Proceedings of the 2nd International Conference on Business and Policy Studies
DOI: 10.54254/2754-1169/9/20230353

Analysis on Text Mining in Stock Market Applications

Linjie Jin[1,a,*]

[1] *Camford Royal School, Beijing, China, 100000*
a. linjie_jin@163.com
**corresponding author*

Abstract: Stocks are a financial product with high risk but high reward and flexible trading that many investors prefer. If an investor can accurately predict the price of a stock, he or she will be rewarded handsomely. Stock prices, on the other hand, are influenced by a variety of factors, including macroeconomic conditions, market conditions, major socioeconomic events, investor preferences, and company business decisions. As a result, stock price forecasting has become the focus and difficulty of research in a variety of fields. Stock price prediction entails gathering news and commentaries, analyzing historical data, and determining the impact of news events on investor sentiment and stock price trends. The purpose of this paper is to provide an introduction to the application of text mining in the stock market, including commonly used text mining and prediction models, as well as to highlight problems in the field and suggest some future directions for improvement or research. Finally, many unresolved issues are raised in order to contribute to future research in this area.

Keywords: text mining, stock market prediction, SVM, LSTM, NB, machine learning

1. Introduction

The stock market has a great impact on individuals, businesses, and even the global economy, and because it is so influential, stock market forecasting is always a very hot topic. However, stock market prediction is a challenging task because the stock market is a dynamic market and the data it generates is massive. This data can be a treasure trove for investors, and analyzing it can help them improve the accuracy of their predictions and allow them to gain more benefits. Also benefiting from the rapid development of technology, various models and algorithms can be implemented. Therefore, text mining has become an important part in predicting stock prices.

A model that can simultaneously use text mining techniques for fast content analysis of news and economic techniques for predicting financial stock fluctuations would be an effective tool for analysts and investors. According to previous studies, it is known that there is a strong correlation between stock price shocks and stock news [1]. In text mining, news, articles, and commentaries about financial markets have received a great deal of attention. Various models have examined different sources of datasets such as Twitter, Yahoo Finance, etc., and there are also researchers who analyze financial markets by region or country. Meanwhile, there are researchers who focus on analytical methods such as time series analysis, and in recent years some researchers have analyzed the trend of investor sentiment and stock prices as a way to improve prediction accuracy, such as in the references [2-3].

DOI: 10.54254/2754-1169/9/20230353

In this paper, we will briefly introduce data mining and prediction models. With the development of computer technology, various models have emerged, such as Support Vector Machine (SVM), Decision Tree (DT), Artificial Neural Network (ANN), and Nave Bayes (NB). Due to space limitation, SVM, NB, and LSTM are selected as representatives. Finally, this paper will also present several shortcomings of existing research and future research directions as a contribution to the field.

2. Text Mining Task

Information is taken from text using text mining technologies. Text preparation and information extraction are the two main components of text mining technology. Text preprocessing is fundamental to the text mining task because it primarily transforms unstructured data into a computer-readable matrix of document words. Text preprocessing is the foundation of text mining, and the objective of various features and methodologies is to minimize data loss. Uysal and Gunal published a paper discussing the impact of preprocessing on text classification, in which they pointed out that an appropriate combination of preprocessing tasks can significantly improve classification accuracy across domains and languages, while even an inappropriate combination may reduce classification accuracy. From this, they argue that the importance of the preprocessing step in text classification is equal to that of the feature extraction, feature selection, and classification steps [4]. Feature selection is also an important step in text mining, which involves removing redundant and unimportant parts and retaining important features. Yang and Pedersen compared five methods: CHI, IG, MI, DF, and TS, where CHI and IG lost the least accuracy in removing items [5].

Time is precious for the stock market, which is so volatile, and the introduction of machine learning algorithms for performing text mining tasks greatly reduces human intervention and the large amount of time spent processing text, reducing time and increasing efficiency, so text mining is a big help for stock market forecasting.

3. Stock Market Predictions

Stock market forecasting is also the prediction of the future value of a company and to solve this problem, two approaches have emerged: technical analysis and fundamental analysis [6]. Technical analysis uses mathematical indicators of stock price construction for analysis and forecasting, such as historical market data, while fundamental analysis uses information such as news to make forecasts for example macro information such as inflation rate, unemployment rate, interest rates, etc. The text mining approach to stock market forecasting uses both technical analysis and fundamental analysis and the results show that the combination of both approaches is effective and robust [6].

The stock market forecasting problem is quite complex, as was already indicated, and models lately have tended to be composite models. In his research, Ji mentioned that some of the existing literature mixes statistical econometric models with machine learning models, or that some literature employs more than two machine learning models concurrently to help increase the accuracy of stock price prediction. These composite models perform more effectively overall than single models [7].

4. Representative Models

In Kumar's survey, he counted 89 papers between 2000 and 2016 and counted the methods used in the articles and concluded that the most commonly used method was support vector machines, followed by NB, neural networks, decision trees, linear regression, etc [8]. Figure 1 shows the statistics of model usage in Kumar's literature [8]. So, in the following, Support Vector Machine (SVM) and Nave Bayes (NB) will be introduced. In addition to this, there is the commonly used long-term memory model (LSTM).

DOI: 10.54254/2754-1169/9/20230353

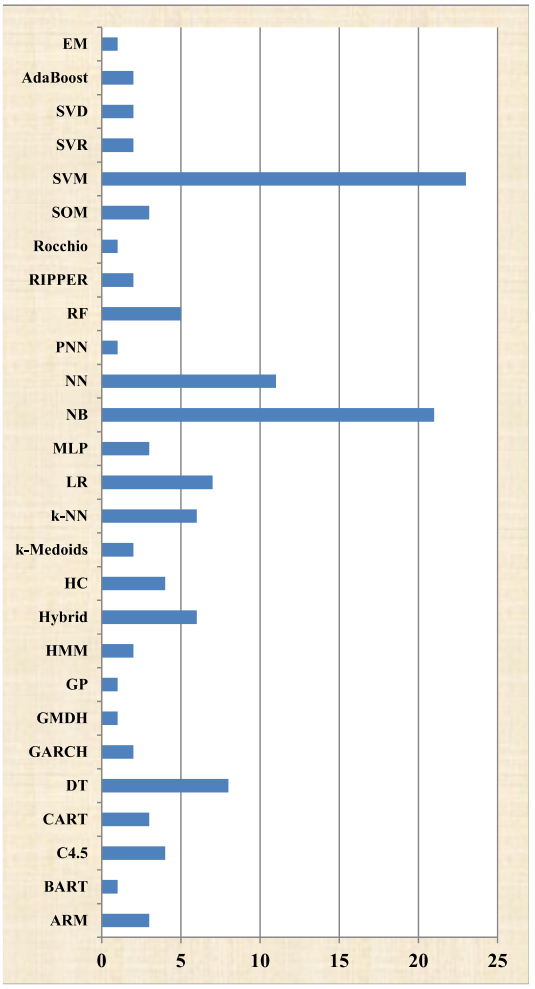

Figure 1: Bar graph of model usage in Kumar's paper.

4.1. SVM

A support vector machine is a model proposed by Vapnik to perform classification tasks by constructing hyperplanes [9]. It can analyze data and identify patterns so that it can perform classification and regression analysis of the data. SVM can classify linearly separable data into two categories and can also make data linearly separable by projecting nonlinear data to higher dimensions using kernel functions (sigmoid, radial, and polynomial). Joachims and Burges show in their own paper that SVM performs well when classifying text linearly [10-11]. SVM performance is determined by the kernel function and user-defined parameters used. For example, Samad used SVM to classify text data in his 2018 article, thereby classifying news into positive or negative financial news, with a radial basis function used in the kernel, with a prediction accuracy of 68%. The authors concluded that the complexity of textual data makes it necessary to use better feature extraction techniques [12]. There is also Sun's article in 2020 that used SVM to help process data for subjective and objective judgments and also for extremity judgments [13]. In Seker's study, SVM and k-nearest neighborhood (KNN) were applied. Seker affirmed the impact of time series analysis methods on the closing price of the stock market and its relevance to the economic news in the case of Turkey [14].

There are too many examples of using SVMs to list them all here. The previous literature amply illustrates the importance of SVM in this field.

DOI: 10.54254/2754-1169/9/20230353

4.2. NB

Naive Bayes is based on the principle of Bayes' theorem. It makes the assumption that variables are independent and feature sets are completely independent from each other. Its high predictive power and ease of use make it very popular, and it is used frequently. Li compared Naive Bayes and SVM when processing information and finally chose NB because both models produced similar results, but NB was less complex in comparison [3]. Sun used NB to classify textual data when determining whether the extracted information was too extreme [13]. However, Bayesian models are not only NB but also the Latent Dirichlet Allocation model (LDA), the Bayesian structural time series model (BST), etc. Such a wide variety gives it a wide range of applicability and can largely satisfy the purpose of the researcher.

4.3. LSTM

An enhanced version of the recurrent neural network model is the LSTM neural network (RNN). Time series modeling frequently uses LSTM because it addresses the issue of gradient vanishing and gradient exploding in RNN by using forget gates, input gates, and output gates. The LSTM is made up of many neurons. The data first passes through the forget gate in each neuron. The next neuron to be updated is not impacted by the information that is decided to be forgotten by the forget gate. The input gate chooses which information can be added in the second stage. Two functions, the sigmoid function and the tanh function, process the input of the local neuron and the output of the preceding neuron to create two results. The two outcomes are then used to decide which information needs to be updated. The outcomes are then saved to the output gate. The output gate ultimately determines which outcome can be produced. Similar to before, the outcome of one neuron's output gate will be input to the following neuron for the subsequent round of processing.

The LSTM can be predicted well. Li's study illustrates the superiority of LSTM models in determining the predictive potential of investor sentiment, particularly when the model input contains predictive information, when compared to logistic regression, support vector machines, and naive bayes [3]. In order to increase the accuracy of forecasts, numerous composite models have evolved in recent years, and LSTM-based prediction models are among the most well-liked of them. LSTM has also demonstrated a good ability to be integrated with other models. Ji's model employs Doc2Vec to train and extract text feature vectors, SAE to reduce the text vectors' dimensionality, and LSTM to forecast future stock values. The Doc-W-LSTM composite model outperforms the three baseline approaches in terms of predictive power [7]. Additionally, Ray's proposed BST-LSTM, which incorporates the Bayesian structural time series model (BST) and also makes use of LSTM to depict nonlinear relationships in the data, is included in the composite model. Ray evaluated many hybrid models and came to the conclusion that BST-LSTM performs significantly [15]. Because LSTM has strong prediction ability and overcomes the gradient exploding and gradient vanishing concerns of RNN, many researchers have employed it.

5. Limitations and Future Directions

In reviewing the literature, some limitations are also found. Most of the existing studies focus on stock market data and investor sentiment, but they neglect to consider the fact that different countries have different economic conditions and political systems, and the impact of government policies on stock prices cannot be ignored. For example, China has a high political priority and an official issuance may usher in a lot of volatility, or some stocks that are on a growth trend are likely to have been restricted due to policy issues. Thus, the political system is also a point of concern. Meanwhile, stock markets in different industries may react differently to news sentiment. For example, the stock markets in traditional energy and the stock market in the pharmaceutical industry have different stock

DOI: 10.54254/2754-1169/9/20230353

price fluctuations for news and policies. Some studies only look at stock price indices like the S&P 500.Perhaps separate analysis can make the prediction more accurate. In addition, the information obtained from text mining may not be fully representative of investors. There is no restriction on financial websites or software that only experts or investors can post comments. Spammers, junk information or bystanders may gather in the comments section. Secondly, people are likely to say and do different things, especially in models that extract information from comments, and investors' comments do not always match their behavior. Existing research does not take into account differences between individuals. For example, the number of stocks held by investors may affect their mindset and sensitivity to news. Investors with a few stocks may be less sensitive to negative news and less eager to sell their stocks than those with a large number of stocks. Existing research treats investors as the same regardless of the number of stocks held, which may affect the precision.

This paper also provides some ideas and directions that will hopefully help future research in this field. Different news may have different effects because there is no absolute objectivity and news can be subjective. In the future, it may be possible to try to distinguish articles or news published by the government, experts, enthusiasts, etc., and process them separately, for example, by adding coefficients to achieve the difference in the degree of expertise or influence. Existing studies have tried to reduce the impact of semantic bias as much as possible, but many text expressions with emoji have another meaning, so in the future, we can try to extract and analyze emoji at the same time, which can effectively reduce the problems caused by various semantic meanings and ambiguities. The timeliness of information is also a very important factor, and the relationship between the time of information extraction and stock price fluctuations can be studied in the future. For example, how the news before the opening affects the opening price, and in the opening also make predictions and analyze how it affects the closing price.

6. Conclusion

This paper introduces text mining techniques, explains how operations such as preprocessing and feature selection can help with text mining, and affirms the usefulness of text mining for stock market prediction. Stock market prediction uses technical analysis and fundamental analysis, introducing the differences between them and pointing out that researchers in recent years tend to use composite models and that the combination of multiple models can help improve the accuracy of prediction. The commonly used models, SVM, NB, and LSTM, are also briefly introduced, and the reasons for their popularity are analyzed in the context of historical literature. This paper aims to contribute to future research in this area.

References

[1] *Azizi, Z., Abdolvand, N., Asl, H.G., Harandi, S.R., 2021. The Impact of Persian News on Stock Returns Through Text Mining Techniques 18.*

[2] *Kulkarni, R., Amidwar, S., Muthya, M., 2019. STOCK PREDICTION THROUGH NEWS SENTIMENT ANALYSIS. J. Archit. 5.*

[3] *Li, Y., Bu, H., Li, J., Wu, J., 2020. The role of text-extracted investor sentiment in Chinese stock price prediction with the enhancement of deep learning. Int. J. Forecast. 36, 1541–1562. https://doi.org/10.1016/j.ijforecast.2020.05.001.*

[4] *Uysal, A.K., Gunal, S., 2014. The impact of preprocessing on text classification. Inf. Process. Manag. 50, 104–112. https://doi.org/10.1016/j.ipm.2013.08.006.*

[5] *Yang, Y., Pedersen, J.O., n.d. A Comparative Study on Feature Selection in Text Categorization 9.*

[6] *Picasso, A., Merello, S., Ma, Y., Oneto, L., Cambria, E., 2019. Technical analysis and sentiment embeddings for market trend prediction. Expert Syst. Appl. 135, 60–70. https://doi.org/10.1016/j.eswa.2019.06.014.*

[7] *Ji, X., Wang, J., Yan, Z., 2021. A stock price prediction method based on deep learning technology. Int. J. Crowd Sci. 5, 55–72. https://doi.org/10.1108/IJCS-05-2020-0012.*

DOI: 10.54254/2754-1169/9/20230353

[8] *Kumar, B.S., Ravi, V., 2016. A survey of the applications of text mining in financial domain. Knowl.-Based Syst. 114, 128–147. https://doi.org/10.1016/j.knosys.2016.10.003.*

[9] *V.N. Vapnik, Statistical Learning Theory, JohnWiley & Sons, NewYork, 1998.*

[10] *Joachims, T. (1998). Text categorization with support vector machines: Learning with many relevant features. In Proceedings of ECML (pp. 137–142). Heidelberg: Springer.*

[11] *Burges, C. J. C. (1998). A tutorial on support vector machines for pattern recognition.Data Mining and Knowledge Discovery, 2(2), 121–167.*

[12] *Abd Samad, P.H.D., Mutalib, S., Abdul-Rahman, S., 2019. Analytics of stock market prices based on machine learning algorithms. Indones. J. Electr. Eng. Comput. Sci. 16, 1050. https://doi.org/10.11591/ijeecs.v16.i2.pp1050-1058.*

[13] *Sun, Y., Liu, X., Chen, G., Hao, Y., Zhang, Z. (Justin), 2020. How mood affects the stock market: Empirical evidence from microblogs. Inf. Manage. 57, 103181. https://doi.org/10.1016/j.im.2019.103181.*

[14] *Seker, S.E., Mert, C., Al-NAAMI, K., Ozalp, N., Ayan, U., 2014. TIME SERIES ANALYSIS ON STOCK MARKET FOR TEXT MINING CORRELATION OF ECONOMY NEWS 6, 23.*

[15] *Ray, P., Ganguli, B., Chakrabarti, A., 2021. A Hybrid Approach of Bayesian Structural Time Series With LSTM to Identify the Influence of News Sentiment on Short-Term Forecasting of Stock Price. IEEE Trans. Comput. Soc. Syst. 8, 1153–1162. https://doi.org/10.1109/TCSS.2021.3073964.*