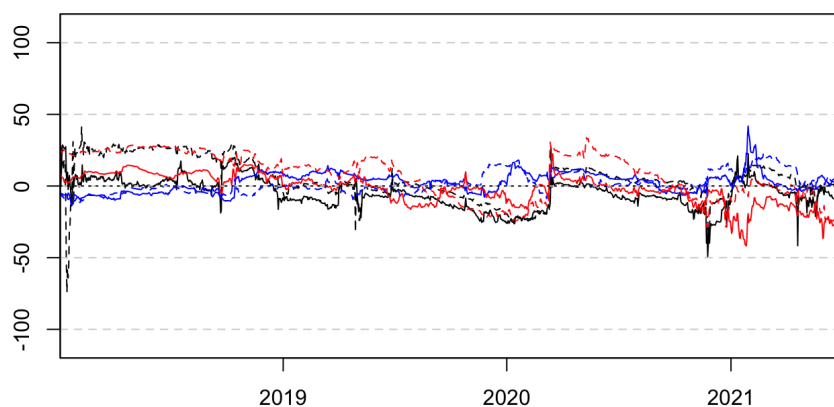


Figure 5.11: ETH Net total, long-term, and short-term directional connectedness measures

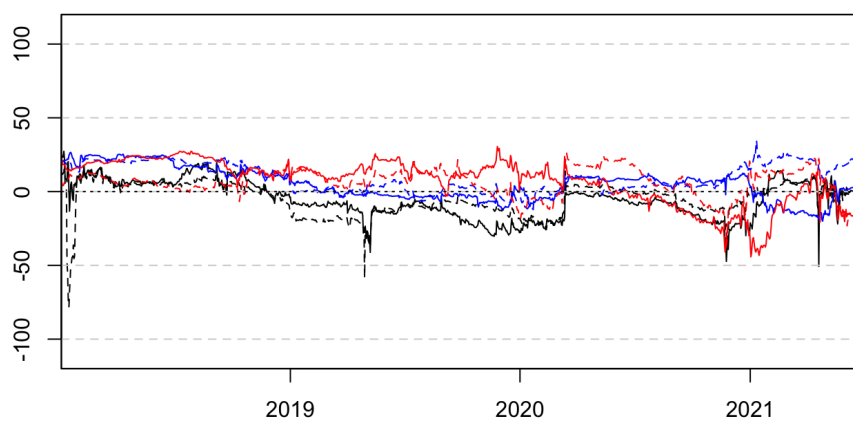


Figure 5.12: XRP Net total, long-term, and short-term directional connectedness measures

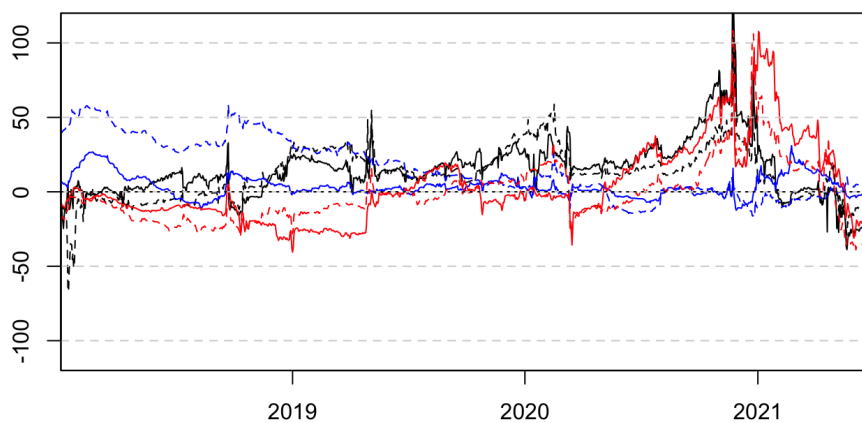


Figure 5.13: DOGE Net total, long-term, and short-term directional connectedness measures

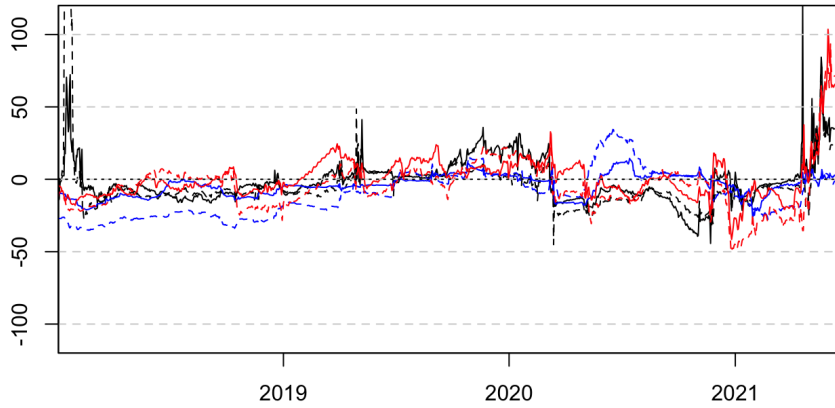
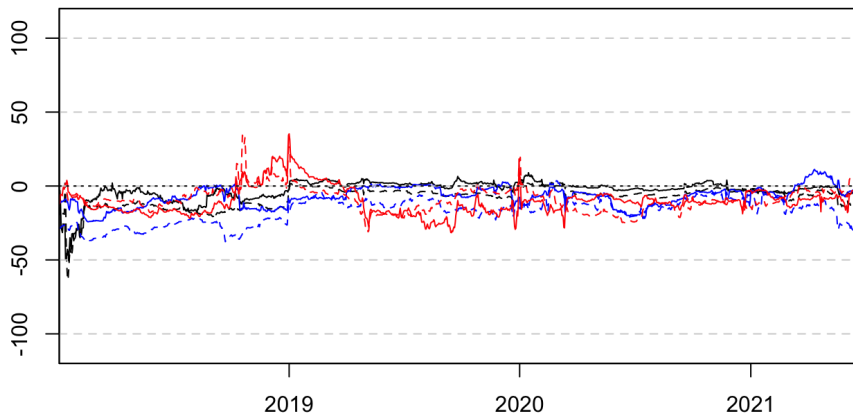


Figure 5.14: SENTIMENT Net total, long-term, and short-term directional connectedness measures

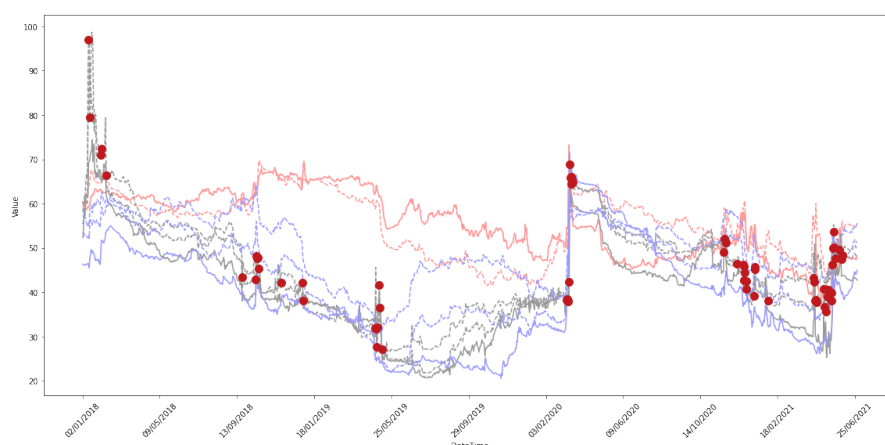


The influence of 'whales' in the cryptocurrency market is a significant aspect of this study. Whales, typically defined as entities or individuals holding large quantities of cryptocurrency (notably those holding over 1,000 Bitcoin ²), can have a profound impact on market dynamics. Their trading behaviors, such as large-scale buying or selling, can substantially alter market supply and demand, leading to notable price shifts. Additionally, given their influential position within the cryptocurrency community, their public sentiment, especially when shared via social media platforms, can influence overall market sentiment.

²<https://www.ledger.com/academy/glossary/whale>

The interconnectedness between whale activities and market sentiments is specifically investigated in this analysis. By utilizing a connectedness measure marked by periods of increased whale activity, a correlation between these activities and fluctuations in market connectedness is observed (see Figure 5.15).

Figure 5.15: Connectedness Spikes and Their Associated Whales Activities



Note: Results are based on a TVP-VAR model with a lag length of order two (BIC) and a 100-step ahead generalized forecast error variance decomposition. Black, red, and blue lines stand for the total, long-term, and short-term connectedness measures, respectively. Solid lines represent negative connectedness measures while dashed lines represent positive connectedness measures. Red markers represent whales' activities.

This suggests that whale actions are closely linked to shifts in market sentiment and activity. Further research, including regression analysis that differentiates between positive (buying) and negative (selling) whale activities, provides deeper insights into how these significant market players influence the broader cryptocurrency landscape.

5.3 Conclusion

This research explores how the cryptocurrency market reacts to changes, particularly highlighting that market sentiment often directly feels the impact of sudden shifts within its network. This key finding points out the vital role that investor sentiment has in responding to fluctuations and unexpected events, influencing and being influenced by the overall behavior of the cryptocurrency market.

The p-values derived from the Granger causality tests over 250-day rolling windows assess the influence of sentiment on the volatility of key cryptocurrencies like Bitcoin, Ethereum, Ripple, and Dogecoin. Reverse Granger causality tests, which explore the potential influence of volatility on sentiment, suggest a feedback loop: market changes affect sentiment, which in turn can impact future market fluctuations. This bidirectional influence indicates that market participants' collective sentiments are both a response to and a driver of price movements.

The study also highlights that cryptocurrency market volatility is sensitive to various external shocks and varies over time. The Total Connectedness Index (TCI) demonstrates a moderate level of interconnectedness within the market, with clear distinctions between cryptocurrencies that primarily receive sentiment-driven shocks and those that transmit them.

The study takes a close look at the connections within the cryptocurrency market, illustrating that shifts in sentiment are immediate reactions to outside shocks and can signal the market's sensitivity and ability to bounce back from such events. This information is very useful for people across the financial world, like investors, analysts, and policymakers, as it gives them a deeper insight into what drives

market trends and how sentiment analysis could help foresee or lessen the effects of widespread market changes.

This research aims to close the gap in applying sentiment analysis between traditional markets and the unique world of cryptocurrencies. It also seeks to add to the conversation on behavioral finance by exploring how the psychology of investors, as shown through social media sentiment, impacts the volatile and speculative cryptocurrency market.

Appendix

Appendix 1: TVP-VAR explanation

The TVP-VAR is represented as follows,

$$\begin{aligned} x_t &= \Phi_t x_{t-1} + \epsilon_t & \epsilon_t &\sim N(0, \Sigma_t) \\ vec(\Phi_t) &= vec(\Phi_{t-1}) + v_t & v_t &\sim N(0, R_t) \end{aligned}$$

where x_t , x_{t-1} , and ϵ_t represent $N \times 1$ dimensional vectors and Φ_t and Σ_t are $N \times N$ dimensional matrices. Furthermore, $vec(\Phi_t)$ and v_t are $N^2 \times 1$ dimensional vectors and R_t is an $N^2 \times N^2$ dimensional matrix.

An empirical Bayes prior is applied where the priors, $vec(\Phi_0)$ and Σ_0 , are equal to the estimation results of a constant parameter VAR estimation based on the full dataset.

$$vec(B_0) \sim N(vec(B_{OLS}), R_{OLS})$$

$$S_0 = S_{OLS}.$$

The Kalman Filter estimation relies on forgetting factors ($0 \leq \kappa_i \leq 1$) which regulates how fast the estimated coefficients vary over time. If the forgetting factor is set equal to 1 the algorithm collapses to a constant parameter VAR. Since it is assumed that parameters are not changing dramatically from one day to another, κ_2 is set equal to 0.99:

$$vec(\Phi_t)|x_{1:t-1} \sim N(vec(\Phi_{t|t-1}), R_{t|t-1})$$

$$vec(\Phi_{t|t-1}) = vec(\Phi_{t-1|t-1})$$

$$R_t = (1 - \kappa_2^{-1})R_{t-1|t-1}$$

$$R_{t|t-1} = R_{t-1|t-1} + R_t$$

The multivariate EWMA procedure for Σ_t is updated in every step, while κ_1 and κ_2 are set equal to 0.99 based on the sensitivity results provided by Koop and Korobilis (2014). Furthermore, Koop and Korobilis (2014) fix the forgetting factors, as well, even if the forgetting factors can be estimated by the data, as in Koop and Korobilis (2013). The main reason to fix the parameters is twofold (i) it increases computational burden substantially and (ii) the value added to the forecasting performance is questionable.

$$\hat{e}_t = x_t - \Phi_{t|t-1}x_{t-1}$$

$$\Sigma_t = \kappa_1 \Sigma_{t-1|t-1} + (1 - \kappa_1) \hat{e}_t' \hat{e}_t$$

$vec(\Phi_t)$ and R_t are updated by

$$vec(\Phi_t)|x_{1:t} \sim N(vec(\Phi_{t|t}), R_{t|t})$$

$$vec(\Phi_{t|t}) = vec(\Phi_{t|t-1}) + R_{t|t-1}x'_{t-1}(\Sigma_t + x_{t-1}R_{t|t-1}x'_{t-1})^{-1}(x_t - \Phi_{t|t-1}x_{t-1})$$

$$R_{t|t} = R_{t|t-1} + R_{t|t-1}x'_{t-1}(\Sigma_t + x_{t-1}R_{t|t-1}x'_{t-1})^{-1}(x_{t-1}R_{t|t-1})$$

Finally, the variances, Σ_t , are updated by the EWMA procedure

$$\hat{e}_{t|t} = x_t - \Phi_{t|t}x_{t-1}$$

$$\Sigma_{t|t} = \kappa_1 \Sigma_{t-1|t-1} + (1 - \kappa_1) \hat{e}_{t|t}' \hat{e}_{t|t}$$

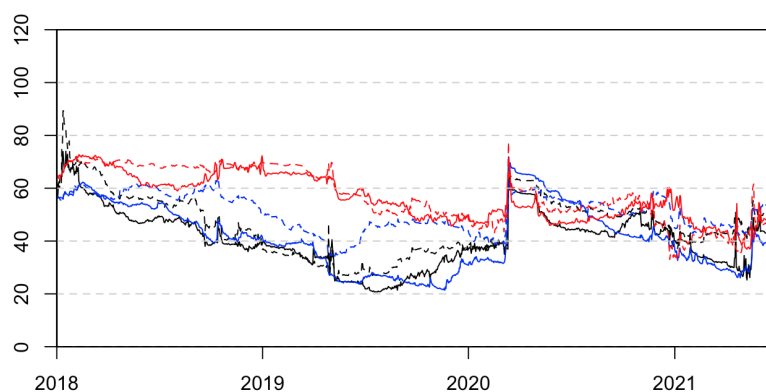
Appendix 2: TVP-VAR different rolling-window results Choosing a 250-day rolling window for the TVP-VAR model was a strategic decision aimed at capturing a comprehensive view of market dynamics while maintaining the model's responsiveness to recent changes. This period roughly equates to one calendar year of trading days, offering a balanced dataset that reflects various market conditions without being overly influenced by short-term fluctuations. This is particularly important in the volatile cryptocurrency market, where understanding longer-term trends amidst daily volatility is crucial.

Furthermore, the 250-day window strikes a balance between computational efficiency and data richness. This choice allows for a practical yet robust approach to modeling market dynamics, providing insights that are both meaningful and timely, which is essential for fast-moving markets like cryptocurrencies.

In addition to the primary analysis using a 250-day rolling window, results from analyses using shorter windows of 100, 150, and 200 days are shown below.

100 days TCI:

Figure 5.16: Total, long-term and short-term connectedness at 100 days rolling window



Note: Results are based on a TVP-VAR model with a lag length of order two (BIC) and a 100-step ahead generalized forecast error variance decomposition. Black, red, and blue lines stand for the total, long-term, and short-term connectedness measures, respectively. Solid lines represent negative connectedness measures while dashed lines represent positive connectedness measures.

100 days NET connectedness:

Figure 5.17: BTC Net total, long-term, and short-term directional connectedness measures at 100 days rolling window

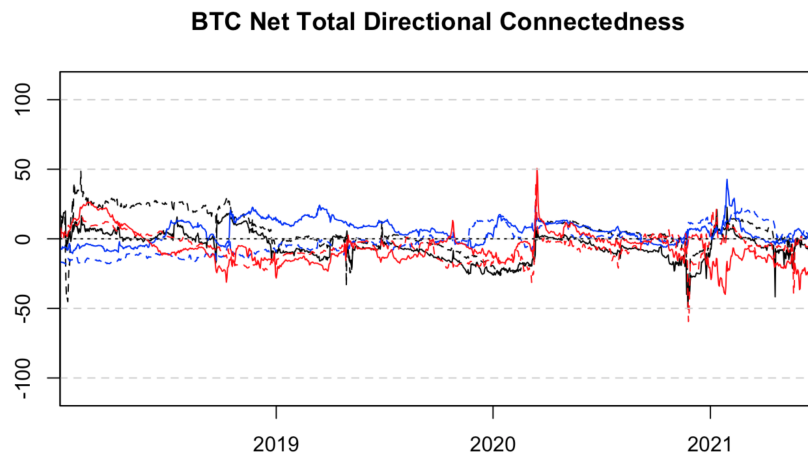


Figure 5.18: ETH Net total, long-term, and short-term directional connectedness measures at 100 days rolling window

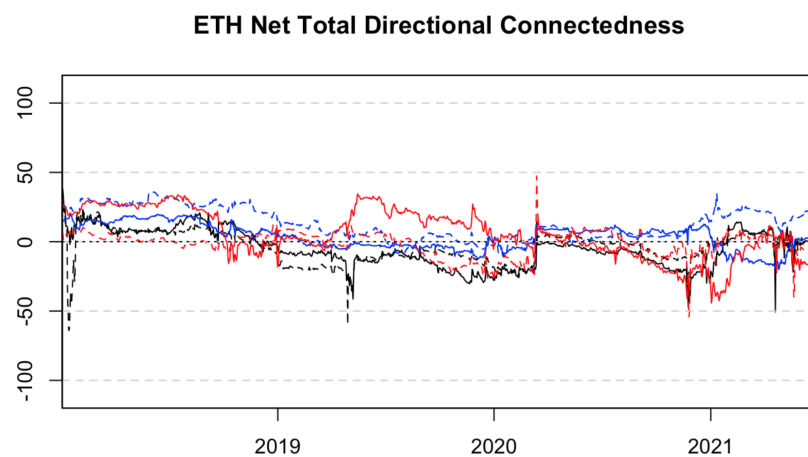


Figure 5.19: XRP Net total, long-term, and short-term directional connectedness measures at 100 days rolling window

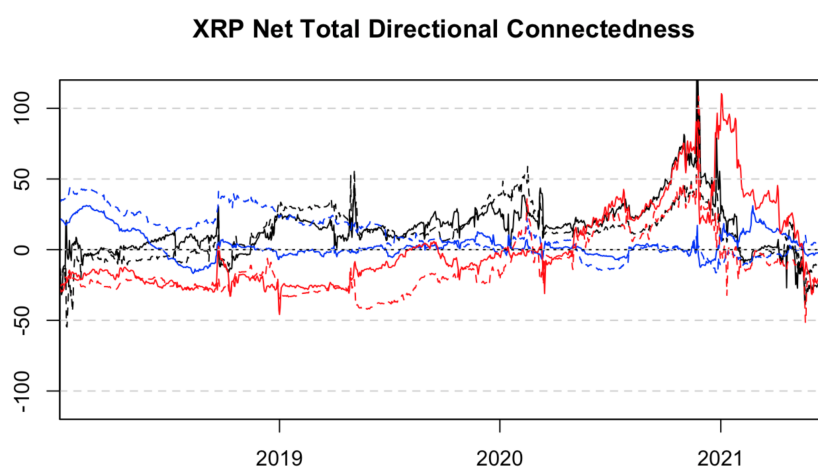
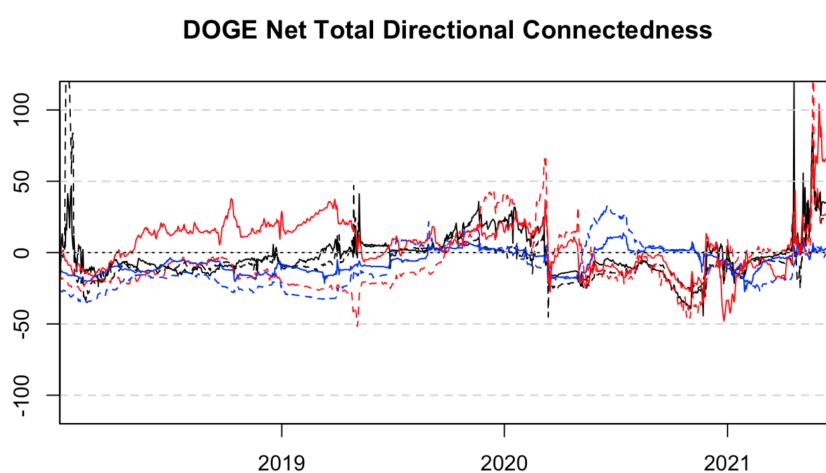
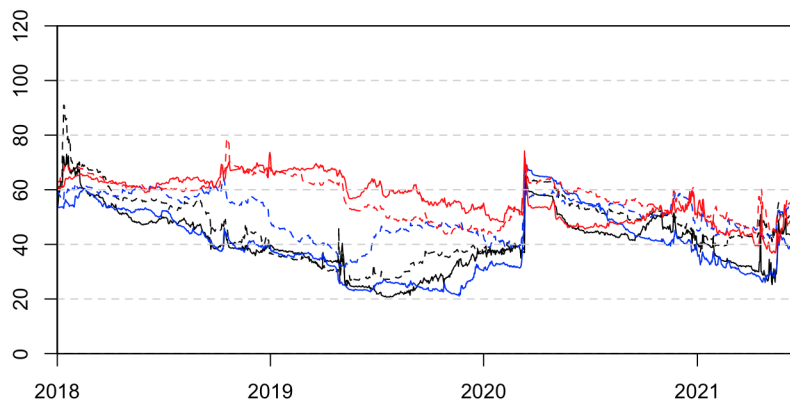


Figure 5.20: DOGE Net total, long-term, and short-term directional connectedness measures at 100 days rolling window



150 days TCI:

Figure 5.21: Total, long-term and short-term connectedness at 150 days rolling window



Note: Results are based on a TVP-VAR model with a lag length of order two (BIC) and a 100-step ahead generalized forecast error variance decomposition. Black, red, and blue lines stand for the total, long-term, and short-term connectedness measures, respectively. Solid lines represent negative connectedness measures while dashed lines represent positive connectedness measures.

150 days NET connectedness:

Figure 5.22: BTC Net total, long-term, and short-term directional connectedness measures at 150 days rolling window

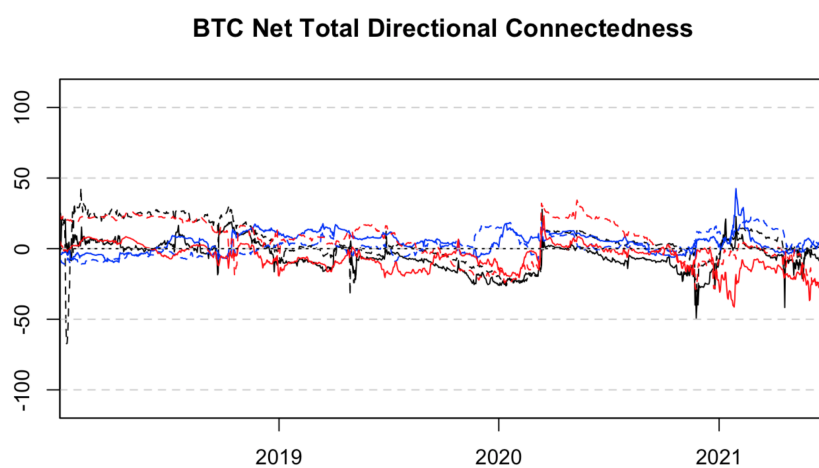


Figure 5.23: ETH Net total, long-term, and short-term directional connectedness measures at 150 days rolling window

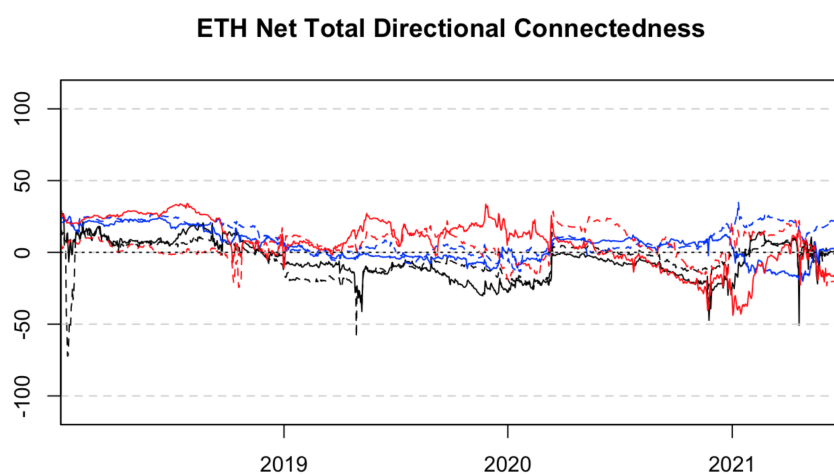


Figure 5.24: XRP Net total, long-term, and short-term directional connectedness measures at 150 days rolling window

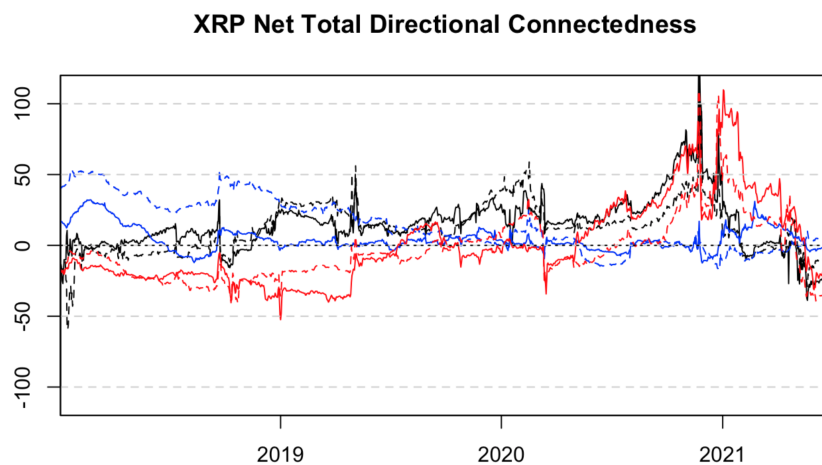
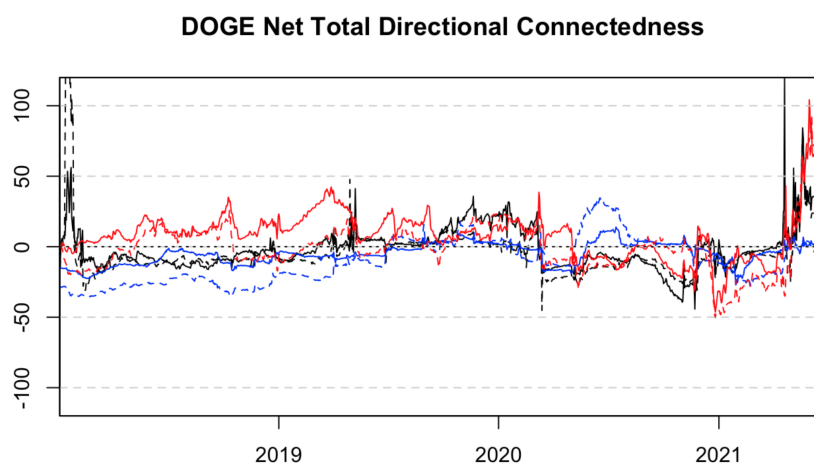
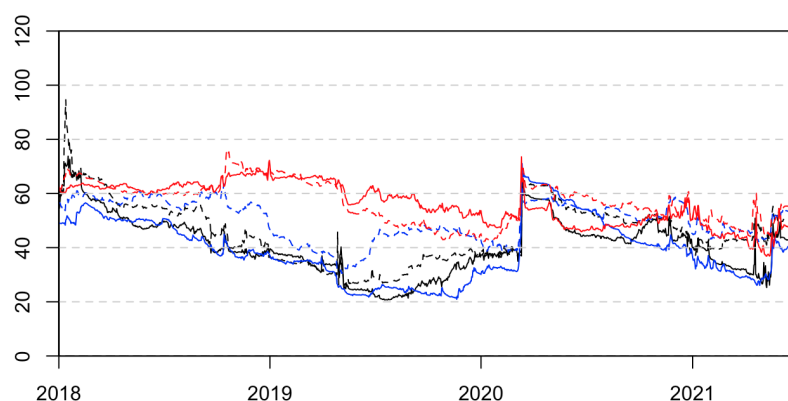


Figure 5.25: DOGE Net total, long-term, and short-term directional connectedness measures at 150 days rolling window



200 days TCI:

Figure 5.26: Total, long-term and short-term connectedness at 200 days rolling window



Note: Results are based on a TVP-VAR model with a lag length of order two (BIC) and a 100-step ahead generalized forecast error variance decomposition. Black, red, and blue lines stand for the total, long-term, and short-term connectedness measures, respectively. Solid lines represent negative connectedness measures while dashed lines represent positive connectedness measures.

200 days NET connectedness:

Figure 5.27: BTC Net total, long-term, and short-term directional connectedness measures at 200 days rolling window

BTC Net Total Directional Connectedness

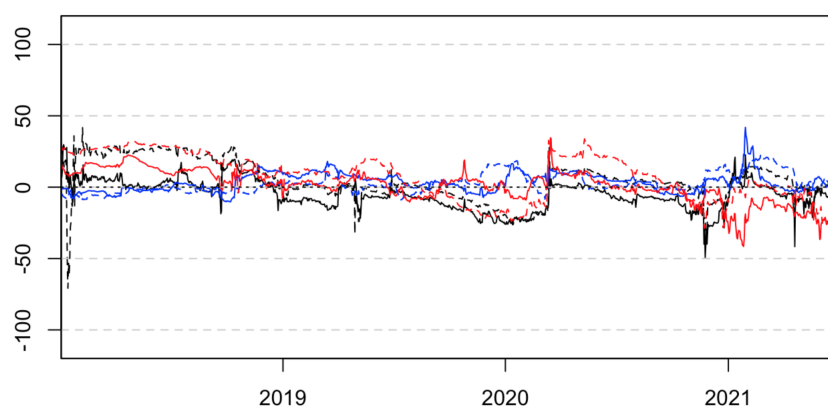


Figure 5.28: ETH Net total, long-term, and short-term directional connectedness measures at 200 days rolling window

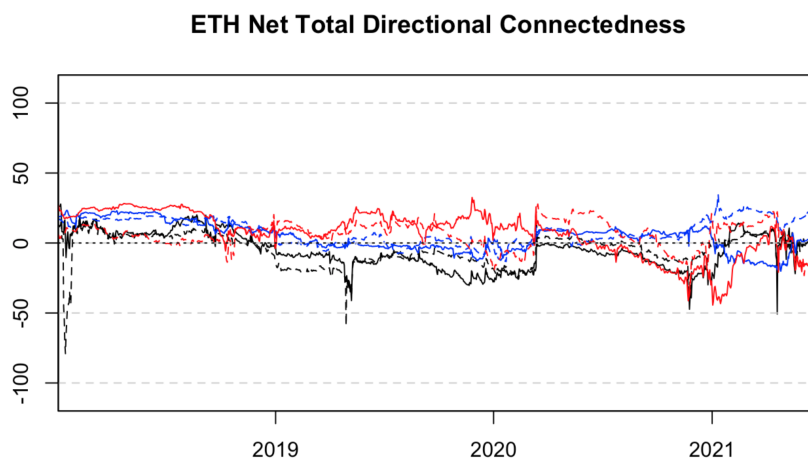


Figure 5.29: XRP Net total, long-term, and short-term directional connectedness measures at 200 days rolling window

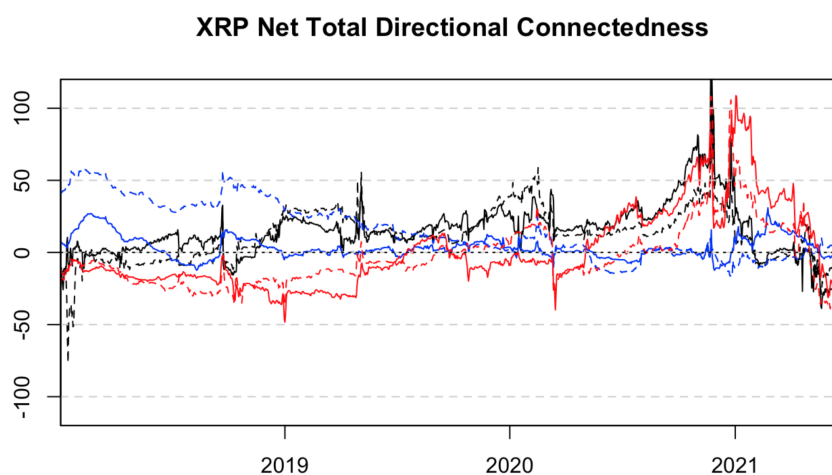
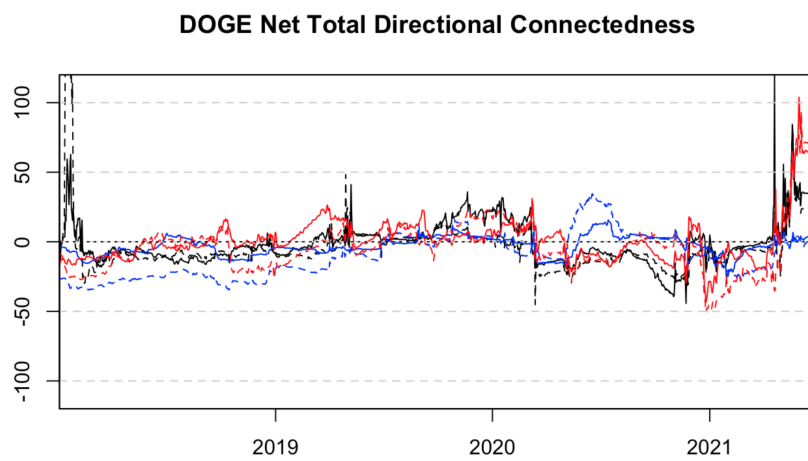


Figure 5.30: DOGE Net total, long-term, and short-term directional connectedness measures at 200 days rolling window



CONCLUSION

6.1 Summary of Results

6.1.1 GameStop Short Squeeze

This research paper investigates the role played by the social media platform Reddit in the events surrounding the GameStop (GME) share price rally in early 2021. Specifically, it analyzes the impact of discussions on the r/WallStreetBets subreddit on the intraday price dynamics of the video game retailer GameStop's stock.

To accurately capture sentiment expressed on Reddit, which often involves unique slang, emojis, and offensive language, a customized lexicon for sentiment analysis is designed in this paper. This "Reddit lexicon" is built upon an existing dictionary (VADER) by adding 130 Reddit-specific terms like "diamond hands" and updating sentiment scores through a manual annotation process. Creating this lexicon tailored to the Reddit community allowed for more precise extraction of

sentiment signals.

Applying the lexicon to a dataset of 10.8 million comments from r/WallStreetBets during January-February 2021, compelling evidence was found that sentiment extracted from this forum impacted GME's intraday price movements. Specifically, results showed the "net sentiment" measure (positive minus negative sentiment) had a positive, statistically significant effect on GME's 5-minute, 10-minute, and 30-minute open-to-open returns.

Crucially, the research discovered that the strength of these sentiment-price relationships was highly time-varying. The impact was much more pronounced during periods when GME shares were rallying compared to periods of price declines. This suggests the r/WallStreetBets crowd was effective in amplifying bullish sentiment to drive the share price higher, but had limited ability to stop selloffs once bearish momentum took over.

While confirming the role of the Reddit forum in initiating the GameStop short squeeze, the findings demonstrate this impact was temporary. The online discussion appeared to aggravate price dynamics temporarily, but fundamental drivers ultimately overwhelmed sentiment effects during the downswing. The results shed light on the interconnected nature of social media and financial markets in the modern investing landscape.

6.1.2 Crypto Sentiment Lexicon

This research explores advanced machine-learning techniques and develops a custom cryptocurrency sentiment lexicon. A key focus is creating effective sentiment analysis tools to deeply analyze content and viewpoints expressed in investor discussions related to the cryptocurrency domain.

Several machine learning models are developed to accurately predict sentiment labels (positive, neutral, negative) in cryptocurrency social media text data. These models adopt a feature-based methodology, extracting key textual features and training algorithms such as logistic regression, random forest, and XGBoost on manually labeled datasets. Logistic regression emerged as the most effective model, outperforming others across configurations involving training data annotated by ChatGPT and human labelers. Notably, ChatGPT annotations provided comparable model performance to manual annotations, highlighting ChatGPT's reliability for scaled training data creation.

2,000 previously unseen messages manually annotated are used to compare the optimal model with other sentiment analysis tools. The model exhibited an impressive 88.7% accuracy rate, significantly surpassing established sentiment analysis tools such as Vader (42.85% accurate) and TextBlob (43.15% accurate). This substantial accuracy improvement sets a new high standard for robust sentiment extraction from text data.

A unique cryptocurrency sentiment lexicon was also developed, containing over 10,000 terms and variants specific to the cryptocurrency market context. Reflecting Loughran and McDonald (2011) approach, this comprehensive lexicon enables

sufficient capturing of emotional language and terminology in cryptocurrency discussions across various grammatical forms. The specialized lexicon plays a vital role in accurately measuring sentiments expressed through the language and unique vocabulary of the cryptocurrency domain.

In summary, this investigation yields significant contributions by developing highly accurate machine-learning models tailored for cryptocurrency sentiment prediction and an extensive domain-specific lexicon. These advanced tools facilitate an in-depth analysis of how market sentiment, influenced by social media discourse, interacts with the volatile dynamics of the cryptocurrency market.

6.1.3 Sentiment Analysis and Cryptocurrency Market Dynamics

This research explores the complex relationship between cryptocurrency market trends and investor sentiment, utilizing advanced econometric techniques like time-variant Granger causality and asymmetric time-varying parameter vector autoregression (TVP-VAR) frequency connectivity. The study relies on a comprehensive dataset spanning January 1, 2018, to June 30, 2021, incorporating web-scraped social media data from platforms like Reddit and Twitter, as well as high-resolution trading data for major cryptocurrencies such as Bitcoin, Ethereum, Ripple, and Dogecoin.

Through the application of Granger causality tests, the dynamic relationship between market sentiment and cryptocurrency volatility is uncovered. The results indicate that past sentiment values can significantly predict current market volatil-

ity for certain cryptocurrencies, with p-values falling below the 0.01 critical value. However, this predictive power fluctuates, as evidenced by cases where p-values approach 1, suggesting that past sentiment becomes a weak predictor of current volatility. These volatile p-value patterns can be attributed to external shocks, regulatory changes, and the inherent volatility cycles within the cryptocurrency market.

To capture the evolving nature of this relationship, a time-varying parameter vector autoregressive (TVP-VAR) model is employed. The resulting average connectedness tables reveal moderate connectivity among the network variables, with a clear asymmetry positive shocks exhibit stronger connectedness than negative shocks. Regardless of the shock nature, sentiment, Ethereum, and Dogecoin emerge as net receivers, while Ripple and Bitcoin act as net transmitters of both positive and negative shocks.

The dynamic connectedness analysis further highlights the sensitivity of total connectedness to external shocks and specific events, with notable peaks observed during periods of market turbulence. The study also unveils the dominance of long-term negative connectedness until early 2020, followed by a brief period where positive connectedness surpassed negative values during the COVID-19 outbreak. By late 2020 and early 2021, positive correlations regained dominance, reflecting investors' risk-averse tendencies.

Furthermore, the research explores the influential role of "whales" entities or individuals holding large quantities of cryptocurrency. Correlations between whale activities and fluctuations in market connectedness suggest their considerable

influence on overall market sentiment and dynamics.

In conclusion, this study provides a comprehensive examination of the dynamic interplay between market sentiment, external factors, and cryptocurrency volatility, utilizing advanced econometric techniques and a unique dataset. The findings offer valuable insights for investors, researchers, and stakeholders in the rapidly evolving cryptocurrency landscape.

6.2 Future Research and Outlook

The comprehensive findings from a series of studies on the impact of social media sentiment on financial markets offer a promising foundation for future research endeavors. These studies have illuminated the complex dynamics between online sentiment and market behavior, presenting several opportunities for further exploration and innovation in both academic and practical realms.

A critical area for future research lies in the expansion and refinement of sentiment analysis tools. The development of the Reddit-specific lexicon and the cryptocurrency sentiment lexicon in these studies has opened the door for more nuanced sentiment analysis in these unique domains. However, substantial scope for enhancing these tools to capture even finer shades of sentiment and linguistic expression is seen. Furthermore, the creation of similar lexicons for other social media platforms could significantly broaden the understanding of the role of online communities in shaping financial market trends. Such advancements would not only contribute to academic knowledge but also provide practical tools for market analysts and investors.

The successful application of ChatGPT for dataset labeling in research suggests a rich potential for leveraging AI and machine learning technologies in financial market analysis. The development of more sophisticated AI-driven models for sentiment analysis could be explored in future research. An exciting opportunity to investigate the potential of AI and machine learning in forecasting market trends is also present, which could revolutionize the way market predictions are made and implemented.

Important policy and regulatory implications are also highlighted by the research, particularly in the context of the growing influence of social media and decentralized finance on financial markets. Future studies should delve into the ramifications of these findings for financial regulation, focusing on aspects of market stability and investor protection. The understanding of the regulatory challenges posed by the increasing decentralization and digitalization of financial markets and the development of frameworks that can adapt to these evolving environments are included.

Comparative analyses between traditional and emerging financial markets, such as cryptocurrencies, could yield valuable insights. Focusing on contrasting market behaviors and investor sentiments across different asset classes could be the focus of future research, uncovering similarities and differences that could inform investment strategies and regulatory policies.

Given the global nature of financial markets, the adoption of a global perspective in future research is imperative. This includes studying the impact of cross-country

sentiments and international events on market dynamics. In an increasingly interconnected financial world, understanding how sentiments in one market can influence others is crucial for global financial stability and cooperation.

In conclusion, the research presented in these papers significantly propels the understanding of the complex relationship between social media sentiment and financial market dynamics. The future outlook for this field of study is filled with promising avenues for further exploration and discovery, offering exciting opportunities for academic researchers and market practitioners alike in this rapidly evolving field.

BIBLIOGRAPHY

Abad, D. and Pascual, R. (2010).

Switching to a temporary call auction in times of high uncertainty.

Journal of Financial Research, 33(1):45–75.

Abolafia, M. Y. (2001).

Making markets: Opportunism and restraint on Wall Street.

Harvard University Press.

Aghion, P., Van Reenen, J., and Zingales, L. (2013).

Innovation and institutional ownership.

American economic review, 103(1):277–304.

Ahmad, K., Han, J., Hutson, E., Kearney, C., and Liu, S. (2016).

Media-expressed negative tone and firm-level stock returns.

Journal of Corporate Finance, 37:152–172.

Akyildirim, E., Corbet, S., Cumming, D., Lucey, B., and Sensoy, A. (2020).

Riding the wave of crypto-exuberance: The potential misuse of corporate blockchain announcements.

Technological Forecasting and Social Change, 159:120191.

Al Guindy, M. (2021).

Corporate twitter use and cost of equity capital.

Journal of Corporate Finance, (forthcoming).

An, Z., Chen, C., Naiker, V., and Wang, J. (2020).

Does media coverage deter firms from withholding bad news? evidence from stock price crash risk.

Journal of Corporate Finance, 64.

Antweiler, W. and Frank, M. (2004).

Is all that talk just noise? the information content of internet stock message boards.

Journal of Finance, 59(3):1259–1294.

Bahdanau, D., Cho, K., and Bengio, Y. (2014).

Neural machine translation by jointly learning to align and translate.

arXiv preprint arXiv:1409.0473.

Bajo, E. and Raimondo, C. (2017).

Media sentiment and ipo underpricing.

Journal of Corporate Finance, 46:139–153.

Baker, H. K., Filbeck, G., and Ricciardi, V. (2017).

Financial Behavior: Players, Services, Products, and Markets.

Oxford University Press.

Google-Books-ID: OVOeDgAAQBAJ.

Baker, M. and Wurgler, J. (2006).

Investor sentiment and the cross-section of stock returns.

Journal of Finance, 61(4):1645–1680.

Bartov, E., Faurel, L., and Mohanram, P. S. (2018).

Can twitter help predict firm-level earnings and stock returns?

The Accounting Review, 93(3):25–57.

Baruník, J., Kočenda, E., and Vácha, L. (2016).

Asymmetric connectedness on the us stock market: Bad and good volatility spillovers.

Journal of Financial Markets, 27:55–78.

Baruník, J. and Křehlík, T. (2018).

Measuring the frequency dynamics of financial connectedness and systemic risk.

Journal of Financial Econometrics, 16(2):271–296.

Baumol, W. J. and Benhabib, J. (1989).

Chaos: significance, mechanism, and economic applications.

Journal of economic perspectives, 3(1):77–105.

Behrendt, S. and Schmidt, A. (2018).

The twitter myth revisited: Intraday investor sentiment, twitter activity and individual-level stock return volatility.

Journal of Banking and Finance, 96:355–367.

Berle, A. A. and Means, G. G. C. (1991).

The modern corporation and private property.

Transaction publishers.

Bikhchandani, S. and Sharma, S. (2001).

Herd behavior in financial markets.

IMF Staff Papers, 47(3).

Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003).

Latent dirichlet allocation.

Journal of machine Learning research, 3(Jan):993–1022.

Bollen, J., Mao, H., and Zeng, X. (2011).

Twitter mood predicts the stock market.

Journal of computational science, 2(1):1–8.

Bommasani, R., Hudson, D. A., Adeli, E., Altman, R., Arora, S., von Arx, S., Bernstein, M. S., Bohg, J., Bosselut, A., Brunskill, E., et al. (2021).

On the opportunities and risks of foundation models.

arXiv preprint arXiv:2108.07258.

Borri, N. (2019).

Conditional tail-risk in cryptocurrency markets.

Journal of Empirical Finance, 50:1–19.

Bouri, E., Jalkh, N., Molnár, P., and Roubaud, D. (2017).

Bitcoin for energy commodities before and after the december 2013 crash: diversifier, hedge or safe haven?

Applied Economics, 49(50):5063–5073.

Brown, G. W. and Cliff, M. T. (2004).

Investor sentiment and the near-term stock market.

Journal of Empirical Finance, 11(1):1–27.

Campbell, J. Y., Lo, A. W., MacKinlay, A. C., and Whitelaw, R. F. (1998).

The econometrics of financial markets.

Macroeconomic Dynamics, 2(4):559–562.

Caporale, G. M., Gil-Alana, L., and Plastun, A. (2018).

Persistence in the cryptocurrency market.

Research in International Business and Finance, 46:141–148.

Carvalho, J. and Plastino, A. (2021).

- On the evaluation and combination of state-of-the-art features in twitter sentiment analysis.
Artificial Intelligence Review, 54:1887–1936.
- Chahine, S., Mansi, S., and Mazboudi, M. (2015).
Media news and earnings management prior to equity offerings.
Journal of Corporate Finance, 35:177–195.
- Chakrabarty, B., Corwin, S. A., and Panayides, M. A. (2011).
When a halt is not a halt: An analysis of off-nyse trading during nyse market closures.
Journal of Financial Intermediation, 20(3):361–386.
- Chan, L. K., Lakonishok, J., and Sougiannis, T. (2001).
The stock market valuation of research and development expenditures.
The Journal of finance, 56(6):2431–2456.
- Chatziantoniou, I., Gabauer, D., and Gupta, R. (2021).
Integration and risk transmission in the market for crude oil: A time-varying parameter frequency connectedness approach.
Technical report, University of Pretoria, Department of Economics.
- Cheah, J. E.-T., Luo, D., Zhang, Z., and Sung, M.-C. (2022).
Predictability of bitcoin returns.
The European Journal of Finance, 28(1):66–85.
- Chen, C. Y.-H., Després, R., Guo, L., and Renault, T. (2019).
What makes cryptocurrencies special? investor sentiment and return predictability during the bubble.
Technical report, IRTG 1792 Discussion Paper.

Chen, H., De, P., Hu, Y. J., and Hwang, B.-H. (2013).

Customers as advisors: The role of social media in financial markets.

Technical report, Working paper.

Chen, H., De, P., Hu, Y. J., and Hwang, B.-H. (2014).

Wisdom of crowds: The value of stock opinions transmitted through social media.

The Review of Financial Studies, 27(5):1367–1403.

Christie, W. G., Corwin, S. A., and Harris, J. H. (2002).

Nasdaq trading halts: The impact of market mechanisms on prices, trading activity, and execution costs.

The Journal of Finance, 57(3):1443–1478.

Collins, H. (2018).

Creative research: the theory and practice of research for the creative industries.

Bloomsbury Publishing.

Conway, B. (2012).

The next hedge fund trade: Social media.

Barron's, April, 17.

Cookson, J. and Niessner, M. (2019).

Why don't we agree? evidence from a social network of investor.

Journal of Finance, 75(1):173–228.

Cookson, J. A. and Niessner, M. (2020).

Why don't we agree? evidence from a social network of investors.

The Journal of Finance, 75(1):173–228.

Corbet, S., Hou, G., Hu, Y., and Oxley, L. (2021).

We reddit in a forum: The influence of messaging boards on firm stability.

Available at SSRN 3776445.

Corbet, S., Larkin, C., Lucey, B., Meegan, A., and Yarovaya, L. (2020).

The impact of macroeconomic news on bitcoin returns.

The European Journal of Finance, 26:1396–1416.

Corbet, S., Lucey, B., Peat, M., and Vigne, S. (2018a).

Bitcoin futures what use are they?

Economics Letters, 172:23–27.

Corbet, S., Meegan, A., Larkin, C., Lucey, B., and Yarovaya, L. (2018b).

Exploring the dynamic relationships between cryptocurrencies and other financial assets.

Economics Letters, 165:28–34.

Corwin, S. A. and Lipson, M. L. (2000).

Order flow and liquidity around nyse trading halts.

The Journal of Finance, 55(4):1771–1801.

Da, Z., Engelberg, J., and Gao, P. (2015).

The sum of all fears investor sentiment and asset prices.

The Review of Financial Studies, 28(1):1–32.

Danbolt, J., Siganos, A., and Vagenas-Nanos, E. (2015).

Investor sentiment and bidder announcement abnormal returns.

Journal of Corporate Finance, 33:164–179.

Daniel, K., Hirshleifer, D., and Subrahmanyam, A. (1998).

Investor psychology and security market under-and overreactions.

the Journal of Finance, 53(6):1839–1885.

- Daniel, K. D., Hirshleifer, D., and Subrahmanyam, A. (2001).
Overconfidence, arbitrage, and equilibrium asset pricing.
The Journal of Finance, 56(3):921–965.
- Das, S. R. and Chen, M. (2007).
Yahoo! for amazon: sentiment extraction from small talk on the web.
Management Science, 53(9):1375–1388.
- Diebold, F. X. and Yilmaz, K. (2012).
Better to give than to receive: Predictive directional measurement of volatility spillovers.
International Journal of Forecasting, 28(1):57–66.
- Diebold, F. X. and Yilmaz, K. (2014).
On the network topology of variance decompositions: Measuring the connectedness of financial firms.
Journal of Econometrics, 182(1):119–134.
- Doran, J. S., Peterson, D. R., and Price, S. M. (2012).
Earnings conference call content and stock price: The case of reits.
The Journal of Real Estate Finance and Economics, 45:402–434.
- Ebadi, A., Xi, P., Tremblay, S., Spencer, B., Pall, R., and Wong, A. (2021).
Understanding the temporal evolution of covid-19 research through machine learning and natural language processing.
Scientometrics, 126:725–739.
- Edmondson, A. C. and McManus, S. E. (2007).
Methodological fit in management field research.
Academy of management review, 32(4):1246–1264.

- Elliott, G., Rothenberg, T. J., and Stock, J. H. (1996).
Efficient tests for an autoregressive unit root.
Econometrica, 64(4):813–836.
- Engelberg, J. (2008).
Costly information processing: Evidence from earnings announcements.
In *AFA 2009 San Francisco meetings paper*.
- Engelberg, J., McLean, R. D., and Pontiff, J. (2020).
Analysts and anomalies.
Journal of Accounting and Economics, 69(1):101249.
- Eve, R. A., Horsfall, S., and Lee, M. E. (1997).
Chaos, complexity, and sociology: Myths, models, and theories.
Sage.
- Fama, E. F. (1970).
Efficient capital markets: A review of theory and empirical work.
The journal of Finance, 25(2):383–417.
- Fang, F., Chung, W., Ventre, C., Basios, M., Kanthan, L., Li, L., and Wu, F. (2021).
Ascertaining price formation in cryptocurrency markets with machine learning.
The European Journal of Finance, pages 1–23.
- Feldman, R., Govindaraj, S., Livnat, J., and Segal, B. (2008).
The incremental information content of tone change in management discussion
and analysis.
- Feng, X. and Johansson, A. (2019).
Top executives on social media and information in the capital market: Evidence
from china.

Journal of Corporate Finance, 58:824–857.

Fischhoff, B., Slovic, P., and Lichtenstein, S. (1977).

Knowing with certainty: The appropriateness of extreme confidence.

Journal of Experimental Psychology: Human perception and performance, 3(4):552.

Frijns, B. and Margaritis, D. (2008).

Forecasting daily volatility with intraday data.

The European Journal of Finance, 14(6):523–540.

Gabauer, D. (2021).

Dynamic measures of asymmetric & pairwise connectedness within an optimal currency area: Evidence from the euro system.

Journal of Multinational Financial Management, 60:100680.

Gentzkow, M., Shapiro, J. M., and Taddy, M. (2019).

Measuring group differences in high-dimensional choices: method and application to congressional speech.

Econometrica, 87(4):1307–1340.

Gkillas, K., Gupta, R., and Pierdzioch, C. (2021).

Forecasting realized volatility of bitcoin returns: Tail events and asymmetric loss.

The European Journal of Finance, 27(16):1626–1644.

Goldstein, M. A. and Kavajecz, K. A. (2004).

Trading strategies during circuit breakers and extreme market movements.

Journal of Financial Markets, 7(3):301–333.

Granger, C. (1969).

- Investigating causal relations by econometric models and cross-spectral methods.
Econometrica, 37(3):44–438.
- Guegan, D. and Renault, T. (2021).
Does investor sentiment on social media provide robust information for bitcoin returns predictability?
Finance Research Letters, 38:116–123.
- Guggilla, C., Miller, T., and Gurevych, I. (2016).
Cnn-and lstm-based claim classification in online user comments.
In *Proceedings of COLING 2016, the 26th international conference on computational linguistics: technical papers*, pages 2740–2751.
- Gurdgiev, C., O’Loughlin, D., Chlebowsky, B., et al. (2019).
Behavioral basis of cryptocurrencies markets: Examining effects of public sentiment, fear, and uncertainty on price formation.
Journal of Financial Transformation, 49:110–121.
- Hartzell, J. C. and Starks, L. T. (2003).
Institutional investors and executive compensation.
The journal of finance, 58(6):2351–2374.
- Hasler, A., Lusardi, A., and Valdes, O. (2021).
Financial anxiety and stress among us households: New evidence from the national financial capability study and focus groups.
FINRA Investor Education Foundation.
- Hautsch, N. and Horvath, A. (2019).
How effective are trading pauses?
Journal of Financial Economics, 131(2):378–403.

Henry, E. and Leone, A. J. (2016).

Measuring qualitative information in capital markets research: Comparison of alternative methodologies to measure disclosure tone.

The Accounting Review, 91(1):153–178.

Hey, J. D. and Morone, A. (2004).

Do markets drive out lemmings or vice versa?

Economica, 71(284):637659.

Hirschey, M., Richardson, V. J., and Scholz, S. (2000).

Stock-price effects of internet buy-sell recommendations: The motley fool case.

Financial Review, 35(2):147–174.

_eprint: <https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1540-6288.2000.tb01418.x>.

Huang, M., Qian, Q., and Zhu, X. (2017).

Encoding syntactic knowledge in neural networks for sentiment classification.

ACM Trans. Inf. Syst., 35(3).

Huang, X., Han, W., Newton, D., Platanakis, E., Stafylas, D., and Sutcliffe, C. (2022).

The diversification benefits of cryptocurrency asset categories and estimation risk: pre and post Covid-19.

The European Journal of Finance, pages 1–26.

Huang, X., Teoh, S. H., and Zhang, Y. (2014).

Tone management.

The Accounting Review, 89(3):1083–1113.

Hutto, C. and 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.

Jalan, A., Matkovskyy, R., and Urquhart, A. (2021).

What effect did the introduction of bitcoin futures have on the bitcoin spot market?

The European Journal of Finance, 27(13):1251–1281.

Jame, R., Johnston, R., Markov, S., and Wolfe, M. C. (2016).

The value of crowdsourced earnings forecasts.

Journal of Accounting Research, 54(4):1077–1110.

_eprint: <https://onlinelibrary.wiley.com/doi/pdf/10.1111/1475-679X.12121>.

Jegadeesh, N. and Wu, D. (2013).

Word power: A new approach for content analysis.

Journal of financial economics, 110(3):712–729.

Ji, Q., Bouri, E., Lau, C. K. M., and Roubaud, D. (2019).

Dynamic connectedness and integration in cryptocurrency markets.

International Review of Financial Analysis, 63:257–272.

Jiang, G., Lee, C. M., and Zhang, Y. (2005).

Information uncertainty and expected returns.

Review of Accounting Studies, 10:185–221.

Jiao, P., Veiga, A., and Walther, A. (2020).

Social media, news media and the stock market.

Journal of Economic Behavior & Organization, 176:63–90.

Jiménez-Zafra, S. M., Martín-Valdivia, M. T., Molina-González, M. D., and Ureña-López, L. A. (2019).

How do we talk about doctors and drugs? sentiment analysis in forums expressing opinions for medical domain.

Artificial intelligence in medicine, 93:50–57.

Kahneman, D., Slovic, P., and Tversky, A. (1982).

Judgment under uncertainty: Heuristics and biases.

Cambridge university press.

Kahneman, D. and Tversky, A. (1979).

Prospect theory: An analysis of decision under risk.

Econometrica, 47(2):263–291.

Kang, J., Lin, S., and Xiong, X. (2022).

What drives intraday reversal? illiquidity or liquidity oversupply?

Journal of Economic Dynamics and Control, 136:104313.

Karaa, R., Slim, S., Goodell, J. W., Goyal, A., and Kallinterakis, V. (2021).

Do investors feedback trade in the Bitcoin and why?

The European Journal of Finance, pages 1–21.

Koop, G. and Korobilis, D. (2013).

Large time-varying parameter VARs.

Journal of Econometrics, 177(2):185–198.

Koop, G. and Korobilis, D. (2014).

A new index of financial conditions.

European Economic Review, 71:101–116.

Koop, G., Pesaran, M. H., and Potter, S. M. (1996).

Impulse response analysis in nonlinear multivariate models.

Journal of Econometrics, 74(1):119–147.

Kraaijeveld, O. and De Smedt, J. (2020).

The predictive power of public twitter sentiment for forecasting cryptocurrency prices.

Journal of International Financial Markets, Institutions and Money, 65:101188.

Le, Q. and Mikolov, T. (2014).

Distributed representations of sentences and documents.

In *International conference on machine learning*, pages 1188–1196. PMLR.

Le, T.-L., Abakah, E. J. A., and Tiwari, A. K. (2021).

Time and frequency domain connectedness and spill-over among fintech, green bonds and cryptocurrencies in the age of the fourth industrial revolution.

Technological Forecasting and Social Change, 162:120382.

Lee, C. M., Ready, M. J., and Seguin, P. J. (1994).

Volume, volatility, and new york stock exchange trading halts.

The Journal of Finance, 49(1):183–214.

Liang, P., Bommasani, R., Lee, T., Tsipras, D., Soylu, D., Yasunaga, M., Zhang, Y., Narayanan, D., Wu, Y., Kumar, A., et al. (2022).

Holistic evaluation of language models.

arXiv preprint arXiv:2211.09110.

Light, J. (2012).

Stocks: how tweet it is; some funds are using social media to find investments.

The Wall Street Journal Eastern Edition, 0(0):B8(1).

Liu, B. and McConnell, J. J. (2013).

The role of the media in corporate governance: Do the media influence managers' capital allocation decisions?

Journal of Financial Economics, 110(1):1–17.

Liu, L. (2019).

Are bitcoin returns predictable?: Evidence from technical indicators.

Physica A: Statistical Mechanics and its Applications, 533:121950.

Loughran, T. and McDonald, B. (2011).

When is a liability not a liability? textual analysis, dictionaries, and 10-ks.

The Journal of finance, 66(1):35–65.

Loughran, T. and McDonald, B. (2013).

Ipo first-day returns, offer price revisions, volatility, and form s-1 language.

Journal of Financial Economics, 109(2):307–326.

Loughran, T. and McDonald, B. (2016).

Textual analysis in accounting and finance: A survey.

Journal of Accounting Research, 54(4):1187–1230.

Lucas Jr, R. E. (1978).

Asset prices in an exchange economy.

Econometrica: journal of the Econometric Society, pages 1429–1445.

Lundberg, S. M. and Lee, S. (2017).

A unified approach to interpreting model predictions.

CoRR, abs/1705.07874.

Ma, T. and McGroarty, F. (2017).

Social machines: How recent technological advances have aided financialisation.

Journal of Information Technology, 32(3):234–250.

Mensi, W., Al-Yahyaee, K. H., and Kang, S. H. (2019).

- Structural breaks and double long memory of cryptocurrency prices: A comparative analysis from bitcoin and ethereum.
Finance Research Letters, 29:222–230.
- Merton, R. C. (1973).
An intertemporal capital asset pricing model.
Econometrica: Journal of the Econometric Society, pages 867–887.
- Nakamoto, S. (2008).
Bitcoin: A peer-to-peer electronic cash system.
Decentralized Business Review, page 21260.
- Nofsinger, J. R. and Sias, R. W. (1999).
Herding and feedback trading by institutional and individual investors.
The Journal of finance, 54(6):2263–2295.
- Nugroho, B. A. (2021).
Spillovers and bivariate portfolios of gold-backed cryptocurrencies and gold during the covid-19 outbreak.
Journal of Islamic Accounting and Business Research, 12(7):1055–1076.
- Odean, T. (1998).
Are investors reluctant to realize their losses?
The Journal of finance, 53(5):1775–1798.
- Oliveira, N., Cortez, P., and Areal, N. (2016).
Stock market sentiment lexicon acquisition using microblogging data and statistical measures.
Decision Support Systems, 85:62–73.
- Pang, B., Lee, L., et al. (2008).

Opinion mining and sentiment analysis.

Foundations and Trends® in information retrieval, 2(1–2):1–135.

Pano, T. and Kashef, R. (2020).

A complete vader-based sentiment analysis of bitcoin (btc) tweets during the era of covid-19.

Big Data and Cognitive Computing, 4(4).

Pele, D. T., Wesselhöfft, N., Härdle, W. K., Kolossiatis, M., and Yatracos, Y. G. (2021).

Are cryptos becoming alternative assets?

The European Journal of Finance, pages 1–42.

Pesaran, H. H. and Shin, Y. (1998).

Generalized impulse response analysis in linear multivariate models.

Economics Letters, 58(1):17–29.

Peters, E. E. (1994).

Fractal market analysis: applying chaos theory to investment and economics, volume 24.

John Wiley & Sons.

Poria, S., Cambria, E., Hazarika, D., and Vij, P. (2016).

A deeper look into sarcastic tweets using deep convolutional neural networks.

arXiv preprint arXiv:1610.08815.

Renault, T. (2017).

Intraday online investor sentiment and return patterns in the u.s. stock market.

Journal of Banking and Finance, 84:25–40.

Ricciardi, V. and Simon, H. (2001).

Behavioral finance: A new perspective for investors and financial professionals.

Retrieved from.

Richards, D. (1990).

Is strategic decision making chaotic?

Behavioral science, 35(3):219–232.

Shefrin, H. (2002).

Beyond greed and fear: Understanding behavioral finance and the psychology of investing.

Oxford University Press.

Shelar, A. and Huang, C.-Y. (2018).

Sentiment analysis of twitter data.

In *2018 International Conference on Computational Science and Computational Intelligence (CSCI)*, pages 1301–1302.

Shen, D., Urquhart, A., and Wang, P. (2019).

Does twitter predict bitcoin?

Economics letters, 174:118–122.

Shleifer, A. and Summers, L. (1990).

The noise trader approach to finance.

Journal of Economic Perspectives, 4:19–33.

Sivasangari, V., Mohan, A. K., Suthendran, K., and Sethumadhavan, M. (2018).

Isolating rumors using sentiment analysis.

Journal of Cyber Security and Mobility.

Sterman, J. D. (1988).

Deterministic chaos in models of human behavior: Methodological issues and experimental results.

System Dynamics Review, 4(1-2):148–178.

Stiassny, A. (1996).

A spectral decomposition for structural var models.

Empirical Economics, 21(4):535–555.

Stone, P. J., Dunphy, D. C., and Smith, M. S. (1966).

The general inquirer: A computer approach to content analysis.

Surowiecki, J. (2005).

The wisdom of crowds.

Anchor.

Tang, D., Qin, B., and Liu, T. (2015).

Document modeling with gated recurrent neural network for sentiment classification.

In *Proceedings of the 2015 conference on empirical methods in natural language processing*, pages 1422–1432.

Taylor, S. E. and Brown, J. D. (1988).

Illusion and well-being: a social psychological perspective on mental health.

Psychological bulletin, 103(2):193.

Tetlock, P. C. (2007).

Giving content to investor sentiment: The role of media in the stock market.

The Journal of finance, 62(3):1139–1168.

Tetlock, P. C. (2015).

Chapter 18 the role of media in finance.

- In Anderson, S. P., Waldfogel, J., and Strömberg, D., editors, *Handbook of Media Economics*, volume 1 of *Handbook of Media Economics*, pages 701–721. North-Holland.
- Tetlock, P. C., Saar-Tsechansky, M., and Macskassy, S. (2008).
More than words: Quantifying language to measure firms’ fundamentals.
The journal of finance, 63(3):1437–1467.
- Trimborn, S. and Härdle, W. K. (2018).
Crix an index for cryptocurrencies.
Journal of Empirical Finance, 49:107–122.
- Tversky, A. and Kahneman, D. (1992).
Advances in prospect theory: Cumulative representation of uncertainty.
Journal of Risk and uncertainty, 5:297–323.
- Von Neumann, J. and Morgenstern, O. (1947).
Theory of games and economic behavior, 2nd rev.
- Wang, X., Liu, Y., Sun, C.-J., Wang, B., and Wang, X. (2015).
Predicting polarities of tweets by composing word embeddings with long short-term memory.
In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1343–1353.
- Wiebe, J., Bruce, R., and OHara, T. P. (1999).
Development and use of a gold-standard data set for subjectivity classifications.
In *Proceedings of the 37th annual meeting of the Association for Computational Linguistics*, pages 246–253.

- Wiesen, T. F., Beaumont, P. M., Norrbin, S. C., and Srivastava, A. (2018).
Are generalized spillover indices overstating connectedness?
Economics Letters, 173:131–134.
- Yang, K.-C. and Menczer, F. (2023).
Large language models can rate news outlet credibility.
arXiv preprint arXiv:2304.00228.
- Yang, Z., Yang, D., Dyer, C., He, X., Smola, A., and Hovy, E. (2016).
Hierarchical attention networks for document classification.
In *Proceedings of the 2016 conference of the North American chapter of the association for computational linguistics: human language technologies*, pages 1480–1489.
- Yao, S., Kong, X., Sensoy, A., Akyildirim, E., and Cheng, F. (2021).
Investor attention and idiosyncratic risk in cryptocurrency markets.
The European Journal of Finance, pages 1–19.
- Zeng, A., Liu, X., Du, Z., Wang, Z., Lai, H., Ding, M., Yang, Z., Xu, Y., Zheng, W., Xia, X., et al. (2022).
Glm-130b: An open bilingual pre-trained model.
arXiv preprint arXiv:2210.02414.
- Zhao, W. X., Zhou, K., Li, J., Tang, T., Wang, X., Hou, Y., Min, Y., Zhang, B., Zhang, J., Dong, Z., et al. (2023).
A survey of large language models.
arXiv preprint arXiv:2303.18223.