Analysis of AI in stock market based on natural language model

Published: 18 July 2025

Citation in BibTeX format

GAITDI 2025: International Conference on Implementing Generative AI into Telecommunication and Digital Innovation
July 18 - 20, 2025
Beijing, China

GAITDI '25: Proceedings of the International Conference on Implementing Generative AI into Telecommunication and Digital Innovation 2025 (July 2025)
https://doi.org/10.1145/3771792.3773855
ISBN: 9798400714924

Analysis of AI in stock market based on natural language model

Bingjie Liu
Jiangxi University of Finance and Economics
Nanchang, Jiangxi Province, China
liubingjie@wbu.edu.cn

Min Hao
School of Economics
Wuhan Business University
Wuhan, Hubei Province, China
2221871709@qq.com

Rong Tan
School of Economics
Wuhan Business University
Wuhan, Hubei Province, China
3356753816@qq.com

Yisong Chen*
College of Computing
Georgia Institute of Technology
Atlanta, GA,, USA
ychen841@gatech.edu

Abstract

Stock is a high-risk and high-yield way of investment and financial management, which occupies an important position in the national economy, and requires investors to have rich knowledge and experience in finance. With the development of the Internet and artificial intelligence, it has attracted many investors, and the integration of artificial intelligence and the financial market has changed the traditional way of securities trading. It has become a popular phenomenon for investors to express their stock opinions on stock forums, which produces a lot of stock comments, and the investor sentiment contained in this information often affects the trend of the stock market. This paper through python web crawler technology, climb Weibo related topic text, using the characteristics of a large number of text semantic expression, on the basis of traditional machine learning methods, through the in-depth study of word frequency-inverse document frequency (TF-IDF) feature extraction method, realize the category of multiple comment corpus text emotion accurate analysis, we found that the public for AI application in the stock market in anticipation at the same time also have some doubts. To provide a certain development direction for artificial intelligence in the future stock market.

CCS Concepts

• **Social and professional topics**; • **Professional topics**; • **Computing and business**; • **Economic impact**;

Keywords

Stocks, artificial intelligence, sentiment analysis, TF-IDF

ACM Reference Format:
Bingjie Liu, Min Hao, Rong Tan, and Yisong Chen. 2025. Analysis of AI in stock market based on natural language model. In *International Conference on Implementing Generative AI into Telecommunication and Digital Innovation (GAITDI 2025), July 18–20, 2025, Beijing, China.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3771792.3773855

1 Foreword

As the barometer of China's national economic development, its market development is closely linked with the development of China's market economy.[1] However, because the stock market is essentially a dynamic, non-stable, complex and changeable system, the price fluctuation of the stock market is unpredictable, and even investors go bankrupt. According to the effective market hypothesis proposed by the stock price trend can fully and timely reflect all valuable information, that is, the network information of the financial market reflects the operation of the financial market.[2] With the development of Internet technology, Internet text has become an important channel to obtain information about the stock market. Shareholders are more inclined to obtain stock market information through the network media, and share and discuss with netizens through the network social platform, as well as the expectation of the future market trend. This information can provide reference for investors' decision-making. [3] Especially in small-cap stocks, stocks with higher sentiment scores tend to have higher excess returns.[4] Obviously, in the context of the rapid rise of big data and artificial intelligence, natural language processing (NLP) technology has made a lot of breakthrough results.[5] The emotional fluctuations of investors will obviously cause complex changes to the stock market and then affect the investment decision-making behavior of other investment groups. At present, scholars at home and abroad have carried out extensive research on the field of financial market. Word frequency-inverse document frequency (TF-IDF) has become a common method, a statistic for assessing the importance of a word in a document collection. Due to the limited ability of traditional methods to understand context, researchers are looking at AI technology to improve the accuracy of emotion analysis. Therefore, based on the influence of AI, the text sentiment analysis of mining emotions and opinions from the network text is an important issue in the field of financial investment, which has a very important reference value for optimizing investment decisions, market transactions and the development of artificial intelligence.

2 problem analysis

This study aims to explore the emotional state of investors on the application of AI technology in the stock market through the analysis of network text data and then provide support for the research and

*Corresponding author.

GAITDI 2025, Beijing, China

ACM ISBN 979-8-4007-1492-4/2025/07
https://doi.org/10.1145/3771792.3773855

Table 1: Descriptive statistical analysis

	title	Titlelink	profilepicture	From	content	thumbnAll
count	1019	1019	1019	1019	1019	288
unique	300	300	977	457	452	128
freq	83	83	5	3	5	6t

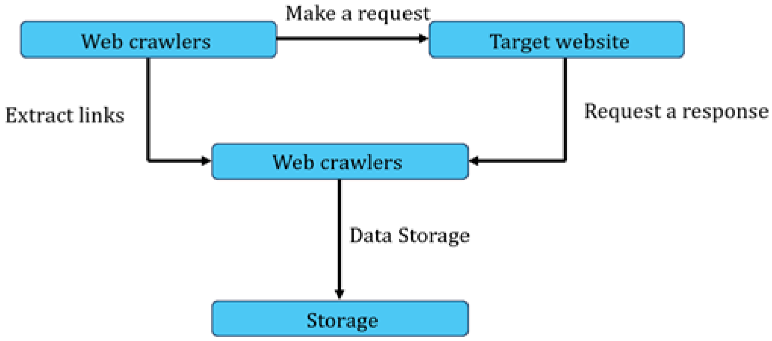

Figure 1: The workflow chart of the web crawler

prediction of the stock market. To achieve this goal, the collected data needs to be processed systematically first. Specifically, the collected effective text data should be visually analyzed to intuitively understand the distribution characteristics and rules of the data; at the same time, the emotional tendency should be analyzed to determine whether the emotions expressed in the text are positive, negative or neutral.

In the data collection stage, this study took ”AI” and ”stock” as the keywords to crawl the relevant comments from the period from January 1, 2024 to February 1,2025 on the official website of Weibo. Weibo is chosen as the data source because it is characterized by extensive user groups, rapid information dissemination and rich topic discussion. It can cover the views and opinions of investors with different backgrounds and experience and provide rich data resources for research.

3 Model building

3.1 Data collection

The research method of this paper is mainly based on python web crawler technology to extract the emotional state of investors on the application of AI technology in the stock market. A total of 1120 pieces of data were collected. First, 1120 pieces of data were screened. As shown in Table 1, 1119 pieces of effective data were determined as 1119, and then the collected data was analyzed by designing relevant emotional words. Its descriptive statistical analysis is shown in Table 1 below.

3.2 Relevant theoretical basis

3.2.1 Natural language model. Web crawler is a program to extract data. It is mainly used to automatically browse the data on the Internet and extract.[6] The basic principle is to write programs to simulate the behavior of a browser, allowing the computer to automatically access the web page and extract the information needed. The workflow chart of the web crawler is shown in Figure 1.

3.2.2 TF-IDF model. TF-IDF is a feature extraction method for text mining and information retrieval. Its basic idea is to use the word frequency (TF) with an entry in the document and evaluate the importance of the entry in the whole document through the inverse document frequency (IDF). The main principle of TF-IDF algorithm is that if a word appears very frequently in that document and appears very frequently in other documents, the word is considered to have good discrimination ability and is suitable for classification.

Word frequency (TF): For the word frequency of an entry in a specific document, the calculation is shown in Equation (1):

$$\boldsymbol{TF}\left(\boldsymbol{t},\boldsymbol{d}\right)=\frac{\boldsymbol{n}\left(\boldsymbol{t},\boldsymbol{d}\right)}{\sum_{\boldsymbol{k}}\boldsymbol{n}\left(\boldsymbol{k},\boldsymbol{d}\right)} \tag{1}$$

In this, $\boldsymbol{n}(\boldsymbol{t},\boldsymbol{d})$ is the occurrence count of word $\boldsymbol{t}$ in document $\boldsymbol{d}$, and is the sum of the occurrence counts of all words in document $\boldsymbol{d}$.

Inverse document frequency (IDF): indicates that a specific entry IDF can be divided by the total number of documents by the number of documents containing the entry, and then the resulting quotient is logarithmic. If fewer articles include the entry, the larger the IDF is, indicating that the entry has a very good article discrimination ability. The calculation is as described in Equation (2):

$$\mathbf{IDF(t,D)}=\mathbf{log}\frac{\mathbf{N}}{\mathbf{d}\in\mathbf{D}:\mathbf{t}\in\mathbf{d}} \tag{2}$$

where $\boldsymbol{N}$ is the total number of documents, $|\mathbf{d}\in\mathbf{D}:\mathbf{t}\in\mathbf{d}|$ is the number of documents containing term $\boldsymbol{t}$. Finally, the calculation formula for TF-IDF is the product of two factors, resulting in a

Table 2: Results of the model evaluation

	MSE	RMSE	MAE
training set	0.0014893566	0.0385921837	0.0247565623
test set	0.0018315075	0.0427961159	0.0264219790

comprehensive feature value, as shown in equation (3):

$$TF - IDF\ (t, d, D) = TF\ (dt) *IDF\ (t, D) \tag{3}$$

TF-IDF takes into account the number of entries in the whole document and their importance in the whole document, so that effective information can be more accurately captured in the emotions of the shareholders. Through TF-IDF feature extraction of documents, differentiated feature vectors can be obtained, which can provide effective data support for subsequent emotion analysis. If the TF-IDF value increases, the higher the word appears in the document, the key to the document, the lower the frequency, the less representative.

3.2.3 Evaluation indicators. We use Micro-F1 values for evaluation, as calculated in equation (4)(5)(6).

$$P = \frac{N_{TP}}{N_{TP} + N_{Fn}} \tag{4}$$

$$R = \frac{N_{TP}}{N_{TP} + N_{FP}} \tag{5}$$

$$F_1 = \frac{2 \times P \times R}{P + R} \tag{6}$$

Where, P is precision, R is recall, N_{TP} is the number of correctly predicted samples, N_{TP} is the number of negative samples incorrectly predicted as positive, and N_{Fn} is the number of positive samples incorrectly predicted as negative.In the UABSA task, an aspect sentiment pair is considered correct only when both the aspect term and the sentiment polarity are correctly predicted. This is done to enhance model performance, reduce model variance, and improve stability and generalization ability.

4 Analysis of text emotional tendencies

4.1 Error analysis

MSE (mean squared error) is the expected value of the difference between the predicted value and the actual value and is used to measure the accuracy of the model. A smaller MSE value indicates a higher accuracy of the model. RMSE (root mean square error) is the square root of MSE and is also often used to evaluate the accuracy of the model. The smaller the value, the higher the accuracy of the model. MAE (mean absolute error) is the mean of absolute error, used to reflect the actual situation of predicted value error. The smaller MAE values indicate the higher accuracy of the model. These indices are frequently used in model evaluation and comparisons.

In this paper, error analysis and test of the natural language model of this problem are conducted. Through calculation, the evaluation results of the training set and test set models are shown in Table 2 below.

According to the MSE data in Table 2, the value is less than 0.002, and the accuracy of the model is high; according to the RMSE data in Table 2 the value is less than 0.05, and the accuracy of the model is high; according to the MAE data in the table, the value is less than 0.03, and the accuracy of the model is high. Based on the analysis results of these three indicators, we can conclude that the constructed model has high reliability and effectiveness in terms of text emotion tendency analysis.

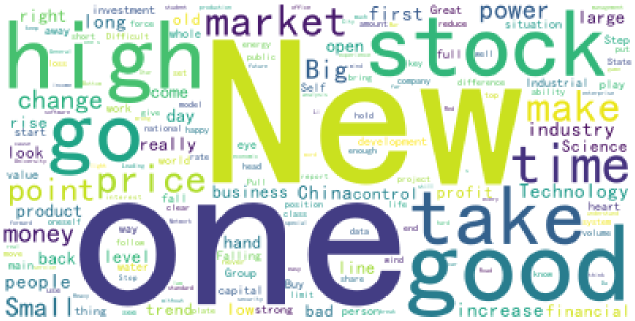

Figure 2: Cloud diagram of words

4.2 Emotional tendency analysis

As can be seen from the word cloud map, the public pays high attention to the application of AI in the stock market, and the words such as "New", "Mark", "Good" and "Product" appear, indicating that the public has a high expectation for the application of AI in the stock market to a certain extent.

In order to further explore the emotional tendency and market reaction degree of the public to the application of AI in the stock market, the distribution of emotional value and quantity is analyzed, and the results are shown in Figure 3.

From Figure 3, less scatter in the zero and negative area means less neutral and negative feedback. The forward area is more dispersed, indicating that the situation is more particularly satisfied. Combined with the analysis of word cloud map, it is likely that although the public has great expectations for the application of AI in the stock market, the public still considers the uncertainty of the stock market and the risks of the application of AI technology.

4.3 For TF-IDF analysis

TF-IDF is a text-weighted model with high accuracy and recall rate, which can be applied to various characteristic choices. The importance of keywords is proportional to the number of appearances in the text and inversely proportional to their frequency in the corpus.

Figure 4 presents the average TF-IDF values for the top 10 most significant entries across the entire text dataset. Bar graphs are arranged from high to low by the average TF-IDF values of feature words, and the higher of the columns indicates the higher importance of entries in the entire text dataset. As can be seen from the

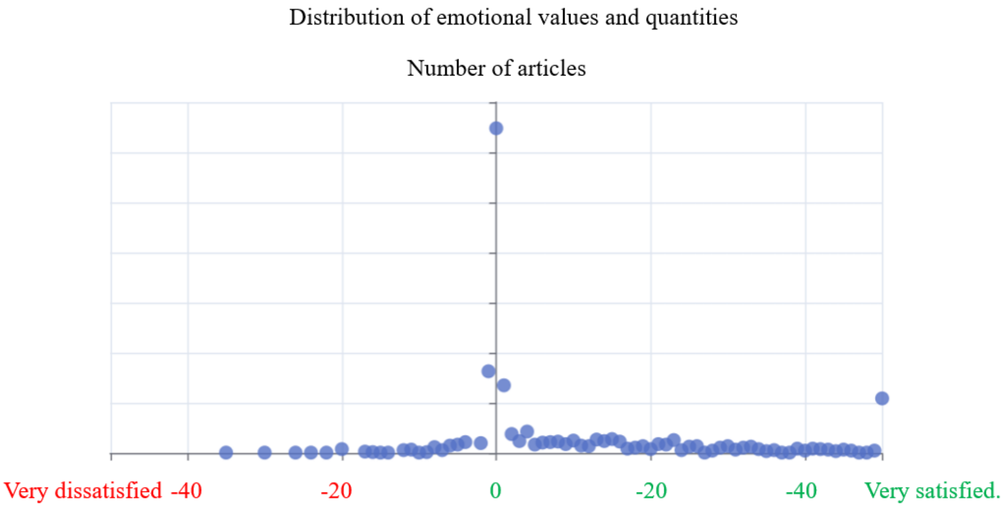

Figure 3: Scatter plot

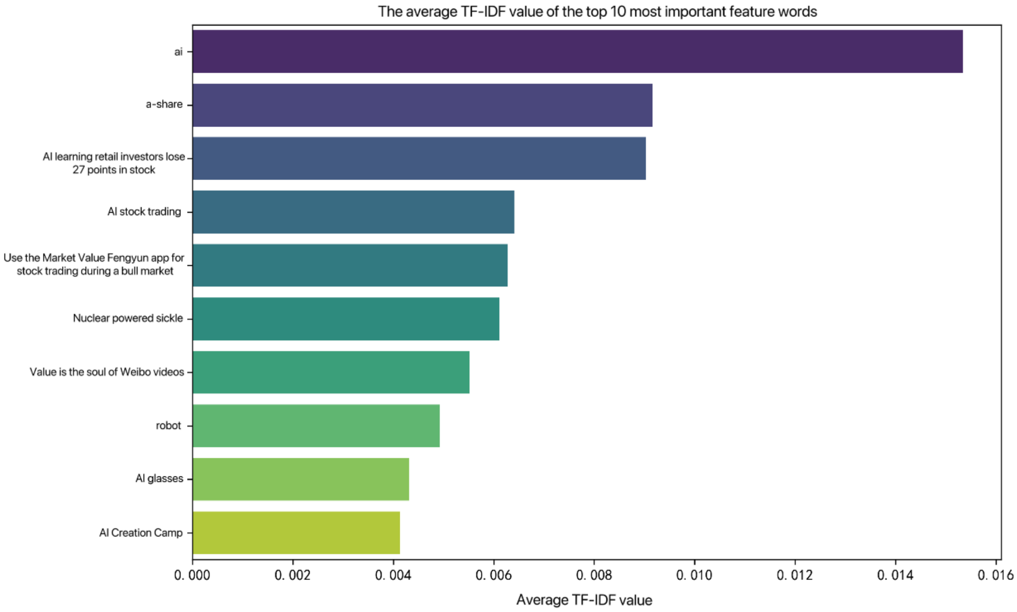

Figure 4: Document entry average TF-IDF values

figure, the ”AI” entry has the highest TF-IDF value, indicating the importance of AI in the stock market.

As can be seen from Figure 5, the scatter is mainly clustered into a relatively concentrated area in the lower left corner, which can be obtained combined with Figure 4, reflecting that the whole text data can be classified into the category ”AI”. It further illustrates the importance of describing the relationship between stocks and AI.

5 Conclusion and suggestion

This study focuses on the application of AI based on natural language models in the stock market. Using Python web crawler technology, we obtained 1119 valid comment data on ”AI” and ”stock” from Weibo between 1 month 1-day 2024- and 2-month 1 day 2025 and conducted in-depth analysis. In terms of model construction, web crawler technology was used to collect data, TF-IDF was used as the feature extraction method, and Micro-F1 value was used for evaluation. Through error analysis, the MSE, RMSE,

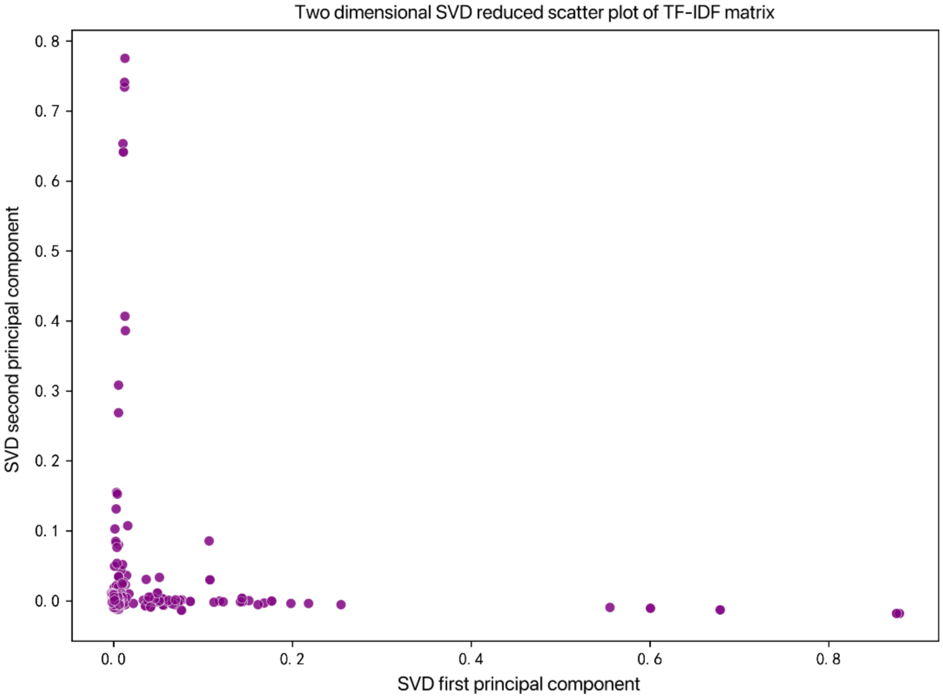

Figure 5: scatter plot of TF-IDF

and MAE values of the training set and test set are all relatively low, indicating that the constructed model has high reliability and effectiveness in text sentiment tendency analysis.

From the results of emotional tendency analysis, the word cloud map shows that the public has a high attention and expectation for the application of AI in the stock market; The distribution of emotional value and quantity shows that although there are many positive feedback, neutral and negative feedback cannot be ignored, reflecting that the public is concerned about the uncertainty of the application of AI while expecting to bring opportunities to the stock market. The TF-IDF value analysis found that the ”AI” entry is the most important in the text data set, and the text data mainly revolves around the ”AI” category, highlighting the important position of AI in the stock market, showing that AI has a good development prospect in the stock market.[7]

Based on the above research conclusions, in order to promote the application effect of AI in the stock market, the following suggestions are put forward:

First, we will strengthen technological research and development. Further improve the accuracy and stability of natural language processing technology in the sentiment analysis of the stock market, optimize the feature extraction methods such as TF-IDF, better capture the relationship between investor sentiment and the trend of the stock market, and provide more accurate decision support for investors.

Second, strengthen risk warning and education. In view of the public’s concerns about the application of AI in the stock market, financial institutions and relevant departments should strengthen the risk warning and education for investors, help investors rationally view the role of AI technology in the stock market, and improve investors’ risk awareness and coping ability.[8]

Third, strengthen multi-source data fusion. Future research can consider integrating more data sources, enriching data dimensions, analysing investor sentiment and market trends more comprehensively, and improving the accuracy of stock market research and forecasting.

References

[1] Liu Xiaojuan, Chen Min, Wang Meili, *et al.* On the status quo of China’s stock market [J]. Technology and Market, 2013,20 (01): 125.
[2] Wang Na, He Dongyue, Liu Lei. Research on the construction of stock market investor sentiment index and its effectiveness —— Emotion analysis based on the post of Oriental wealth stock bar [J]. Price Theory and Practice, 2022,(11):146-151.DOI:10.19851/j.cnki.cn11-1010/f. 2022.11. 353.
[3] Banerjee S ,Aggarwal D ,Sengupta P .Do stock markets care about ESG and sentiments? Impact of ESG and investors’ sentiment on share price prediction using machine learning[J].Annals of Operations Research,2025,(prepublish):1-40.
[4] Suzuki M ,Ishikawa Y ,Teraguchi M , *et al.*Sentiment works in small-cap stocks: Japanese stock’s sentiment with language models[J].International Journal of Information Management Data Insights,2025,5(1):100318-100318.
[5] Xu Fengmin, Ma Jierao, Jing Kui. ESG perspective and stock market pricing —— Evidence from AI language models and news texts [J]. The Contemporary Economic Science, 2023,45(06):29-43.DOI:10.20069/j.cnki.DJKX. 202306003.
[6] Zeng Jianrong, Zhang Yangsen, Zheng Jia, etc. Implementation technology and application of web crawler for multiple data sources [J]. Computer Science, 2019,46 (5): 304-309.
[7] Banerjee S ,Aggarwal D ,Sengupta P .Do stock markets care about ESG and sentiments? Impact of ESG and investors’ sentiment on share price prediction using machine learning[J].Annals of Operations Research,2025,(prepublish):1-40.
[8] Huang Z ,Liao B ,Hua C , *et al.*Leveraging ChatGPT for enhanced stock selection and portfolio optimization[J].Neural Computing and Applications,2025,(prepublish):1-17.