Application of Word Frequency Analysis in Stock Price Prediction

Fan Mo[1,a,*]

[1]Department of Computer Science, University of Manchester, Manchester, UK
a. fan.mo-4@manchester.ac.uk
**corresponding author*

Abstract: The improvement of hardware performance and the rapid development of AI technology have brought the possibility of using machine learning and big data models to predict markets and stocks. However, some models that only consider past trend and data of a stock has met some limit when facing real-time events which may cause market sentiment fluctuations. This paper introduced a model, trying to take real-time news as a factor to predict the stock price of a stock. It uses keyword extraction and word frequency analysis to obtain the sentiment from the news, which innovatively match the sentiment of words with the rate of return of a specific stock and its frequency. And the model combines it with past stock data to predict the stock price. The result shows that for some emerging industry that are sensitive with the news, such as the Internet industry and electronic based company, the accuracy of the model that considering news has improved. For traditional industries, the accuracy improvement of this model is not obvious.

Keywords: stock prices, AI, word frequency analysis

1. Introduction

Market expectations and stock forecasting are research topics in economics and marketing studies. In the field of quantitative training, a good method for stock forecasting, will significantly reduce market uncertainty and bring benefits. By reviewing past research, it's found that most work of stock predicting was usually based on fundamental analysis and technical analysis [1]. It's pointed out that three aspects should be considered in fundamental analysis: analyzing macroeconomics indicators such as GDP (Gross Domestic Product) and CPI (Consumer Price Index), analyzing the whole industry of the stock and analyzing the financial status and its operations of a company [2]. Another research for fundamental analysis shows that by using the rate of return, the accuracy of classifying companies into "good" and "poor" can be improved to 74.6%, which greatly reduces uncertainty for investors and helps them make better stock forecasts and profit from them [3].

In recent year, the advancement of computer hardware and AI has made it possible to build stock models based on a significant amount of historical transaction data. When machine learning is applied, an important factor to evaluate the model is that whether various factors and related effects that affect stock prices can be accurately discovered and learned by the model [4]. Some classic learning models, such as logistic regression and support vector machine, can be well used for stock forecasting [5]. In recent years, with the development of deep learning, some LSTM models based on past data can predict indicators of stocks and their volatility with considerable accuracy [6]. For the more unpredictable Chinese stock market, the model using LSTM algorithm can even improve the accuracy of the return forecast from 14.3% to 27.2% [7]. Other algorithm based on deep learning like MLP (Multilayer Perceptron), RNN (Recurrent Neural Networks) and CNN (Convolutional Neural Network) can also be applied in it, and research find that some algorithm like CNN, it is even possible to adapt the model to different stocks with some common inner factors and dynamics [8-9]. In short, stock forecasting by machine learning and using data from past has great potential.

But in addition to a significant amount of historical data on stock, the stock price is frequently significantly influenced by what is happening in the real world. Here is a clear illustration. Figure 1 shows the stock price of Nvidia spanning 30 April to 7 July in 2023 [10].

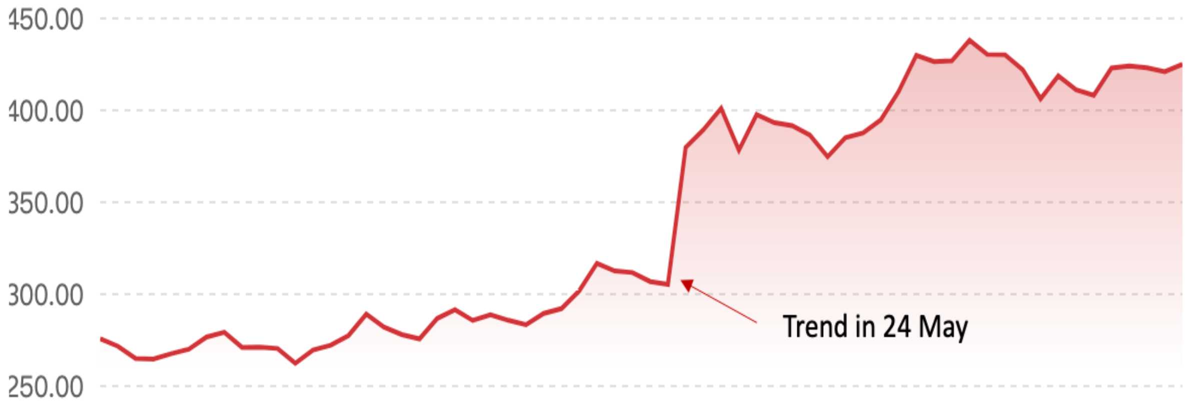


Figure 1: Graph of Nvidia stock price from 2023.4.30 –2023. 7.7.

According to an article reported by Consumer News and Business Channel (CNBC) in May 25, the first-quoter report was released by Nvidia, and the revenue was far above the general estimate and the prediction [11]. At the same day, the company also released their new graphic chip for the needs of AI computing. Compared with previous products, the performance of the new chip has been greatly improved, which will adapt to the high calculation demands for building large model of AI in the future, such as Chat GPT. These events finally sent its stock soaring 24% on that day [11]. It’s an obvious example for how the real-time news and information affected the market sentiment, and finally make a big difference to the trend of the stock price. In other words, models only based on previews data often perform unwell when market sentiment fluctuates [9], and looking back at previous studies, there are few models to predict stock that can combine past stock data with sentiments from real-time news, and research on the impact of current events on stocks is limited. Based on these facts and observations, the motivation and method of this research can be determined.

2. Dataset

This article introduced a model build for predicting stocks, which considers real-time news as a factor of impacting the market and stock price. The research introduced by this article is based on two dataset, “S&P 500 Stock Data” and “All the News” from Kaggle, they respectively contain various stock price and index, and news in the time series [12-13]. The model uses TF-IDF algorithm and word frequency to extract keywords from past news, calculate word sentiment based on stock return and finally predict stock prices based on sentiment contains in words from news [14].

For evaluating the result, another model with same hyper-parameter was used, which is trained with the news sentiment removed. Since there’s only 1 variable “news sentiment” in two training sets, by comparing the Mean Squared Error (MSE) of the two models, it can be judged whether the accuracy of the model improves after considering real-time news, and finally evaluate the performance of the model.

3. Methodologies

3.1. Assumptions

Using news for stock price forecasting, is essentially measures the sentiment embedded in the news and establishes a relationship between stock returns and sentiments in news for the corresponding period [15]. First, some assumptions were made to support this research and build this relationship.

As 1. The news in a period contains the information of sentiments that can influence the price of a stock on that period, the sentiment from each article can be measured as a value S.

As 2. S is proportional to the return rate of the stock on that period, a high value of S represents the high expectation from the market that the stock price will go up.

As 3. The sentiment of an article a S_a is influenced by some of the words $w_{1,}w_{2,\dots,}w_i$ from all the words in article W_a independently.

$$\{w_{1,}w_{2,\dots,}w_i\} \epsilon W_a \quad (1)$$

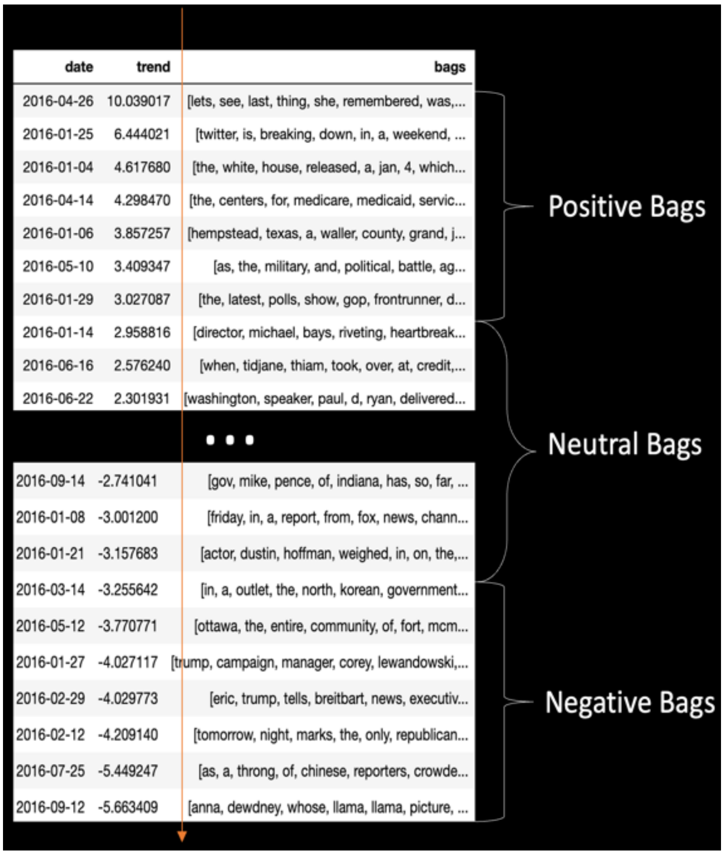
$$S_a = f(w_{1,}w_{2,\dots,}w_i) \quad (2)$$

These assumptions are based on experience that how news and information affected the market, and these do not violate the logic and even intuition of economics and marketing [16]. Sentiment information is not emotions in the human sense, like happy angry or sad, it’s positive or negative information to influence the price contained in news or articles. And it’s noticed that for As 3., the only feature considered for each word in an article is the word frequency, which does not consider the contextual relationship between words and text, so some information in the article will be lost. This method is also called “Bag of Words”, and it will significantly improve the efficiency of text processing and reduce time cost [16-17].

3.2. Word Distribution

According to assumptions, sensitive articles and sentiment words are different for each stock. In other words, after an information was released, it may have a great impact on the price of some stocks, but it will not affect other stocks. So, for a stock model being established, we need to filter out its related words.

From two datasets, stock yield data combined with news from a long period can be obtained. According to As 2., the news from a period with high or low stock return, may contains more positive or negative sentiment information than news from the period that the stock price is smooth. By sorting the stock return of groups of days, a data frame contains price fluctuations with “word bags” from news can be obtained as shown in Figure 2.



date	trend	bags
2016-04-26	10.039017	[lets, see, last, thing, she, remembered, was,...
2016-01-25	6.444021	[twitter, is, breaking, down, in, a, weekend, ...
2016-01-04	4.617680	[the, white, house, released, a, jan, 4, which...
2016-04-14	4.298470	[the, centers, for, medicare, medicaid, servic...
2016-01-06	3.857257	[hempstead, texas, a, waller, county, grand, j...
2016-05-10	3.409347	[as, the, military, and, political, battle, ag...
2016-01-29	3.027087	[the, latest, polls, show, gop, frontrunner, d...
2016-01-14	2.958816	[director, michael, bays, riveting, heartbreak...
2016-06-16	2.576240	[when, tidjane, thiam, took, over, at, credit,...
2016-06-22	2.301931	[washington, speaker, paul, d, ryan, delivered...
		• • •
2016-09-14	-2.741041	[gov, mike, pence, of, indiana, has, so, far, ...
2016-01-08	-3.001200	[friday, in, a, report, from, fox, news, chann...
2016-01-21	-3.157683	[actor, dustin, hoffman, weighed, in, on, the,...
2016-03-14	-3.255642	[in, a, outlet, the, north, korean, government...
2016-05-12	-3.770771	[ottawa, the, entire, community, of, fort, mcm...
2016-01-27	-4.027117	[trump, campaign, manager, corey, lewandowski,...
2016-02-29	-4.029773	[eric, trump, tells, breitbart, news, executiv...
2016-02-12	-4.209140	[tomorrow, night, marks, the, only, republican...
2016-07-25	-5.449247	[as, a, throng, of, chinese, reporters, crowde...
2016-09-12	-5.663409	[anna, dewdney, whose, llama, llama, picture, ...

Figure 2: Graph of sorted data frame with stock return and news.

In this data frame, the label "trend" is calculated by the formula of the percentage of stock return, which reflets the fluctuate of the price on that period, and the label "bag" contains all the word appeared in news from that period. By sorting these bags and divided them into 3 types as shown in Figure 2, it can be deduced that a word bag in "Positive Bags" may contains more positive information than bags in "Neutral Bags", and so as bags in "Negative Bags". By this inference, and As 3., it's obvious that the frequency of related words in "Positive Bags" and "Negative Bags", are higher than the frequency in "Neutral Bags". Figure 3 shows an example of distributions of keyword "apple" for Apple stock, compared with words "banana", which not has a strong correlation with Apple stock.

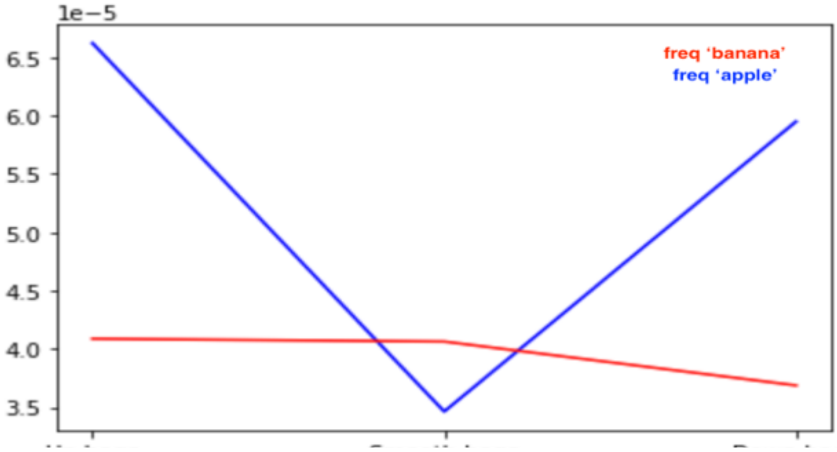


Figure 3: Graph of distributions of word "apple" and "banana" in 3 types of bags.

3.3. Word Filter

The comparation method mentioned above is not rigorous, since different words may naturally have different frequency, and it's meaningless for comparing the variance of distributions of different words. To measure how close the word related to the stock, a value Δ for each word was established.

$$\Delta_w = \frac{freq(w)_{positive}+freq(w)_{negative}-2freq(w)_{neutral}}{freq(w)_{positive} + freq(w)_{negative}+freq(w)_{neutral}} \tag{3}$$

The value Δ_w measures the difference of the frequency of word w in 3 types of bags and divided by their addition. In short, Δ_w shows the variance of the distribution of word w, and compared with its general frequency. Δ will remain at a low value when a neutral word is considered, which has a smooth distribution both in Positive, Negative and Neutral bags. However, this formula can also get some extreme value, these values are mostly generated by words with very low frequency, and only appears in 1 type of bags. Table 1 shows some typical words and their Δ, when the model of Apple stock has been established.

Table 1: Comparison table of Δ of related words and non-related words.

irrelated word	Δ	Related word	Δ
“and”	0.001	“aapl”	0.704
“are”	0.017	“macbook”	0.323
“they”	0.021	“iphone”	0.352
“maché”	-2	“apple”	0.217

Obviously, the value Δ of some words that are more relevant to Apple stock, compared with values of some neutral rare words, are not in the same order of magnitudes. By setting a range to the Δ of words, words that has a stronger association with the stock can be chosen:

$$R = \{ w \mid \Delta_w \in [\alpha_1, \alpha_2] \} \tag{4}$$

R is a set of words that relevant to the stock, chosen by $\left[\alpha_1, \alpha_2\right]$, which are the lower and upper bound to limit Δ_w.

3.4. TF-IDF Algorithm

Although 80% of the useless words filtered out from news, there's also many words remained, which will also bring some noise to the model. Another algorithm called TF-IDF was used to get keywords from a word set.

tf, the term frequency of a word w in article a is given by:

$$tf_{\ w} = freq(w)_{\ a} \tag{5}$$

idf, the inverse document frequency of a word w in a corpus C with n documents $A_{\ 1\ldots}A_n$ is given by:

$$C_{contains\ w} = \{ A_{\ i} \mid w \ \in \ A_{\ i} \ , \ A_{\ i} \ \in C \} \tag{6}$$

$$n' \ = \ number\ of\ articles\ in\ C_{contains\ w} \tag{7}$$

$$idf_{\ w} = \log\left(\frac{n}{n'}\right) \tag{8}$$

And the formula of calculating the $tf - idf$ value for the word w is shown as:

$$tf - idf_{\ w} = \frac{tf_{\ w}}{idf_{\ w}} \tag{9}$$

This algorithm shows that based on a corpus, each word from a text has a $tf - idf$ value, which measures the importance of the word in this text [18]. For words with low frequency in a text, the importance of it is low, and this also occurs when a word with high frequency in all of texts from the corpus, such as prepositions and conjunctions. In other word, some keywords from the text may have high $tf - idf$ values. By calculating and sorting each $tf - idf$ values of each word from each bag in news dataset, and keeps top n words remained for each bag, a word set K with keywords from original news dataset can be obtained.

3.5. Word Dictionary

Finally, when dealing with a news dataset with fluctuate data of a particular stock, with methods introduced before, a dictionary D can be given by:

$$D = \{\, w \mid w \in \ V \cap K \,\} \tag{10}$$

It is believed that words in D contains sentiment information, and the stock price is sensitive when news with some words in D released.

3.6. Calculate Sentiment

According to As 2 & 3., it can be deduced that the sentiment of words in D, appears in news in a period, is proportional to the yield of the stock. For instance, the word “rise” may more likely appears in news from days which has a high stock reward. Based on this and data of the stock “trend”, the sentiment s_w of word w in D can be established as:

$$T = \{t_{\ 1},t_2 \ \dots t_n\} \tag{11}$$

$$s_w = \Sigma\, T \tag{12}$$

Where $t_{\ 1},t_2 , \dots t_n$ is the stock “trend” which reflects the percentage of the stock return on a period, some of them are positive, some are negative. For all t in set T, it is chosen from periods when the word w appears in news. In short, s_w can be seen as an average return from stock when w appears. For positive words, the addition of values in T might be high, compared with negative words, which may contain more negative value in T. Finally, words in D is divided into “Positive” and “Negative” classes, by comparing s_w with 0. To get sentiment value S for an article a, a credible formula is given as:

$$S_{\ a} = \frac{pos-neg}{pos+neg} \tag{13}$$

Where pos is the number of positive words appears in news, and neg is the number of negative words. Combined with two words classes in dictionary D, this simple formula is taken to assign sentiment values to news in dataset [19].

3.7. Model Training and Evaluation

The model was trained by the Random Forest, which takes “Open”, “High”, “Low”, “Close”, “Volume” from stock data as features. To improve the accuracy, these features use the value of the past three days. Additionally, another feature S contains the sentiment of real-time news is considered. The right part of Figure 4 shows how it was built.

For evaluating the model, another model with same hyper-parameter can be build, which does not read S and only takes features from previous stock data from training set. By predicting from the same test set, two Mean Squared Error (MSE) can be obtained. Figure 4 visually shows the process of evaluating the model.

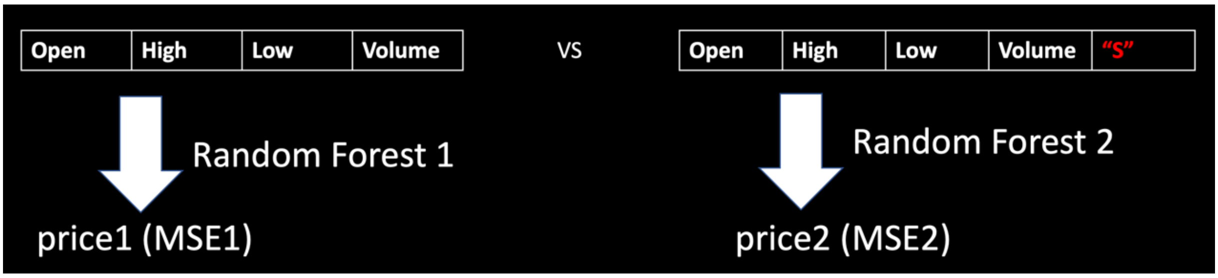


Figure 4: The process of building the model and model evaluation.

Obviously, A formula for computing the percentage difference of two number x and y can be given as:

$$diff(\%) = \frac{(x-y)}{y} \times 100\% \tag{14}$$

This formula can be used to compare two MSE obtained from two models, and evaluating how well the accuracy has improved when real-time news is considered. A value I was created to measure its improvement.

$$I = \frac{(MSE1-MSE2)}{MSE2} \times 100\% \tag{15}$$

It’s obvious that a positive value of I represents the model performs better when news and real-time information are considered. Similarly, a negative value means that the accuracy of the model drops, and the news considered is more like noise.

4. Result and Discussion

The result is shown as Table 2, which illustrates how MSE of the model reduced when sentiment of real-time news was considered in 6 stocks.

Table 2: Stocks and its corresponding I from model evaluation.

Stock	MSFT	AAPL	NVDA	DE	NUE	DHI
I	+6.29%	+4.07%	+2.21%	-4.78%	-5.47%	+0.72%

6 stocks are selected, and the above method is used to build the model. Companies of first 3 stocks are Microsoft, Apple and Nvidia, and all of them are based on Internet and high-tech industry. The remaining 3 stocks are also representative, and the companies corresponding to them are John Deere, Nucor Corporation and D.R. Horton. These companies are leading companies in agriculture, steel and

iron, and real estate industries. In short, the results of these three stocks reflects the performance of this model in traditional industries.

According to the blank, the result shows that for some technology-based company, the model improves the accuracy in some stage. Especially for Microsoft and Apple, when the two models use the same number of decision trees in their forest, the parameter S helps the advanced model reduces its MSE by about 6% and 4% respectively.

But this model is not suitable for all stocks. Based on results from representative stocks, it can be found that the accuracy reduced in agriculture industry and steel and iron industry. The feature S is more like a noise for building the model in these industries, which finally leads to its poor performance. For DHI stock, which is a proxy for the real estate sector, the accuracy of models incorporating S improved marginally.

From this result, it can be inferred that when the news is considered as a factor, the accuracy of this model has improved in the stock prediction related to the high-tech industry. However, in traditional industry, the news sentiment is useless when a model is built. What's more, this result also proves that for Internet companies and high-tech based companies, their stock prices may be more susceptible to real-time events, and more sensitive to positive or negative information contained in the news. Compared with it, traditional and basic industries are more stable in the face of market sentiment and expectation. Interestingly, in people's common sense, the real estate industry is also sensitive to news, so the accuracy of the model has also been slightly improved, which is a little different from other industries. To explain the underperformance of traditional industry models, it may also be possible that the sentiment information affecting traditional industries is more hidden in the news and cannot be well perceived by this model.

5. Conclusion and Reflection

5.1. Summary

In conclusion, when predicting stock price with sentiments from real-time news, this model improves the accuracy when predicting stocks based on Internet and high-tech industry, but it has a poor performance when the stock is based on the traditional industry. This result is consistent with the example given at the beginning of the article about Nvidia stock being affected by real-time news. This can also prove the importance of real-time information for technology and their development.

At the same time, this can also prove that the stock prices of technology industries such as the Internet industry are more sensitive to real-time news, while for information and sentiment from news to affect traditional industries, they may more implicit, and have more complex dynamic, so in the end, it shows that the stock prices of these traditional industries, such as agriculture and steel, are not sensitive to news.

5.2. Reflection

In fact, it is just a simple attempt for predicting future stock price with real-time news. Combining MSE with absolute value of stocks, it can be said that this model does not bring a significant improvement in the accuracy of stock price prediction. For processing news, "Bag of Words" and TF-IDF algorithm it used will lose information based on the relationship between words, and some better NLP methods like Word2Vec or BERT can be used. Instead of building corpus for each stock, it can also use some classical and complete corpus. Finally, some subjective factors above like assumptions and formulas, and limited calculation capacity, may also leads bias and non-generalizable conclusion.

References

[1] *Shah, D., Isah, H., & Zulkernine, F. (2019) Stock market analysis: a review and taxonomy of prediction techniques. International Journal of Financial Studies, 7.*

[2] *Hu, Y., Liu, K., Zhang, X., Su, L., Ngai, E. W. T., and Liu, M. (2015). Application of evolutionary computation for rule discovery in stock algorithmic trading: A literature review. Applied Soft Computing, 36, 534–551.*

[3] *Dutta, A., Bandopadhyay, G., and Sengupta, S. (2012) Prediction of stock performance in the Indian stock market using logistic regression. International Journal of Business and Information, 7(1), 105.*

[4] *Shah, V. H. (2007) Machine learning techniques for stock prediction. Foundations of Machine Learning| Spring, 1(1), 6-12.*

[5] *Jiang, W. (2021) Applications of deep learning in stock market prediction: recent progress. Expert Systems with Applications, 184, 115537.*

[6] *Kyoung-Sook, M., and Hongjoong, K. (2019) Performance of deep learning in prediction of stock market volatility. Economic Computation & Economic Cybernetics Studies & Research, 53(2).*

[7] *Chen, K., Zhou, Y., and Dai, F. (2015) A LSTM-based method for stock returns prediction: A case study of China stock market. In 2015 IEEE international conference on big data (big data), 2823-2824.*

[8] *Hiransha, M., Gopalakrishnan, E. A., Menon, V. K., and Soman, K. P. (2018) NSE stock market prediction using deep-learning models. Procedia computer science, 132, 1351-1362.*

[9] *Schumaker, R. P., and Chen, H. (2009) A quantitative stock prediction system based on financial news. Information Processing & Management, 45(5), 571–583.*

[10] *Nvidia. Available at: https://www.msn.cn/zh-cn/money/watchlist?tab=Related&id=a1mou2&ocid=ansMSNMoney11&duration=3M&relatedQuoteId=a1yv52&relatedSource=MlAl&src=b_secdans, last accessed on 2023/8/5*

[11] *CNBC. Available at:https://www.cnbc.com/2023/05/25/nvidia-on-track-for-record-high-driven-by-ai-chip-demand.html, last accessed on 2023/8/5*

[12] *S&P 500 stock data. Available at: https://www.kaggle.com/datasets/camnugent/sandp500, last accessed on 2023/8/5*

[13] *All the News. Available at: https://www.kaggle.com/datasets/snapcrack/all-the-news, last accessed on 2023/8/5*

[14] *Qaiser, S., and Ali, R. (2018) Text mining: use of TF-IDF to examine the relevance of words to documents. International Journal of Computer Applications, 181(1), 25-29.*

[15] *Ke, Z. T., Kelly, B . T., and Xiu, D. (2019) Predicting Returns With Text Data. National Bureau of Economic Research.*

[16] *Lin, J. H., Zhang, Y. F., Chen, L. Y., Deng. Y.M. (2022) News Sentiment and Machine Learning Investment Strategy. China Journal of Econometrics, 2(4),881-908.*

[17] *Qader, W. A., Ameen, M. M., and Ahmed, B. I. (2019) An overview of bag of words; importance, implementation, applications, and challenges. In 2019 international engineering conference (IEC), 200-204.*

[18] *Aizawa, A. (2003) An information-theoretic perspective of tf–idf measures. Information Processing & Management, 39(1), 45–65.*

[19] *Fu W. B., Zeng H. (2022) Can Non-Punitive Supervision Restrain the Management's Tone Manipulation? Empirical Evidences Based on Annual Report Texts. Contemporary Finance & Economics, 0(3): 89-101.*