International Research Journal on Advanced Engineering and Management
https://goldncloudpublications.com
https://doi.org/10.47392/IRJAEM.2024.0203

e ISSN: 2584-2854
Volume: 02
Issue: 05 May 2024
Page No: 1504-1508

Web Application for Company Stock Price Prediction

Hitesh Chaudhari [1], Deepesh Bhede [2], Saurabh Thorat [3], Ayush Gharge [4], Deepa Ekhande [5]
[1,2,3,4]UG - Dept. of Information Technology, PCE, New Panvel, India.
[5]Professor- Information technology department, PCE, New panvel, India.
Email Id: *hchau20ite@student.mes.ac.in[1], dbhede20it@student.mes.ac.in[2], sthorat20ite@student.mes.ac.in[3], agharge20ite@student.mes.ac.in[4], deepaekhande@mes.ac.in[5]*

Abstract

This study contributes to predicting stock prices. In this paper, we have provided a novel machine learning and deep learning technique for stock prediction, which capitalizes on the complex emotional patterns of investors that can be extracted from Twitter data. By joining two separate datasets, such as Twitter sentiments and the associated stock prices, we have applied complex algorithms and evaluated this method, which classifies a user's sentiment into a positive and negative one. It was possible to address the complexity of investor behavior within the highly unstable stock market environment. One of the most important problems that are encountered by developers is gaining the possibility to properly capture stock price fluctuations in the temporal dimension. It implies the necessity to utilize such reliable evaluation metrics as precision-recall factors for classification purpose and root mean squared error for regression purpose. by using these metrics, the performance of the predictive model can be evaluated in detail, which allows to draw meaningful permanent conclusions within the context of how well it reflects the complex relationship between investor sentiment and market dynamics. The results of the evaluation conducted as part of the process of validating the proposed hypothesis show that, indeed, the predictive accuracy improves significantly from the use of machine learning and deep learning algorithms if paired with Twitter users' emotional sentiment data. This combination enhances the predictive capacity of the model and provides meaningful insights into how the changes in stock prices are impacted by the sentiment of investors. Moreover, in the process of this study, the explanation and detail consideration of the algorithms for feature engineering and data processing were provided, which stands crucial for increasing the predictive model's accuracy The study has a valuable implication for both the financial analysis and technical spheres due to the resolution of these methods and presents a model for further research activities in stock prediction. To conclude, this interdisciplinary strategy that combines sentiment analysis and stock prediction not only advances our understanding of the investor behavior but also brings the opportunity for novel applications in the emerging field at the juncture of technology and finance. As a result, the current research sheds light on the intricate connection between the dynamics of the stock market and investor sentiment, thus offering new insights that can better inform predictive modeling research in the financial industry.

Keywords: *Natural Language Processing (NLP), Machine learning (ML), Deep learning (DL).*

1. Introduction

The thesis consists of three stages: sentiment analysis, stock price prediction, and machine learning algorithms. The first stage uses sentiment analysis on a Twitter dataset to classify investor tweets about stock [1]. This process requires data pre-processing to filter out irrelevant texts. The second stage predicts stock prices using regression techniques, based on past stock prices. The final stage uses machine learning algorithms and deep learning to predict stock prices. All algorithms are evaluated for accuracy, and the most accurate algorithm is selected for future price prediction. The thesis's primary objective is to develop an intelligent model that provides investors with detailed information on acquiring assets in the stock market. The model must be robust and require hyper parameter tuning at the end of the training phase [2].

2. Literature Survey

This section views state-of-the-art efforts to identify stock-market prices using sentimental analysis and deep learning on tweets text data [3]. The study by Yichuan Xu and Vlado Keselj IEEE (2019) uses the attention based Lstm variant for the prediction.Their approach involves combining of finance tweets sentiment and stock technical indicators to gains better performance from this modified LSTM.Accuracy of this model is 56% but future improvements could increase the number of percentage. Mudinas, Zhang, and Levene (2019) classified sentiments from news and tweets into eight categories (such as fear and anger) [4]. Only a small number of sentimental emotions were somewhat correlated with subsequent stock movements. Without taking into account emotional factors, technical factor-based predictions typically yield better results. Khedr & Yaseen (2017) also came to the same conclusion. To determine the weight for each token and categorize gathered financial news into positive and negative sentiment categories, they employed the TF-IDF and N-gram methods [5]. The K-NN classifier yielded 59.18% accuracy for the sentiment attributes method and 89.80% accuracy for the sentiment and historical stock data method. The author did not, however, demonstrate the accuracy attained using just historical data [6]. Technical aspects appear to be the deciding elements in stock prediction while extracting sentiment attributes have harmful effects on results. Financial articles from 2016-01-01 to 2020-04-01 were used by Kabbani & Usta (2022) to forecast the daily stock trend. Given the speed at which opinions are shifting, only the trends for today and tomorrow are anticipated. High correlation characteristics with the article's emotion scores were chosen for the final data set following correlation analysis [7]. With the usage of random forest, gradient boosting machine algorithm, and linear regression, the model achieved an average accuracy of 63.58%. A model was presented by Weng, Ahmed, and Megahed (2017) to forecast stock movement one day in advance. In contrast to sentiment analysis, the author's forecast methods included market data, Wikipedia traffic, Google news counts, and a range of technical indications. The model only took into account changes in Apple shares between May 1, 2012, and June 1, 2015. combining information from multiple sources, the accuracy is 85.8%, indicating that increased data can improve the prediction. The study by R.Satishkumart JETIR (2020) shows the stock price forecast system combines the successful use of long short-term memory for technical analysis and sensitive analysis for fundamental analysis to provide accurate results. Sentiment analysis was conducted using the selected keyword [8]. The user who is not familiar with the stock market might find this system handy. Individuals with varying levels of trading experience can utilize this approach to forecast stock prices in the future. In the future, the technology will allow for intra-day prediction and enhance sentiment analysis to obtain an impact factor for fundamental analysis free of sarcasm. The study by Hatefi Ghahfarrokhi examines how SM data may forecast Tehran Stock Exchange (TSE) factors by taking into account the closing prices and daily returns of three different stocks [9]. A three-month period of StockTwits was collected. A learning-based and lexicon-based approach was put forth to get information from internet forums. Furthermore, since the current Persian lexicons are unsuitable for SA, a bespoke sentiment lexicon was developed. Following the creation and computation of daily sentiment indices based on comments, novel predictor models using multi-regression analysis were put forth [10]. The investigation also took into account the quantity of comments and the reliability of the individuals. Results indicate that characteristics of TSE stocks affect how predictable they are. It is demonstrated that mood and comment volume may be useful for estimating the daily return, and that the trust coefficients of the three stocks react differently [11].

3. Methodology

In this section, we provide an in-depth exploration of the techniques employed in our project which involves preprocessing, sentiment analysis, model development, model training and performance measurement [12].

3.1. Preprocessing

We first convert the tweets CSV file into a Pandas dataframe. Since the stock market is not open on the weekends and specific holidays, such as Good Friday and Thanksgiving, we dropped these days [13]. Following are the steps we will perform for the preprocessing the data using the NLTK:

- Remove HTML entities
- Substitute @mentions, urls, etc. with whitespace using regular expressions
- Substitute any non-alphabetic whitespace.
- All the words in lowercase.
- Removing stop words.
- Perform stemming of words with lemmatization

After the preprocessing using the Natural Language Toolkit, we will perform the sentiment analysis on each tweet. To generate the sentiment score we shall use **VADER sentiment analysis.** Empty columns were ignored, and each tweet is classified as positive (value of +1) and negative (value of 0). We then averaged the individual sentiment values so that a single sentiment value was present for every day. We then converted the stock data CSV file into a Pandas dataframe and dropped the columns that weren't needed. We will join the stock data dataframe and the sentiment data frame into a single dataframe [14].

3.2. Sentiment Analysis Model: Vader

The sentiments from the tweets are calculated from the dataset that contains twitter data of daily tweets. This data contains keywords such as #TSLA. The sentiments of tweets are calculated using the twint library of Python. This library is responsible for providing simple APIs for sentiment analysis with regards to NLP related research.

3.3. Model Development

Random Forest is an ensemble method that fits multiple decision trees on different datasets to improve predictive accuracy. It has a large number of parameters that can be optimized, including the number of trees and features to consider. XGBOOST is a method of combining weak learners to a strong learner, with participants that were incorrectly classified weighted more heavily than those correctly classified. LSTM, or Long Short-Term Memory, is an ensemble method that uses memory blocks in the recurrent hidden layer to store the temporal state of the network and gates to control information flow. Bidirectional long-short-term memory (Bi-LSTM) and Stacked Bidirectional long-short-term memory (Stacked Bi-LSTM) is a process that allows input to flow in both directions, preserving future and past information. These methods are used for regional features and have different computational costs.

3.4. Model Training

When we got a sentimental score for each single day then we converted the stock data CSV file into a Pandas dataframe and dropped the columns that weren't needed. We will join the stock data dataframe and the sentiment data frame into a single dataframe and pass to a model for training.

3.5. Performance Measurement

For performance measurement root mean squared error (RMSE) and R2 score is used. The results indicate that incorporating emotional sentiment of users improves the overall performance of the system [15].

3.6. Sample Dataset Used

We will use two datasets for this project.

Stock Price Data: Daily opening prices of the last 5 years for 5 different companies (Amazon, Apple, Facebook, Google, Microsoft, Netflix, Tesla, etc). This data was collected from Yahoo! Finance.

Tweet Data of Stocks: Twitter data containing approximately 30-50 tweets per day. will be collected using the Twitter API.

The tweets included the username and the text of the tweet. The tweets will be split based on their given date, since our project is focused on predicting daily opening stock prices in Figure 1.

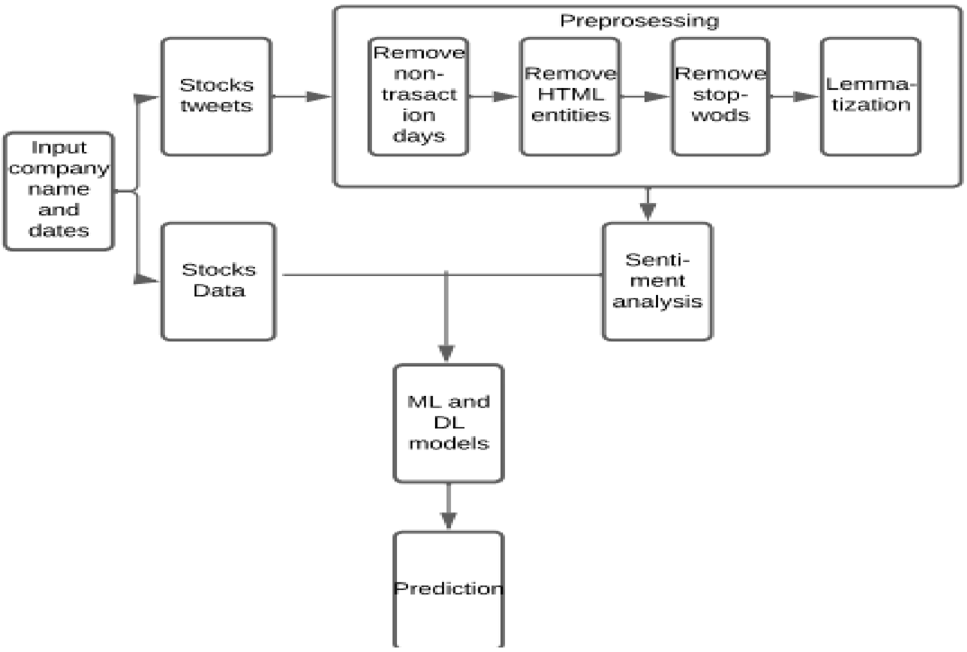

Figure 1 System Architecture

4. Result and Analysis

In this result section we show the accuracy obtained by the multiple machine learning and deep learning algorithms.We use the apple tweets data and stocks data for result.Among the all models Stacked BI-LSTM gives the highest accuracy and Random forest regression model gives the lowest accuracy.For training and testing purpose we split the dataset into two parts.First 90% use for training and remaining is used for testing purpose.

Table 1 Accuracy for APPLE

Models	Accuracy
Random Forest	66.68%
XgBoost	78.89%
LSTM	80.11%
BI-LSTM	83.56%
Stacked BI-LSTM	89.43%

For performance measurement root mean squared error (RMSE) and R2 score is used in Table 1.

Conclusion

Through this project, we have showcased how the sentiment analysis of tweets can be utilized to give investors meaningful insights regarding the stocks of different companies. Using a Python-based library, twint, to collect the data, the project has employed several machine learning techniques such as Random Forest, XGBOOST, and Linear Regression to achieve detailed and accurate results. Utilizing LSTM, BI-LSTM, and Stack BI-LSTM have enabled the Deep Learning Model to process the data tiresomely into something that is straightforward and readable while doing it from both directions. Our project is intended to provide investors with refined points so that they can truly invest in a way they would like to see. Hence, the overall project indicates the researcher’s knowledge of combining sentiment analysis and highly advanced machine learning techniques which offers valuable insights for investment community.

Acknowledgements

We would like to sincerely thank each and every contributor for their insightful contributions to this study report. Their combined knowledge and efforts greatly improved the caliber and substance of this work.

References

[1]. De Gooijer JG, Hyndman RJ (2005) 25 years of IIF time series forecasting: a selective

review. Soc Sci Electron Publ 22(3):443–473.

[2]. J. Jagwani, M. Gupta, H. Sachdeva, and A. Singhal, “Stock Price Forecasting Using Data from Yahoo Finance and Analyzing Seasonal and Nonseasonal Trend,” in 2018 Second International Conference on Intelligent Computing and Control Systems (ICICCS), Madurai, India, Jun. 2018, pp. 462–467, doi: 10.1109/ICCONS.2018.8663035.

[3]. Y. Lei, K. Zhou, and Y. Liu, “Multi-Category Events Driven Stock Price Trends Prediction,” in 2018 5th IEEE International Conference on Cloud Computing and Intelligence Systems (CCIS), Nanjing, China, Nov. 2018, pp. 497–501, doi:10.1109/CCIS.2018.8691392

[4]. B. Jeevan, E. Naresh, B. P. V. kumar, and P. Kambli, “Share Price Prediction using Machine Learning Technique,” in 2018 3rd International Conference on Circuits, Control,Communication and Computing (I4C), Bangalore, India, Oct. 2018, pp. 1–4, doi: 10.1109/CIMCA.2018.8739647.

[5]. M. Usmani, S. H. Adil, K. Raza, and S. S. A. Ali, “Stock market prediction using machine learning techniques,” in 2016 3rd International Conference on computer and Information Sciences (ICCOINS), 2016, pp. 322–327.

[6]. J. Du, Q. Liu, K. Chen, and J. Wang, “Forecasting stock prices in two ways based on LSTM neural network,” in 2019 IEEE 3rd Information Technology, Networking, Electronic and Automation Control Conference (ITNEC), Mar. 2019, pp. 1083–1086, doi: 10.1109/ITNEC.2019.8729026.

[7]. S. E. Gao, B. S. Lin, and C.-M. Wang, “Share Price Trend Prediction Using CRNN with LSTM Structure,” in 2018 International Symposium on Computer, Consumer and Control (IS3C), Dec. 2018, pp. 10–13, doi: 10.1109/IS3C.2018.00012.

[8]. T. Gao, Y. Chai, and Y. Liu, “Applying long short term memory neural networks for predicting stock closing price,” in 2017 8th IEEE International Conference on Software Engineering and Service Science (ICSESS), Beijing, China, Nov. 2017, pp. 575–578, doi:

[9]. Dutta U, Hanscom JS, Zhang R, Han T, Lehman Q, Lv R, Mishra S. 2021. Analyzing twitter users’ behavior before and after contact by the Russia’s internet research agency. Proceedings of the ACM on Human-Computer Interaction 5(CSCW1):1-24

[10]. Hao Z, Chen-Burger Y-HJ. 2022. An investigation into influences of tweet sentiments on stock market movements. In: Agents and multi-agent systems: technologies and applications. Singapore: Springer Nature Singapore. 87-97

[11]. García-Méndez S, De Arriba Pérez F, Barros-Vila A, González-Castaño FJ. 2022. Detection of temporality at discourse level on financial news by combining Natural Language Processing and Machine Learning. Expert Systems with Applications 197:116648

[12]. Ahmar AS, Del Val EB. 2020. SutteARIMA: short-term forecasting method, a case: Covid-19 and stock market in Spain. Science of the Total Environment 729:138883

[13]. Biau G, GC, Koehler L, Wattelle P-H. 2018. Random forests and decision trees: a comparison. Journal of Machine Learning Research 19:1-25

[14]. Kinyua JD, Mutigwe DJ, Cushing C, Poggi M. 2021. An analysis of the impact of president trump’s tweets on the djia and S & P 500 using machine learning and sentiment analysis. Journal of Behavioral and Experimental Finance 29:100447

[15]. Maqsood H, Mehmood M, Maqsood M, Yasir S, Afzal F, Aadil MM, Selim I, Muhammad K. 2020. A local and global event sentiment based efficient stock exchange forecasting using deep learning. International Journal of Information Management 50:432-451 Paper ID: 2404234_ Web Application for Company Stock Price Prediction.