The Impact of Market Sentiment and Financial Reports on Stock Prices: A Machine Learning Approach

Fenghui Zhao

School of Accounting, Southwestern University of Finance and Economics, Chengdu, 611130, China
zhaofenghui749@gmail.com

Abstract. Machine learning will be employed in this paper to investigate the joint effects of market sentiment and corporate financial indicators on short-term stock prices. Natural language processing is used to collect and process sentiment data from financial news and social media, and at the same time, financial indicators are obtained from corporate reports. Logistic Regression, Random Forest, XGBoost, SVM and LSTM are all contained in the same experiment. Combine the financial and sentiment features to construct a prediction model, then evaluate the model. All the models will have the same data processing and evaluation methods in the experiment. According to the above analysis, in addition to financial indicators, sentiment information can also help predict. Random Forest and XGBoost are the best models among those that have been evaluated. Therefore, the data of sentiment can provide us with more immediate information on the market's anticipation of changes. Combining the sentiment and financial information of a company can help predict changes in stock prices and offer some support for quantitative investment.

Keywords: Stock Price Prediction; Machine Learning; Market Sentiment; Financial Indicators; XGBoost.

1 Introduction

Both the company's economic indicators and investor mood affect the equity market. Through its main indicators of profit, funds, operations, etc., the financial statements help investors understand how financially healthy a company is. At the same time, the sentiment in news articles, social media and online forums also shows the short-term expectations and views of market participants [1]. With the development of digital information, text data have gradually begun to be used for stock price prediction. The above data can help us sense how investors and the market are feeling at any given time, and thus extend the content of traditional financial indicators.

Recently, some research has begun to use machine learning and sentiment analysis in the problem of financial forecasting to improve the accuracy of traditional methods. Random Forest and XGBoost are good ensemble methods that work well in general, and they can better model the non-linear and interaction effects of many financial

© The Author(s) 2026
L. C. Yoong (ed.), *Proceedings of the 2026 2nd International Conference on Financial Innovation and Marketing Management (FIMM 2026)*, Advances in Economics, Business and Management Research 408,
https://doi.org/10.2991/978-94-6239-801-6_146

indicators than linear models [2]. In addition, BERT and other Transformer models are very good at extracting contextual sentiment information from text and can help us better understand the mood of the market and investors' behaviour. Although there have been some achievements, most studies have concentrated on either financial indicators or sentiment data individually. Therefore, the potential benefits of combining both types of information have not been fully realised yet, and people do not know how an integrated model can better predict short-term stock price fluctuations [3].

In the early days, most of the methods for predicting stock prices were based on statistics and time series, such as ARIMA and ARCH; they were not suitable for capturing the irregular changes and dynamics of the financial market [4]. The models cannot address the many relationships among various financial indicators, and are therefore less accurate in the face of rapid changes in the market. Recently, many kinds of machine learning algorithms, such as SVM, Random Forest, XGBoost, LSTM, etc., have also shown good results in the analysis of complex financial data [2, 5]. The above methods can build a model of the non-linear relationship, learn from a large amount of historical data, obtain information on the dependence of time, and are more flexible and suitable for short-term stock price prediction.

To solve the problems mentioned above, this paper puts forward a new way to combine financial indicators of companies with sentiment features derived through Natural Language Processing (NLP). Combine the numerical financial data and the text of the sentiment information to gain a comprehensive understanding of the reasons for changes in stock prices. Under the same experimental conditions, many models are used in this paper, such as Logistic Regression, Random Forest, XGBoost, SVM, BERT and LSTM, to examine how well the combination of financial and sentiment features can forecast short-term stock prices. This way can be used to compare the traditional and advanced types of machine learning methods, and it will be observed that combining sentiment data with other financial indicators has good results.

2 Methodology

2.1 Dataset and Preprocessing

Financial indicators and the market's mood will be used together to predict changes in stock prices over a short period. EPS, net profit, revenue growth, ROE, current ratio and debt-to-asset ratio are financial indicators that can be obtained from company reports and Yahoo Finance [4]. The above indicators can help us know how financially strong a company is and how well it operates. At the same time, sentiment information is also obtained from financial news, social media platforms and investor forums to learn about investors' attitudes and short-term market outlooks [1, 6].

Address missing values and outliers in the financial data first to improve the quality of the financial data. Finally, the different scales of the financial indicators are normalised. Tokenization, stop-word removal and lemmatisation are carried out on the text data to prepare them for analysis. NLP technology and BERT models are used to get the sentiment scores, and contextual information in the text is extracted [7]. Finally,

the time intervals of the financial and sentiment features are normalised to form an all-sample dataset for the prediction model [8].

2.2 Model Design

Some machine learning models have been employed in this paper, including but not limited to Logistic Regression, Decision Tree, Random Forest, XGBoost, SVM, BERT and LSTM. Logistic Regression is the basis model because it is simple and easy to interpret. Decision Trees are easy to interpret, and Random Forest and XGBoost can increase the accuracy of prediction by using many trees and reducing variance [2,9].

SVM has no linearity of data, BERT can extract the context information of sentiment expression in language, and LSTM is used to model the temporal dependence of stock price data. The old ARCH model is still relatively simple, but one can learn to grasp the level of market fluctuation to some extent [10].

Table 1 shows a small number of the models, such as Logistic Regression, Random Forest, XGBoost, KNN and boosting methods; the other models will be discussed in the text to save space here.

2.3 Experimental Design

Financial and sentiment features are combined, and then the data are divided into training and test sets by time. This way, the order of the data can be preserved; that is to say, past information will be used to predict future prices and prevent information leakage among the datasets. Cross-validation is used to optimise hyperparameters and improve the generalisation ability of the model; that is to say, it works well on many divisions of the data.

MSE and RMSE are used to measure the degree of difference between the predicted and actual stock prices; that is, how far apart they are. Compare them to determine how much the addition of the financial and sentiment indices can enhance the predictive power of the model; at the same time, all types of machine learning algorithms (Logistic Regression, Decision Tree, Random Forest, XGBoost, SVM, BERT, LSTM, etc.) will be used. ARCH models for time series are also a type of conventional method that can be used to study fluctuations in the market.

Table 1 shows a small number of models, such as Logistic Regression, Random Forest, XGBoost, KNN, and boosting methods, and the rest are introduced in the text to provide a general introduction to the model performance and feature contribution [3, 11, 12].

3 Results

3.1 Model Performance

Table 1 shows the training and cross-validation errors of all the models.

Table 1. Predictive performance of different models

Model	Training error	Cross-validation error	Test error
Linear regression	4.1450	4.5205	3.2092
Random forest	2.4318	9.4861	31.4330
XGBoost	1.1538	9.9722	32.0254
KNN	4.4712	9.9683	29.1450
Boosting	1.7337	9.6999	31.7220

3.2 Result Analysis

Based on the above results, the traditional model of Logistic Regression and a linear baseline have higher prediction errors and are unable to consider the non-linear characteristics of financial markets. Many things affect the stock market at any given time, such as changes in investor sentiment and economic fluctuations; these are usually not considered simple linear models. Random Forest, XGBoost and Boosting are all ensemble methods that can achieve good accuracy and stable prediction for this reason. Many Decision Trees are used to model the complex relations of attributes and reduce variance [2]. Among all the models that were tested, XGBoost was the top performer; therefore, it had good boosting and regularization features and performed well without overfitting [9].

Based on the above analysis, the model combining financial indicators and market sentiment performs better than the model using only financial data. Sentiment features provide some information about short-term changes in investors' expectations and the market that are not included in the traditional financial indicators. Together, they can better predict the fluctuations in stock prices in the short run [6,9]. Table 1 shows the performance of some models, such as Logistic Regression, Random Forest, XGBoost, KNN and boosting methods, and the text will introduce the rest of the models to provide a comprehensive overview of model performance. Based on the above results, it is suggested that more data should be used to enhance the accuracy of forecasts and inform quantitative investment.

4 Conclusion

A model of machine learning has been built to predict the stock price tomorrow based on changes in market sentiment and financial indicators. The combination of numerical financial data and textual sentiment information is used to better understand the reasons for changes in market behaviour. Both kinds of features are better than a model based only on financial indicators for the prediction according to the experiments. Among the models that were tested, XGBoost performed best; it is an ensemble learning method that uses boosting and regularization to solve the problem of overfitting and thus can extract complex patterns in the data.

The research shows that sentiment information can reveal the present outlook and response of the market by investors, etc., which are not covered by the conventional financial indicators. In addition to the above financial indicators, some other indices from alternative data can also be added to help the forecasting model better respond to changes in the market. Therefore, it is suggested to collect a large number of data from all parts of the market to predict stock prices.

In the future, more data sources will be added to the research, such as analyst reports, trends in social media, other types of alternative information, etc., and at the same time, advanced deep learning methods, such as graph neural networks and attention-based models, will also be used. Therefore, it will be more feasible to make forecasts and support the decision for quantitative investment. Expanding the framework in the above directions can also provide some reference for the market development of different sectors at different times and offer practical ideas for research and investment.

References

1. Brown, G.W., Cliff, M.T.: Investor sentiment and the stock market. Journal of Empirical Finance 11(1), 1–27 (2004)
2. Chen, T., Guestrin, C.: XGBoost: A scalable tree boosting system. In: Proc. 22nd ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining, pp. 785–794 (2016)
3. Li, X., Wu, P., Wang, W.: Incorporating stock price volatility and investor sentiment for stock price prediction: A deep learning approach. IEEE Access 8, 123260–123271 (2020)
4. Fama, E.F., French, K.R.: Common risk factors in the returns on stocks and bonds. Journal of Financial Economics 33(1), 3–56 (1993)
5. Fischer, T., Krauss, C.: Deep learning with long short-term memory networks for financial market predictions. European Journal of Operational Research 270(2), 654–669 (2018)
6. Tetlock, P.C.: Giving content to investor sentiment: The role of media in the stock market. Journal of Finance 62(3), 1139–1168 (2007)
7. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: BERT: Pre-training of deep bidirectional transformers for language understanding. In: Proc. NAACL-HLT, pp. 4171–4186 (2019)
8. Mittermayer, M.A., Knolmayer, G.F.: Text mining techniques for news-based stock prediction. Business & Information Systems Engineering 48(6), 405–414 (2006)
9. Zhang, L., Wu, Y.: Financial report analysis and stock prediction based on XGBoost. Computational Economics 58(3), 857–876 (2021)
10. Engle, R.P.: Autoregressive conditional heteroskedasticity with estimates of the variance of United Kingdom inflation. Econometrica 50(4), 987–1002 (1982)
11. Rögnvaldsson, T., Óskarsdóttir, K., Stefánsson, S.: Machine learning for stock market prediction: A review. Applied Computing and Informatics (2020)
12. Zhou, Y., Wang, H.: Ensemble learning with news sentiment for stock trend prediction. Applied Intelligence 52(8), 9102–9116 (2022)

Open Access This chapter is licensed under the terms of the Creative Commons Attribution-NonCommercial 4.0 International License (http://creativecommons.org/licenses/by-nc/4.0/), which permits any noncommercial use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.

