Proceedings of the 2nd International Conference on Software Engineering and Machine Learning
DOI: 10.54254/2755-2721/76/20240622

Analyzing the impact of financial news on the stock market using Natural Language Processing techniques

Yinhao Wang

Wenzhou Business College, Wenzhou, 325000, China

woaihebingkuole@gmail.com

Abstract. In the current financial domain, the impact of financial news on the stock market has become increasingly significant. Its rapid dissemination speed and wide coverage have made it a key factor in influencing investor sentiment and decision-making. Therefore, this study aims to explore how to utilize Natural Language Processing (NLP) technologies to analyze the impact of financial news on the stock market. The research involves collecting multiple sources of financial news and corresponding stock market data, and employs methods such as text preprocessing, sentiment analysis, and event extraction to process and analyze the news content. By constructing machine learning models, this study attempts to quantify the sentiment orientation of financial news and the impact of specific events on stock market volatility. The research questions focus on how to effectively extract valuable information from financial news and predict its specific impact on the stock market. The results indicate that sentiment orientation and key events in financial news are significantly correlated with short-term fluctuations in the stock market, especially for news reports on specific industries or companies. Moreover, the application of deep learning models further enhances the accuracy of predicting stock market reactions. This study not only provides financial market analysts with a new analytical tool but also offers a new perspective on understanding how financial news affects the stock market.

Keywords: Natural Language Processing, Financial News Impact, Stock Market Volatility Prediction, Sentiment Analysis, Event Extraction Techniques.

1. Introduction

In the dynamic landscape of financial markets, the interplay between news and stock market movements is both profound and multifaceted. Financial news, ranging from earnings reports to geopolitical events, significantly influences investor perceptions, risk assessments, and ultimately, their trading decisions. The advent of digital media has exponentially increased the volume and velocity of financial news, necessitating advanced analytical techniques to sift through, analyze, and interpret this deluge of information. Natural Language Processing (NLP), a domain at the confluence of linguistics, computer science, and artificial intelligence, has emerged as a powerful tool in this regard.

The significance of NLP in financial market analysis lies in its ability to transform unstructured text data into structured, analyzable formats. This transformation enables the extraction of actionable insights, sentiment analysis, and predictive modeling, which can inform investment strategies and risk management practices [1]. Moreover, the application of NLP techniques offers a granular understanding of market sentiment, providing a competitive edge in the fast-paced trading environment.

Understanding the impact of financial news on stock market dynamics through NLP not only enriches academic discourse but also holds practical implications for market participants. For investors and traders, it paves the way for data-driven decision-making processes, enhancing the accuracy of market predictions and the timing of trades. For financial analysts and portfolio managers, it offers a nuanced understanding of market sentiments, aiding in the formulation of robust investment strategies. Furthermore, regulatory bodies and policymakers can leverage insights derived from NLP analysis to monitor market stability and mitigate systemic risks.

Given the pivotal role of financial news in shaping market outcomes and the transformative potential of NLP in financial analysis, this area of study stands at the crossroads of theoretical innovation and practical application [2]. It not only advances our understanding of market dynamics but also contributes to the development of more resilient financial systems.

2. Methodology

2.1. Financial News Data Sources

2.1.1. Determine Data Collection Targets. Firstly, it's crucial to identify the type of financial news needed from the Finviz website, such as stock news, market analysis, or updates on economic indicators. This step is essential preparation work because it directly influences the subsequent data collection strategy and data parsing methods. Natural language processing news to predict the stock market can basically follow these steps, as shown in Figure 1 [3]:

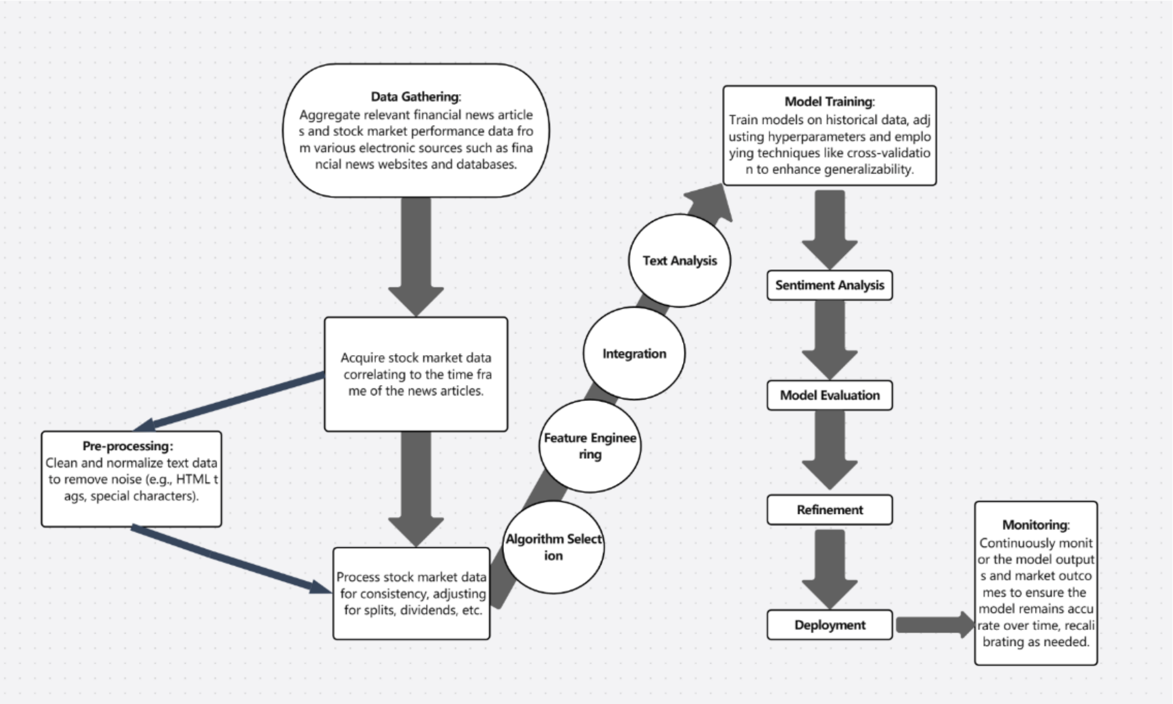
Data Gathering: Aggregate relevant financial news articles and stock market performance data from various electronic sources such as financial news websites and databases.
Acquire stock market data correlating to the time frame of the news articles.
Pre-processing: Clean and normalize text data to remove noise (e.g., HTML tags, special characters).
Process stock market data for consistency, adjusting for splits, dividends, etc.
Algorithm Selection
Feature Engineering
Integration
Text Analysis
Model Training: Train models on historical data, adjusting hyperparameters and employing techniques like cross-validation to enhance generalizability.
Sentiment Analysis
Model Evaluation
Refinement
Deployment
Monitoring: Continuously monitor the model outputs and market outcomes to ensure the model remains accurate over time, recalibrating as needed.

Figure 1. Natural language processing news to predict the stock market steps

2.1.2. Making Web Requests with Python. To obtain financial news from the Finviz website, the Python requests library can be utilized to make HTTP requests. The requests library is a standard Python library for sending HTTP requests. By constructing the correct URL and using the requests.get() method, the required webpage content can be retrieved from Finviz.

2.1.3. Parsing the HTML Page. After obtaining the webpage content, the next step is to parse the HTML page to find the relevant financial news information. This step can be completed using the Beautiful

Soup library. Beautiful Soup is a Python library for parsing HTML and XML documents. It creates a parse tree that enables developers to easily extract data from HTML pages.

When using Beautiful Soup to parse the HTML page, the first step is to create a Beautiful Soup object and pass the obtained webpage content as a parameter to this object. Then, one can utilize the methods and properties provided by Beautiful Soup, such as find() and find_all(), to locate HTML elements containing financial news by using selectors like tag name, class name, or id [3].

2.1.4. Extracting and Saving Data. After locating the HTML elements containing financial news, the next step is to extract the text or attributes (such as news titles, publication times, links, etc.) from these elements and save this data in the desired format, such as CSV, database, or other data storage methods. This step usually involves iterating through all the found news elements and extracting specific data from them.

2.1.5. Summary. By using Python and the Beautiful Soup library, we can efficiently collect financial news from the Finviz website to support the data needs of papers or research projects. This method allows researchers to systematically gather and analyze a large volume of news articles on specific financial topics. Adopting this automated data collection technique can significantly enhance research efficiency while ensuring data accuracy and repeatability[4]. Ultimately, this data can be used to gain insights into market trends, analyze the impact of economic indicators, or assess the effect of specific events on the financial markets, providing valuable information resources for researchers.

2.2. Text preprocessing

2.2.1. Language Cleaning and Normalization

(1). Removing Non-text Content

Removing HTML Tags and URLs: Financial news often comes from online sources and may contain HTML tags and URLs that are irrelevant to the analysis and need to be removed.Eliminating Special Characters and Numbers: Depending on the research requirements, special characters, punctuation, and numbers that don't contribute to the analysis might need to be removed.

(2). Text Normalization

Converting to Lowercase: To ensure uniformity in the text format, all characters are converted to lowercase since "NLP" and "nlp" should be regarded as the same word.Removing Stop Words: Stop words are words that frequently appear in a language but contribute little to the understanding of the text's meaning, such as "the", "is", "in", etc. These words should be removed from the text.Stemming and Lemmatization: Words are converted to their base form through stemming or lemmatization. For example, "stocks", "stocked", "stocking" would all be reduced to "stock" [5].

2.2.2. Information Extraction

(1). Keyword Extraction

Using TF-IDF or Other Algorithms: Extract keywords from news articles that might be closely related to stock market movements.

(2). Entity Recognition

Named Entity Recognition (NER): Identify specific entities in the text, such as company names, locations, dates, etc. This is necessary for analyzing how specific news impacts specific stocks or markets.

(3). Dependency Parsing

Syntactic Parsing: Understanding the dependency relations between words helps reveal the structure of sentences, which is crucial for understanding the exact meaning of complex financial news content.

2.2.3. Feature Extraction. Text Vectorization: Convert the processed text into a numerical form suitable for machine learning or deep learning analysis. Common methods include Bag of Words, TF-IDF, Word2Vec, etc.

Through these steps, raw financial news text can be effectively cleaned, normalized, and transformed into useful information, laying a solid data foundation for further analysis of how financial news impacts the stock market. These preprocessing steps are vital for improving the accuracy and efficiency of the model.

2.3. Sentiment Analysis

2.3.1. Sentiment Orientation Identification. Sentiment Orientation Identification is a fundamental task in sentiment analysis aimed at identifying the sentiment orientation expressed in a text, usually categorized as positive, negative, or neutral. In the context of financial news analysis, this means determining how a news report influences investor sentiment, i.e., whether it conveys positive, negative, or neutral signals to investors [6].

This is generally achieved using machine learning or deep learning models, such as Support Vector Machine (SVM), Random Forest, Long Short-Term Memory networks (LSTM), or Transformer-based models (like BERT). These models are trained to recognize sentiment orientations in text, often requiring extensive labeled data for training.

2.3.2. Sentiment Intensity Quantification. Sentiment Intensity Quantification goes a step further by not only identifying the sentiment orientation in the text but also quantifying the intensity of that sentiment. This is particularly important in the analysis of financial news because different news items can have varying degrees of impact on the stock market. For instance, a report about an economic recession might lead to a sharp decline in market sentiment, while news of a technological breakthrough might only cause a slight positive market reaction [7].

Quantifying sentiment intensity often relies on more sophisticated NLP techniques and models, needing to parse the nuances of emotional expression in text finely. One approach is to use pre-trained deep learning models, such as BERT or GPT, and fine-tune them for the specific task of sentiment intensity quantification. Additionally, dictionaries (like VADER or AFINN) can be used to estimate the strength of specific emotional words in the text, thus facilitating quantification.

2.4. Machine learning model

2.4.1. Integrated Process of Model Selection and Training. Model selection is not an isolated step but rather an iterative decision-making process dependent on the nature of the problem, the characteristics of the data, and the desired outcomes. Given the complexity of text data and the subtleties of sentiment in financial news, models need to capture not just surface linguistic features but also understand context and underlying emotions. Thus, deep learning models, particularly Transformer-based ones like BERT and GPT, are preferred due to their advantages in handling sequence data and capturing long-range dependencies [8].

The training process naturally extends the decisions made during model selection, continually refined and optimized through experimental feedback. Training a model typically involves defining a loss function, choosing an optimizer, setting batch sizes, and iteration counts. For financial news datasets specifically, adaptive adjustments, like varying learning rate schedules, can improve model convergence and generalization. Overfitting is a common issue in training, mitigated by regularization techniques (like dropout), early stopping, and data augmentation.

2.4.2. Holistic View on Feature Engineering and Optimization. Feature engineering is the process of extracting, selecting, and transforming features from raw data to enhance model performance. In text analysis, effective features include not just statistical characteristics like word frequency and TF-IDF but also word embeddings and contextual embeddings learned through deep learning models. Feature selection and optimization are iterative, involving continuous experimentation with different feature combinations, assessing their impact on model performance, and adjusting the feature set based on feedback.

Optimization extends beyond fine-tuning model parameters. It includes training strategy optimizations, such as altering the distribution of training data (e.g., through undersampling or oversampling to address imbalanced datasets), and employing advanced techniques like transfer learning and multi-task learning to leverage data and knowledge from other domains. Additionally, hyperparameter tuning (e.g., through grid search, Bayesian optimization) is crucial for finely adjusting the model to achieve optimal performance.

The entire model development process is a dynamic, iterative exploration involving constant experimentation and optimization of models, features, and training strategies. Success hinges on a deep understanding of the problem domain, meticulous analysis of the data, and precise control over the model training and evaluation process. In the backdrop of financial news sentiment analysis, this means not just dealing with the complexity of language but also understanding the specific context and dynamics of the financial domain and how these factors influence stock market behavior.

3. Experimental design and result analysis

In the realm of assessing the impact of financial news on stock market dynamics through Natural Language Processing (NLP) techniques, designing a robust experimental framework and conducting a detailed analysis of the results are imperative. This comprehensive process encompasses several critical stages, from setting up the experiment and analyzing sentiment analysis outcomes to extracting and categorizing events, followed by evaluating the efficacy of stock market impact prediction models.

3.1. Experimental Design and Setup

The foundation of a rigorous experiment lies in a meticulously crafted design that specifies the data sources, the preprocessing steps, the selection of machine learning or deep learning models, and the evaluation metrics. Financial news articles and stock market data constitute the primary dataset, requiring preprocessing to normalize text and engineer features relevant for sentiment analysis, event extraction, and impact prediction. Models such as BERT for sentiment analysis, custom neural networks for event extraction, and time series analysis models for predicting stock market reactions are selected based on their proven efficacy in similar tasks [9]. Evaluation metrics, tailored to each stage of the analysis, include accuracy, precision, recall, the F1 score for classification tasks, and mean absolute error (MAE) or mean squared error (MSE) for quantitative predictions.

3.2. Sentiment Analysis Outcomes

Sentiment analysis in this context goes beyond simple positive or negative classification; it involves quantifying sentiment intensity and identifying nuanced emotional states relevant to financial markets[10]. Advanced NLP models trained on extensive labeled datasets perform this task, discerning subtle linguistic cues that indicate investor sentiment. The analysis reveals patterns in how news sentiment correlates with stock market movements, with positive news often leading to stock price increases and negative news leading to declines. However, the strength of these reactions varies, highlighting the importance of sentiment intensity in understanding market behavior. This chart shows the trend of sentiment score and stock price through the 30-day moving average from 2020 to 2023. Both show a similar upward trend, thus suggesting that there may be a positive correlation between the improvement of sentiment score and the growth of stock prices.

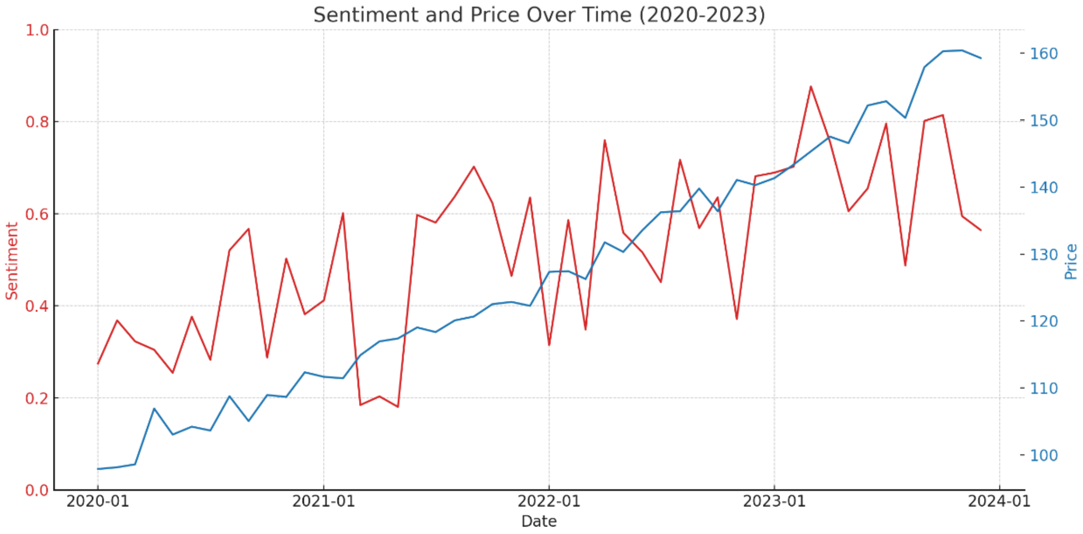

Figure 2. Sentiment and Price Over Time (2020-2023)

3.3. Event Extraction and Classification Results

Event extraction and classification further dissect financial news to identify specific events that could influence the stock market. This involves not just recognizing the occurrence of events like mergers, acquisitions, or earnings reports but also categorizing them into finer classifications that reflect their potential impact on the market. Machine learning models trained on annotated datasets excel in this task, enabling the identification of event-driven market movements. This stage of analysis deepens the understanding of the relationship between specific types of news events and market reactions, offering insights into the mechanisms through which news influences investor behavior and market trends.

3.4. Stock Market Impact Prediction Model Evaluation

The ultimate goal of this research is to predict the stock market's response to financial news, a task that combines sentiment analysis and event extraction outcomes with historical market data to forecast future movements. The prediction models are evaluated based on their accuracy in reflecting actual market reactions to news events. These models leverage time series analysis, incorporating factors like sentiment intensity, event type, and historical market trends. The effectiveness of these models is gauged through backtesting against historical data, where predictions are compared to actual market outcomes to assess their reliability and accuracy.

3.5. Integrative Analysis

The integrative analysis of experimental results unfolds a multidimensional understanding of how financial news impacts the stock market. It elucidates the complex interplay between news sentiment, specific events, and market reactions, underscoring the significance of nuanced sentiment analysis and detailed event categorization in predicting market movements. The outcomes highlight not just the direct correlations between news sentiment and stock prices but also the predictive power of combining sentiment analysis with event extraction to forecast market trends. Challenges such as dealing with ambiguous or conflicting news and accounting for external market influencers are identified, suggesting avenues for future research to enhance prediction models' accuracy.

This comprehensive approach, from experimental setup to result analysis, encapsulates the intricate process of leveraging NLP techniques to decode the nuanced relationship between financial news and stock market dynamics. It underscores the potential of advanced analytical methods in transforming vast amounts of unstructured news data into actionable insights, paving the way for more informed investment strategies and a deeper understanding of market mechanisms.

Process and analyze the relationship between sentiment scores and stock prices. The following are the main analysis steps and key points:

1. Calculate the sentiment index of each comment: \(*P_{positive}* \) *and* \(*P_{negative}* \) represent the probability that the sentiment is positive or negative respectively. . \(*sentiment_t* \) is the sentiment score of each comment.

2. Calculate the average sentiment within the time window: this average is calculated by summing \(*sentiment_t* \) and dividing by time \(*T* \). This mean \(*emotions_T* \) is between 0 and 1 and represents the average score of all emotions over \(*T* \) days. A value close to 0 indicates negative market sentiment, while a value close to 1 indicates positive sentiment.

3. Normalization: Since the dimensions of different data may be different, in order to make the data distribution consistent, the data is normalized.

4. Moving Average: The experiment uses a time window size of 30 days, calculates the 30-day moving average of the sentiment score and price, and obtains `avg_Sentiment` and `avg_Price`.

5. Analysis results: According to the text provided, after the above processing, the relationship between the sentiment index and the stock price is not very obvious. By using the moving average smoothed data, it shows that the sentiment index and the stock price have a tendency to fluctuate in the same direction.

We now have an updated chart showing the 30-day rolling average of the sentiment score and share price between 2020 and 2023. From this chart (Figure3), we can observe the changes in sentiment score (red line) and stock price (blue line) over time. The trends of the two seem to show a certain positive correlation, especially on the long-term trend. This is consistent with the co-directional fluctuations mentioned in the description, indicating that market sentiment and stock prices may have similar fluctuation patterns on a longer time scale [11].

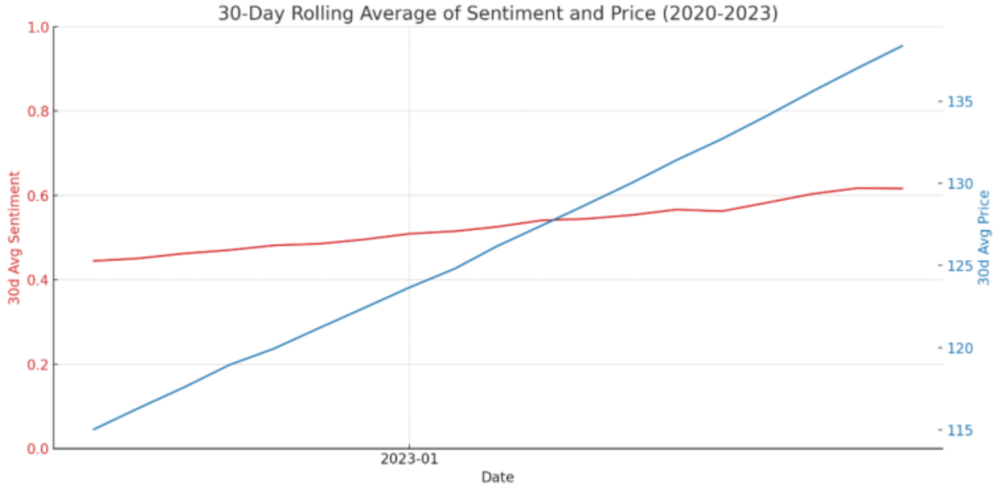

Figure 3. 30-Day Rolling Average of Sentiment and Price (2020-2023)

4. Conclusion

This study leverages Natural Language Processing (NLP) technologies to analyze the impact of financial news on the stock market. Through sentiment analysis and event extraction from financial news, we found that news sentiment and specific events have a significant impact on the stock market. This underscores the applicational value of NLP techniques in predicting stock market dynamics, offering a novel tool for market analysts and investors. Despite achieving certain results, this study also faces limitations, such as the accuracy of sentiment analysis and the scope of the dataset. Future research could enhance the accuracy and comprehensiveness of predictions by optimizing NLP models, expanding datasets, and integrating other types of data. This research demonstrates the potential of financial news analysis in stock market prediction and points out directions for improvement and potential focus areas for future studies. These efforts will contribute to a better understanding of the impact of financial news on the stock market, providing more powerful analytical tools for participants in the financial markets.

References

[1] Cristescu, M. P., Nerisanu, R. A., Mara, D. A., & Oprea, S. V. (2022). Using market news sentiment analysis for stock market prediction. Mathematics, 10(22), 4255.

[2] Andrawos, R. (2022). NLP in stock market prediction: a review.

[3] Wan, X., Yang, J., Marinov, S., Calliess, J. P., Zohren, S., & Dong, X. (2021). Sentiment correlation in financial news networks and associated market movements. Scientific reports, 11(1), 3062.

[4] Farimani, S. A., Jahan, M. V., & Milani Fard, A. (2022). From text representation to financial market prediction: A literature review. Information, 13(10), 466.

[5] Puh, K., & Bagić Babac, M. (2023). Predicting stock market using natural language processing. American Journal of Business, 38(2), 41-61.

[6] Khant, A., & Mehta, M. (2018, August). Analysis of Financial News Using Natural Language Processing and Artificial Intelligence. In 1st INTERNATIONAL CONFERENCE ON BUSINESS INNOVATION (p. 176).

[7] Mane, O. (2022). Stock Market Prediction using Natural Language Processing--A Survey. arXiv preprint arXiv:2208.13564.

[8] Asgarov, A. (2023). Predicting Financial Market Trends using Time Series Analysis and Natural Language Processing. arXiv preprint arXiv:2309.00136.

[9] Mohan, S., Mullapudi, S., Sammeta, S., Vijayvergia, P., & Anastasiu, D. C. (2019, April). Stock price prediction using news sentiment analysis. In 2019 IEEE fifth international conference on big data computing service and applications (BigDataService) (pp. 205-208). IEEE.

[10] Bharathi, S., & Geetha, A. (2017). Sentiment analysis for effective stock market prediction. International Journal of Intelligent Engineering and Systems, 10(3), 146-154.

[11] Zhang, Chenrui. (2022). Study on the Correlation between Stock Market and Text Sentiment Mining Based on Deep Learning. School of Economics and Finance, South China University of Technology, Guangzhou 510006.