Analysis of Chinese Financial Discourse Based on Topic Clustering and Emotional Evolution

by

Ruoheng Du

An honors thesis submitted in partial fulfillment

of the requirements for the degree of

Bachelor of Science

Business and Economics Honors Program

NYU Shanghai

May 2024

Professor Marti G. Subrahmanyam	Professor Li Guo
Professor Christina Wang	
Professor Ye Jin	
Faculty Advisers	Thesis Adviser

Preface

The intersection of social media text analysis and financial markets analysis has become a emerging field of study, offering valuable insights into investor focus and sentiment. This study delved into this intersection, leveraging advanced techniques to analyze the dynamics of financial discussions online.
By examining a rich dataset spanning from 2017 to 2023, this study explored ten key topics in financial discussions, shedding light on how sentiment trends and topic preferences evolve over time. Through the integration of topic modeling, sentiment analysis, and stock return prediction techniques, this study aimed to provide a comprehensive understanding of the relationship between social media discussions and stock market dynamics.
This study represented a significant endeavor to deepen the understanding of the role of social media in financial decisions, hoping to contribute to the broader discourse on the use of social media texts in financial analysis and provide valuable insights for investors, researchers, and practitioners alike.

Acknowledgements

I would like to express my sincere gratitude to Professor Li Guo, my thesis advisor, for her guidance and support throughout the past year of research. Her insights and advice have been invaluable in shaping this work.
I am also grateful to Professor Christina Wang, Professor Ye Jin, and Professor Marti G. Subrahmanyam, our faculty advisors, for organizing many individual meetings and providing valuable feedback. I would also like to thank the speakers invited by them in the seminars every week, which have enriched my knowledge of finance, economics, and marketing. Special thanks to our teaching assistant Xinyi Yang for her continuous support and assistance.
I would like to acknowledge my fellow classmates in the Honors Programs for sharing this enriching journey with me, and all the professors and friends I have met during my undergraduate studies for their support.
Lastly, I would like to express my deepest gratitude to my parents for their unconditional love and unwavering support throughout my academic and personal life.

Abstract

This study examined the dynamics of financial discussions on the social media platform and their implications for stock market. Analyzing data ranging from 2017 to 2023 from Weibo stock super topic, a popular Chinese social media platform, this study employed topic modeling, sentiment analysis, and stock return prediction techniques to understand how discussions had evolved over time, particularly in the context of the COVID-19 pandemic.

There are several key findings. Firstly, a clear upward trend was observed in the stock super topic discussion frequency, indicating a renewed focus on financial topics as the effects of the pandemic waned and life returned to normal.

Secondly, through BERTopic, ten key topics were found. Topic modeling analysis highlighted the dominance of the Medical Care topic in discussions, with the relative importance of other topics such as Technology topic and Stock Indexes topic evolving over time. Additionally, the post-COVID period witnessed a shift in focus, with declines in discussions related to Medical Care topic and increases in interest in Artificial Intelligence topic.

Thirdly, through FinBERT, social media posts were classified into three categories: positive, neutral, and negative. Sentiment analysis revealed an overall decline in positive sentiment over time, with the Medical Care topic and the Revenues topic standing out for their consistently low and high positive sentiment, respectively. Besides, the fluctuation in sentiment, notably the post-COVID rebound in Medical Care topic's sentiment and the decline in Revenues topic's sentiment, highlighted the dynamics of market sentiment and the changing attitudes of investors.

Lastly, stock return prediction analysis demonstrated the predictive power of sentiment analysis, particularly in forecasting market movements. Through the analysis of Pearson correlation, coefficient p-value, and RMSE reduction, the Revenues topic emerged as a significant predictor of market performance across major stock indexes, including Shanghai Composite Index, Shenzhen Composite Index, and CSI 300 Index.

Overall, this research contributed to the growing literature on social media text analysis and financial markets. It underscored the importance of sentiment analysis in forecasting stock market movements and highlighted the potential of social media texts to inform financial decision-making. Future research could further explore the influence of expanding the dataset to include more diverse topics and utilizing financial models pre-trained on Chinese social media texts to gain more robust and accurate insights into market dynamics and investor sentiment.

Keywords

Topic Modeling; Sentiment Analysis; Stock Return; Weibo Super Topic

Contents

1 Introduction

Analyzing financial discourse has become increasingly important in understanding evolving trends in the stock returns. Traditionally, the random walk theory has dominated financial theory, suggesting that the future returns cannot be predicted, and there are no discernible patterns or trends, while later, according to the Efficient Market Hypothesis, stock returns are largely driven by new information rather than the past performances [1, 2, 3, 4]. This means that when new information becomes available, it can shape investors' perceptions, expectations, and overall sentiment towards a particular stock or the market as a whole. Positive news, such as strong earnings reports or favorable economic indicators, tends to generate optimism and enthusiasm among investors, leading to increased buying activity and rising stock returns. Conversely, negative news, such as poor financial results or geopolitical tensions, can evoke fear and pessimism, prompting investors to sell off their holdings, leading to lower stock returns. These emotional responses to new information can create short-term fluctuations and even longer-term trends in stock returns. As individual investors interpret and react to news based on their emotions, their collective actions contribute to market volatility and can amplify or dampen the impact of new information on stock returns. Moreover, emotions can influence investment decisions beyond the immediate reaction to news. They can impact risk perception, investment horizons, and the willingness to take on or avoid certain investment opportunities. Therefore, individual sentiments play a significant role in investment decisions and stock returns movements, resulting in a growing interest in analyzing individual sentiments to understand their influence on stock returns.

In recent years, the global sentiment has undergone unprecedented changes due to the COVID-19 pandemic. Research focusing on the period within the first six months after the onset of the pandemic has revealed that the outbreak of COVID-19 resulted in a drastic decline in public sentiments [5, 6]. In other words, the outbreak of COVID-19 triggered panic and uncertainty on a global scale, leading to a sharp decline in emotions. Faced with the rapid spread and unpredictability of the pandemic, many people experienced fear, anxiety, and sadness, which were reflected in social media and other communication channels. These significant findings indicate that the pandemic has indeed exerted a profound influence on sentiment, which could potentially have implications for stock returns. As a result, the objective of this study was to

investigate the impact of the pandemic on sentiments and explore its potential effects on stock returns.

This study took Weibo, a popular social media platform in China, as a valuable source of data. Weibo allows users to create and join communities called super topics dedicated to specific interests, including the stock super topic that this study used. By leveraging the textual data from Weibo stock super topic, this study aimed to uncover the key topics of discussion among stock investors, track the trends in sentiment changes, and explore the potential impact of these sentiments on stock returns. Specifically, this study addressed the following research questions:

1. What are the key topics discussed in the Weibo stock super topic, and how frequently are they mentioned?

2. What are the attitudes (sentiments) of people towards these key topics, and how are they changing?

3. Are changes in the sentiments of the key topics related to changes in stock returns?

To achieve these objectives, the study consisted of two main phases. In the first phase, advanced topic modeling technique BERTopic was used to identify key topics in discussion posts. This allowed for the identification and analysis of the popularity and evolution of topics over time. In the second phase, finance-related sentiment analysis tool FinBERT was employed to determine the attitudes related to these discussions. Analyzing the expressed sentiments in the posts helped in understanding how these sentiments influence investor decisions and stock market trends.

This study stood out for its novel approach to understanding the dynamics of China's financial landscape in the wake of significant disruptions produced by the COVID-19 outbreak. Firstly, it utilized a finance-centric sentiment analysis model trained on financial texts, which had a better understanding of financial terminologies and its relationship to emotions, allowing for a more detailed analysis of investor sentiments. Secondly, the study focused on using Weibo, an easily accessible mainstream social media platform in China, as the data source, which ensured a diverse range of opinions and perspectives were captured in the analysis. Lastly, this study examined social media posts from both pre-COVID to post-COVID time periods, providing a comprehensive study of recent public views and allowing for an exploration of how the COVID-19 pandemic impacted investor sentiments. Overall, by analyzing discussions within Weibo super topics, the study offered a unique perspective on investor sentiments and discussion trends,

which could be valuable for investors and others interested in the dynamics of stock market.

This study was structured as follows: Section 2 reviewed related works in the field, highlighting the existing literature on topic modeling and sentiment analysis in financial discourse and the use of social media texts for stock return analysis. Section 3 detailed the methodology, including the analysis of Weibo stock super topic and the use of finance-focused advanced models. Section 4 presented the results of the analysis, including the identified key topics, trends in sentiment changes, and stock return prediction analysis. Finally, section 5 concluded the study by summarizing the findings and discussing their implications for understanding the dynamics of stock market.

2 Literature Review

The literature on stock market analysis has long recognized the importance of understanding the impact of different topics and sentiments on stock returns. This section provided an overview of key studies in three main areas: the impact of key topics on stock returns, the effect of social media sentiment on stock returns, and the impact of the pandemic on sentiment.

2.1 Impact of Key Topics on Stock Returns

Previous researches have highlighted the significance of key topics discussed in corporate press releases and media outlets on stock returns. Feuerriegel, Ratku, and Neumann (2016) conducted a research analyzing the effects of topics found in corporate press releases on stock returns in the German market [7]. They identified 40 topics in ad hoc announcements and found significant variations in their effects on abnormal stock returns. While some topics showed no impact on stock returns, others, such as drug testing, exhibited a considerable effect. However, this research was limited by its focus on the German market, which might not reflect the dynamics of the Chinese market. Additionally, as corporate press releases represented the official statements from companies, they might not capture investor reactions comprehensively. In a similar vein, Biktimirov, Sokolyk, and Ayanso (2021) examined the pandemic-related topics discussed in the printed edition of the Wall Street Journal throughout 2020 [8]. They identified 15 topics and their research revealed that the hype surrounding a topic, rather than the sentiment, was positively correlated with stock returns, with the debt market and financial markets

in particular being strongly positively correlated. However, similar to the previous research, this research was limited by its focus on the U.S. market. Moreover, their research only utilized articles published in 2020, which might not provide a comprehensive view of sentiment changes related to COVID-19 across different periods.

In conclusion, these researches underscore the importance of identifying key topics and their associated sentiment in influencing stock returns. While their findings provide valuable insights, further research is needed in this study to explore the broader implications of key topics on stock market dynamics, particularly in the context of the Chinese stock market.

2.2 Effect of Social Media Sentiment on Stock Returns

Investor sentiments play a crucial role in influencing stock returns. Several researches argued that investor sentiments particularly affected stocks that were difficult to value and costly to arbitrage. Baker and Wurgler (2006) predicted that when sentiment was low, subsequent returns were relatively high for small stocks, young stocks, high volatility stocks, unprofitable stocks, non-dividend-paying stocks, extreme growth stocks, and distressed stocks [9]. Conversely, when sentiment was high, these categories of stocks earned relatively low subsequent returns. Building on this, Zhou (2018) discussed various measures of investor sentiments based on market, survey, and text and media data, highlighting their ability to explain such stock returns [10]. Moreover, Zhou suggested that it was important to focus on aggregating measures over various sources so that investor sentiments could be analyzed thoroughly through technical methods such as statistically modeling.

Subsequent researches delved into the relationship between social media sentiment and stock market performance, highlighting the expanding role of sentiment analysis in stock market research. McGurk, Nowak, and Hall (2020) examined the relationship between investor sentiments and stock returns using textual analysis on social media posts from Twitter [11]. Their research found a positive and significant effect of investor sentiments on abnormal stock returns. The limitations of this research were evident due to its dependence on Twitter data, which was not applicable to the Chinese context, and its narrow emphasis on examining the connection between investor sentiments and stock returns exclusively based on varying company sizes. Such specific details of companies are often confidential, making it challenging to obtain. Consequently, general investor sentiments are less likely to be easily influenced by company-specific

information and more prone to be impacted by general information that can affect the industry as a whole, which was a crucial aspect has been overlooked in this research.

In a similar vein, Bollen, Mao, and Zeng (2011) investigated the correlation between sentiments derived from Twitter posts and the Dow Jones Industrial Average (DJIA) over time [12]. Their study revealed that some specific public sentiment dimensions could significantly improve the accuracy of predicting daily changes in DJIA closing values. However, this research was limited due to its early timeframe as well as the reliance on Twitter data, which restricted its ability to reflect the current state of the Chinese market. Additionally, the research's sentiment analysis approach, utilizing lexicons and rule-based matching to label sentiment, lacked contextual understanding, potentially leading to inaccuracies in comprehending the sentiments expressed in social media posts.

Turning to the Chinese market, Wang, Xiang, Xu, and Yuan (2022) examined the impact of sentiments from an online message forum on stock returns [13]. Using a controlled experiment, they found a significant causal effect of social media sentiments on same-day stock returns, particularly driven by messages with positive sentiments. However, the study's reliance on Eastmoney Guba, while informative, might limit its generalizability due to the platform not being among the most commonly used. Additionally, the use of sentiment classification tools from Baidu and iFLYTEK, while powerful, might not fully grasp the meanings of financial terminologies, potentially impacting the accuracy of sentiment analysis in the financial domain.

Overall, while these researches shed light on the relationship between social media sentiments and stock returns, in this study, it is imperative to understand the impact of recent events, such as the COVID-19 pandemic, on investor sentiments, which provides valuable insights into the evolving stock market and its influence on stock returns.

2.3 Impact of the COVID-19 on Social Media Sentiment

Wang et al. (2022) leveraged over 654 million social media posts covering more than 100 countries to develop a global dataset on sentiment expressions, using the advanced natural language processing technique Bidirectional Encoder Representations from Transformers (BERT) [5]. Their research during the initial COVID-19 outbreak (January 1 to May 31, 2020) demonstrated two compelling applications. They consistently found that the COVID-19 outbreak led to a sharp decline in global sentiments, followed by an asymmetric and slow recovery. They also discov-

ered that lockdown policies had a moderate or negligible impact on sentiments, with significant variations among countries. This research proves that the COVID-19 outbreak had a significant and lasting impact on global sentiment expressions, including in China, which highlighted the importance of understanding such sentiment changes in predicting the effects of major events on stock market trends. However, this research primarily focused on social media posts from the first wave of COVID-19 and did not fully explore the long-term impact of the pandemic on sentiments and stock market trends.

Similarly, Xie et al. (2021) investigated the public response to COVID-19 on Weibo by collecting 719,570 Weibo posts and analyzing them using text mining techniques, including Latent Dirichlet Allocation topic modeling and lexicon-based sentiment analysis [6]. They found that in response to the COVID-19 outbreak, Weibo users engaged in various discussions, including learning about COVID-19, showing support for frontline staffs and workers, and expressing concerns about economic and life restoration. However, their research was limited by its short research period, covering posts only from January 1, 2020 to June 30, 2020, and the use of dictionary-based sentiment analysis, which might not fully capture the context of the posts.

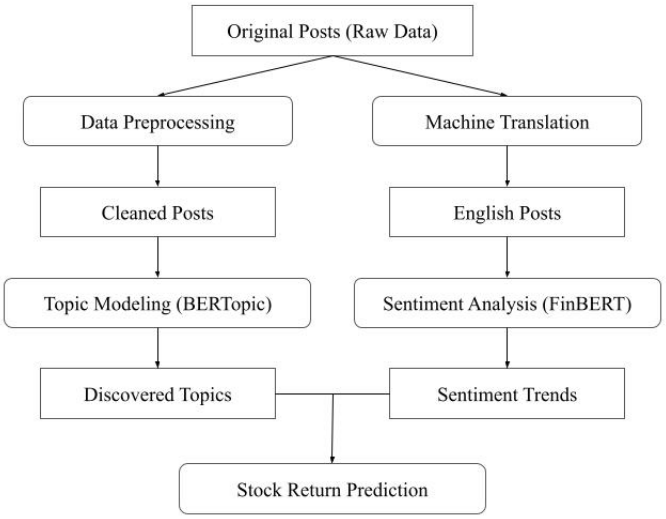
In conclusion, understanding these sentiment changes is crucial for predicting the impact of global events, such as the COVID-19 pandemic, on stock market trends. However, to provide a more comprehensive understanding, future research in this study would consider longer-term sentiment trends and advanced sentiment analysis techniques.

3 Methodology

This section provided a comprehensive methodological framework for the study, detailing the data collection process, preprocessing steps, topic modeling and sentiment analysis techniques employed, as well as the machine translation approach used for translating Chinese text to English.

3.1 Methodological Framework

Figure 1: Methodological Framework.



As shown in Figure 1, the methodology of this study comprised several key steps. First, original posts were obtained using web scraping, serving as the raw data. Next, the original posts underwent data preprocessing to eliminate text noise and stop words, and underwent text segmentation. Additionally, the original posts were translated into English using machine translation. Subsequently, the BERTopic model was applied to the cleaned Chinese data for topic modeling, while the FinBERT model was used for sentiment analysis on the English posts to reveal sentiment trends. Finally, based on the different sentiment trends identified for various topics, stock return predictions were conducted to assess the impact of sentiment trends of specific topics on stock returns.

3.2 Data Collection

Weibo is a social media platform where people in mainland China share information and interact with each other. Posts on Weibo are recorded in the form of short texts, with a length not exceeding 140 characters. The super topics on Weibo refers to the topic hubs where Weibo users can gather and discuss specific topics of interest. In the context of this study, stock super topic was used. It is the place where investors discuss and share information about stock market, such as investment strategies and other relevant company news.

Although integrated tools like Octoparse are available for automatic crawling of web data, they can only support a limited quantity of posts displayed in the software, and fail to collect data for over 3 years, including the periods before, during, and after the COVID-19. To address this limitation, this study developed a web crawler based on Python Selenium to collect data. The web crawler simulated users logins and searched for all posts in the stock super topic. The Python requests library was used to send HTTP requests and parsed the returned JSON data to obtain Weibo content. Additionally, the BeautifulSoup library was used to parse the content of the posts, extracting the published Weibo text content and other relevant information such as timestamps.

The figure below represents a sample post from the dataset.

Figure 2: A Sample Post.

date	content
2023-01-05 14:49:34	白酒+新能源赛道集体大涨，不是我的菜！今天亏钱已成定局，节后账户三连阳失败！股票

3.3 Data Preprocessing

In the data preprocessing stage, three main steps were employed. First, the Python regular expression package was used to remove noise from the dataset. This step involved eliminating irrelevant information commonly found in Weibo posts, such as "forward images," as well as characters indicating the associated stock super topic that not relevant to investors' actual posts. Additionally, emojis, punctuation, and numbers were removed.

Second, a stop list was applied. Chinese contains many meaningless words, such as discourse markers, that are frequently used but lack semantic meaning, and these words are typically removed in the analysis. This study utilized a common Chinese stop word list developed by Harbin Institute of Technology.

Finally, segmentation was performed. Unlike English, where words are separated by spaces, Chinese text does not have spaces between words. Therefore, segmentation is a crucial step in Chinese text mining. This study used the popular Python Jieba package for segmentation.

After these preprocessing steps, the dataset was reduced to 191,460 unique posts. The segmented sample post is shown below.

Figure 3: A Segmented Sample Post.

date	content	segmented_content
2023-01-05 14:49:34	白酒+新能源赛道集体大涨，不是我的菜！今天亏钱已成定局，节后账户三连阳失败！股票	'白酒' '新能源', '赛道', '集体', '大涨', '菜', '亏钱', '已成定局', '节后', '账户', '连阳', '失败'

3.4 Topic Modeling

In this study, the BERTopic model was employed for topic modeling, which derives from the Bidirectional Encoder Representations from Transformers (BERT) model and is a useful tool for discovering latent topics in collections of documents. BERTopic extends traditional topic modeling approaches by extracting coherent topic representations through a class-based variation of TF-IDF [14]. Specifically, BERTopic generates document embedding using the pre-trained transformer-based language model, clusters the embedding, and generates topic representations using the class-based TF-IDF procedure. Prior research in the financial domain compared the application of three topic modeling methods—Latent Dirichlet allocation, Top2Vec, and BERTopic—and found that BERTopic performed the best [15]. Therefore, BERTopic was selected for this study to identify topics in the posts related to stock market discussions.

3.5 Sentiment Analysis

In this study, sentiment analysis was conducted using the FinBERT model, specifically designed for analyzing sentiment in financial texts. Financial sentiment analysis presents challenges due to the specialized financial terminologies and limited labeled data available for training in the financial domain. As a result, other general models, such as those trained on commercial product reviews, are often ineffective in this context. However, FinBERT, based on the Bidirectional Encoder Representations from Transformers (BERT) model, addresses these challenges by leveraging the pre-trained language model and fine-tuning on financial domain-specific corpora. It had demonstrated superior performance compared to other machine learning methods in financial sentiment analysis [16].

3.6 Machine Translation

In this study, machine translation was employed to translate Chinese posts into English. This step is necessary because the sentiment analysis model for financial texts in Chinese is not as mature as its English counterpart, due to the availability of more labeled text for training in English. Therefore, to leverage an advanced English sentiment analysis model, the posts need to be translated into English. For this purpose, this study used Youdao Translation, and a sample translated post is provided below.

Figure 4: A Translated Sample Post.

date	content	segmented_content	translated_content
2023-01-05 14:49:34	白酒+新能源赛道集体大涨，不是我的菜！今天亏钱已成定局，节后账户三连阳失败！股票	'白酒', '新能源', '赛道', '集体', '大涨', '菜', '亏钱', '已成定局', '节后', '账户', '连阳', '失败'	Liquor new energy track collective rise is not my food today's loss of money is a foregone conclusion after the festival account three consecutive positive failed stocks

4 Results

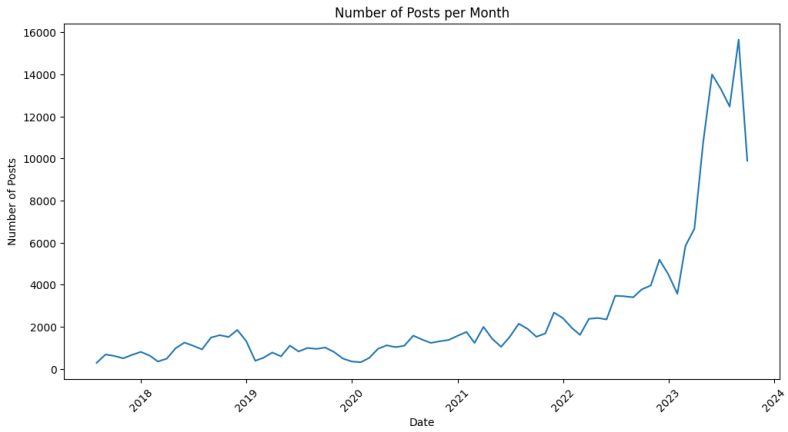
4.1 Data Overview

In this study, the data was divided into three stages based on the COVID-19 timeline in mainland China. The pre-COVID stage spanned from July 2017 (the earliest post month in the dataset) to December 2019. This is because the earliest official announcement from mainland China was a document issued by the Wuhan Municipal Health Commission on December 30, 2019 [17]. As a result, this period is significant as it represents the time before the majority of people became aware of the COVID-19 outbreak and began discussing its potential impacts. The during-COVID stage covered the period from January 2020 to December 2022, coinciding with the active management and response to the pandemic. The post-COVID stage covered from January 2023 to September 2023, which is the latest post month in the dataset. This is because that on December 27, 2022, China announced its decision to lift its epidemic control measures, and on January 8, 2023, China officially lifted immigration restrictions, marking a significant transition point for people to gradually return to pre-pandemic life [18]. This chronological division allows for a detailed analysis of how topics and sentiments evolved across different phases

of the COVID-19 pandemic in China.

Figure 5 illustrates the monthly distribution of posts. Overall, there is no apparent seasonal variation in the data. A clear upward trend can be observed, with a significant increase in posts in 2023. This could suggest that as the pandemic ends and life is gradually returning to normal, people may now be paying more attention to and engaging in a wider range of financial topics and market activities, leading to an increase in discussion frequency. However, due to the lack of publicly available data on the number of followers for specific super topics, it is challenging to determine the exact reason for this sharp increase.

Figure 5: Number of Posts per Month.



4.2 Topic Modeling Results

The topic modeling process involved several steps to uncover latent themes within a collection of documents. Firstly, embedding was extracted from the posts using the pre-trained language model. The embedding was then dimensionality-reduced to facilitate clustering. Next, the clustering algorithm was applied to group similar posts together, forming distinct topics. Tokenization of these topics allowed for the extraction of representative words. This comprehensive approach, implemented through the BERTopic framework, enabled the identification and exploration of meaningful topics within the posts.

4.2.1 Identified Topics

Table 1: Identified Topics.

Topic	Label	Count
0	General Market	117630
1	Medical Care	2071
2	Artificial Intelligence	1568
3	Technology	1439
4	Revenues	1085
5	Stock Indexes	945
6	Speculators	676
7	Past Performance	613
8	Growth Stock	540
9	Future Performance	510
10	Renewable Energy	488

Table 1 displays the identified topics by BERTopic, their labels, and the count of posts associated with each topic. The topics encompass a variety of subjects, including General Market, Medical Care, Artificial Intelligence, Technology, Revenues, Stock Indexes, Speculators, Past Performance, Growth Stock, Future Performance, and Renewable Energy. Among these topics, General Market topic encompasses the general discussions about the Chinese stock market and constitutes approximately 60% of the total dataset, representing the overall theme of the posts. There are ten topics representing more specific themes, each capturing a distinct aspect of the discussions within the dataset: 1. Medical Care topic represents the posts related to healthcare and medicine discussions; 2. Artificial Intelligence topic includes discussions about artificial intelligence, large models, and computing power; 3. Technology topic encompasses discussions related to various technologies, particularly storage service technology; 4. Revenues topic refers to discussions about net profits, year-on-year performance, and dividends of the companies; 5. Stock Indexes topic symbolizes discussions related to various stock indexes in China; 6. Speculators topic is about speculators and their speculative activities in financial markets; 7. Past Performance topic involves discussions about past stock returns, and past buying and selling activities; 8. Growth Stock topic represents discussions about growth stocks and their potentials; 9. Future Performance topic includes discussions about future stock returns and preparatory actions; 10. Renewable Energy topic includes discussions about photovoltaics, electricity, lithium batteries, wind power, and other related contents.

4.2.2 Topic Relative Frequency Over Time

Figure 6: Relative Frequency of Topics (Excluding General Market) Over Time.

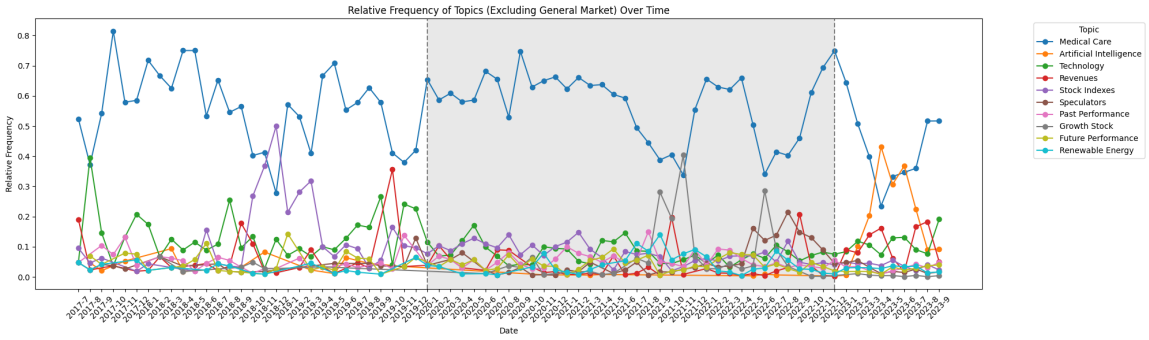
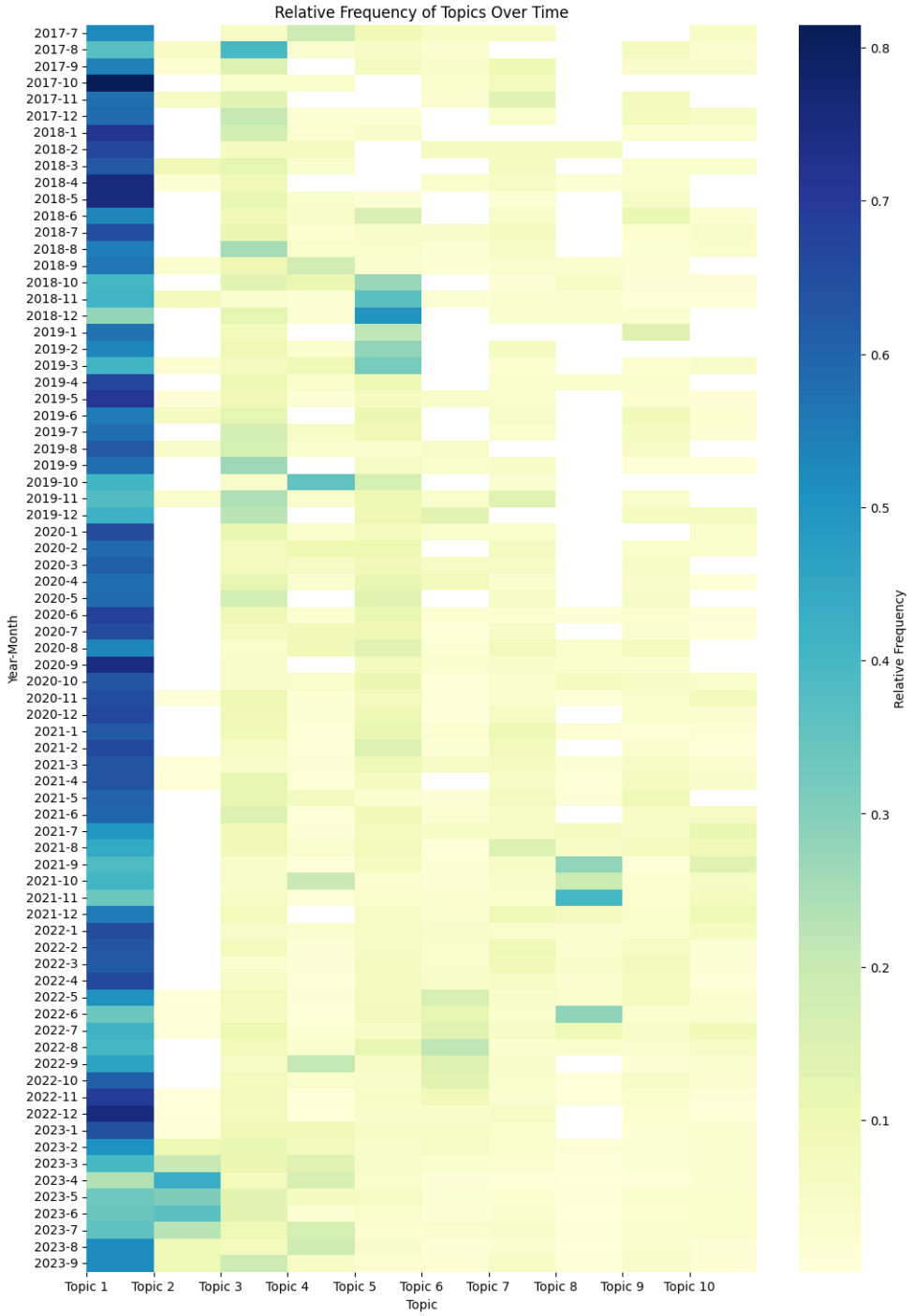


Figure 6 illustrates the monthly relative frequency of posts for the ten specific topic (excluding General Market topic). The y-axis represents the relative frequency, calculated as the percentage of posts for a specific topic relative to the total number of posts for all ten topics in a given month. The x-axis denotes the timeline from July 2017 to September 2023, with the gray shaded area indicating the period during the COVID-19 pandemic (from January 2020 to December 2022).

It can be found that the deep blue line representing the Medical Care topic consistently remains higher than the other lines, indicating sustained high discussion levels regarding related aspects. In contrast, discussions related to the other nine topics exhibit similar and relatively low frequencies, with only a few minor peaks, indicating that they are less prominent topics of discussion compared to the Medical Care topic.

Figure 7 below is the heatmap of the relative frequency of the ten specific topics over time, spanning from July 2017 to September 2023. It reveals a dynamic pattern of topic prevalence over time. Each row represents a specific month, while each column corresponds to one of the ten topics. The color intensity, ranging from dark blue to white, indicates the relative frequency of each topic in a particular month, with darker shades indicating higher frequency and lighter shades indicating lower frequency. This visualization allows for a clear understanding of how the discussions around these topics have evolved over the years, highlighting periods of increased or decreased interest in specific aspects.

Figure 7: Heatmap of Relative Frequency of Topics (Excluding General Market) Over Time.



The dark intensity of the blocks in the first column shows that the Medical Care topic (Topic 1 in the figure) consistently maintains a high level of discussion, reflecting its continued significance. For other topics, the intensity of their blocks varies, reflecting fluctuations in their discussion frequency.

During the latter half of 2017, the Technology topic (Topic 3 in the figure) dominates, sur-

passing even the Medical Care topic in August 2017. Later, From October 2018 to March 2019, the Stock Indexes topic (Topic 5 in the figure) takes the lead, overtaking Medical Care topic in December 2018 as the most discussed topic among investors. Then, in October 2019, the Revenues topic (Topic 4 in the figure) sparks heated discussions and become the second most popular topic at the time. Following that, starting from March 2020, although neither topic has particularly high frequencies, the Technology topic and Stock Indexes topic (Topic 5 in the figure) alternates as the second most discussed topics. Afterwards, from September to November 2021, the Growth Stock topic (Topic 8 in the figure) is highly discussed, surpassing Medical Care topic in November 2021 to become the top topic. Later on, in the second half of 2022, Growth Stock topic, Speculators topic (Topic 6 in the figure), and Revenues topic take turns to become the second most discussed topics. Eventually, from March to July 2023, the Artificial Intelligence topic (Topic 2 in the figure) surges in discussion frequency, becoming the second most discussed topic and surpassing the Medical Care topic twice in April 2023 and June 2023 to become the topic of highest discussion.

This indicates a dynamic shift in popular topics over time and changes in investor focuses. Throughout different periods, the relative importance of various topics evolves, with Technology topic and Stock Indexes topic repeatedly emerging as the second most discussed topics, each surpassing the dominance of the Medical Care topic at least once. This provides insights into investors' interest and priorities amid evolving conditions, highlighting the adaptability of investor attention to changing circumstances.

4.2.3 Topic Relative Frequency in Different Stages

Table 2: Relative Frequency of Topics (Excluding General Market) in Three Stages.

Label	Pre-COVID	During-COVID	Post-COVID	Average
Medical Care	0.56	0.57	0.43	0.52
Artificial Intelligence	0.04	0.01	0.20	0.08
Technology	0.13	0.08	0.11	0.11
Revenues	0.06	0.04	0.11	0.07
Stock Indexes	0.13	0.08	0.04	0.08
Speculators	0.04	0.05	0.03	0.04
Past Performance	0.05	0.06	0.03	0.05
Growth Stock	0.04	0.07	0.00	0.04
Future Performance	0.05	0.04	0.02	0.04
Renewable Energy	0.03	0.04	0.03	0.03

Table 2 compares the relative frequency of each topic in three stages: Pre-COVID, During-COVID, and Post-COVID. It also includes the average relative frequency for each stage.

As the average relative frequency of each topic has shown, Medical Care topic has the highest frequency of 0.52, followed by Technology topic of 0.11, with Artificial Intelligence topic and Stock Indexes topic tied for third place. However, except for Medical Care topic and Artificial Intelligence topic, the changes in frequency for other topics are relatively small, with the frequency differences between the three stages not exceeding 0.1.

Further analysis of the Medical Care topic reveals minimal changes between the pre-COVID and during-COVID stages, but a significant decrease post-COVID, indicating a substantial decrease in investor interest related to healthcare discussions after the pandemic. Conversely, the Artificial Intelligence topic shows minor differences between pre-COVID and during-COVID stages, but a significant increase post-COVID, suggesting a substantial increase in investor interest related to artificial intelligence discussions after the pandemic. This may reflect a shift in focus after the end of the pandemic. During the pandemic, healthcare is one of the key areas of concern due to the significant impact of the pandemic on individuals' health and public healthcare systems. However, when the pandemic is gradually brought under control, people may have started to focus on other areas such as technological advancements. Artificial intelligence, as an emerging field, may have attracted increased interest and discussions, as it holds significant potential for future development, particularly in addressing similar crises and improving quality of life.

In conclusion, the analysis of topic modeling reveals a dynamic landscape of discussion topics among investors on Weibo. While the Medical Care topic consistently maintains high discussion frequency level, other topics such as Technology topic and Stock Indexes topic have emerged as the second most discussed topics at different periods, surpassing the dominance of the Medical Care topic. Furthermore, the post-COVID period sees a significant decrease in discussions related to Medical Care topic, while the Artificial Intelligence topic witnesses a notable increase. This shift in focus post-pandemic suggests a changing narrative among investors, potentially influenced by broader societal factors. These findings highlight the importance of tracking topic trends to understand market sentiments and predict stock return movements. The next sections will delve into sentiment analysis results and evaluate the effectiveness of sentiment trends in predicting stock returns, providing further insights into the relationship between social media

discussions and stock market performance.

4.3 Sentiment Analysis Results

The FinBERT sentiment analysis provided insights into the overall sentiment expressed in the posts texts by giving sentiment labels for each post. The sentiment labels could indicate the positivity or negativity associated with each text, where -1 represented negative sentiment, 0 represented neutral sentiment, 1 represented positive sentiment.

4.3.1 Overall Sentiment Distribution

Figure 8: Overall Sentiment Distribution by Topic.

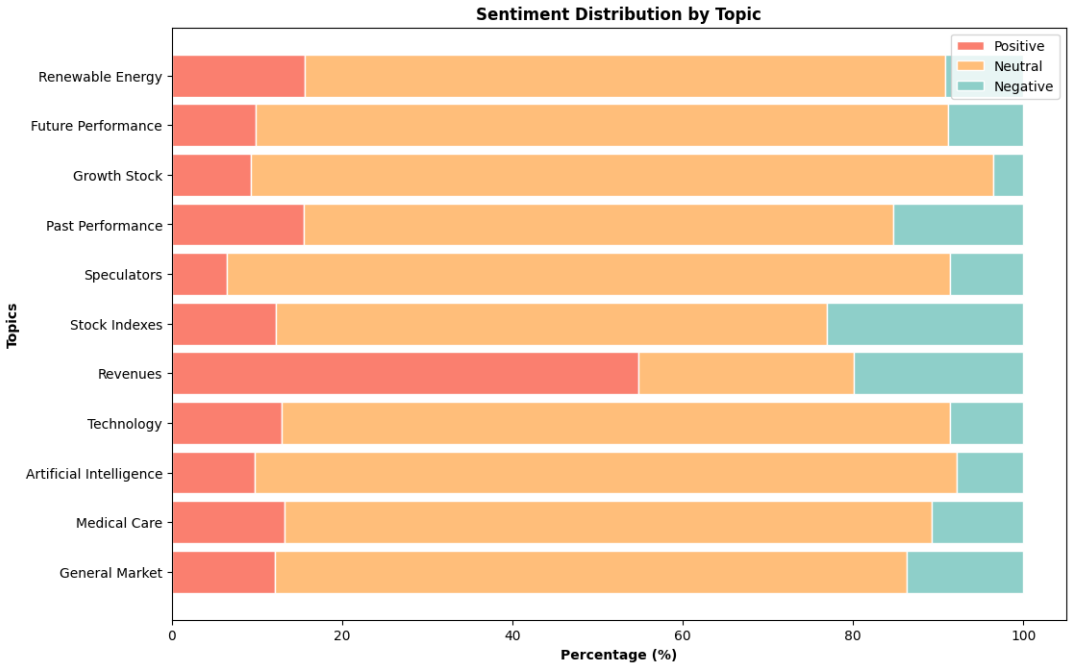


Figure 8 consists of eleven bars of equal length. Each bar represents either the General Market topic or a specific topic, and is divided into three proportions. The red proportion represents the percentage of positive sentiment, the orange proportion represents the percentage of neutral sentiment, and the blue proportion represents the percentage of negative sentiment. For General Market topic, the majority of posts are neutral, followed by negative sentiments, then positive sentiments.

The cross-sectional comparison reveals variations in sentiment distribution across topics compared to the General Market topic. On the one hand, most specific topics have lower positive

sentiment ratios compared to the General Market topic, which indicates a relatively cautious or conservative investor attitude. However, notable is the Revenues topic, which exhibits a significantly higher positive sentiment ratio of over 50% compared to other topics and the General Market topic. This could reflect investors' positive views on company revenues and profitability, considering these factors positively impact stock performance. On the other hand, some certain topics have higher negative sentiment ratios compared to the General Market topic, which indicates more concerns or negative perceptions among investors towards those specific topics. For instance, the Stock Indexes topic shows a relatively higher negative sentiment ratio, which could reflect uncertainties or negative expectations regarding stock indexes.

4.3.2 Topic Positive Sentiment Percentage Over Time

Figure 9: Heatmap of Positive Sentiment Percentage of Topics (Excluding General Market) Over Time.

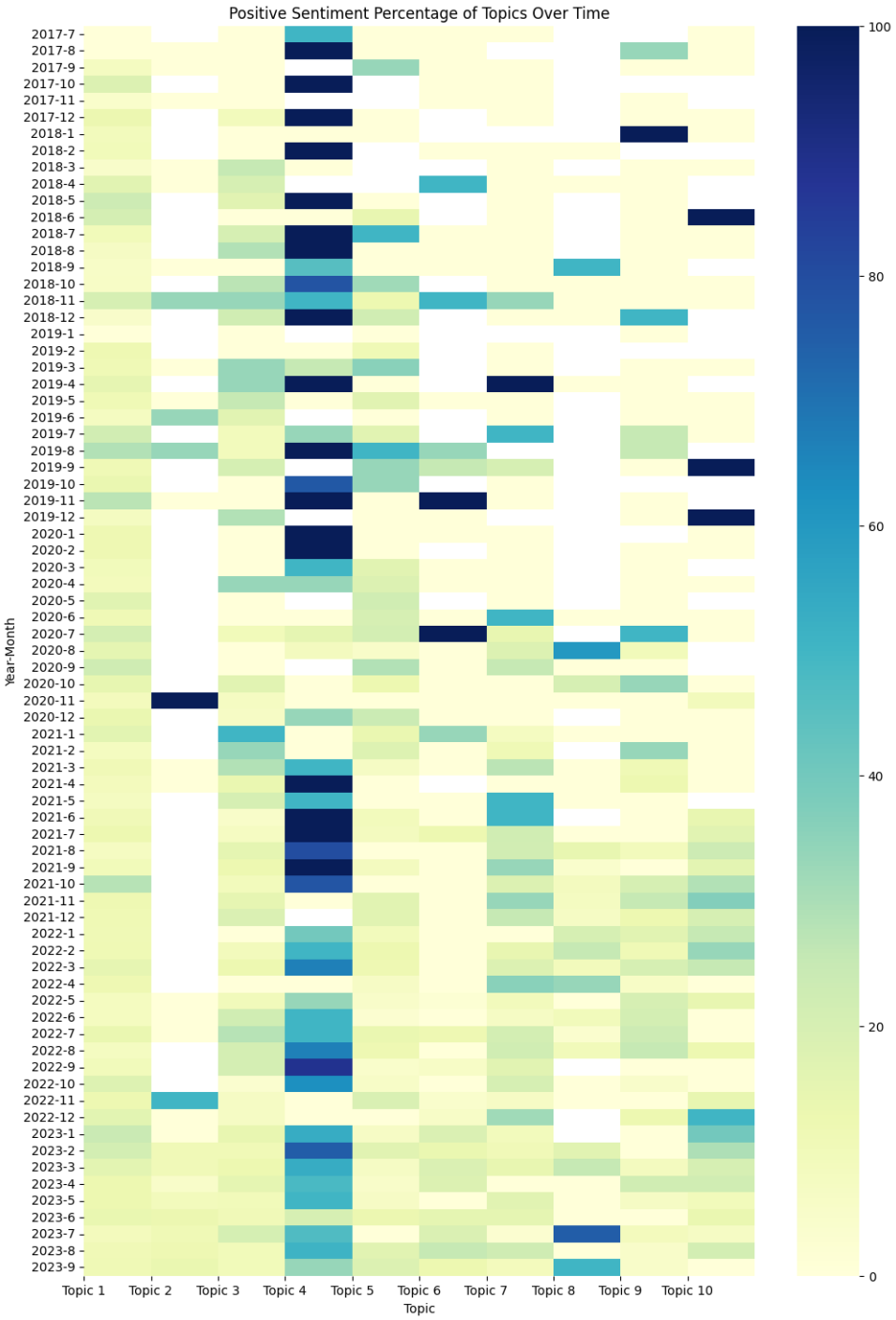


Figure 9 illustrates the sentiment trends for different topics over time. The x-axis represents the timeline from July 2017 to September 2023, while the y-axis displays the ten specific topics analyzed. The color intensity indicates the percentage of positive sentiment, with darker shades indicating higher percentage of positive sentiment and lighter shades indicating lower percentage

of positive sentiment.

Several notable patterns emerge. Firstly, the number of dark-colored plaids representing high percentage of positive sentiment is greater before the pandemic compared to during and after it. This suggests that before the pandemic, many topics exhibit more instances of high positive sentiment percentage before the pandemic. However, during and after the pandemic, only some topics demonstrate a few instances of high positive sentiment percentage, indicating an overall decreasing trend in positive sentiment percentage.

Secondly, the Revenues topic stands out with consistently high positive sentiment, especially during the pandemic after 2021, surpassing all other topics. This trend suggests a sustained interest or positive outlook regarding revenue-related discussions, potentially reflecting economic or financial optimism. In contrast, the Medical Care topic consistently maintained below 50% positive sentiment, indicating a predominantly neutral or negative sentiment surrounding discussions related to medical care. This trend may reflect the complexities and challenges associated with healthcare topics, particularly during times of uncertainty such as the pandemic.

In addition to the Revenues topic and Medical Care topic, before the pandemic, several topics exhibit instances of 100% positive sentiment, including Renewable Energy topic, Past Performance topic, Future Performance topic, and Speculators topic. However, during the pandemic, the occurrence of such high positive sentiment instances decreased significantly, with only Speculators topic and Artificial Intelligence topic showing isolated instances. Post-pandemic, there was a notable peak in positive sentiment for Growth Stock topic, with the positive sentiment percentage exceeding 60%. This surge suggests a renewed interest or positive outlook toward growth stocks among investors in social media discussions.

4.3.3 Topic Positive Sentiment Percentage in Different Stages

Table 3: Positive Sentiment Percentage of Topics (Excluding General Market) in Three Stages.

Label	Pre-COVID	During-COVID	Post-COVID
Medical Care	12.83	12.58	14.25
Artificial Intelligence	17.24	16.67	9.53
Technology	14.35	13.77	11.83
Revenues	64.56	64.77	50.14
Stock Indexes	19.59	10.77	9.36
Speculators	16.67	3.09	14.91
Past Performance	6.67	17.71	14.29
Growth Stock	7.14	8.60	23.08
Future Performance	9.43	11.93	6.47
Renewable Energy	16.67	14.23	17.46

Table 3 presents the percentage of positive sentiment for each topic across three stages: Pre-COVID, During-COVID, and Post-COVID. Firstly, Medical Care topic exhibits the lowest sentiment during the pandemic, but its positive sentiment percentage rebounds afterward, suggesting a shift from negative sentiment during the pandemic to optimism after the pandemic. Secondly, Revenues topic has a very high positive sentiment percentage before the pandemic and increases during the pandemic. However, it drops significantly after the pandemic, although it still remains higher than all other topics. This could be influenced by the release of financial statements each quarter. Before the pandemic, there might have been more focus on a company's profitability and revenue, hence showing higher positive sentiment, while after the pandemic, due to the impact of the pandemic on the economy, there might have been a change in expectations regarding corporate earnings, leading to a decrease in the positive sentiment.

What's more, the Artificial Intelligence topic and Technology topic have higher positive sentiment percentages before the pandemic but experience declines during and after it. This decline may reflect a change in sentiment towards the tech sector during the pandemic. Similarly, Stock Indexes topic exhibits a downward trend, which may reflect investors' concerns about market performance or uncertainty about the economic outlook. Besides, Speculators topic experiences a sharp decline in positive sentiment percentage during the pandemic but recovers afterward, which may be attributed to investors' conservative attitudes towards stock market changes and risks during the pandemic.

In conclusion, the sentiment analysis results highlight a clear trend of decreasing positive sentiment over time across various topics. The Medical Care topic consistently has lower positive

sentiment percentage compared to other topics, while the Revenues topic stands out with consistently high positive sentiment percentage. Furthermore, the analysis reveals a shift in sentiment during and after the pandemic. The Medical Care topic's positive sentiment percentage, initially low during the pandemic, rebounds with optimism afterward, possibly reflecting a change in focus or sentiment towards healthcare as the pandemic situation improves. In contrast, the Revenues topic's positive sentiment percentage is high before the pandemic, increases during it, and then drops significantly afterward. This change in sentiments underscores the complex interplay between external events, such as the pandemic, and investors' sentiments in financial discussions. By tracking the sentiment trends, investors can gain valuable insights into changing market dynamics and adjust their investment decisions accordingly.

4.4 Stock Return Prediction Analysis

The stock return prediction analysis forecasted the future returns of stocks based on the sentiments of the topics extracted from social media posts. This section explored the effectiveness of topic modeling and sentiment analysis in predicting stock returns by taking the positive sentiment percentage of each topic as factors, highlighting the potential impact of investor sentiment and discussion topics on stock market performance.

In this study, the three main stock indices in China—Shanghai Composite Index, Shenzhen Composite Index, and CSI 300—were chosen as the predictive variables to investigate the relationship between the positive sentiment ratios of the ten specific topics and stock returns. Three criteria were used: Firstly, the Pearson correlation was employed to study the linear relationship between the positive sentiment percentage of each topic and each stock index. A correlation value greater than 0 indicated a positive correlation, less than 0 indicated a negative correlation, and 0 indicated no linear correlation. Secondly, a multiple linear regression was conducted using the stock index's own returns and the positive sentiment percentage of a specific topic as independent variables. The p-value of the coefficient for the positive sentiment percentage of the topic was then examined. If the p-value was greater than 0.1, it suggested that the independent variable was not statistically significant, indicating that the relationship between the topic's positive sentiment percentage and the stock return might be due to randomness. If the p-value was less than or equal to 0.1, it meant that the independent variable was partially significant statistically, indicating that it might have some impact on the dependent variable. Lastly, the im-

provement in predictive performance of this multiple linear model, as measured by the decrease in the Root Mean Squared Error (RMSE), was compared to an autoregressive model of the stock index that involved only their past returns as predictors. If the decrease in RMSE was significant, it indicated that the inclusion of this topic's positive sentiment percentage as independent variable had a positive impact on the model's predictive performance, making the prediction results more accurate. If the RMSE did not change much, it suggested that the positive sentiment percentage of the topic had a weak predictive ability for stock returns.

4.4.1 Shanghai Composite Index

Table 4: Results of Shanghai Composite Index.

Labels	Person Correlation	Coefficient P-value	RMSE Reduction
General Market	0.24	0.03	0.12
Medical Care	-0.11	0.35	0.01
Artificial Intelligence	0.20	0.09	0.00
Technology	0.02	0.88	2.86
Revenues	-0.26	0.02	2.54
Stock Indexes	-0.04	0.76	0.02
Speculators	0.10	0.38	0.79
Past Performance	-0.05	0.68	0.40
Growth Stock	0.06	0.60	3.22
Future Performance	0.15	0.20	0.54
Renewable Energy	-0.02	0.87	0.03

The Shanghai Composite Index is one of the major stock indices in China, and is regarded as a benchmark for the Chinese stock market. It tracks the performance of all A-shares and B-shares listed on the Shanghai Stock Exchange, representing a wide range of industries and sectors in China. The index is widely followed by domestic and international investors as an indicator of the overall performance of the Chinese stock market.

Table 4 presents the results of the analysis for the Shanghai Composite Index. Each row represents a different topic, and the columns provide the information of Pearson correlation, p-value of the coefficient, and RMSE reduction. Among all the topics, only the Revenues topic satisfies all three criteria by showing a significant negative correlation of -0.26 with the Shanghai Composite Index. Its inclusion in the model yields a p-value of 0.02, indicating statistical significance. Additionally, incorporating the positive sentiment percentage of this topic reduces the RMSE by 2.54. These findings suggest that the sentiment surrounding revenues in the social media discussions may have a predictive relationship with the performance of the Shanghai

Composite Index.

4.4.2 Shenzhen Composite Index

Table 5: Results of Shenzhen Composite Index.

Labels	Person Correlation	Coefficient P-value	RMSE Reduction
General Market	0.28	0.01	0.06
Medical Care	0.03	0.83	0.27
Artificial Intelligence	0.11	0.34	0.03
Technology	-0.01	0.93	5.33
Revenues	-0.21	0.07	2.60
Stock Indexes	0.00	0.98	0.00
Speculators	0.13	0.27	0.00
Past Performance	-0.04	0.71	0.43
Growth Stock	-0.09	0.45	5.12
Future Performance	0.03	0.77	0.43
Renewable Energy	-0.03	0.77	0.06

The Shenzhen Composite Index is a stock market index of the Shenzhen Stock Exchange in China. It tracks the performance of A-share stocks listed on the Shenzhen Stock Exchange. This provides a comprehensive representation of the Shenzhen stock market, and is widely used as a benchmark for investors and analysts to evaluate the performance of stocks listed on the Shenzhen Stock Exchange.

Table 5 illustrates the analysis results for the Shenzhen Composite Index. The table includes Pearson correlation, coefficient p-value, and RMSE reduction for each topic. Similar to the Shanghai Composite Index, only the Revenues topic meets all three criteria. It exhibits a significant negative correlation of -0.21 with the Shenzhen Composite Index, a p-value of 0.07, indicating partial statistical significance, and an RMSE reduction of 2.6, suggesting that incorporating the positive sentiment percentage of the Revenues topic improves the prediction accuracy of this stock index. This implies that sentiment related to revenues in the social media discussions may be predictive of the performance of the Shenzhen Composite Index.

4.4.3 CSI 300 Index

Table 6: Results of CSI 300.

Labels	Person Correlation	Coefficient P-value	RMSE Reduction
General Market	0.30	0.01	0.60
Medical Care	-0.05	0.66	0.03
Artificial Intelligence	0.19	0.09	0.11
Technology	-0.01	0.93	3.93
Revenues	-0.29	0.01	3.46
Stock Indexes	-0.05	0.65	0.43
Speculators	0.11	0.34	0.21
Past Performance	-0.06	0.59	1.06
Growth Stock	0.03	0.78	4.27
Future Performance	0.10	0.37	0.64
Renewable Energy	-0.05	0.68	0.13

The CSI 300 Index is a capitalization-weighted stock market index consisting of top 300 A-share stocks listed on the Shanghai Stock Exchange or Shenzhen Stock Exchange. It represents a comprehensive reflection of the performance of the Chinese stock market.

Table 6 presents the results of the analysis for the CSI 300 Index. Similar to the Shanghai Composite Index and the Shenzhen Composite Index, each row represents a different topic, and the columns provide the information of Pearson correlation, p-value of the coefficient, and the RMSE reduction. Among all the topics, only the Revenues topic satisfies all three criteria. It exhibits a significant negative correlation with the CSI 300 Index, indicated by a Pearson correlation value of -0.29. Moreover, incorporating the positive sentiment percentage of the Revenues topic into the model results in a p-value of 0.01, considerably lower than the threshold of 0.1. Additionally, the RMSE is reduced by 3.46, suggesting that sentiment surrounding revenues in the social media discussions may possess predictive potential regarding the performance of the CSI 300 Index.

In summary, through the analysis of Pearson correlation, coefficient p-value, and RMSE reduction, it is evident that the positive sentiment percentage of the Revenues topic is consistently significant across all three major stock indexes. This suggests that the sentiment surrounding revenue-related discussions in the social media posts has a notable impact on stock return prediction. This finding underscores the potential for sentiment analysis, particularly regarding revenue-related discussions, to provide valuable insights for stock market forecasting. It implies that monitoring sentiment trends related to revenue discussions could be a useful tool for

investors seeking to understand and predict stock market movements.

5 Conclusion

The analysis of social media discussions on financial topics revealed a dynamic landscape influenced by various factors, including the evolving global context and shifting investor interests. A clear upward trend in the stock super topic discussion frequency was observed, with a significant increase in 2023. This surge suggested a renewed focus on financial topics as the effects of the pandemic waned and life returned to normal. In this context, this study delved into the key findings from topic modeling, sentiment analysis, and stock return prediction to elucidate the changing dynamics of financial discussions and their implications for stock market performance.

Firstly, the BERTopic topic modeling analysis unveiled the General Market topic and ten other distinct topics in financial discourse, with the Medical Care topic consistently dominating discussions. However, the relative discussion frequencies of various topics evolved over time, with Technology topic and Stock Indexes topic emerging as prominent themes during different periods. Moreover, there was a notable shift in focus in the post-COVID period, as reflected by the decline in discussions related to the Medical Care topic and the surge in interest in the Artificial Intelligence topic.

Secondly, the FinBERT sentiment analysis revealed an overall decline in positive sentiment over time. The Medical Care topic consistently received lower positive sentiment percentage compared to other topics, while the Revenues topic stood out with significantly high positive sentiment percentage, indicating investors' confidence on its resilience. The fluctuation in sentiment, particularly the rebound in Medical Care topic's sentiment and the decreasing trends in Revenues topic's sentiment after the pandemic, underscored the dynamics of investor attitudes on market performance.

Lastly, stock return prediction analysis, done through Pearson correlation, coefficient p-value, and RMSE reduction, underscored the predictive power of sentiment analysis, particularly in forecasting stock return movements. The consistent significance of the positive sentiment surrounding the Revenues topic across major stock indexes highlighted its potential as a leading indicator of market performance.

Overall, this study contributed to the growing body of literature on the intersection of social media text analysis and financial markets analysis, particularly in the context of the pre- and post-COVID-19 era. By integrating topic modeling, sentiment analysis, and stock return prediction techniques, this study provided insights into the dynamics of financial discussions on social media and their implications for stock market performance, both before and after the COVID-19 pandemic. The findings underscored the importance of incorporating sentiment analysis into investment strategies, especially during periods of market volatility, and highlighted the potential for social media texts to inform financial decision-making in a rapidly changing environment.

The limitations of this study included the size of the dataset and the methods used for modeling and translation. Firstly, the dataset used was limited in size, potentially impacting the comprehensiveness and accuracy of the analysis. A larger dataset with more posts could provide more representative results. Secondly, the models or the translation method used in the study had limitations and might not fully capture or translate all subtle differences or contexts present in social media discussions. This could affect the accuracy and comprehensiveness of the topic modeling and sentiment analysis.

Looking ahead, future research could explore the following directions. Expanding the dataset to include more diverse topics and discussions could provide a more comprehensive analysis of sentiment trends in financial discussions on social media. This could provide a better understanding of investors' sentiments and their impact on stock market dynamics. Additionally, utilizing pre-trained Chinese financial models could reduce the reliance on translation and the limitation caused by domain-specific terminologies. This approach could improve the accuracy and reliability of sentiment analysis, especially in capturing subtle meanings and contexts in financial discussions on social media. Continuing to innovate in this field could lead to more robust and accurate insights into market dynamics and investors' sentiments, ultimately enhancing investment strategies and decision-making processes.

References

[1] E. F. Fama, "Efficient capital markets: A review of theory and empirical work," *The Journal of Finance*, vol. 25, no. 2, pp. 383–417, 1970.

[2] E. F. Fama, L. Fisher, M. C. Jensen, and R. Roll, "The adjustment of stock prices to new information," *International Economic Review*, vol. 10, no. 1, pp. 1–21, Feb. 1969. [Online]. Available: http://ideas.repec.org/a/ier/iecrev/v10y1969i1p1-21.html

[3] E. F. Fama, "Efficient capital markets: Ii," *The Journal of Finance*, vol. 46, no. 5, pp. 1575–161, 1991.

[4] ——, "The behavior of stock-market prices," *The Journal of Finance*, vol. 38, no. 1, pp. 34–105, 1965. [Online]. Available: http://dx.doi.org/10.2307/2350752

[5] J. Wang, Y. Fan, J. Palacios, Y. Chai, N. Guetta-Jeanrenaud, N. Obradovich, C. Zhou, and S. Zheng, "Global evidence of expressed sentiment alterations during the covid-19 pandemic," *Nature Human Behaviour*, vol. 6, no. 3, pp. 349–358, 2022. [Online]. Available: https://doi.org/10.1038/s41562-022-01312-y

[6] R. Xie, S. K. W. Chu, D. K. W. Chiu, and Y. Wang, "Exploring public response to covid-19 on weibo with lda topic modeling and sentiment analysis," *Data and Information Management*, vol. 5, no. 1, pp. 86–99, 2021. [Online]. Available: https://doi.org/10.2478/dim-2020-0023

[7] S. Feuerriegel, A. Ratku, and D. Neumann, "Analysis of how underlying topics in financial news affect stock prices using latent dirichlet allocation," in *2016 49th Hawaii International Conference on System Sciences (HICSS)*. Koloa, US: IEEE, New York, 5-8 January 2016, pp. 1072–1081.

[8] E. N. Biktimirov, T. Sokolyk, and A. Ayanso, "Sentiment and hype of business media topics and stock market returns during the covid-19 pandemic," *Journal of Behavioral and Experimental Finance*, vol. 31, p. 100542, Sep. 2021. [Online]. Available: https://doi.org/10.1016/j.jbef.2021.100542

[9] M. Baker and J. Wurgler, "Investor sentiment and the cross-section of stock returns," *The Journal of Finance*, vol. 61, no. 4, pp. 1645–1680, 2006. [Online]. Available: https://doi.org/10.1111/j.1540-6261.2006.00885.x

[10] G. Zhou, "Measuring investor sentiment," *Annual Review of Financial Economics*, vol. 10, pp. 239–259, 2018. [Online]. Available: https://doi.org/10.1146/annurev-financial-110217-022725

[11] Z. McGurk, A. Nowak, and J. C. Hall, "Stock returns and investor sentiment: Textual analysis and social media," *Journal of Economics and Finance*, vol. 44, pp. 458–485, 2020. [Online]. Available: https://doi.org/10.1007/s12197-019-09494-4

[12] J. Bollen, H. Mao, and X. Zeng, "Twitter mood predicts the stock market," *Journal of Computational Science*, vol. 2, no. 1, pp. 1–8, Mar. 2011. [Online]. Available: https://doi.org/10.1016/j.jocs.2010.12.007

[13] X. Wang, Z. Xiang, W. Xu, and P. Yuan, "The causal relationship between social media sentiment and stock return: Experimental evidence from an online message forum," *Economics Letters*, vol. 216, p. 110598, Jul. 2022. [Online]. Available: https://doi.org/10.1016/j.econlet.2022.110598

[14] M. Grootendorst, “Bertopic: Neural topic modeling with a class-based tf-idf procedure,” 2022.

[15] W. Chen, F. Rabhi, W. Liao, and I. Al-Qudah, “Leveraging state-of-the-art topic modeling for news impact analysis on financial markets: A comparative study,” *Electronics*, vol. 12, no. 12, p. 2605, 2023. [Online]. Available: https://doi.org/10.3390/electronics12122605

[16] D. Araci, “Finbert: Financial sentiment analysis with pre-trained language models,” 2019.

[17] S. V. Lawrence, *COVID-19 and China: A chronology of events (December 2019-January 2020)*. Congressional Research Service, 2020.

[18] J. Ge, “The covid-19 pandemic in china: from dynamic zero-covid to current policy,” *Herz*, vol. 48, pp. 226–228, 2023. [Online]. Available: https://doi.org/10.1007/s00059-023-05183-5