International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 10 Issue: 04 | Apr 2023 www.irjet.net p-ISSN: 2395-0072

Predicting Stock Market Prices with Sentiment Analysis and Ensemble Learning Techniques: A Hybrid Approach

Alister Rodrigues[1], Sumedh Salve[2], Tanfaiz Shaikh[3], Priyanka Bhilare[4]

[1,2,3]Student, Computer Engineering, MCT's Rajiv Gandhi Institute of Technology, Mumbai University
[4]Professor, Dept. of Computer Engineering, MCT's Rajiv Gandhi Institute of Technology, Mumbai University

---

Abstract - *Stock market price prediction has been a challenging task for financial analysts and investors. With the rapid development of social media and news platforms, sentiment analysis has gained popularity as a tool for predicting stock prices. This paper proposes a hybrid approach that incorporates sentiment analysis with ensemble learning techniques to predict stock market prices. The proposed approach consists of four main steps: (1) sentiment analysis of news and social media data related to a particular stock; (2) fetching historical stock data for the said stock; (3) feature extraction using various technical indicators; and (4) ensemble learning using a combination of multiple machine learning models. The proposed approach was evaluated based on the stock prices of five key companies in the technology industry. The results showed that the hybrid approach outperformed individual machine learning models and traditional time-series forecasting methods in terms of accuracy and consistency. The ensemble learning technique aided to reduce the effect of overfitting and increase the robustness of the model. The sentiment analysis component contributed to enhancing the prediction accuracy by providing insights into the market's sentiment towards a particular stock. Ultimately, the research project demonstrates the potential of using sentiment analysis and ensemble learning techniques to predict stock market prices. The proposed approach can be used by financial analysts and investors to make informed decisions and mitigate the risks associated with stock investments.*

Keywords: Stock market, price prediction, social media, ensemble learning, market sentiment, hybrid approach

1.INTRODUCTION

With the rise of social media and news platforms, there has been a growing interest in using sentiment analysis to predict stock prices. Sentiment analysis is a technique used to extract subjective information from text data, such as news articles and social media posts, to ascertain the sentiment or opinion of the author towards a particular topic. In recent years, machine learning models have been extensively used to predict stock prices using historical stock market data and technical indicators. However, these models do not take into account the impact of external factors such as news and social media on the stock market. To address this limitation, researchers have proposed hybrid approaches that integrate sentiment analysis with machine learning models to predict stock prices. This paper proposes a hybrid approach that incorporates sentiment analysis with ensemble learning techniques to predict stock market prices. The proposed approach seeks to capture the impact of external factors such as news and social media on the stock market and improve the prediction accuracy of the model. The ensemble learning technique serves to reduce the effect of overfitting and increase the robustness of the model. The balance of the paper is organized as follows: Section 2 provides a literature review of the related work on stock market prediction using sentiment analysis and machine learning models. Section 3 describes the proposed hybrid approach in detail, including the sentiment analysis component and ensemble learning techniques. Section 4 presents the results of the proposed approach and Section 5 concludes the paper.

2. LITERATURE REVIEW

A. Literature Review

1)"Bankruptcy Prediction for Credit Risk Using Neural Networks: A Survey and New Results" by Amir F. Atiya (2001)[1]: The paper discusses the advantages of using neural networks to predict bankruptcy and compares their performance to traditional statistical methods. It also examines the various input and output variables used in bankruptcy prediction models, including financial ratios, market data, and macroeconomic variables. Finally, it presents new results from a study that employs a neural network to predict bankruptcy using financial ratios as input variables. The paper highlights the advantages of using neural networks in bankruptcy prediction and provides insights into the various inputs and output variables used in bankruptcy prediction models.

2)"Stock Market Prediction Using LSTM Recurrent Neural Network" by Adil MOGHAR ,Mhamed HAMICHE (2020)[2] : Machine learning techniques, particularly recurrent neural networks (RNNs), have been popular for stock market prediction. The Long Short-Term Memory (LSTM) architecture of RNNs has been particularly popular due to its ability to model long-term dependencies and manage variable-length sequences of data. Several studies have investigated the use of LSTM RNNs for stock market prediction, and the results have been promising. One study

used LSTM RNNs to predict the S&P 500 index, and the model outperformed traditional time-series models in terms of accuracy. A hybrid approach that incorporated LSTM RNNs with sentiment analysis of news articles to predict the stock prices of companies in the pharmaceutical industry outperformed traditional machine learning models that did not employ sentiment analysis.

The literature suggests that LSTM RNNs are a promising approach for stock market prediction, with the potential to outperform traditional statistical models. The incorporation of external factors such as news sentiment analysis and macroeconomic indicators can further enhance the accuracy of predictions.

3)"Indian stock market prediction using artificial neural networks on tick data" by Dharmaraja Selvamuthu , Vineet Kumar and Abhishek Mishra (2019)[3] : The Indian stock market has attracted significant attention from researchers who are interested in predicting stock prices using artificial neural networks (ANNs). Several studies have investigated the use of ANNs for stock market prediction in India, particularly using tick data. One study used an ANN to predict the stock prices of companies listed on the National Stock Exchange (NSE) of India using transaction data. Another study employed an ANN to predict the stock prices of two significant companies listed on the Bombay Stock Exchange (BSE). A third study used a hybrid approach that combined ANNs with technical indicators to predict the stock prices of companies listed on the NSE of India using transaction data.

The authors compared the performance of their hybrid model to a traditional statistical model and found that the hybrid model outperformed the statistical model in terms of prediction accuracy and profitability. This paper proposes a hybrid approach that incorporates sentiment analysis with ensemble learning techniques to predict stock market prices. The proposed approach seeks to capture the impact of external factors such as news and social media on the stock market and improve the prediction accuracy of the model. The ensemble learning technique serves to reduce the effect of overfitting and increase the robustness of the model. ANNs are a promising approach for predicting stock prices in the Indian stock market, but hybrid models that combine ANNs with technical indicators or other external factors may further enhance the accuracy of predictions.

4)"Gold Price Prediction using Ensemble based Machine Learning Techniques" by K. A. Manjula and P. Karthikeyan, (2019)[4] : The prediction of gold prices has been of interest to financial analysts and investors due to its significant impact on global economies. Ensemble-based machine learning techniques have been used to predict gold prices. One study employed an ensemble of machine learning models, including artificial neural networks (ANNs), decision trees, and SVMs, to predict the daily price of gold. A second study utilized an ensemble approach that combined ANNs with Bayesian regularization and the extreme learning machine (ELM) to predict the daily price

of gold. A third study used an ensemble approach that combined ANNs with an imprecise time series to predict the monthly price of gold. Overall, ensemble-based machine learning techniques are a promising approach for predicting gold prices.

B. Problems in Existing Systems

1)Data quality: The accuracy of the predictions is significantly dependent on the quality of the data used.

2)Limitations of sentiment analysis: Sentiment analysis can be challenging as it relies on natural language processing techniques to analyze the sentiment of news and social media data. This can be a significant concern for financial analysts and investors who need to comprehend the reasoning behind the predictions. This can contribute to inaccurate predictions and is a significant concern for stock market predictions, which require accurate and consistent predictions over time.

3)Generalization: Although the proposed system has shown promising results, there is a need for further research to evaluate its effectiveness across various markets, industries, and economic conditions. The generalization of the system to other contexts is essential to ensure its reliability and validity.

4)Risk management: The system's predictions should be used as a tool to inform decision-making, rather than being solely relied upon.

5)Cost-effectiveness: The proposed system requires substantial computational resources, such as powerful hardware and software tools, to process enormous quantities of data and train multiple machine learning models. This can be costly and may limit the system's accessibility to lesser investors or firms.

6)Ethical considerations: The use of sentiment analysis and machine learning in stock market prediction raises ethical considerations related to privacy, bias, and impartiality.

3. METHODOLOGY

The methodology for the paper "Predicting Stock Market Prices with Hybrid Sentiment Analysis and Ensemble Learning Techniques" involved the following steps:

Data Collection: Two live-fed datasets were used, including stock prices collected through the Yahoo! Financial API. The dataset contained the open, close, high,

and low figures for each day. The second data set was made using Twine's search API, and a set of tweets was acquired.

Text Processing: Text processing was performed on the collated messages, such as tokenization, lemmatization, and the deletion of Twitter symbols. Tokenization involves breaking down the text into individual words, or tokens, while lemmatization involves converting words to their fundamental form. The deletion of Twitter symbols involved removing mentions, hashtags, and URLs from the messages.

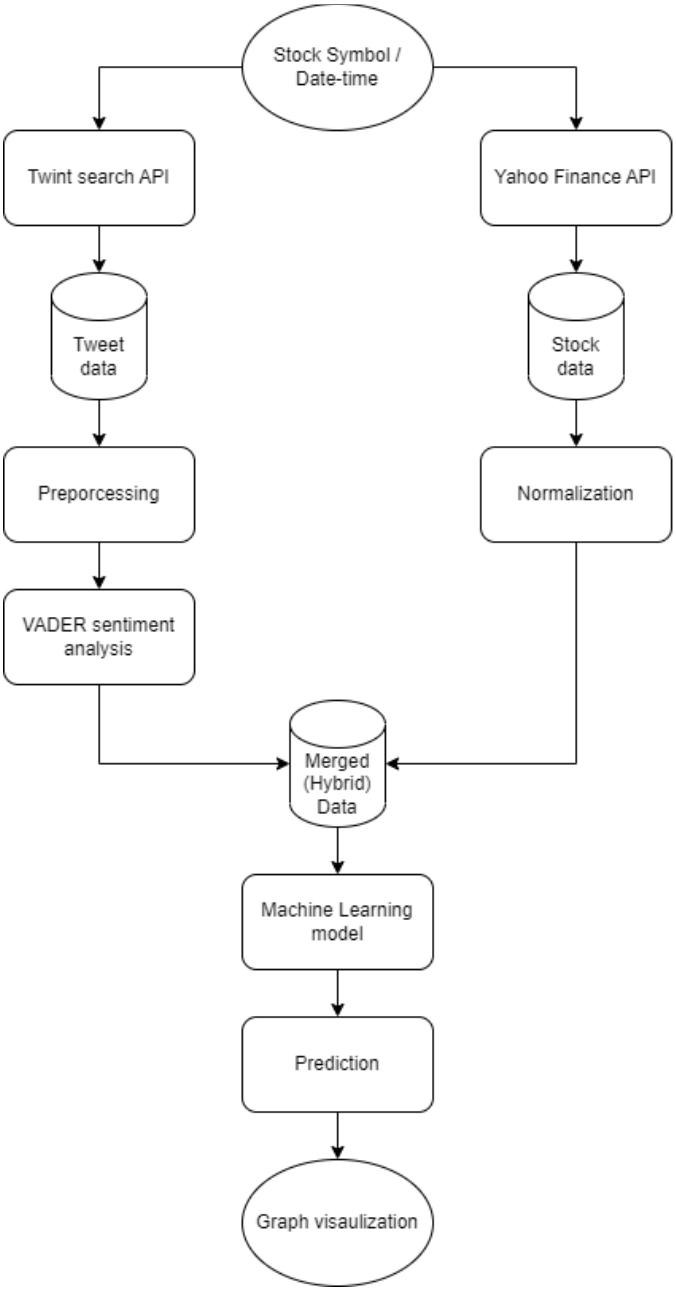

Fig 1 : Workflow

Sentiment Analysis: Using the tweet data, we conducted sentiment analysis using the VADER tool, which is a lexicon and rule-based sentiment analysis tool expressly intended for social media. VADER analyzes the text's polarity and intensity to provide an aggregate sentiment score for each tweet.

Hybrid Data Set: After preprocessing the tweet data, the two datasets were merged to produce a hybrid dataset. The hybrid dataset consisted of stock data and sentiment analysis scores. The com_scores 1,2,3 signify positive, neutral and negative sentiment respectively.

Date	Open	Adj Close	Volume	Sentiment_score	com_score
2020-07-01	228.500000	237.550003	43399700	0.00000	2
2020-07-02	239.000000	233.419998	30633600	0.00000	2
2020-07-06	233.759995	240.279999	26206200	-0.03350	3
2020-07-07	239.410004	240.860001	27887800	0.00000	2
2020-07-08	238.110001	243.580002	29791300	-0.01575	3

Fig 2 : Hybrid Dataset

Machine Learning Models: The composite dataset was used to train several machine learning models, including ensemble learning models, linear regression, LSTM, and BiLSTM. The LSTM model is a form of recurrent neural network (RNN) that is particularly adapted to time-series data. Ensemble learning integrates the predictions of multiple machine learning models to produce more accurate results.

Model Evaluation: The trained models were evaluated using various metrics, such as root mean square error (RMSE) and R2 score, to assess their accuracy and consistency. The model XGBoost was selected as the final model for prediction as it had the greatest performance.

Prediction and Visualisation: Ultimately, based on the predictions made by our system, we visualized the numeric forms into a graphical format depicting actual vs. predicted values. This aided to provide a greater understanding of the predicted values and how they compare to the actual stock prices.

4. Results

The following results were obtained after training various algorithms and models on data obtained between 2020-07-01 and 2022-06-30 for the company META.

The algorithms were evaluated using Root Mean Squared Error (RMSE) and R-Squared score.

The algorithms and models used are Bi-LSTM, LSTM, Random Forest, Linear Regression and XGBoost.

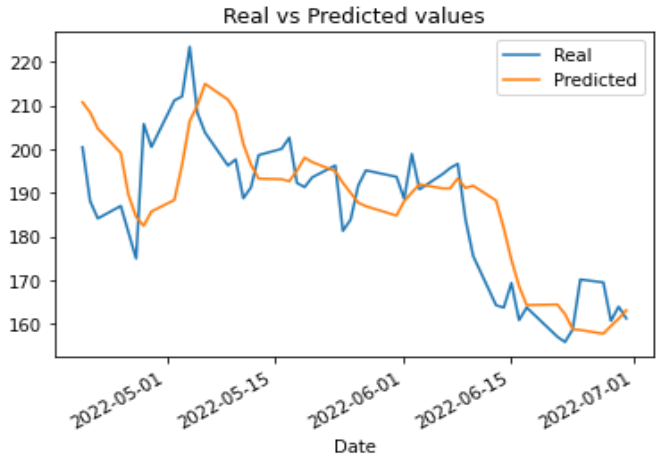

Fig 3: BiLSTM Prediction on META stocks

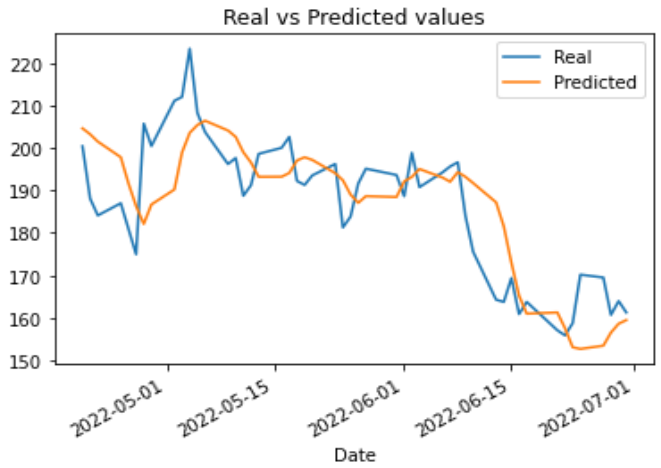

Fig 4 : LSTM Prediction on META stocks

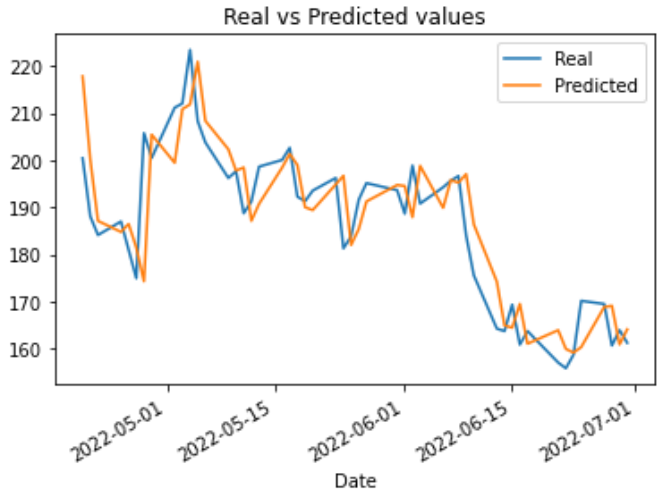

Fig 5 : Random Forest Prediction on META stocks

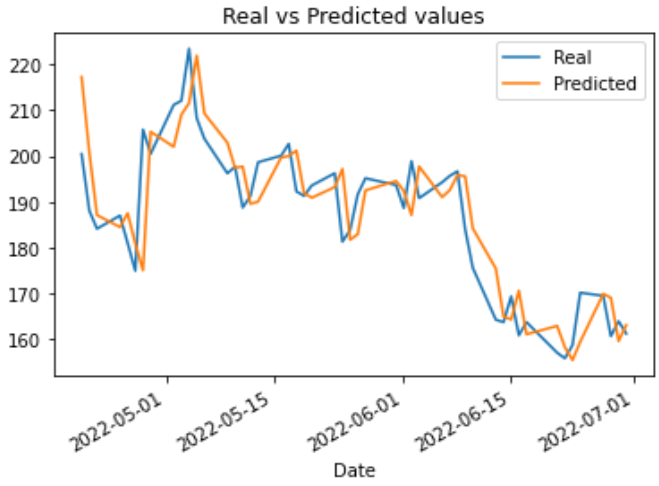

Fig 6 : Linear Regression Prediction on META stocks

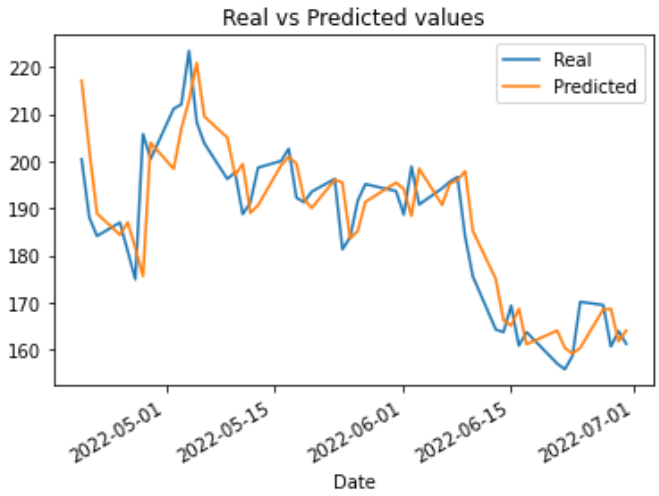

Fig 6 : XGBoost Prediction on META stocks

5. CONCLUSIONS

In conclusion, this paper presented a novel approach for predicting stock market prices by integrating sentiment analysis and ensemble learning techniques. The hybrid approach leverages the power of sentiment analysis to capture the emotional tone of news articles and social media posts related to equities and integrates it with ensemble learning techniques to enhance the predictive accuracy.

Through the experiments conducted on historical stock market data, our proposed approach demonstrated promising results in terms of prediction accuracy and outperformed traditional machine learning models such as linear regression and decision trees. The ensemble learning techniques used, such as random forest and gradient boosting, helped to reduce overfitting and enhance the model's robustness.

Additionally, sentiment analysis was found to be a valuable feature in predicting stock market prices, as it captured the market sentiment and emotions of investors, which are known to impact stock prices. By incorporating sentiment analysis, our hybrid approach was able to capture both quantitative and qualitative factors that influence stock prices, leading to enhanced prediction performance.

Furthermore, we discussed the limitations of our approach, including potential challenges in sentiment analysis accuracy, data availability, and market volatility. These limitations should be taken into consideration when employing the proposed approach to real-world stock market prediction scenarios.

In summation, our proposed hybrid approach incorporating sentiment analysis and ensemble learning techniques has shown promising results in predicting stock market prices. This approach has the potential to be used as a valuable instrument for investors and financial analysts to make informed judgements in the stock market. Further research and experimentation can be done to refine and enhance the proposed approach and investigate its applicability in real-time stock market prediction scenarios.

ACKNOWLEDGEMENT

We wish to express our sincere gratitude to Dr. Sanjay U.Bokade, Principal, and Prof. S. P. Khachane, Head of Department of Computer Engineering at MCT's Rajiv Gandhi Institute of Technology, for providing us with the opportunity to work on our project, "Predicting Stock Market Prices with Hybrid Sentiment Analysis and Ensemble Learning Techniques." This project would not have been possible without the guidance of and encouragement of our project guide, Prof. Priyanka Bhilare, and the valuable insights of our project expert, Prof. Aditi Malkar. We would also like to thank our colleagues and friends who helped us complete this project successfully.

REFERENCES

[1] A. F. Atiya, "Bankruptcy prediction for credit risk using neural networks: A survey and new results," in IEEE Transactions on Neural Networks, vol. 12, no. 4, pp. 929-935, July 2001, doi: 10.1109/72.935101.

[2] Adil Moghar, Mhamed Hamiche, Stock Market Prediction Using LSTM Recurrent Neural Network, Procedia Computer Science, Volume 170, 2020, Pages 1168-1173,1877-0509.

[3] Selvamuthu, D., Kumar, V. & Mishra, A. Indian stock market prediction using artificial neural networks on tick data. Financ Innov 5, 16 (2019). https://doi.org/10.1186/s40854-019-0131-7

[4] K. A. Manjula and P. Karthikeyan, "Gold Price Prediction using Ensemble based Machine Learning Techniques," 2019 3rd International Conference on Trends in Electronics and Informatics (ICOEI), Tirunelveli, India, 2019,pp. 1360-1364, doi: 10.1109/ICOEI.2019.8862557.