Research Proposal

Applicant's Name:

Peng Yifeng （彭一锋）

Proposed Topic:

Research on the Impact of Cross Platform Social Media Signals on Financial Market Returns

Abstract:

This research investigates the impact of social media signals from multiple platforms—Twitter, StockTwits, and Seeking Alpha—on financial market returns. By employing advanced data analysis methods, including Principal Component Analysis (PCA), Factor Analysis, and Independent Component Analysis (ICA), alongside cutting-edge large language models (LLMs), we aim to extract and compare sentiment and attention signals across platforms. Our study addresses the gap in existing literature by focusing on cross-platform differences and their implications for market-relevant information. We hypothesize that while sentiment signals predict positive next-day returns, attention signals forecast negative returns due to potential overreactions. The research also explores how platform-specific characteristics influence the informativeness of these signals. Through rigorous data collection and analysis, we expect to provide insights into the nuanced role of social media in financial markets, offering valuable guidance for investors and policymakers. This study not only enhances our understanding of social media's influence on market dynamics but also demonstrates the potential of artificial intelligence in financial analysis.

Introduction:

Research Background

The term "social media" was officially included in Merriam-Webster's dictionary in 2004, but the concept of social media influencing financial markets dates back to the 17th-century Amsterdam Stock Exchange clubs, where the "South Sea Bubble" of September 1720 was also fueled by the lively discussions among investors in London coffeehouses. The impact of social networks on trading behavior and markets has long been a hot topic in academia.

Over the past two decades, social media has experienced explosive growth and has become an integral part of people's daily lives. An increasing number of individuals now consider social media as their primary source of news. This trend has not only transformed social interactions but also profoundly influenced the dissemination of information within financial markets.

Social media has been playing an increasingly significant role in financial markets. Investors frequently express their views on securities on social media, while companies also utilize these platforms to disclose information and interact with investors. However, it was not until the recent trading frenzy driven by social media, particularly the GameStop phenomenon in 2021 [1], that the role of social media in investing garnered widespread attention. These events have raised important questions about the role of social platforms in trading and information dissemination, prompting a series of new research initiatives.

Traditionally, most research on investor social media has focused on single platforms, such as Twitter, StockTwits, and Seeking Alpha [2]. However, these platforms differ significantly in terms of user demographics, content formats, and information dissemination methods. For instance, Twitter, Instagram, and Facebook are known for their short messages and rapid spread, whereas Seeking Alpha is characterized by in-depth analytical articles. This diversity implies that each platform may generate unique informational content, potentially exerting different influences on the market.

Attention and sentiment signals on social media have different predictive roles for market returns [3]. Attention generally has a stronger predictive power for negative returns on the next trading day, while investor sentiment tends to predict positive returns. This distinction underscores the importance of considering both social media sentiment and attention simultaneously, as well as the necessity of differentiating between various investor social platforms.

With the increasing prominence of social media in financial markets, it has not only become a crucial channel for information dissemination but also a significant factor

influencing market dynamics. Understanding how these platforms generate and transmit market-related information is of great importance to investors and policymakers. This study aims to fill the research gap in cross-platform social media influence, providing new perspectives on the role of social media in financial markets.

Research Aims

Although the importance of social media in financial markets has been increasing, existing research primarily focuses on the impact of individual platforms, often overlooking the significant differences between platforms, including variations in user demographics, content forms, and information dissemination methods. This approach of studying single platforms limits researchers' comprehensive understanding of the overall influence of social media.

Specifically, the limitations of existing research are manifested in the following aspects:

Platform Specificity: Most studies focus solely on a specific social media platform, failing to consider the interactive effects and comprehensive impacts of information dissemination across different platforms. This may lead to an underestimation or misinterpretation of the overall influence of social media in financial markets.

Separation of Sentiment and Attention: Many studies do not analyze the combined impact of sentiment and attention signals on market returns simultaneously. Sentiment is often associated with positive returns, while attention is linked to negative returns. Ignoring either of these factors may result in inaccurate predictions of market dynamics.

Speed and Depth of Information Dissemination: The speed and depth of information dissemination vary significantly across different platforms. Twitter is known for its rapid dissemination and short messages, whereas Seeking Alpha provides more in-depth analysis. These differences may affect the timeliness and accuracy of investor decisions, yet there is a lack of systematic research on this issue.

Given these limitations, this study raises the following questions:
How do different social media platforms generate and transmit market-related information, known as "social signals"?
How do the expressions of sentiment and attention signals across different platforms influence market returns?
How can large language models be utilized to more accurately analyze and compare sentiment and attention signals across platforms?

By exploring these questions, this study aims to provide new insights into the role of social media in financial markets and offer more comprehensive information analysis tools for investors and policymakers.

Research Significance

This study not only holds significant theoretical reference value but also demonstrates substantial practical significance. By exploring the role of cross-platform social media in financial markets, it aims to fill existing research gaps and provide new perspectives and tools for investors and policymakers.

Theoretical Contributions:

A New Perspective on Cross-Platform Analysis: Most existing research is limited to a single social media platform, overlooking the differences between various platforms. This study, by analyzing major platforms such as Twitter, StockTwits, Seeking Alpha, Weibo, and Eastmoney, reveals the unique characteristics of these platforms in generating market-related information. This will provide new insights into the impact of social media for the academic community and promote further cross-platform research.

Complexity of Social Signals: By simultaneously analyzing emotional and attention signals, we can gain a more comprehensive understanding of how social media influences market returns. This dual-signal analysis helps resolve contradictions regarding the roles of sentiment and attention in existing literature.

Practical Applications:

Investment Decision Support: The results of this study can help investors better utilize social media information for decision-making. By identifying the impact of sentiment and attention signals across different platforms, investors can adjust their investment strategies to enhance returns.

Policy Formulation Reference: For regulatory bodies, understanding how social media influences market dynamics is crucial for developing effective policies. The data and analysis provided by this study can serve as an important basis for policy formulation, aiding regulators in better managing market risks.

Application of Large Language Models: By incorporating large language models for data analysis, this study demonstrates how advanced technologies can improve the accuracy of sentiment analysis. This not only enhances the technical depth of the research but also provides a practical case for the future application of artificial intelligence in the financial sector.

Literature Review

In recent years, the influence of social media on financial markets has grown significantly, attracting the attention of numerous scholars. Existing research primarily focuses on how social media serves as a channel for information dissemination, affecting investor behavior and market dynamics [4][5].

Firstly, social media is regarded as a crucial platform for investor communication and information dissemination. Investors utilize social platforms to share their views on financial markets or products, publish market forecasts, and offer investment advice. These platforms differ in their user base and information characteristics. For instance, Twitter is known for its rapid dissemination of concise information, while Seeking Alpha provides in-depth analytical articles. This diversity results in the unique nature of market-related information generated by different platforms.

Secondly, research has found that emotional and attention signals on social media have different predictive effects on market returns. Emotional signals typically foretell positive returns, whereas attention signals are associated with negative returns [3]. This distinction underscores the importance of considering both social media sentiment and attention simultaneously. However, most studies are limited to single platforms, failing to comprehensively explore cross-platform influences.

Moreover, the structural characteristics of social media (such as character limits, content moderation, and user real-name systems) and the differences among user groups (such as professional investors versus ordinary users) also influence the transmission of information and market reactions. For instance, after StockTwits introduced character limits in 2019, its sentiment signals became more predictive of next-day stock returns, indicating that platform-specific features can significantly impact the content of information [6].

In summary, existing literature highlights the significant role of social media in financial markets, but most studies are limited to the analysis of single platforms, failing to fully uncover the differences in information across platforms and their comprehensive impact on the market. This study aims to fill this gap by conducting cross-platform analysis, providing new insights into understanding the role of social media in financial markets.

In the study of the impact of social media on financial markets, communication theory provides a crucial perspective for understanding how platform characteristics influence information dissemination. Communication theory emphasizes that the properties of the medium affect the content of the information and its dissemination effects [7]. In the context of social media, the structural features of different platforms and the differences in user demographics may lead to unique characteristics in the content of

information and market-related messages.

Medium Characteristics: Different social media platforms possess distinct medium characteristics, which directly influence the manner in which information is disseminated. For instance, Twitter is renowned for its brevity and rapid dissemination of information, whereas Seeking Alpha offers more in-depth analytical articles. These differences can lead to variations in the market impact of information generated on different platforms.

User Demographic Differences: The differences in user demographics across platforms also affect the spread of information and market reactions. The information needs and interaction patterns between professional investors and ordinary users vary, which may result in different interpretations and reactions to the same event across various platforms.

Information Interactivity: The interactivity of social media allows information to spread rapidly and spark widespread discussions. This multidimensional mode of information dissemination differs from the one-way communication of traditional media, making social media's influence on financial markets more complex and profound.

Although existing research has revealed the significant role of social media in financial markets, most studies have focused on individual platforms, failing to fully uncover the differences in information across different platforms and their combined impact on the market. Additionally, how to utilize emerging technologies such as large language models to more accurately analyze cross-platform information dissemination remains an open question. This study aims to fill this gap through cross-platform analysis, providing new insights into the role of social media in financial markets.

Research Hypothesis

In existing research, the impact of social media on financial markets has garnered significant attention; however, the majority of studies focus on a single platform, overlooking the substantial differences between various platforms. This limitation restricts our comprehensive understanding of the overall influence of social media. Therefore, this study aims to fill this gap by conducting a cross-platform analysis to explore the role of social media in financial markets. The following specific research questions and hypotheses are proposed to address this gap:

How does the correlation between social media sentiment and market returns manifest?

Hypothesis 1: Overall, positive sentiment signals on social media predict positive returns in the market the following day, while negative sentiment signals predict negative returns. This hypothesis is based on a review of existing literature, which suggests that sentiment signals typically contain information related to returns.

How do sentiment signals on different platforms affect market returns?

Hypothesis 2: There are significant differences in the predictive power of sentiment signals on market returns across platforms such as Twitter, StockTwits, and Seeking Alpha. Specifically, due to differences in user demographics and platform characteristics, long-form analyses on Seeking Alpha may provide deeper emotional insights, making them more accurate in predicting market returns. Conversely, Twitter, with its short-form content, may have an advantage in immediacy but lack depth.

What role do large language models play in cross-platform sentiment analysis?

Hypothesis 3: Utilizing large language models can enhance the precision of cross-platform sentiment analysis by better understanding complex linguistic structures and identifying subtle emotional changes, thereby improving the accuracy of sentiment signals in predicting market returns.

These hypotheses aim to uncover how social media sentiment signals influence market returns and how differences between platforms affect this relationship. By validating these hypotheses, this study not only seeks to enrich the existing literature but also to provide investors with more precise tools for information analysis to optimize investment decisions.

Methodology

Design of Research

In this study, we intend to employ a quantitative research methodology to explore the impact of sentiment and attention signals generated on social media platforms on financial market returns through systematic analysis. The research design aims to gain a deeper understanding of how different social media platforms, through their unique user bases and platform characteristics, influence the generation and dissemination of market information.

The study is grounded in communication theory, emphasizing how the media characteristics of different social media platforms affect the content of information and its market influence. We hypothesize that significant differences exist in the market influence of information generated across platforms due to variations in platform features (such as differences in platform interaction and user demographics). By conducting cross-platform analysis, we will identify and compare sentiment and attention signals on various social media platforms to investigate how these signals influence market returns. Specifically, we will analyze the differences in the manifestation of these signals across platforms and assess the role of large language models in enhancing the accuracy of cross-platform sentiment analysis.

Data Acquisition

In this study, we will collect data from multiple social media platforms to comprehensively analyze the role of social media in financial markets. To ensure the accessibility of data and the rigor of the research, we have selected the following primary data sources:

1. Twitter:

Data Source: We plan to use Twitter data provided by Social Market Analytics (SMA), a company specializing in providing sentiment information for professional investors.

Data Type: We will obtain snapshots at 4:00 PM each day, including the number of tweets related to each company and the average sentiment over the past 24 hours.

Data Range: We will analyze data from 2012 to 2023 to ensure a sufficiently long time span to observe trends and changes.

2. StockTwits:

Data Source: Detailed message-level data is obtained from the StockTwits platform. StockTwits primarily focuses on the financial markets, where users can tag specific companies using "cashtags."

Data Type: This includes all single-company messages from 2012 to 2023. The sentiment scores calculated using the MarketLex algorithm, which is based on user-defined tags ("bullish" or "bearish"), will be utilized.

Sample Limitation: The analysis is restricted to messages mentioning a single company to ensure the accuracy of sentiment assignment.

3. Seeking Alpha:

Data Source: Utilizing article-level sentiment data provided by Ravenpack 1.0, articles with a relevance score above 75 are retained.

Data Type: Employing Event Sentiment Scores (ESS), which range from -1 to 1, where 0 indicates neutral sentiment, and positive (negative) values denote positive (negative) sentiment.

Data Scope: Similarly, data from 2012 to 2023 is analyzed to maintain consistency with other platforms.

4. Traditional News Media and Corporate Announcements:

To control the impact of other information sources on the market, we will collect news reports from traditional media outlets (such as The Wall Street Journal) and news feeds from Dow Jones Newswires, along with their sentiment scores. Additionally, we will gather the dates of 8-K filings and earnings announcements, which are sourced from the SEC Analytics Suite on WRDS and the IBES database.

5. Market Return Data:

Daily abnormal returns are calculated using the CRSP database by subtracting the market-weighted market return from the company's daily return.

By integrating these diverse data sources, this study aims to comprehensively analyze how social media platforms generate and disseminate market-related information. The data from these platforms not only cover different characteristics of information dissemination but also reflect the behaviors of various user groups, providing rich data support for our research.

Data Analysis

In this study, we will employ a variety of data analysis methods to delve into the impact of emotional and attentional signals on market returns within social media platforms. To enhance the precision and innovativeness of our analysis, we have incorporated Large Language Models (LLMs) as a novel tool, combining them with traditional analytical methods to provide more comprehensive insights.

1. Data Preprocessing

- Text Cleaning and Standardization:

  Initially, the text data collected from Twitter, StockTwits, and Seeking Alpha is

cleaned. This involves removing noise characters, stop words, and irrelevant information to ensure data consistency. Regular expressions and natural language processing (NLP) tools are employed for text standardization, ensuring consistent terminology and formatting across different platforms.

- Sentiment Signal Extraction:
  Traditional NLP Methods: Traditional NLP techniques, such as sentiment lexicons and machine learning classifiers, are utilized for preliminary sentiment analysis of the text. Open-source tools like VADER or TextBlob are used to identify positive, negative, and neutral sentiments within the text.
  Application of Large Language Models: LLMs , such as OpenAI's GPT series, are introduced for deeper sentiment analysis. LLMs can comprehend complex language structures and identify subtle emotional shifts. Pre-trained large language models are employed to classify sentiment in text from each platform, generating more precise sentiment scores. These models, trained on extensive datasets, learn language patterns and can accurately identify sentiments in various contexts.

- Attention Signal Extraction:
  Volume Analysis: Calculate the message volume for each company across various platforms as the foundation for attention signals. Utilize the API interfaces of each platform to obtain daily message volumes and compute the proportion relative to the total message volume to standardize the attention signals.
  Topic Modeling: Employ topic modeling techniques (such as Latent Dirichlet Allocation, LDA) to identify the main topics within the text. This aids in understanding which topics attract more user attention and how they influence market returns.

2. Applications of Large Language Models

In this study, large language models (LLMs) are introduced as an innovative tool to enhance the accuracy of social media sentiment and attention signal analysis. The powerful natural language processing capabilities of LLMs enable them to identify subtle emotional changes and thematic focus points within complex financial texts.

- Model Selection and Configuration:
  We opted for advanced open-source large language models such as the Meta Llama3 series, which have been pre-trained on extensive corpora and exhibit exceptional zero-shot learning abilities. To enhance their performance in the financial domain, we employed instruction tuning techniques. This involved converting a small portion of supervised financial sentiment analysis data into instruction data for fine-tuning the LLMs. During application, we integrated a retrieval-augmented module to provide LLMs with additional contextual information from reliable external sources, thereby improving the accuracy of sentiment predictions. Special attention was given to the model's ability to handle

numerical data and contextual understanding in financial texts, ensuring its effectiveness in sentiment analysis.

- Sentiment Signal Analysis:
  Using LLMs for sentiment classification of social media text, generating sentiment labels or quantitative sentiment scores. By comparing with traditional NLP methods, it is possible to verify the advantages of LLMs in identifying complex linguistic structures and subtle emotional changes. Additionally, the performance differences of LLMs across different platforms can be evaluated to reveal how platform-specific characteristics influence the generation of sentiment signals.

- Attention Signal Analysis:
  Utilizing LLMs to automatically identify key topics and focal points in text, to understand which topics attract more user attention on different platforms. By integrating retrieval-augmented modules, LLMs can obtain supplementary information from external sources, enhancing their understanding of short financial news and social media posts.

- Performance Evaluation and Optimization:
  By comparing with traditional sentiment analysis models and other LLMs (such as Gemini and Qwen), the improvement in accuracy and F1 score of the model is evaluated. The expected results of this study show that through instruction fine-tuning and retrieval augmentation, LLMs exhibit significant advantages over traditional methods in processing financial text, especially in scenarios requiring numerical understanding and contextual grasp.

3. Data Analysis Methods

- Principal Component Analysis (PCA):
  PCA is employed to decompose data from various platforms in order to extract common and specific components. By comparing these components, we can identify the uniqueness of each platform in generating market-related information. The anticipated outcomes of this study include: the first principal component (PC1) of PCA is used to explain the commonality of signals across different platforms. For attention signals, PC1 is capable of explaining a significant portion of the variability, whereas for emotional signals, the explanatory power of PC1 is comparatively weaker.

- Independent Component Analysis (ICA):
  Independent Component Analysis (ICA) is utilized to decompose mixed signals into mutually independent non-Gaussian signals. Unlike PCA, which focuses on maximizing variance, ICA emphasizes the independence of signals. The study intends to use ICA to identify unique emotional and attention signals on each platform, which may reflect specific user behaviors or information dissemination

patterns.

- Regression Analysis:
  This study employs a multiple regression model to assess the predictive power of sentiment and attention signals on next-day abnormal returns. Specifically, abnormal returns are treated as the dependent variable, with sentiment and attention signals as independent variables, while controlling for traditional media information, corporate announcements, lagged returns, and volatility.

By integrating these various data analysis techniques, this research not only enables more precise extraction of common and specific components from social media but also deepens our understanding of how these signals influence market returns. This multi-layered methodology provides a more comprehensive perspective, unveiling the intricate role of social media in financial markets.

Variable Definition

This study meticulously defines and describes several key variables used to analyze the impact of sentiment and attention signals on market returns across social media platforms. Each variable's construction process is carefully designed to ensure it accurately reflects the research objectives.

- Sentiment Signal:
  Definition: The sentiment signal measures the emotional inclination in social media posts, ranging from -1 (extremely bearish) to +1 (extremely bullish).
  Construction Method: For Twitter, we utilize daily sentiment snapshots provided by Social Market Analytics; for StockTwits, we employ custom sentiment scores calculated by the MarketLex algorithm; for Seeking Alpha, we use the Event Sentiment Score (ESS) from Ravenpack.
  Measurement: On each platform, the average sentiment score is computed for all posts or articles about a particular company from 4:00 PM the previous day to 4:00 PM the current day.

- Attention Signal:
  Definition: Attention signals reflect the level of discussion intensity regarding a particular company on social media platforms.
  Construction Method: By calculating the number of posts for each company within a specific time period and normalizing it as a proportion relative to the total number of messages on the platform.
  Measurement: $Attention_{i,t} = \frac{Messages_{i,t}}{\sum_i Messages_{i,t}}$ where $Messages_{i,t}$ represents the number of messages for company i on day t.

- Control Variables:
  Traditional Media Information: This includes the number and sentiment scores of news reports from The Wall Street Journal and Dow Jones Newswires. Data provided by Ravenpack is used to aggregate daily company-specific news sentiment.
  Company Announcements: The collection of 8-K filing dates and earnings announcement dates. These data are sourced from the SEC Analytics Suite on WRDS and the IBES database to control for the impact of company-specific events on market returns.
  Lagged Returns and Volatility: The previous day's market returns and volatility, calculated using the CRSP database, are used as control variables to mitigate the interference of overall market volatility on the prediction of abnormal returns.

Robustness Analysis

Model validation and robustness testing are crucial steps in ensuring the reliability and validity of the analysis results. We employ a variety of methods to verify the accuracy of the model and conduct robustness tests to ensure that the results are not influenced by specific assumptions or data selection.

- Cross-Validation:
  We use cross-validation techniques to assess the predictive performance of the regression model. Specifically, we divide the dataset into training and testing sets, repeatedly training and testing the model to reduce the risk of overfitting. Cross-validation helps us evaluate the model's performance on different subsets of data, thereby ensuring its good generalization ability.

- Leave-One-Out Cross-Validation (LOOCV):
  In the case of small sample sizes, we employ the leave-one-out cross-validation method for validation. Each time, we remove one observation from the dataset, use the remaining data for model training, and then predict the removed observation. This process is repeated until all observations have been removed once.

- Sensitivity Analysis:
  By assessing the model's response to different assumptions and parameter changes, we can verify the reliability of the results and identify potential influencing factors.

  Steps of Sensitivity Analysis:
  1. Baseline Model Construction: First, a baseline regression model is constructed to evaluate the impact of sentiment and attention signals on market returns. This model includes both main variables and control variables, such as information from traditional news media, company announcements, lagged returns, and volatility.
  2. Parameter Variation Testing: Key parameters are gradually adjusted in the

baseline model to observe changes in the results. For example, changing the method of extracting sentiment signals (such as using different platforms or different large language models) to assess the impact of these changes on market return predictions.

3. Alternative Variable Testing: Different datasets or alternative variables are used for testing. For instance, replacing sentiment signals with average sentiment over different time windows, or using different methods to measure attention signals, to check the impact of these alternative variables on the results.

4. Model Complexity Testing: The number of variables in the model is increased or decreased to evaluate the impact of model complexity on the robustness of the results. By comparing the differences between simplified models and complex models, significant variables affecting the results are identified.

- Alternative Model Testing:

  Alternative model testing is a crucial step in ensuring the robustness of research results. By employing different models and algorithms, it is possible to verify whether the primary results are dependent on specific statistical methods or assumptions. The following outlines the specific steps and algorithms to be used in alternative model testing:

  Steps in Alternative Model Testing:

  1. Baseline Model Construction: This step is identical to the sensitivity analysis described above.

  2. Alternative Model Selection: Multiple alternative models are selected to verify the robustness of the results. These models include but are not limited to:

  Nonlinear Regression Models: Such as logistic regression or Poisson regression, used to handle nonlinear relationships.

  Time Series Models: Such as ARIMA (Auto Regressive Integrated Moving Average) models, used to capture temporal dependencies in the data.

  Machine Learning Models: Such as Random Forests or Support Vector Machines (SVM), used to identify complex patterns and nonlinear relationships.

  3. Parameter Estimation and Comparison: Parameters are estimated for each alternative model, and the results are compared with those of the baseline model. Particular attention is paid to changes in key parameters (such as sentiment and attention signal coefficients) to determine if the results are consistent.

  4. Result Interpretation and Validation: The results of each alternative model are analyzed to interpret the impact of key variables on market returns under different models. If the results of the alternative models are consistent with the baseline model, this strengthens the credibility of the research conclusions.

If the results are inconsistent, further analysis is conducted to identify the potential causes of the discrepancies, such as data characteristics, assumption conditions, or algorithm limitations.

Expected Result

In this study, we aim to uncover how sentiment and attention signals on social media platforms influence financial market returns through their analysis. Our research questions and hypotheses provide clear directions for the anticipated outcomes.

1. Relationship between Sentiment and Market Returns
Anticipated Result 1: We expect that sentiment signals will significantly predict positive abnormal returns on the following day. This expectation is based on findings from existing literature, which indicates that positive sentiment is generally associated with positive market reactions. The common components of sentiment signals across multiple social platforms should exhibit consistent positive predictive abilities.

2. Relationship between Attention and Market Returns
Anticipated Result 2: Attention signals are expected to predict negative abnormal returns on the following day. This negative correlation likely reflects the market's overreaction to highly attention-grabbing events, leading to partial price corrections after initial volatility. We anticipate this phenomenon to be observable across all platforms, though its intensity may vary depending on platform characteristics.

3. Cross-Platform Differences
Anticipated Result 3: Sentiment and attention signals on different social media platforms exhibit significant differences in predicting market returns. We expect that Seeking Alpha, due to its higher proportion of in-depth analysis articles, may perform better in predicting returns based on sentiment signals, while Twitter, with its rapid dissemination characteristics, may have an advantage in predicting returns based on attention signals.

4. Application Effects of Large Language Models
Expected Result 4: The introduction of large language models for sentiment analysis should enhance the precision of cross-platform comparisons. We anticipate that the sentiment signals identified by LLMs will more accurately reflect market trends compared to traditional NLP methods, thereby enhancing the predictive power of the signals.

By validating these expected results, this study aims not only to enrich the existing literature but also to provide actionable insights for investors and policymakers to better utilize social media information for decision-making. These results will unveil the profound impact of social media on financial markets and drive further in-depth research in related fields.

Research Timeline

This study is planned to be completed within approximately six months, covering all stages from project initiation to report writing:

Month 1: Project Initiation and Data Collection

Weeks 1-2: Project kick-off meeting to clarify research objectives and establish the methodological framework.

Weeks 3-4: Acquire and organize data from Twitter, StockTwits, and Seeking Alpha platforms, including sentiment and attention signals. Collect traditional news media information and corporate announcements to ensure all necessary control variable data is in place.

Month 2: Data Preprocessing and Initial Analysis

Weeks 5-7: Conduct data cleaning and standardization to ensure consistency across platforms.

Week 8: Utilize large language models (LLMs) for sentiment signal extraction, combined with traditional NLP techniques for initial analysis. Complete the extraction and initial analysis of attention signals.

Month 3: In-depth Data Analysis

Weeks 9-10: Apply Principal Component Analysis (PCA), Factor Analysis, and Independent Component Analysis (ICA) for in-depth signal analysis.

Weeks 11-12: Construct a baseline regression model to evaluate the impact of sentiment and attention signals on market returns. Perform cross-validation to verify the model's accuracy.

Month 4: Model Validation and Robustness Testing

Weeks 13-14: Conduct sensitivity analysis to assess the robustness of the model using Bootstrap resampling and Monte Carlo simulations.

Week 15: Perform alternative model testing by validating results with nonlinear regression, time series analysis, and machine learning methods.

Week 16: Summarize and document the findings from the model validation and robustness testing.

Month 5: Result Analysis and Discussion

Week 17: Analyze the differences in sentiment and attention signals across platforms and their impact on market returns.

Weeks 18-20: Draft the results section, discussing the findings and their theoretical and practical implications. Prepare the initial draft for internal review.

Month 6: Report Writing and Submission

Weeks 21-22: Revise the report based on feedback, refining the literature review,

methodology, and results sections.
Week 23: Write the introduction, conclusion, and abstract, and perform overall formatting and layout adjustments.
Week 24: Conduct a final review and submit the complete research report.

Feasibility Analysis

Through a comprehensive feasibility analysis in terms of technical feasibility, innovativeness, and budget feasibility, we ensure the smooth advancement and successful completion of the research.

- Technical Feasibility

Algorithms and Models: This study employs a variety of data analysis methods, including Principal Component Analysis (PCA), Factor Analysis, Independent Component Analysis (ICA), and Large Language Models (LLMs). These methods possess robust capabilities in handling large-scale social media data and complex emotional signals. PCA and Factor Analysis are well-established in financial research, capable of accurately extracting common and specific components from signals. The open-source LLMs selected for this study can be directly utilized after filling out relevant information on the corresponding websites.

Data Acquisition and Processing: We obtain data from platforms such as Twitter, StockTwits, and Seeking Alpha, which offer good accessibility. Using professional tools like Social Market Analytics and Ravenpack, we efficiently collect and process large volumes of social media sentiment and attention signals. Additionally, traditional news media information and company announcements provide necessary control variables for our research.

- Innovation

Cross-Platform Analysis: Most existing research focuses on a single social media platform. In contrast, this study reveals the uniqueness of different platforms in generating market-related information through cross-platform comparison. This approach not only fills a gap in the current literature but also provides new insights into the role of social media in financial markets.

Application of Large Language Models (LLMs): The introduction of LLMs for sentiment analysis represents a significant innovation in this study. LLMs are capable of identifying complex language structures and subtle emotional changes, enhancing the accuracy of cross-platform sentiment signal comparison. This provides a practical case for future applications of artificial intelligence in the financial field.

- Budget, Time, and Personnel Feasibility

Resource Allocation: The primary resources required for this study include data acquisition costs, computational resources, and software licensing fees. As we utilize publicly available or commercially accessible data sources, and the required computational resources can be flexibly obtained through cloud services, the budgetary demands for this study are relatively low.

Time and Personnel: The research project is expected to be completed within six months, with a reasonable and compact time schedule. It is believed that researchers will be able to overcome the challenges in the study and advance each stage efficiently and on time.

Reference

[1] Umar, Z., Gubareva, M., Yousaf, I., & Ali, S. (2021). A tale of company fundamentals vs sentiment driven pricing: The case of GameStop. Journal of Behavioral and Experimental Finance, 30, 100501.

[2] Wang, G., Wang, T., Wang, B., Sambasivan, D., Zhang, Z., Zheng, H., & Zhao, B. Y. (2015, February). Crowds on wall street: Extracting value from collaborative investing platforms. In Proceedings of the 18th ACM conference on computer supported cooperative work & social computing (pp. 17-30).

[3] Cookson, J. A., Lu, R., Mullins, W., & Niessner, M. (2024). The social signal. Journal of Financial Economics, 158, 103870.

[4] Delfanti, A. (2021). The financial market of ideas: A theory of academic social media. Social Studies of Science, 51(2), 259-276.

[5] Ren, J., Dong, H., Padmanabhan, B., & Nickerson, J. V. (2021). How does social media sentiment impact mass media sentiment? A study of news in the financial markets. Journal of the Association for Information Science and Technology, 72(9), 1183-1197.

[6] Oliveira, N., Cortez, P., & Areal, N. (2013). On the predictability of stock market behavior using stocktwits sentiment and posting volume. In Progress in Artificial Intelligence: 16th Portuguese Conference on Artificial Intelligence, EPIA 2013, Angra do Heroísmo, Azores, Portugal, September 9-12, 2013. Proceedings 16 (pp. 355-365). Springer Berlin Heidelberg.

[7] McLuhan, M. (1975). McLuhan's Laws of the Media. Technology and culture, 16(1), 74-78.