Date of publication xxxx 00, 0000, date of current version xxxx 00, 0000.

Digital Object Identifier 10.1109/ACCESS.2024.Doi Number

*Modeling and Optimization of News-Stock Price Correlation Based on Topic Influence Selection*

Huaxi Liu1,#, Zonghan Jiang2,#, Desheng Li3,#, Lin Gao4,#,\*

1 School of Information Science and Technology, Guangdong University of Foreign Studies, Guangzhou, 510006 China

2 School of Information Science and Technology, Guangdong University of Foreign Studies, Guangzhou, 510006 China

3 College of Photonic and Electronic Engineering, Fujian Normal University, Fuzhou, 350007 China

4 School of Software, Dalian University of Technology, Dalian, 116620, China

Corresponding author: Lin Gao (e-mail: notch@mail.dlut.edu.cn).

#co-first authors

ABSTRACT In this study, we investigate and quantify the impact of different topics of news on stock market dynamics. We classified 49,000 news headlines for Dow Jones Industrial Average (DJIA) from 2008 to 2016 into distinct categories (e.g., politics, technology, entertainment) and categorized each into three sentiment categories (negative, neutral, positive). Granger causality tests were employed to quantify the relationships between news categories and subsequent stock market fluctuations, while time-period segmentation analysis was conducted to validate the robustness of the Granger causality results. Our findings reveal that specific news categories significantly influence stock market trends at 1 to 3 days lag. Notably, negative entertainment news exhibits a substantial impact (P-value < 0.02), while negative technology-related news (mean P-value < 0.18) and positive health fitness news (mean P-value < 0.17) also demonstrate statistically significant correlations. These results underscore the heterogeneous effects of news sentiment across domains on financial markets, providing actionable insights for investors and policymakers.

INDEX TERMS Stock market, news topics, sentiment analysis

1. INTRODUCTION

With the increasing complexity of financial markets and the rapid speed of information dissemination, the dynamic correlation between news texts and stock prices has gradually become a research hotspot in both academia and industry. A large body of research has shown that the sentiment orientation, topic distribution, and the influence of news events drive stock price fluctuations to some extent. News not only influences the stock market through sentiment but also, to a large extent, shapes market trends through the variation and dissemination of topics (Ren, 2023). However, existing studies mainly focus on sentiment analysis or time series modeling, with less in-depth exploration of the influence selection of news topics and their deeper correlations with stock prices. In particular, the lack of attention to the heterogeneous impact of different topics on stock prices limits the ability of existing models to understand market dynamics (Wang and Li, 2025; Minaee et al., 2021).

The core goal of this study is to explore the correlation between news topics and stock prices, and to analyze the impact of news sentiment and topics on the market. Specifically, this study aims to identify influential news topics through a staged modeling framework and then analyze the relationship between these topics, news sentiment, temporal factors, and stock price fluctuations. The significance of the study lies in three aspects: first, in the theoretical aspect, we propose a framework for quantifying "topic influence" based on topic-stock price causality, providing new perspectives for current research on topic heterogeneity analysis; second, in the methodological aspect, we combine natural language processing techniques, statistical methods (such as Granger causality test), and deep learning models (such as FinBERT or LSTM) to analyze news text and stock market data; finally, in the practical application aspect, this research helps to identify news topics and sentiment that have a significant impact on stock prices, providing valuable insights for understanding market dynamics. This study has the following innovations. First, in the quantification of topic influence, we propose a weight distribution method based on topic-stock price causality (using Granger causality test), overcoming the limitations of traditional topic modeling and enabling a more accurate identification of key topics related to stock price movements. Second, we utilize the FinBERT model for news sentiment analysis and combine it with time series analysis methods (such as CEEMDAN and SC-LSTM) to explore the potential links between sentiment, topics, temporal factors, and stock prices.This research aims to provide innovative theories and methods for financial text analysis and the study of its relationship with stock prices, especially in news topic selection and influence analysis, offering new perspectives in the field. We hope that this study will provide financial market participants with methods and tools to better understand the relationship between news information and stock prices, and provide a solid theoretical foundation for future financial market analysis.

1. RELATED WORK
2. Subject classification

Subject Classification is a process of systematically organizing and categorizing information resources according to their subject or disciplinary field of content. Subject classification methods can be roughly divided into four categories according to the amount of data and explanatory needs: Rule based, Traditional machine learning, Topic model, and Deep learning (Table 1.).

The rule-based classification method uses manual rule making for text classification, which is highly interpretative when the data volume is small. However, a study that classified emotions using expert constructed word lexicons showed that the reliance on manual work led to the inefficiency of this method (Atmadja & Purwarianti, 2015). So, the research of this method nowadays starts to focus on combining other algorithms to reduce human resources. For example, Cui et al. (2019) constructed a classifier based on regular expressions to replace the work of experts and improve classification performance.

Traditional machine learning methods are more based on statistical methods to deal with medium-sized labeled data. Common algorithms for topic classification include: Naive Bayes (McCallum & Nigam, 1998), Support Vector Machine (SVM) (Joachims, 1998), Term Frequency-Inverse Document Frequency (TF-IDF) (Baeza-Yates & Ribeiro-Neto, 1999; Wang & Manning, 2012), Chi-Square Test (Yang & Pedersen, 1997), Mutual Information (MI) (Forman, 2003), and Maximum Entropy Model (Berger et al., 1996). These algorithms have different characteristics and different scenarios. In recent years, in order to optimize the algorithm structure to adapt to a wider range of scenarios, more and more research began to try to combine different algorithms to improve the classification ability. By combining Bayesian methods with mutual information, Nurfikri et al. (2018) proposed a news topic classification model that exhibits superior performance. There is also a study that combines SVM algorithm with information gain feature selection method to optimize traditional topic classification (Rizaldy & Santoso, 2017).

Topic modeling method is an algorithm applied to unsupervised topic discovery scenarios. This algorithm is mostly used to mine potential topics and can also be used for topic classification. LDA, as one of the currently popular topic modeling methods, Omrani et al. (2023) used it to classify true and false news. Their bilingual model demonstrated higher accuracy and F1 score than previous studies. Nonnegative Matrix Factorization (NMF), which was proposed by Lee and Seung in the 1999 issue of Nature, can directly extract themes from data. (2003). On the basis of NMF, Tang et al. (2011) proposed a new method for data clustering and classification, which is superior to traditional classification algorithms.

With the arrival of the era of big data, the amount of data is becoming larger and larger, and the data structure is becoming more and more complex. Deep learning is powerful in processing long sequence text, especially suitable for text analysis tasks in big data environment (Vaswani et al., 2017). Minaee et al. (2021) gave a comprehensive overview of the application of deep learning in text classification, including different model architectures such as Convolutional Neural Network (CNN), Recurrent Neural Network (RNN), Long Short-Term Memory (LSTM), Gated Recurrent Unit (GRU) and Transformer, and discussed their advantages and challenges when dealing with large-scale text data. Faced with the drawback of low interpretability in deep learning (black box problem), there are currently many studies attempting to construct a more transparent machine learning system to alleviate it (Doshi Velez & Kim, 2017).

Table 1. Comparison of Common Topic Classification Algorithms.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Method Type** | **Data requirements** | **Interpretability** | **Applicable scenarios** | **Example References** |
| Rule based | low | high | simple structured data | |  | | --- | | Atmadja & Purwarianti, 2015; Cui et al., 2019 | |
| Traditional machine learning | medium | medium | small and medium-sized labeled data | McCallum & Nigam, 1998；Joachims, 1998 |
| Topic model | low | medium | unsupervised subject discovery | Omrani et al., 2023；Tang et al., 2011 |
| Deep learning | high | low | large scale complex text | Minaee et al., 2021；Vaswani et al., 2017 |

*Note: This table compares four common topic classification algorithms, highlighting their usage scenarios and features.*

1. Stock forecast and the impact of news

Stock prediction holds significant importance for both retail investors and professional analysts. With the development of technologies such as machine learning, new methods for predicting stock prices have emerged. In algorithmic stock price forecasting, techniques are generally categorized into predictive technologies and clustering-based technologies.

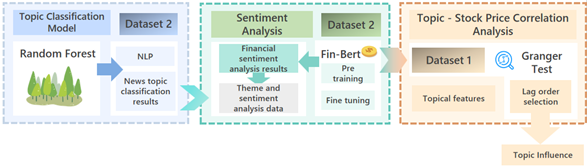
Ticknor (2013) designed an ANN model to capture characteristics from input stock variables through variance networks. Rout et al. (2014) developed a Computationally Efficient Functional Link ANN, enhancing generalization ability and incorporating technical indicators based on the previous work. Shrivas and Sharma (2018) employed multiple machine learning algorithms including SVM, but faced significant limitations regarding datasets. Hadavandi et al. (2010) created a hybrid approach for stock price forecasting by combining Genetic Fuzzy Systems with ANN. Vargas et al. (2017) utilized deep learning to predict market trends and directions. Xu et al. (2018) developed a Convolutional Neural Network that captures critical information from stock data and used Long Short-Term Memory Networks (LSTM) to learn context relationships within financial news for predicting stock market trends. Hsieh et al. (2011) proposed an integrated system ABC-RNN for forecasting, decomposing time series using wavelet transformation, processing selected input features with RNNs, and optimizing network weights and biases using the Artificial Bee Colony (ABC) algorithm, demonstrating promising research prospects for such methodologies. In the latest research of stock forecast. Zhu et al. (2024) introduced a new hybrid neural network model based on LSTM, demonstrating the superiority of the proposed model over benchmark models. Wang and Li (2025) proposed a stock return prediction method named NS-FEDformer, which enhances the model’s ability to extract sequence features.

Numerous studies have already shown that specific themes or public sentiment can affect the stock prices of related companies. With the advancement of technology, methods for predicting stock prices by analyzing expert comments, public sentiment, and market news have been developing. For different themes, the model needs to identify information related to stocks within the text. In more recent research, Ren (2023) used the P-P algorithm which combines BERT (Bidirectional Encoder Representations) with a local sentiment word database to analyze the relationship between sentiment words and stock prices. Cui and Huang (2024) proposed a method that uses lightweight language models to extract media news sentiment features and employs a GRU (Gated Recurrent Unit) deep learning network to predict stock prices based on historical data. Chakraborty (2024) introduces a method combining a two-stage convolution-LSTM neural network with LLM (Large Language Model, designed to understand and generate human-like text)for comprehensive stock consultation analysis. The model leverages the advantages of LSTM in analyzing time-series data and LLM in handling and understanding textual information.

In addition, numerous studies have identified and verified the impact of news on the stock market, particularly on investor reactions. Michael T. Kiley (2004) suggested that the market's response to economic news (e.g., unemployment rate announcements) varies depending on the context, especially whether investors' expectations have already incorporated the information beforehand. When news exceeds market expectations (i.e., unexpected information), short-term fluctuations may occur in the market, while delayed reactions are more commonly observed during long-term adjustments. Jonathan L. Rogers et al. (2009) found a significant delayed reaction of company-specific news (such as earnings announcements) on the market, particularly when biases exist within the information environment. Paul C. Tetlock (2007) emphasized that such delays might stem from irrational investor behavior or limitations in information processing capabilities. Therefore, our research not only needs to assess the immediate effects of news on the following day’s stock market to account for the impact of unexpected news but also must consider the long-term effects caused by delayed reactions.

Recent studies increasingly combine social media, financial news, and deep learning for market analysis and prediction. Belcastro et al. analyzed tweet activity (frequency, likes, retweets) to explore its link with cryptocurrency prices and predict price fluctuations using text analysis. Cantini, R et al. proposed TM-FID, which leverages news text and visual content via fine-tuned BERTweet and ViT models for misinformation detection and topic modeling, and introduced metrics to assess topic quality and cross-attention effectiveness. In another work, they fine-tuned BERT on few labeled tweets to identify COVID-19-related misinformation by topic. Krauss, C et al. developed a simple equal-weight ensemble of deep neural networks, gradient boosting, and random forests trained on the S&P 500 for one-day-ahead forecasting, improving returns and demonstrating the effectiveness of ensemble models in financial prediction.

1. Related Models and Methods



1. Flow chart of our work

Figure 1 depicts the staged modeling framework. To achieve this goal, this study proposes a staged modeling framework. First, in the topic influence selection stage, we use TF-IDF algorithm to extract potential topics from the news corpus and employ Random Forest model to classify daily news on various topics. Second, we employ the FinBert-based sentiment classification model to perform sentiment analysis on daily news, which is fine-tuned according to the features of our dataset. This process ultimately generates a quantitative distribution dataset containing daily news topics across various categories and their corresponding sentiment polarity quantitative distribution, serving as the input dataset for the subsequent Granger causality test. Third, through topic-price correlation analysis (such as Granger causality test), select the topics with significant influence. Next, the rationality of Granger causality test was tested through time segment analysis, and finally the degree of influence of 7 topics of news on the stock market was obtained.

Our team reckons that the staged modeling framework fits naturally with well-established approaches in both natural language processing (NLP) and financial econometrics. Methodologically, the modular and step-by-step design makes the process clear and easy to follow, with each stage built on mature and widely used techniques. The process of the first stage follows a traditional text mining pipeline, which ensures both interpretability and scalability. In the second stage, we incorporate a domain-specific FinBERT model for sentiment analysis, fine-tuned on our dataset. We have found that this step reflects a growing trend in financial text research, namely the development and application of language models specialized for the financial domain. In the third stage, we employ the Granger causality test to explore how topic-specific sentiment signals are linked to stock price fluctuations over time. We chose this method because it is a standard and trusted tool in econometrics for identifying predictive relationships between time series variables. In fact, many previous studies have successfully used Granger tests for this purpose. For example, Tetlock (2007) employed Granger causality tests to examine whether media pessimism predicts market outcomes such as returns and trading volume. Bollen et al. (2011) showed that social media mood variables derived from Twitter data Granger-cause stock market movements. Based on these precedents, we believe the Granger test is a solid and well-grounded choice for our analysis.

From a theoretical perspective, our framework connects with existing financial models but also brings several new elements. It is rooted in information-based asset pricing theory, which argues that markets respond to new information. However, instead of treating all news as a single undifferentiated signal, we separate it into topics and sentiment polarities, which we believe gives a more precise view of how different kinds of information influence the market. Compared with behavioral finance theories that focus on investor sentiment and psychological biases, our approach is quantitative and data-driven. By relying on a deep learning–based sentiment analysis model rather than surveys or heuristic measures, aligning with the modern behavioral finance trend of using big data and NLP methods.

Unlike traditional sentiment analysis models that only give an overall score for all news combined, our method identifies topic-specific sentiment signals and measures their individual causal impact on stock prices. In our view, this approach greatly improves the explanatory power and practical value of the results.

(修改点1)

1. Data Collection:

Our study investigates the impact of news topics on stock market trends and combines each topics feature with news emotion features to optimize the strategy of predicting DJIA movement through a multimodal analysis framework, utilizing two complementary datasets spanning news headlines and financial markets from Kaggle.

1. Daily Top 25 News Headline Dataset

The primary news corpus (Sun, J, 2016) comprises 2,949 daily records from Reddit's WorldNews Channel (June 8, 2008 - July 1, 2016), featuring the top 25 most engaged news headlines per date and corresponding log return based on DJIA. This curated selection mechanism ensures our analysis focuses on news items with demonstrated public attention significance. In our research, we will use a topic classification model to classify the news in this dataset, and calculate the proportion of each news topic per date to investigate the impact of each news topic on the stock market.

Table 2. Daily News for Stock Market Prediction

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Date** | **DJIA** | **Top1** | **Top2** | **……** | **Top25** |
| 2015/1/13 | -0.00154079 | China has just banned the burqa in its biggest Muslim city | US and EU.. politicians use Charlie Hebdo attack to call for more Internet surveillance -- Fusion | …… | There is a.. lava flow in Iceland the size of Manhattan |
| 2015/1/14 | -0.01064998 | Cameroon.. Army Kills 143 Boko Haram Fighters | Air France.. hands out copies of Charlie Hebdo on flights | …… | Number 2 on Al-Qaeda's Most Wanted List Sells French Fries in a Florida Mall Food Court |
| 2015/1/15 | -0.00612299 | Saudi man.. sentenced 10 years jail and weekly public canings for 5 months. He is guilty of setting up a public online forum for debate and discussion. | Children.. caged to keep the streets clean for the Pope: Police round up orphans and chain them in filth during pontiff's visit to Philippines | …… | Fossil found.. by P.E.I. boy fills gap in reptile evolution |
| 2015/1/16 | 0.01095890 | Saudi Arabia.. publicly beheads a woman in Mecca | Boko Haram.. Appears to Be Using Abducted Girls as Suicide Bombers | …… | Nigeria:.. Satellite images show horrific scale of Boko Haram attack on Baga |

*Note: This table shows the structure and content of the aforementioned data.*

1. News Topic Classification Dataset

To obtain a reliable topic classification model, we employ a labeled dataset (Elkomy et al., 2024) containing 199,707 news instances across 17 distinct thematic categories (e.g., politics, technology, business). This comprehensive taxonomy, developed through expert annotation and NLP validation, serves as the training foundation for our topic classification model which is utilized to classify the Daily Top 25 News Headlines.

Table 3. News Classification and Analysis using NLP

|  |  |  |  |
| --- | --- | --- | --- |
| **Date** | **Headline** | **Content** | **News Categories** |
| 19-09-2023 | Assam CM Himanta Biswa Sarma gets Singapore's top fellowship | Assam CM Himanta Biswa Sarma has been bestowed with Lee Kuan Yew Exchange Fellowship of Singapore, an official release said on Sunday. Sarma was also extended an invitation to visit Singapore as a Lee Kuan Yew Fellow for his "dedicated leadership in public works and development", the statement added. He has become the first Assam CM to receive this fellowship. | ['politics', 'national'] |
| 19-09-2023 | Haryana CM announces ₹50 lakh for family of jawan killed in J&K | Haryana CM Manohar Lal Khattar on Sunday met with the family of Major Aashish Dhonchak, who lost his life in Anantnag encounter in J&K and announced ₹50 lakh compensation. "Major Aashish Dhonchak was a promising young man. He reached the rank of Major in 11 years of his service," CM Khattar said after meeting his family. | ['national'] |
| 19-09-2023 | Man caught on video having 'unnatural sex' with dog in Thane, FIR filed | A man in his mid-fifties was allegedly caught on video while having unnatural sex with a dog in Thane's Mumbra, near Mumbai. The accused, identified as Karim, has been booked, Mumbra Police said, adding that he is currently absconding. The incident came to light after a local citizen recorded the alleged act and shared it with animal activists. | ['national'] |
| 19-09-2023 | Man skydives in US with PM’s pic on Indian map on bday, video out | Union Minister Jyotiraditya Scindia has shared a video showing a man skydiving in US with a picture of PM Narendra Modi on the Indian map on the occasion of the PM's birthday. Sharing the video, Scindia said, "The love and admiration that Indians hold in their hearts for PM Shri Narendra Modi Ji is indeed special." | ['world', 'national'] |

*Note: This table shows the structure and content of the aforementioned data.*

1. Financial Sentiment Dataset

If deep learning methods are employed to analyze financial sentiment in text, the model must first learn from data with reliable annotations, adjusting the parameters of the neural network to achieve effective predictive performance. We identified a specific dataset within the TensorFlow dataset repository—the Hugging Face Financial PhraseBank dataset. The Financial PhraseBank is a dataset designed for financial sentiment analysis and is commonly used to train and evaluate natural language processing models on text classification tasks within the financial domain. After examining the structure and features of the textual data, we found that this dataset closely resembles "Daily News Top" in terms of structural format, sentence length, and relevance to financial content. Therefore, it is highly suitable to use this dataset as a foundation for fine-tuning FinBERT to adapt to the objectives of our research.

The dataset encompasses approximately 5,000 short financial statements. These sentences were annotated by thirteen master's students majoring in finance and three researchers, who considered only the potential positive, neutral, or negative impact of the news on stock prices from an investor’s perspective. A majority voting system was applied to determine the final sentiment label for each piece of news. The reliability of sentiment annotation in this dataset is expected to be significantly higher than those labeled by researchers outside the financial field or by large language models.

The original dataset contains versions with varying degrees of agreement among annotators. For instance, the subset *'sentences\_75agree'* includes only sentences where at least 75% of annotators reached a consensus. However, we selected the subset *'sentences\_allagree'*, which consists of sentences where all annotators agreed. The certainty of labels and the presence of key linguistic features in this subset make it particularly beneficial for model fine-tuning.

Table 4. News and corresponding financial sentiment

|  |  |
| --- | --- |
| **Content** | **Label** |
| In the third quarter of 2010, net sales increased by 5.2 % to EUR 205.5 mn, and operating profit by 34.9 % to EUR 23.5 mn. | positive |
| According to Gran, the company has no plans to move all production to Russia, although that is where the company is growing. | neutral |
| Pretax profit totaled EUR 9.0 mn, down from EUR 36.3 mn in 2007. | negative |

*Note: This table displays the structure formed after extracting the original dataset.*

1. Discussion on data applicability

For a long time, the relationship between news narratives and stock market trends has been a topic of great concern in financial research. To explore this connection, our research utilized data from 2008 to 2016, a period of market turmoil and major economic events such as the global financial crisis and the European sovereign debt crisis. During this period, traditional financial news media, including Reuters, Bloomberg and CNN Finance, were the main sources of market information for investors. These centralized and editorial-controlled sites offer consistent, authoritative and verifiable content, providing an ideal structured environment for analyzing the causal relationship between news sentiment and thematic coverage and stock indices such as the DJIA.

Although this dataset is nearly a decade earlier than the current (2025) financial information landscape, it provides historical validity for establishing baseline models and control experiments. Our team believes that the results derived from this dataset must be interpreted in the historical context in which they emerged. It can well explain the impact of the release of news on different themes between 2008 and 2016 on the changes in the stock market. However, in today's different market environment, this research result needs to be applied more carefully, or newer datasets should be added to enhance the usability of the experimental results.

A notable limitation of this study is the temporal scope of the dataset, which covers the period from June 2008 to July 2016. The landscape of financial information dissemination has evolved significantly since this period, particularly with the rise of social media and real-time news platforms, which have accelerated the speed and altered the dynamics of market-moving information. Consequently, the representative power of this dataset, while substantial for its time, may be limited in capturing the nuances of the current market environment. The patterns of influence and investor reaction times identified in our analysis might differ in today's more interconnected and algorithmically driven markets. Therefore, a crucial direction for future research is to apply and validate this framework using more recent data. Such an extension would not only test the robustness of our findings across different market regimes but also enhance the model's overall relevance and contextual applicability.(修改点2)

1. Topic Classification Modeling
2. Natural Language Processing

First of all, we need to train models that can classify news headlines into topics, the first step is to process the textual data, which will become digital vectors through a series of conversion techniques in natural language, so that they can be processed by machine learning techniques. Some NLP processes are required.

**Text Cleaning**

The text data we have is garbled, missing removed, and each news headline is converted to lowercase, non-alphabetic characters are removed, and word splitting is done to make each sentence in the original text data into a list structure consisting of words, here for word splitting we use punkt tool which is based on the Moses word splitting algorithm. After that, the converted data are deactivated, deactivated words refer to words that are very common in natural language text, they usually do not carry specific meanings, such as "the" , "a" , "an" , "in" and so on. In text categorization, these words may interfere with the training effect of the model, so they need to be removed from the text. After that, stemming is performed on the data to remove the affixes from the words so that the stem can be recognized as a variant of the same word. It aims to group the various variants of a word into the same stem to improve accuracy and reduce the size of the feature space.

**Feature Extraction**

After that, the text is transformed into numerical feature vectors by TF-IDF (Term Frequency-Inverse Document Frequency) which is a statistical method for evaluating the importance of a word in a document or corpus that takes into account the frequency of a word in a particular document and its It takes into account the frequency of a word in a particular document and its prevalence in the whole collection of documents, thus assigning to each word a weight value indicating its relative importance in distinguishing different categories. It consists of two parts: word frequency (TF) and inverse document frequency (IDF).

TF is calculated by the following formula:

Where denotes each document, denotes the number of times word appears in document , so the denominator of the formula is the sum of the number of words in the document, and denotes the number of times word appears in document . This formula solves for the value corresponding to word in its document.

IDF is calculated as follows:

Where denotes the number of documents in which word appears and denotes the total number of documents, this variant from the standard formula smoothes out the encoded value for each word to avoid 0.

Afterwards, the encoded value of each word in a document is obtained by multiplying these two values, thus making the whole text a numerical feature vector.

1. Random Forest Classification

**Principle**

Random Forest is an integrated learning method that is mainly used for classification, regression and other tasks. It improves the accuracy and stability of the model by constructing multiple decision trees and combining their results. The underlying model is the decision tree model, which forms a tree structure by recursively partitioning the dataset into different subsets. Each internal node represents a differentiation by a certain feature, each branch represents a different going direction after differentiation by the node, and each leaf node represents a category prediction.

In the classification task, the splitting criterion of the random forest is to choose the best splitting point so that the Gini impurity is minimized, which is defined as:

where is the number of categories and is the probability that the sample belongs to the ith category.

The model constructs different training sets to train multi-class decision trees by randomly drawing samples from the input data, i.e. Bootstrap sampling. For each decision tree in the forest, instead of considering all the features in the feature selection of the split node, the best features are selected from a subset of the extracted features to reduce the correlation between the trees. A decision tree algorithm is applied to each Bootstrap sample to generate a decision tree. Since different data samples and feature subsets are used for each construction, each tree is unique. Ultimately, the one with the most predictions from all trees in the forest is selected as the classification result by a voting mechanism that combines the predictions of all trees in the forest. For validation, about one-third of the untrained data is used for performance evaluation, i.e., OOB estimation, which is an unbiased estimation method.

1. Model Evaluation

In this project, we compared the performance of different classification models, including the Random Forest model and the BERT deep learning model. The results revealed that, with the same amount of data, the Random Forest model (average accuracy reached 0.88) outperformed BERT (average accuracy reached 0.84) across most metrics, including precision, recall and f1-score. Consequently, we ultimately selected the Random Forest as the classification model.

Table 5. Topic Classification Accuracy By Random Forest

|  |  |  |  |
| --- | --- | --- | --- |
|  | precision | recall | f1-score |
| Health\_\_\_Fitness | 0.85 | 0.90 | 0.87 |
| business | 0.89 | 0.88 | 0.89 |
| entertainment | 0.84 | 0.86 | 0.85 |
| politics | 0.92 | 0.87 | 0.89 |
| science | 0.92 | 0.91 | 0.92 |
| sports | 0.80 | 0.81 | 0.80 |
| technology | 0.91 | 0.90 | 0.90 |

*Note: This table shows the evaluation indicators of the classification model in various topics.*

Table 6. Topic Classification Accuracy By BERT

|  |  |  |  |
| --- | --- | --- | --- |
|  | precision | recall | f1-score |
| Health\_\_\_Fitness | 0.86 | 0.81 | 0.83 |
| business | 0.84 | 0.86 | 0.85 |
| entertainment | 0.80 | 0.86 | 0.83 |
| politics | 0.86 | 0.86 | 0.86 |
| science | 0.87 | 0.88 | 0.88 |
| sports | 0.84 | 0.73 | 0.78 |
| technology | 0.81 | 0.80 | 0.81 |

*Note: This table shows the evaluation indicators of the classification model in various topics.*

1. Misclassification risks

There are many factors that may affect the misclassification risk of random forests. For example, the number of trees, the depth of each tree, the way features are selected, the imbalance of samples, etc. In addition, the risk of misclassification may be related to the variance and bias of the model. Because the random forest improves the generalization ability by reducing the variance, while a single decision tree may have a relatively high variance and is prone to overfitting. Reducing overfitting should help to decrease misclassification errors on the test set.

In this project, we adopted the method of class weight adjustment, assigning higher weights to the minority classes and reducing the probability of the minority classes being misclassified. Meanwhile, we also carried out model tuning and used grid search to find the optimal parameters of the model.

1. Sentiment Analysis Framework

The objective of this framework is to systematically quantify the sentiment polarity (positive/neutral/negative) and intensity associated with each financial news headline. This enables a granular examination of the interaction between emotional tone and news categorization in influencing market movements. By leveraging domain-specific natural language processing (NLP) models, the framework aims to enhance the accuracy of sentiment detection in financial contexts, thereby supporting the downstream task of assessing the correlation between news sentiment and stock price fluctuations.

1. Model Selection

**Model Architecture**

To address the challenges inherent in financial sentiment analysis, this study employs FinBERT, a transformer-based language model pre-trained on financial texts and fine-tuned on corpora such as earnings call transcripts, SEC filings, and Reuters financial news. FinBERT has demonstrated superior performance in interpreting domain-specific language compared to general-purpose sentiment models.

**Basic principles**

BERT uses a bidirectional Transformer encoder to pretrain text, which means it considers both the forward and backward context of a word during the training process. It uses a masking language model (MLM) to randomly mask some vocabulary in the input during the pre training stage and predict them, in order to learn the structure and semantic information of the language. In addition, Bert also performs the next sentence prediction, that is, given two sentences A and B, the model needs to determine whether B is the next sentence of A. This helps to understand the relationship between sentences.

1. Model Customization

**Model Adaptations**

To optimize FinBERT for headline-level sentiment analysis, the following customizations are introduced:

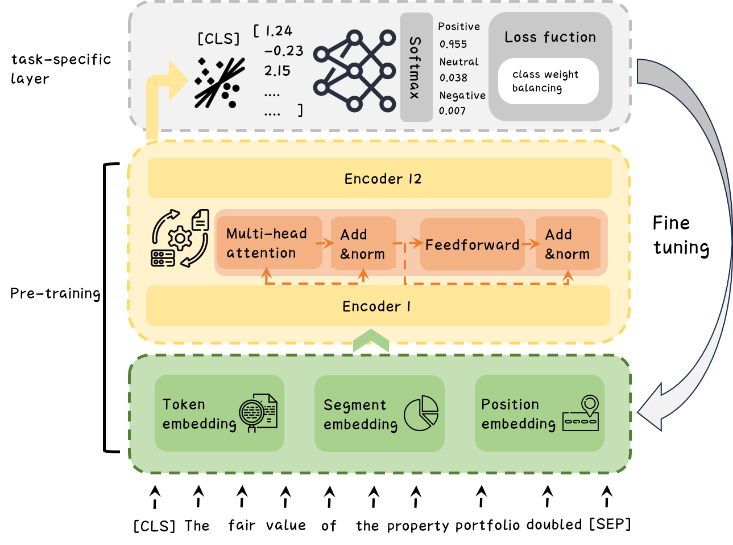
**Input Truncation**: Headlines are truncated or padded to a fixed length of 64 tokens, ensuring compatibility with FinBERT’s input limitations while preserving computational efficiency.

**Entity-Aware Attention Masking**: Named Entity Recognition (NER) is performed using SpaCy to identify and prioritize financial entities such as company names, stock indices, and monetary references. The FinBERT model relies on the Transformer architecture, with its core being the self attention mechanism. This mechanism allows the model to consider the information of all other words in the sentence when processing a word, thereby generating richer word representations. Entity Ware Attention Masking modifies the self attention mechanism so that when calculating attention scores between a word and other words, it additionally considers whether these words belong to pre-defined important entities. The specific implementation involve adding a marker to each word to indicate whether it belongs to the target entity category, and adjusting attention weights based on this marker. Attention masks are adjusted to emphasize these entities, enabling the model to focus on sentiment-bearing components of the headline.

**Reasoning Mechanism Fine-tuning**: The input text is processed by the FinBERT model, which outputs a vector containing probability distributions for three sentiment labels (positive, negative, neutral). These probabilities sum to 1, and the model would typically select the sentiment with the highest probability as the output. During the fine-tuning phase, since FinBert's training is based on company financial reports, analysis reports, and corporate report data—which rarely covers short sentence lengths or news genres—we further fine-tuned FinBert using the Financial PhraseBank dataset to better adapt it to our research task. First, we processed the data by understanding its storage structure and extracting the text along with corresponding labels, then mapped them to numerical values and removed any missing entries. Next, the dataset was split into training, validation, and test sets at a ratio of 81:9:10. We loaded the pre-trained model and tokenizer, which tokenizes the text using the WordPiece algorithm and vocabulary, encoding it into torch tensor format inputs with the structure [input\_ids, token\_type\_ids, attention\_mask, label].

In the downstream task, we obtain the hidden state for each token, i.e., a 768-dimensional vector, from the BERT model. Since financial sentiment classification is a sequence-level classification task focusing on the overall sentiment tendency of a sentence, the next step involves processing the hidden state of the [CLS] token to predict the sentence's financial sentiment (as the [CLS] token is specifically designed during pre-training to capture the global semantic information of the entire sequence). A linear hidden layer is added to map the hidden state to a 3-dimensional output, which is then transformed into a probability distribution via softmax to achieve classification for the three financial sentiment categories.

Subsequently, we set up the training parameters, established checkpoints and save points, defined the learning rate and batch size, and applied an early stopping strategy. Accuracy was selected as the evaluation metric, and training would halt if the metric showed no improvement or declined over three consecutive epochs. Additionally, an appropriate weight decay rate was determined to prevent overfitting. We employed cross-validation within the defined parameter space to identify the optimal parameter combination. Ultimately, the learning rate was set to, the batch size to 32, and the weight decay rate to 0.01.

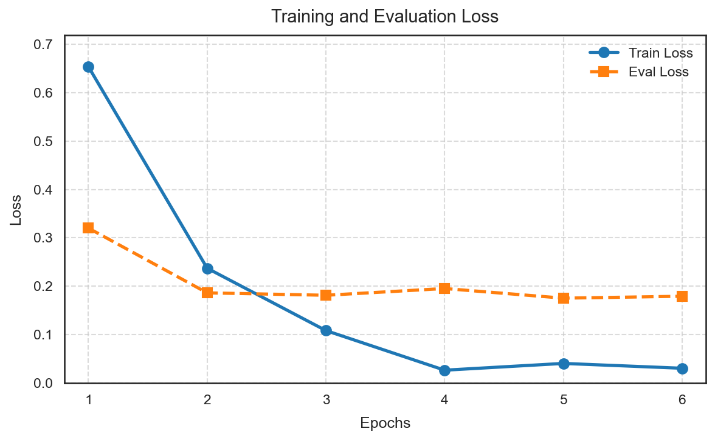


1. Model Architecture Diagram

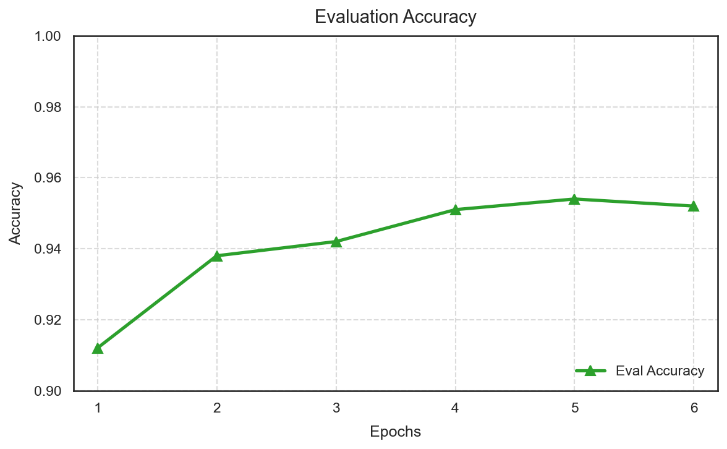
During fine-tuning, FinBert performs forward propagation (calculating loss), backpropagation (updating weights), and evaluates model performance. In each training iteration, the model adjusts its parameters based on the computed loss in order to minimize prediction errors. The cross-entropy loss function was used to calculate the error between predicted outputs and true labels.

The overrepresentation of neutral-class samples in the training data negatively impacts the accuracy and recall of minority classes and skews evaluation metrics due to majority-class dominance. To address this, we incorporate class weights into the cross-entropy loss function, assigning higher weights to underrepresented classes. This encourages the model to better recognize positive and negative financial sentiments by reducing bias toward the majority class.

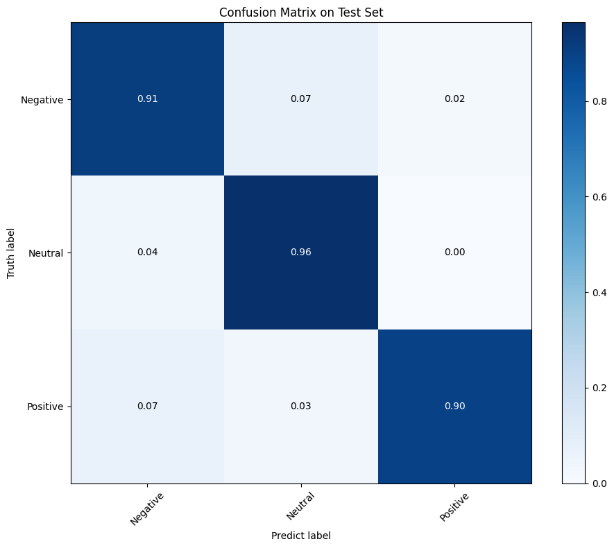
We tracked the changes in loss during fine-tuning and the accuracy on the validation set, observing that the model fits the financial text data well—the training loss gradually decreased over iterations, and the validation loss also showed a downward trend. To prevent overfitting, we employed weight decay and mini-batch training, and selected the optimal model at an appropriate number of iterations by monitoring evaluation metrics on both training and validation sets. The related parameter changes are shown below.



1. Training and Validation Loss Changes Plot



1. Validation Set Accuracy Changes Plot



1. The Confusion Matrix of The Test Set for The Optimal Model

The average accuracy achieved when directly using the FinBert model for prediction was approximately 93%, which increased to 98% after fine-tuning. This indicates that the fine-tuned model can effectively extract features specific to the news genre and accurately capture financial sentiment tendencies.

1. Time series analysis model
2. Data preparation
3. Data preprocessing

Using a topic classification model, the top 25 news items are categorized into 21 different topics (including three emotional categories for each of the seven thematic classifications). We count each theme for the day and remove columns other than count, date, and label. Due to the label column in the dataset being the DJIA for the current day, it is clearly meaningless for our prediction target. So we move the data in the Label column up one unit so that the Label column is a prediction for the next day. And the last column will be deleted due to the lack of predictable labels.

1. Stationarity test

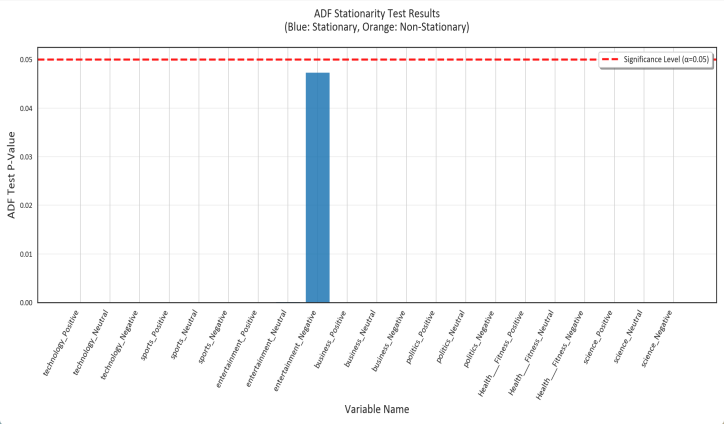
Due to the traditional Granger causality test requiring all variables to be stationary sequences, directly modeling non-stationary time series may lead to spurious regression problems. To ensure the effectiveness of causal inference, this article first conducts an augmented Dickey Fuller (ADF) test on all variables to identify their stationarity features.

ADF test is an extension of DF test, used to test **whether a time series has unit roots**. If the sequence contains unit roots, it is a non-stationary sequence with random walk characteristics. The ADF test constructs a regression model that includes lagged differences to test the null hypothesis that the sequence has a unit root, in order to determine whether it is stationary.

The ADF test considers the following linear model forms:

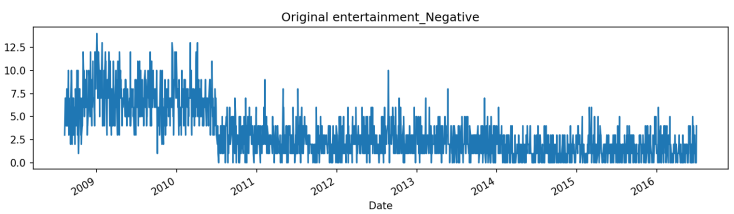
Among them, represents the first-order difference of time series , is a constant term, is the time trend term (if any), is the coefficient related to the original time series lagged by one period, is the coefficient of the lagged difference term, and H is the error term.

In this experiment, we conducted an Augmented Dickey Fuller (ADF) test on the dataset involved to evaluate the stationarity of each time series. The results showed that all categories of time series data rejected the unit root hypothesis (p-value<0.05), indicating that these sequences are stationary. However, it is worth noting that the ADF test p-value of time series data labeled with "entertainment-N egative" is relatively high, approaching the significance level (α=0.05). Nevertheless, it is still classified as a stationary sequence.



1. ADF Test Results Presented in P-value Format

To further understand the time series characteristics of the "entertainment-Negative" label, we have plotted its time series line graph. The results showed that the sequence had multiple periodic highs between 2009 and 2011, followed by an overall flattening of fluctuations. It is worth noting that the emergence of these short-term peaks coincides highly with several international events with widespread dissemination effects in terms of time. For example, the sudden death of Michael Jackson (2009) and the concentrated outbreak of the News International phone hacking scandal (2011) both sparked intense discussions worldwide at the time. Such events, due to their strong narrative tension and emotional impact, may significantly increase the intensity of negative emotional expression in "entertainment" related texts in a short period of time.



1. Time Series Chart of ‘entertainment-Negative’
2. Granger Causal Relation Test
3. Basic principles

The Granger causality test, proposed by Clive Granger, is used to determine causal relationships between time series variables. The basic assumption of Granger causality test is that if time series X has predictive ability for the future values of time series Y, then X is called the Granger cause of Y. This causal relationship does not imply a direct causal connection, but rather refers to the information contained in X that helps improve the accuracy of predicting Y. In our experiment, Y represents the rise and fall of DJIA (Label), while X represents the count of each news topic.

Granger's mathematical expression takes into account the following linear model form:

Autoregression:

Union-regression:

Where ,represent intercept terms; p,q represents the maximum lag order of Y and X respectively; , represent the parameters to be estimated; , represents the error term.

1. Inspection standards

By comparing the sum of squared residuals (RSS) of these two models, we further use an F-test to determine whether adding the lagged term of X significantly reduces the error variance.

The formula for F-test statistic is as follows:

is the sum of squared residuals of a model without X; is the sum of squared residuals of a model containing X; T is the sample size; K is the total number of parameters in the model containing X; q is the lag order of X.

If the calculated F value is greater than the critical value, reject the null and accept **the alternative that X provides Granger-causal predictive information for Y**. In addition, we also referenced the p-value of the results through the built-in functions of the statsmodels library. The P-value is calculated based on the F-statistic and its degrees of freedom, reflecting the probability of observing the current or more extreme result under the null hypothesis. If the P-value is below the chosen threshold, we reject the null and conclude that **X exhibits statistically significant Granger-causal predictive power for Y**.

1. Backtesting

To verify the feasibility and effectiveness of the Granger causality test results in practical decision-making scenarios, we conducted a backtesting experiment that incorporates common financial performance metrics (including annualized return, Sharpe ratio, and maximum drawdown) along with the p-values of Granger tests across different news categories.

In the backtesting experiment, we proposed the composite news sentiment signal , which integrates the p-values of news categories to design a simulated trading strategy. The concept of is to compress the “relative abnormality” across multidimensional news categories into a single scalar latent factor, representing the cross-topic, evidence-weighted shock intensity of daily news. This enables both dimensionality reduction and noise mitigation. When the composite signal is strong, the strategy takes a long position in the market; conversely, when the signal is negative, the strategy takes a short position. On one hand, this approach consolidates inputs from 21 news categories into a single indicator directly applicable for trading decisions; on the other hand, the weighting scheme emphasizes categories with stronger statistical evidence (p-value < 0.1), thereby concentrating the signal on themes with higher “information content.”

The calculation formula of the composite news sentiment signal is as follows

Where  denotes the set of news categories with high relevance (p-value < 0.1); represents the normalized static factor weight for each strongly associated news category; denotes the rolling z-score standardization of the daily difference in news counts for each category, which reflects the change or “surprise” in news flow and is often more closely aligned with market reactions. The interpretation of and is provided below.

Where represents a “strength score” aggregated across several lags , formed by summing for each news category as evidence of significance.

Through the above formula, the p-values of each news category can be transformed into association weights. Clearly, the smaller the p-value and the more consistent the significance across lags, the larger the value of , indicating stronger evidence.

The calculation formula of the z-score standardized daily difference for news category intensity is as follows

Where denotes the daily difference in news counts for category , i.e., the number of news items on a given day minus the number on the previous day. This value is then standardized by the z-score. This process eliminates differences in volatility scales across categories, ensuring comparability of factor contributions. Essentially, it measures each category’s “daily news increment” as a standardized deviation relative to the past days. In this backtesting experiment, the rolling window length is set to 63 (approximately three months of trading days). To further enhance robustness, we truncate within the interval [-3, 3], thereby mitigating distortions in the composite signal or excessive trading impulses caused by extreme outliers (such as sudden surges in news volume).

Next, we map the composite news sentiment signal through the nonlinear function tanh to obtain a continuous position bounded within [-1,1]. This transformation smooths the signal, compresses exposure, and reduces excessive trading. Furthermore, the position is shifted forward by one trading day for execution, meaning that the signal generated at the close of day t-1 is implemented at the open of day t, effectively avoiding the problem of future information leakage.

For the calculation of strategy daily return , we multiply the daily position () by the target return of the day (). In the final backtesting experiment, we compare this strategy against the passive buy-and-hold benchmark using three evaluation metrics: annualized return, Sharpe ratio, and maximum drawdown, in order to examine whether investment strategies incorporating news category association weights are effective.

The calculation formula of is as follows

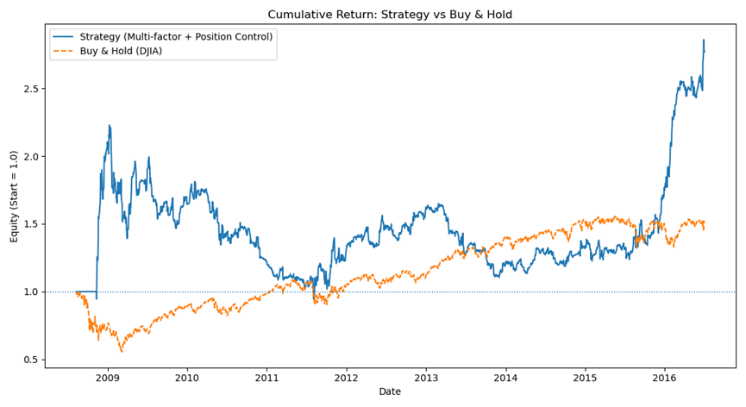
1. Backtesting Result

Based on the multifactor position control strategy constructed from the composite news sentiment signal derived using short-term lags (1 to 3 days) in the Granger causality tests, we find that the strategy demonstrates significant profitability during the sample period (2008–2016). The annualized return of the strategy reaches 13.81%, which is 8.33 percentage points higher than the buy-and-hold strategy of the Dow Jones Industrial Average (5.48%). The excess return is statistically and economically significant. In terms of risk-adjusted performance, the strategy’s Sharpe ratio is 0.5978, clearly outperforming the benchmark of 0.3661, indicating that the strategy can achieve a higher risk premium per unit of risk undertaken.

Table 7: Performance Metrics of Multi-factor Strategy vs Buy-and-Hold Benchmark

|  |  |  |  |
| --- | --- | --- | --- |
| Strategy | annualized interest rate | Sharpe ratio | Maximum drawdown |
| Multi-factor & Position Control | 13.81% | 0.5978 | 57.66% |
| Buy & Hold Benchmark strategy | 5.48% | 0.3661 | 44.43% |

Note: This table presents the performance comparison between the multi-factor position control strategy and the buy-and-hold benchmark (DJIA) over the period 2008-2016.

Although the multifactor strategy exhibits strong return performance, its risk profile is not negligible. The maximum drawdown of the multifactor position strategy reaches 57.66%, higher than the 44.43% of the benchmark strategy, reflecting the inherent volatility risk of active investment strategies. However, considering that the return-to-risk ratio of the strategy (0.24) is significantly superior to that of the benchmark (0.12), this level of risk remains within an acceptable range. Furthermore, the cumulative return curve shows that the strategy experienced substantial periods of return surges in 2009 and 2016, ultimately achieving nearly 200% cumulative returns, far exceeding the 50% of the benchmark strategy.

1. Cumulative return curve of Multi-factor Strategy & Buy-and-Hold Benchmark

At the same time, through separate diagnostics by lag, we identified important patterns in factor timeliness. In the short-term lag analysis (1 to 3 days), the lag-3 configuration performs best, with an annualized return of 20.20%, a Sharpe ratio of 0.7848, and maximum drawdown controlled at 50.27%, demonstrating the optimal balance of risk and return. This finding supports the behavioral finance hypothesis of delayed information transmission, suggesting that the market’s response to specific factor information is not completed instantaneously.

Table 8: Lag Analysis OF Strategy Performance Metrics

|  |  |  |  |
| --- | --- | --- | --- |
| Lag period | annualized interest rate | Sharpe ratio | Maximum drawdown |
| 1 day | 10.57% | 0.4619 | 71.99% |
| 2 day | 5.10% | 0.3162 | 68.93% |
| 3 day | 20.20% | 0.7848 | 50.27% |

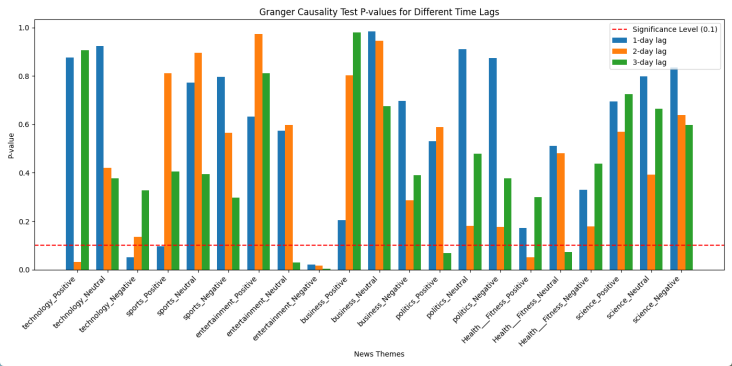
Note: This table shows performance metrics when using only significant factors at 1, 2, and 3-day lag periods respectively.

1. Result

**Interpretation disclaimer.** The Granger causality test evaluates whether past values of one series improve the out-of-sample prediction of another; it therefore identifies *predictive power* rather than structural or mechanistic causation. All references to “Granger causality” in what follows should be understood strictly in this predictive sense.

1. Short-term Lag Analysis

Based on the analysis of the Granger causality test results (Table 4), we found that only a few specific news topic categories showed statistically significant causal relationships with stock prices in the short-term lag period (1 to 3 days). Among them, negative entertainment news is particularly prominent, showing significant causal effects at all short-term lag orders (P values below the significance level of 0.1), indicating that this type of news has sustained and stable predictive power for stock prices, regardless of whether market reactions occur at a lag of 1, 2, or 3 days.



1. Granger Causality Test P-value in 1-3 days lag

As the lag period changes, the influence of other news types also exhibits time dependency: at a one-day lag, negative technology news and positive sports news significantly impact stock prices; at a two-day lag, positive technology news and positive health/fitness news begin to show significant influence; by the three-day lag period, neutral entertainment news, positive political news and neutral health/fitness news demonstrate significant predictive power. The table below shows how the P-value of each type of news story changes with a lag of one to three days.

Table 9. Granger Causality Test P-value in 1-3 days lag (Specific Value)

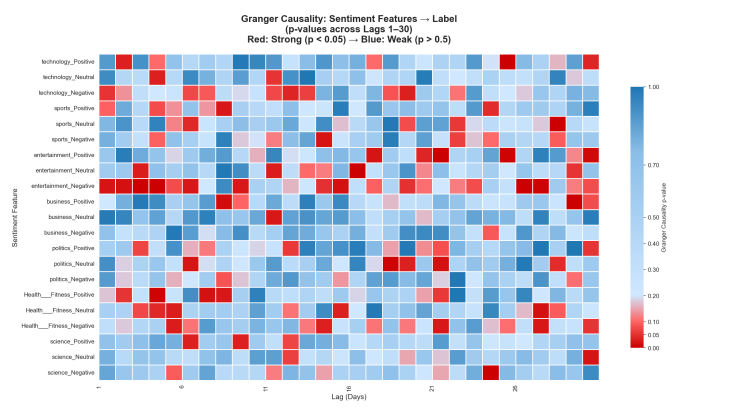
|  |  |  |  |
| --- | --- | --- | --- |
| P-value | 1 day-lag | 2-days lag | 3-days lag |
| technology\_Positive | 0.876104 | 0.031550 | 0.906803 |
| technology\_Neutral | 0.924824 | 0.419807 | 0.377896 |
| technology\_Negative | 0.051940 | 0.134430 | 0.328475 |
| sports\_Positive | 0.095757 | 0.812576 | 0.405505 |
| sports\_Neutral | 0.772292 | 0.896448 | 0.393636 |
| sports\_Negative | 0.796910 | 0.566184 | 0.296847 |
| entertainment\_Positive | 0.632492 | 0.973374 | 0.812422 |
| entertainment\_Neutral | 0.574215 | 0.598356 | 0.030358 |
| entertainment\_Negative | 0.019841 | 0.015754 | 0.003886 |
| business\_Positive | 0.203674 | 0.803667 | 0.980734 |
| business\_Neutral | 0.984237 | 0.946383 | 0.674934 |
| business\_Negative | 0.697014 | 0.286932 | 0.389535 |
| politics\_Positive | 0.530892 | 0.589097 | 0.069207 |
| politics\_Neutral | 0.911093 | 0.179702 | 0.477872 |
| politics\_Negative | 0.874113 | 0.177430 | 0.377974 |
| Health\_\_\_Fitness\_Positive | 0.172495 | 0.051448 | 0.299896 |
| Health\_\_\_Fitness\_Neutral | 0.512327 | 0.481971 | 0.072877 |
| Health\_\_\_Fitness\_Negative | 0.330134 | 0.178914 | 0.438121 |
| science\_Positive | 0.695592 | 0.570320 | 0.725958 |
| science\_Neutral | 0.799579 | 0.392506 | 0.664569 |
| science\_Negative | 0.835826 | 0.639599 | 0.598555 |

Note: This chart shows the Granger test results (P-value) of various types of new*s from 1 to 3 days lag*

This temporal evolution pattern reveals the time dynamics of different types of news affecting stock prices, suggesting that investors' reaction speeds vary for different news content. The persistent influence of negative entertainment news may reflect the market's continued sensitivity to this type of information, while other news types may require more time to be fully absorbed and reflected by the market, and may even have short-term cyclical effects. These findings provide important insights into the temporal relationship between information dissemination and market reaction, offering potential value for investment decisions and market predictions.

1. Segmentation Analysis Result

Although short-term lag analysis (1 to 3 days) reveals the dynamic correlation between some news topics and stock prices, especially in terms of the predictive ability of negative entertainment news to remain significant in the short term, such analysis may not be sufficient to capture potential impact mechanisms over longer time spans. To comprehensively examine the lagged effect of news sentiment on stock prices, we further expanded the analytical framework by extending the lag order to 1 to 30 days. By constructing a p-value heatmap for Granger causality test, the system observes the significant evolution patterns of various news topics at different time windows, aiming to identify whether there are causal relationships that are not significant in the short term but have predictive power in the long term, or to evaluate whether certain effects are strengthened, weakened, or delayed over time. This method helps to distinguish between instantaneous reactions and cumulative effects, thereby revealing the dynamic temporal characteristics of the impact of news sentiment on financial markets.



1. Granger Causality Test P-value in two separate periods

Conversely, business and political negative sentiment emerged as significant drivers in Part 2, suggesting that external contextual factors (e.g., market phases, policy shifts) may modulate the dominance of specific news themes over time. Notably, categories like political positive sentiment and health/fitness neutral sentiment exhibited significance in only one subset, implying their effects may be context-dependent rather than universally robust. These findings align with—and extend—the time-dependent patterns identified earlier: while some news types (e.g., negative entertainment) demonstrate remarkable temporal consistency, others exhibit instability not only across lag periods but also across data subsets, likely due to unobserved variables or sampling limitations. Methodologically, this dual-layered analysis—combining multi-lag and dataset-splitting approaches—highlights the necessity of integrating both temporal and contextual dimensions when modeling news-driven market behaviors. For investors, these results reinforce the importance of adaptive strategies that account for both the *persistence* of certain news effects (e.g., negative entertainment) and the *ephemeral* nature of others, which may dominate only under specific conditions or timeframes.

In summary, this study advances a nuanced framework for understanding how news sentiment interacts with financial markets—one that balances universal patterns with context-driven variability.

1. Conclusion

This study focuses on the complexity and conditional impact of news topics on stock price dynamics. Through multiple lag analysis and dataset segmentation methods, it reveals the unique effects of different types of news on the stock market in different time periods. Especially prominent is negative entertainment news, whose impact on stock prices shows significant consistency and stability across all test lags, reflecting the market's continued sensitivity to such information.

Further research has found that as the lag period changes, the impact of different types of news on stock prices exhibits a time-dependent characteristic. For example, negative technology news and positive sports news have a significant impact on stock prices in the short term (one day); In the mid-term (two days) and long-term (three days), other types of news such as positive technology, health/fitness, and political news also begin to show their influence. In addition, through segmentation analysis of the dataset, we observed that the effects of certain news topics were only significant in specific subsets, indicating that these effects may be moderated by external contextual factors such as market stages or policy changes.

This study not only deepens our understanding of how news sentiment affects financial market changes, but also provides an innovative methodological framework to guide investors in identifying and selecting high impact news topics, thereby supporting more accurate investment decisions. The research findings emphasize the importance of adaptive strategies, considering both the persistence of certain news effects and the temporary nature of other news effects and their dominant role under specific conditions or time periods. Ultimately, this study provides more accurate stock price prediction tools for financial market participants and lays a solid theoretical foundation for future financial decisions.

REFERENCES

1. Adham Elkomy. (2024). News Classification and Analysis using NLP [Dataset]. Retrieved from https://www.kaggle.com/datasets/adhamelkomy/news-classification-and-analysis-using-nlp.
2. Anonymous. (2024). arXiv:2410.12807 [q-fin.ST]. Retrieved from https://doi.org/10.48550/arXiv.2410.12807
3. Atmadja, A. R., & Purwarianti, A. (2015). Comparison on the rule based method and statistical based method on emotion classification for Indonesian Twitter text. In 2015 International Conference on Information Technology Systems and Innovation (ICITSI) (pp. 1-6). Bandung, Indonesia. https://doi.org/10.1109/ICITSI.2015.7437692
4. Baeza-Yates, R., & Ribeiro-Neto, B. (1999). Modern information retrieval. Addison-Wesley.
5. Berger, A. L., Pietra, S. A. D., & Pietra, V. J. D. (1996). A maximum entropy approach to natural language processing. Computational Linguistics, 22(1), 39-71.
6. Bollen J, Mao H, Zeng X. Twitter mood predicts the stock market[J]. Journal of computational science, 2011, 2(1): 1-8.
7. Belcastro, L., Carbone, D., Cosentino, C., Marozzo, F., & Trunfio, P. (2023). Enhancing cryptocurrency price forecasting by integrating machine learning with social media and market data. *Algorithms*, *16*(12), 542.
8. Cantini, R., Cosentino, C., Kilanioti, I., Marozzo, F., & Talia, D. (2023, October). Unmasking covid-19 false information on twitter: A topic-based approach with bert. In *International conference on discovery science* (pp. 126-140). Cham: Springer Nature Switzerland.
9. Cantini, R., Cosentino, C., Kilanioti, I., Marozzo, F., & Talia, D. (2025). Unmasking deception: a topic-oriented multimodal approach to uncover false information on social media. *Machine Learning*, *114*(1), 13.
10. Cui, M., Bai, R., Lu, Z., Li, X., Aickelin, U., & Ge, P. (2019). Regular expression based medical text classification using constructive heuristic approach. IEEE Access, 7, 147892-147904. https://doi.org/10.1109/ACCESS.2019.2946622
11. Cui, T., & Huang, F. R. (2024). Stock prediction based on sentiment analysis large models: A combined GRU and ALBERT prediction model. Dongyue Tribune, 45(2), 113–123. https://doi.org/10.15981/j.cnki.dongyueluncong.2024.02.013
12. Doshi-Velez, F., & Kim, B. (2017). Towards a rigorous science of interpretable machine learning. arXiv preprint arXiv:1702.08608. https://doi.org/10.48550/arXiv.1702.08608
13. Forman, G. (2003). An extensive empirical study of feature selection metrics for text classification. Journal of Machine Learning Research, 3(Mar), 1289-1305.
14. Hadavandi, E., Shavandi, H., & Ghanbari, A. (2010). Integration of genetic fuzzy systems and artificial neural networks for stock price forecasting. Knowledge-Based Systems, 23(8), 800–808. https://doi.org/10.1016/j.knosys.2010.04.010
15. Hsieh, T. J., Hsiao, H. F., & Yeh, W. C. (2011). Forecasting stock markets using wavelet transforms and recurrent neural networks: An integrated system based on artificial bee colony algorithm. Applied Soft Computing, 11(2), 2510–2525. https://doi.org/10.1016/j.asoc.2010.09.010
16. Joachims, T. (1998). Text categorization with support vector machines: Learning with many relevant features. In Proceedings of the 10th European Conference on Machine Learning (ECML) (pp. 137-142). Springer.
17. C. Krauss, X. A. Do, and N. Huck, “Deep Neural Networks, Gradient-Boosted Trees, Random Forests: Statistical Arbitrage on the S&P 500,” *European Journal of Operational Research (on ScienceDirect)*, Jun. 01, 2017. https://doi.org/10.1016/j.ejor.2016.10.031 (accessed Sep. 08, 2025).
18. D. D. Lee and S. H. Sebastian, “Learning the parts of objects by nonnegative matrix factorization,” *Nature*, vol. 401, no. 6755, pp. 788–791, 1999, doi: https://doi.org/10.1038/44565.
19. A. Mccallum and K. Nigam, “A Comparison of Event Models for Naive Bayes Text Classification,” 1998. Available: https://cdn.aaai.org/Workshops/1998/WS-98-05/WS98-05-007.pdf
20. S. Minaee *et al.*, "Deep Learning for Text Classification: A Survey," *ACM Comput. Surv.*, vol. 54, no. 3, Art. no. 61, pp. 1-40, Mar. 2021, doi: 10.1145/3439728.（找不到）
21. S. Minaee, N. Kalchbrenner, E. Cambria, N. Nikzad, M. Chenaghlu, and J. Gao, "Deep Learning-Based Text Classification: A Comprehensive Review," *ACM Comput. Surv.*, vol. 54, no. 3, Art. no. 62, pp. 1-40, Apr. 2021, doi: 10.1145/3439726.
22. Nurfikri, F. S., Mubarok, M. S., & Adiwijaya. (2018). News topic classification using mutual information and bayesian network. In 2018 6th International Conference on Information and Communication Technology (ICoICT) (pp. 162-166). Bandung, Indonesia. https://doi.org/10.1109/ICoICT.2018.8528806
23. Omrani, P., Ebrahimian, Z., Toosi, R., & Akhaee, M. A. (2023). Bilingual COVID-19 fake news detection based on LDA topic modeling and BERT transformer. 2023 6th International Conference on Pattern Recognition and Image Analysis (IPRIA), Qom, Iran. pp. 01-06. https://doi.org/10.1109/IPRIA59240.2023.10147179
24. Ren, K. W. (2023). Research on the co-movement of concept stock sentiment and stock prices based on text mining [Doctoral dissertation, Nanjing University of Posts and Telecommunications]. https://doi.org/10.27251/d.cnki.gnjdc.2023.000784
25. Rizaldy, A., & Santoso, H. A. (2017). Performance improvement of Support Vector Machine (SVM) with information gain on categorization of Indonesian news documents. In 2017 International Seminar on Application for Technology of Information and Communication (iSemantic) (pp. 227-232). Semarang, Indonesia. https://doi.org/10.1109/ISEMANTIC.2017.8251874
26. Rout, A. K., Biswal, B., & Dash, P. K. (2014). A hybrid FLANN and adaptive differential evolution model for forecasting of stock market indices. International Journal of Knowledge-Based Intelligent Engineering Systems, 18(1), 23–41. https://doi.org/10.3233/KES-140294
27. N. Aydin and G. Yurdakul, “Assessing countries’ performances against COVID19 via WSIDEA and machine learning algorithms,” *Applied Soft Computing*, vol. 97, p. 106792, 2020, doi: <https://doi.org/10.1016/j.asoc.2020.106792>.(未找到)
28. J. Sun, "Daily News for Stock Market Prediction, Version 1," Kaggle, Aug. 2016. [Online]. Available: <https://www.kaggle.com/aaron7sun/stocknews>. [Accessed: Sep. 07, 2025].
29. Ticknor, J. L. (2013). A Bayesian regularized artificial neural network for stock market forecasting. Expert Systems with Applications, 40(14), 5501–5506. https://doi.org/10.1016/j.eswa.2013.04.001
30. Tang, J., Ceng, X., & Peng, B. (2011). New methods of data clustering and classification based on NMF. 2011 International Conference on Business Computing and Global Informatization, Shanghai, China, pp. 432-435. https://doi.org/10.1109/BCGIn.2011.114
31. Tetlock P C. Giving content to investor sentiment: The role of media in the stock market[J]. The Journal of finance, 2007, 62(3): 1139-1168.
32. Vargas, M. R., de Lima, B. S., & Evsukoff, A. G. (2017). Deep learning for stock market prediction from financial news articles. In Proceedings of IEEE International Conference on Computational Intelligence and Virtual Environments for Measurement Systems and Applications (pp. 60–65).
33. A. Vaswani *et al.*, “Attention is All you Need,” *Advances in Neural Information Processing Systems*, vol. 30, 2017, Available: https://proceedings.neurips.cc/paper\_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aaPaper.pdf
34. J. Wang and Y. X. Li, "Research on Stock Return Prediction with NS-FEDformer Model," *Computer Engineering and Applications*, vol. 61, no. 9, pp. 334–342, 2025.
35. S. Wang and C. Manning, “Baselines and Bigrams: Simple, Good Sentiment and Topic Classification,” Association for Computational Linguistics, 2012. Available: https://aclanthology.org/P12-2018.pdf
36. Xu, B., Zhang, D., Zhang, S., Li, H., & Lin, H. (2018). Stock market trend prediction using recurrent convolutional neural networks. In Proceedings of CCF International Conference on Natural Language Processing and Chinese Computing (pp. 166–177).
37. Xu, W., Liu, X., & Gong, Y. (2003). Document clustering based on non-negative matrix factorization. In Proceedings of the 26th annual international ACM SIGIR conference on Research and development in information retrieval (pp. 267-273). ACM. https://doi.org/10.1145/860435.860485
38. Yang, Y., & Pedersen, J. O. (1997). A comparative study on feature selection in text categorization. In Proceedings of the Fourteenth International Conference on Machine Learning (ICML) (pp. 412-420). Morgan Kaufmann Publishers Inc.
39. Zhu, R., Zhong, G.-Y., & Li, J.-C. (2024). Forecasting price in a new hybrid neural network model with machine learning. Expert Systems with Applications, 249(Part B), 123697. https://doi.org/10.1016/j.eswa.2024.123697

Huaxi Liu was born in 2003 in Dongguan, Guangdong Province, China. He is currently an undergraduate student majoring in Software Engineering at Guangdong University of Foreign Studies and is expected to receive the Bachelor of Engineering degree in 2026.

His research interests focus on natural language processing, data mining, and sentiment analysis with applications in public opinion and information analysis. He has conducted research on identifying sources of controversy in contemporary mainstream films by leveraging sentiment analysis techniques to examine audience reviews.

Lin Gao was born in 2003 in Tianjin, China, and entered Dalian University of Technology in 2022 to pursue a Bachelor's degree in Software Engineering. Research interests include machine learning, deep learning, and cross-domain applications of natural language processing. He has conducted research on the design of knowledge graphs and the implementation of query systems based on Graph RAG, with relevant works documented in software copyright registrations.

**Jiang Zonghan** a Bachelor of Computer Science candidate (expected graduation: July 2026) at Guangdong University of Foreign Studies (GDUFS). Research interests focus on Natural Language Processing (NLP) and image processing and have gained practical experience in data analysis and data mining.

**Li Desheng** was born in 2004 in Quanzhou, Fujian Province, China. He entered Fujian Normal University in 2023 to pursue a Bachelor's degree in Electronic Information Engineering and is expected to graduate in 2026. His research interests mainly include natural language processing, image processing, and machine learning. He has rich practical experience in UI interface design and cross disciplinary artificial intelligence projects, and has obtained Huawei HCIP certification.