Turkish Journal of Computer and Mathematics Education *Vol.11 No.01 (2020), 878-885*
DOI: https://doi.org/10.17762/turcomat.v11i1.13585
Research Article

Forecast of Stock Market using Machine Learning Strategies

Bhanu Prakash Dubey
Department of Computer Science & Information Technology, Graphic Era Hill University, Dehradun Uttarakhand India 248002

Abstract
Accurately predicting the stock market is one of the things that investors are most interested in since it may help them make money in the economy. Given that such markets are significantly influenced by volatility and news, it is difficult to predict stock prices, which are solely dependent on market timing. Owing to this difficulty and volatility, it is vital to evaluate stock forecasting using historical data as well as external variables such as investor behaviour, social media, and financial news. Thus, this study recommends using regression and machine learning algorithms to estimate the price of equities on upcoming days based on investor sentiment. The experiment is conducted on Yahoo Finances and combines Twitter repository data on investor sentiment. In the subsequent step, the concept of sentiment analysis is applied to the monitoring of Twitter user tweets. The tweets are then sorted into positive and negative groups based on the sentiment score. In addition, machine learning algorithms are used to forecast Yahoo Finance stock values. To solve this issue, we propose reducing the complexity of time sequence models by employing regression approaches that integrate a hybridized concept of sentiment analysis and machine learning algorithms, which may result in higher accuracy. The testing results validate the best linear regression prediction accuracy and demonstrate an overall system performance enhancement.
Keywords: finance, forecast, machine learning, predictions, volatility

Introduction

Financial market frequently refers to the exchange of monetary assets between persons. A crucial element of the financial sector is the stock exchange, where commodities and stocks are traded. This offers numerous opportunities for major corporations to invest and earn substantial sums of money. In addition to corporations, stock traders invest their money in the stock market by selling or holding stocks. In order to acquire profits and make money, traders must track down stock prices that are anticipated to rise and sell stocks that are anticipated to fall [1]. Traders' decisions on the purchase and sale of stocks depend heavily on stock movements, which must be precisely anticipated. On the one hand, this decision relies heavily on social media research, while on the other hand, investors cannot rely just on financial news websites because these websites provide a vast volume of data. As a result, it is difficult for traders to precisely predict the stock market because all financial markets are highly dependent on external sources such as social networking platforms and business news. Thus, an autonomous system that might make decisions on behalf of investors by analysing previous stock movements is required. Using the foundations of machine learning, a large number of researchers created this automated method. Machine Learning offers algorithms that glean insights from financial news and leverage vast amounts of past data to forecast future stock prices. Two main determinants of a trader's decision [2] are the sentiments associated to social media posts and the corresponding financial information of equities. So, these elements must be addressed while developing a stock market prediction system. The stock market has always displayed a high degree of volatility, despite being largely recognized as the most prolific source of revenue from investments. The stock is so volatile because the activities of the company's directors have a direct effect on its price and value. The level of competition on the market does not assist business decisions because it is often unpredictable. Consequently, the investor's objective is to construct a portfolio that can buy and sell equities as necessary to achieve a balance of return components. In order for this to occur, it is necessary for investors to comprehend mathematical models so that their decision-making capacity is supported by statistical data. This resulted in the invention of Portfolio Optimization Problem (POP), a strategy for assigning complex tasks based on historical events and then presenting them to the most suitable investors [3]. The impact of social media platforms on the stock market is depicted in Figure 1.

DOI: https://doi.org/10.17762/turcomat.v11i1.13585

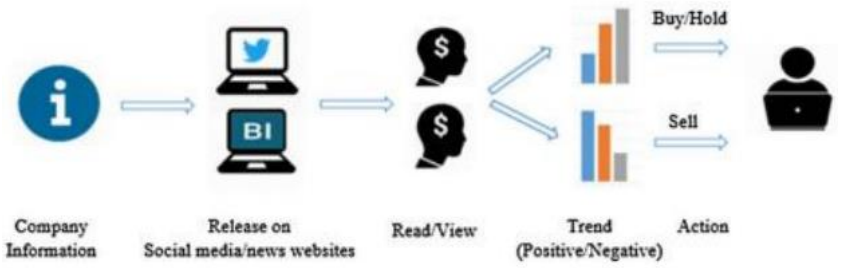


Figure1: Sentimental Impact on Stocks

According to recent studies and research, the volatility of the stock market is due to a variety of factors. The news media's production of information is one such factor. In addition to the facts and statistics derived from news sources, there are additional elements that must be weighed in order to analyze the stock market and give investors with favourable outcomes. The deployment and usage of computational power and intelligence that can process vast volumes of information from not only enterprises but also the nation's financial sectors is the only answer that academics see for all of the aforementioned problems, according to their findings. Only then will the ability to make broader decisions rise, providing investors with better market investment options. The following are the primary effects of stock price fluctuations:

- The mechanism of fluctuations is one of the primary concerns when projecting stock prices. The market prices of stocks are often influenced by a number of factors, including political and economic considerations. All of these factors influence investor behavior, resulting in extreme volatility. Thus, market prices are constantly shifting. This change increases the likelihood of a price fall and has a substantial impact on the occurrence of mistake spaces. Occasionally, these price-related concerns cause investors to disregard economic factors, which has unanticipated effects on the nation's overall economic health
- The sentiment of investors regarding stocks is another factor influencing their pricing. Some investors may believe that investing a small amount of their savings is wise, while others may disagree. While feelings like as enthusiasm and confidence will motivate an investor to continue investing, emotions such as discouragement and fear may discourage an investor from continuing to do so
- A further challenge associated with projecting stock prices is the collection and storage of large volumes of data and related information on numerous companies' stocks. Retaining this information is necessary for future forecasts, but doing so for an extended period of time is arduous in and of itself

In recent years, machine learning has gained enormous popularity in the field of stock prediction by employing ensemble classifiers and hybridization approaches to improve the system's overall accuracy. In addition to machine learning, deep learning has become a prominent method for enhancing system efficiency. This paper proposes the deployment of machine learning algorithms to forecast the stock market by assessing the sentiments related to social media posts on Twitter and the stock prices of a stock listed on the stock market using Yahoo finance (Ex: TCS).

Related Works

The fundamental goal of this paper is to forecast the stock prices of the TCS dataset by analysing Twitter user attitudes. Many authors have used social media and news websites to predict the stock market in existing stock prediction systems. This has prompted the development of models capable of producing outputs with the highest degree of precision and yielding improved price prediction performance. This section of the paper describes systems based on machine learning and sentiment analysis that have been developed by researchers in the past.

It has been thoroughly observed that news information from news outlets has a major impact on stock prices. In addition, news presented on television channels has a considerable impact on the randomness of stock trading patterns. Several studies have proved the significance of economic factors in predicting stock prices and their impact on people's decisions to purchase specific stocks. As a result of the financial behaviour of equities, investors from all over the world are becoming more interested in the financial marketing environment. As evidence, social media has a big impact on the process of predicting stock prices and serves as the foundation

for interactions between investments that occur between investors. In general, these interactions should occur often. In order for technical analysts to obtain early insight into the network impacting investor sentiment, it is essential for them to watch all of the emotions exhibited on the stock market. Using these insights allows for more prudent decision-making.

Authors in [3] studied the concept of sentiment analysis and applied linguistic criteria to characterize user tweets. For word count and feature score, this criterion was applied to algorithms based on machine learning. Classifiers were linked with annotations including Bag of Words and Part of Speech so that the model could achieve the best level of accuracy.

Authors in [4] expanded on his previous work in [3], classifying tweets as positive or negative based on their heterogeneous structure. NB was the classification algorithm applied to this category. The authors in [5] developed a model in that could classify movie reviews as either favourable or negative based on Twitter sentiments. The model was implemented using three ML-based approaches, and their accuracy was subsequently compared. Furthermore to financial news, breaking news is also published online. Authors in [6] analysed businesses that employed workers with 10 years of experience by analysing Twitter sentiments. As a result, it is clear that sentiment analysis is a commonly employed technique that may derive emotions from textual data and unearth new information. Thus, the concept is currently garnering a great deal of interest in text analytics. Many machine learning (ML)-based models that can estimate future stock prices based on historical data are currently employed by stock market traders.

The authors in [7] used sentiment analysis as part of natural language processing (NLP) to poll customer opinion on social media sites. Both tracking box office estimates and market fluctuations utilized the same concept and methodology. In another study [8], the author asserted that less data was available, resulting in less research being conducted on the concept of sentiment analysis. Yet, this study was conducted at a time when the internet was less developed than it is today. As a result, online comments and text storage rose rapidly following the advent of the Internet.

In a different work [9], the author proposed utilizing deep learning techniques to anticipate the prices of stocks. Author claims that RNN and LSTM networks can be employed to classify and forecast prices based on historical past events. Implementation of the LSTM algorithm enhanced accuracy performance. In a subsequent study, the authors in [10] utilized ANN and CNN algorithms to predict price fluctuations in stocks. The layers found in CNNs were essential in determining the price of stocks by solving complex equations. Using CNNs, however, enhanced the accuracy of the same forecast.

For this reason, it is vital to maintain a daily log, and the resulting data is then presented using graphs and charts. After carefully evaluating the properties of stocks on these bar graphs and charts, the decision to buy or sell is taken. The authors of a second survey done by them in [11] concluded that, of all the approaches used to anticipate stock values, technical analysis was applied most frequently. A broad indicator was utilized by technical analysis to determine when to buy and sell specific stock currencies. Unfortunately, it was determined that these strategies did not generate a significant return on investment. As a result, scientists ran tests based on the significance of the material and shifted their focus to fundamental analysis. Yet, recent advancements in the fields of natural language processing and text analysis have been detected, which support the prediction and boost its accuracy.

Methodologies Used

Three phases comprise the entirety of the paper's working execution. Despite the fact that the three phases are distinct from one another, they finally work together to accomplish the thesis. In the first phase, sentiment analysis is utilised. Using a dataset extracted from the Twitter database, investors' tweets about stocks and Yahoo Finance are collected and analysed for sentiment. This analysis of sentiments collects information from textual format and classifies it as either positive or negative feedback using classification techniques. This concept is implemented by focusing on news linked to the companies whose stock values are being researched. To filter out unnecessary texts and only use relevant ones, however, this method requires data pre-processing operations. The second step concerns stock price forecasting. This prediction is based on a distinct Yahoo Finance dataset. Using regression methods, stock fluctuations are forecast. Because it is necessary to compute and anticipate future TCS stock values, this stage is heavily backed by the previous stage's findings. The final

step is the application of machine learning techniques to make the same forecast. After testing each algorithm against a set of metrics for accuracy, the algorithm with the highest accuracy is adopted to predict the future stock price.

A. Analysis of Investor Sentiment using VADER

When a person or investor intends to invest in stocks, social media and news sources are common sources of information. Also, he may gain knowledge from many websites and professional journals. The research advises employing sentiment analysis of Twitter tweets to comprehend how individuals feel about TCS stocks and how they've communicated that attitude through comments and other contextual expressions.

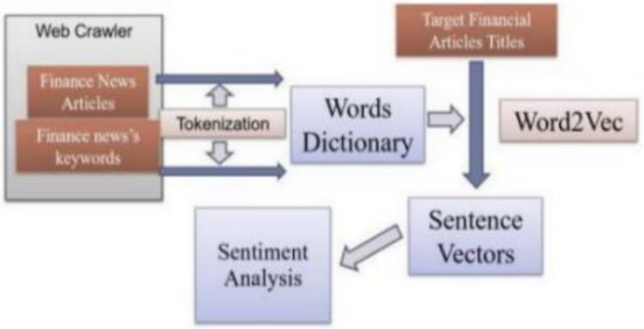


Figure 2: Collection of investor sentiments using VADER

The dataset including daily Twitter tweet data is used to calculate the sentiments of tweets. This data contains keywords like #financialstocks, #stockprices, #stockprediction, and #buystocks. Twitter sentiment is determined using the Python tweepy package. This library must provide user-friendly APIs for sentiment analysis for NLP-related research.

After that, the tokenization technique eliminates punctuation and white space. The tweets are then classified as either positive or negative using classification techniques based on machine learning. The primary objective of sentiment analysis is to collect historical data on investor sentiment towards Yahoo Finance. The sentiment index is calculated daily by integrating user comments and textual expressions of emotion. After acquiring the sentiment tendency, classification methods are employed to generate this sentiment index. The suggested thesis combines the sentiment analysis model with a VADER model that normally does not require a huge dataset in order to achieve high accuracy.

B. Stock Prediction Model

Owing to the unexpected nature of time series functions, the proposed research faces a variety of obstacles. One of the fundamental tenets of time series analysis is that a series of TCS stock prices is not inherently stable. Analysts have observed a periodic shift in the linkages between each model's sequence and its associated values at a certain point in time. To analyse the given sequence and predict its values, the thesis aims to develop a model. Following an investigation of emotions, the second section of the thesis forecasts TCS stock values. The datasets from both sources are integrated and sent into the pre-processing, extraction, and selection phase. The final dataset is delivered to the training and testing phases. During this phase, certain machine learning approaches are often employed. In the final step, results are obtained and evaluated against parametric functions to attain the appropriate level of precision.

Data Implementation

To successfully accomplish the proposed work, we employed machine learning algorithms for stock prediction and ideas of sentiment analysis to extract the emotional inclination of investors that influences their purchasing and selling decisions. This portion of the article describes the approaches employed and the sequence of steps necessary to implement the model.

A. Dataset Used

The algorithm is executed on two datasets acquired from the Twitter database and Yahoo Finance's TCS stock prices. The database gathered from Twitter contains textual investor sentiments and the opinions of individuals who wish to purchase or sell stocks. On the Twitter platform, these sentiments are expressed in the form of text, which is then transformed to numeric values in the form of 0s and 1s, where 0 denotes a negative opinion and 1 represents a favourable opinion. The process of data collecting entails gathering the image source and formatting the obtained data structure. The proposed implementation includes the collection of Yahoo Finance datasets to anticipate stock prices and Twitter as the social media platform for collecting user sentiments. Python is used for this purpose to interact with the Twitter API. Python application receives all inputs required for sentiment analysis, including the start and finish dates of stocks.

	Date	Open	High	Low	Close	Adj Close	Volume
0	01/02/2008	19.462856	19.512857	18.882856	19.107143	12.740701	252686000
1	04/02/2008	19.172857	19.414286	18.774286	18.807142	12.540661	224808500
2	05/02/2008	18.632856	19.142857	18.414286	18.480000	12.322518	285260500
3	06/02/2008	18.690001	18.845715	17.395714	17.428572	11.621424	393318100
4	07/02/2008	17.138571	17.825714	16.752857	17.320000	11.549030	520832900

Figure 3: Dataset Used

B. Data Pre-Processing

When applying machine learning algorithms on unprocessed data, it is necessary to convert the data to a more suitable format. These are the pre-processing processes necessary for tweet conversion:

- Tokens are created from Tweets
- Cashtags are eliminated since they no longer include information that is valuable to deep learning systems
- The URLs are discarded
- Stop words are removed

The code below demonstrates the elimination of irrelevant data from a database:

	Tweets	Adj Close	Volume
0	0.100000	12.740701	252686000
1	0.100000	12.540661	224808500
2	0.100000	12.322518	285260500
3	0.050000	11.621424	393318100
4	0.100000	11.549030	520832900
...	...	...	...
2765	0.050000	155.632523	26192100
2766	0.050000	154.019440	41587200
2767	0.100000	164.544296	61109800
2768	0.047619	165.729218	40739600
2769	0.050000	165.808884	32668100

Figure 4: Data Pre-Processing Stage

C. Workflow of the Proposed Methodology

The operationalization of the model begins with the collection of data from Twitter's repository for sentiment analysis and Yahoo Finance for predicting TCS stock prices. Once the dataset has been collected, it is subjected to data pre-processing, during which the Twitter sentiment texts are filtered using techniques such as tokenization, truncation, removal of stop words, and cleaning of imbalanced data. . Next, the resulting dataset undergoes a process of feature extraction in which only price prediction-relevant features are extracted. This filtered textual data are translated to quantitative numbers using the sentiment score values. Numeric sentiment scores are used to categorise textual opinions as positive or negative. This procedure is conducted utilising VADER as a toolset to provide the final sentiment analysis of text. The resulting dataset is currently combined into the TCS stocks data repository. This dataset's stock prices are labelled, and instructions are supplied to identify high and low stock prices. When data has been labelled, it undergoes the process of data partitioning into training and testing sets. At this point, prediction models based on machine learning are utilised for testing purposes. In this study, we have implemented SVM and KNN as classification algorithms that could predict the stock prices by combining the textual sentiments obtained from Twitter. In the final stage, results generated from the prediction models are evaluated and the final accuracy of the model is concluded.

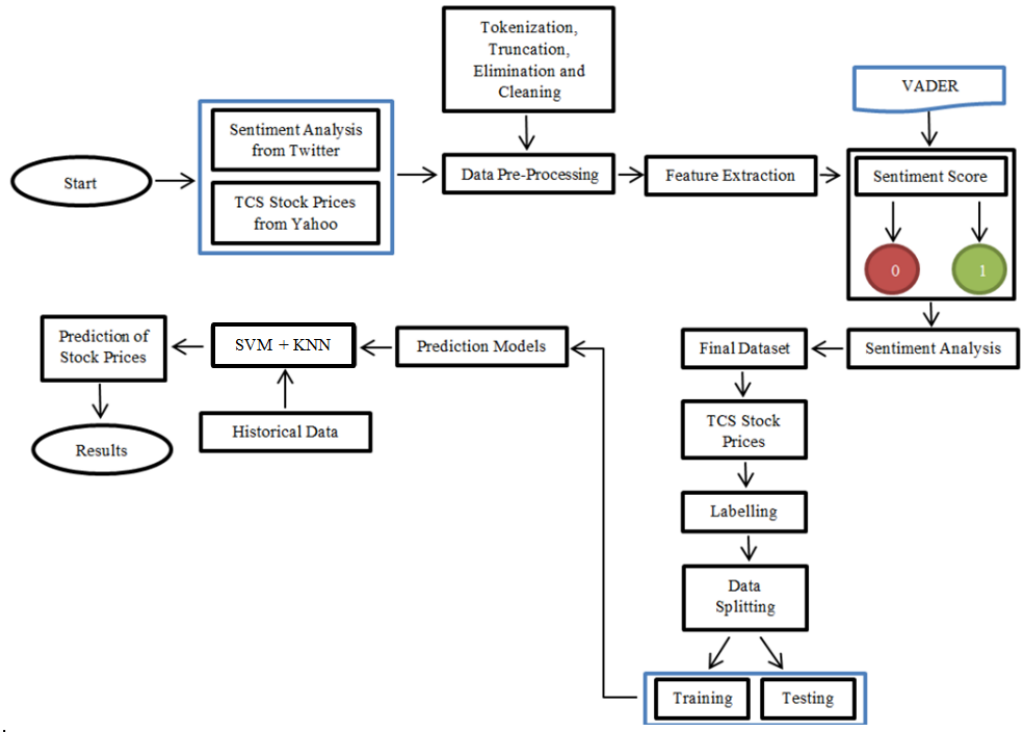
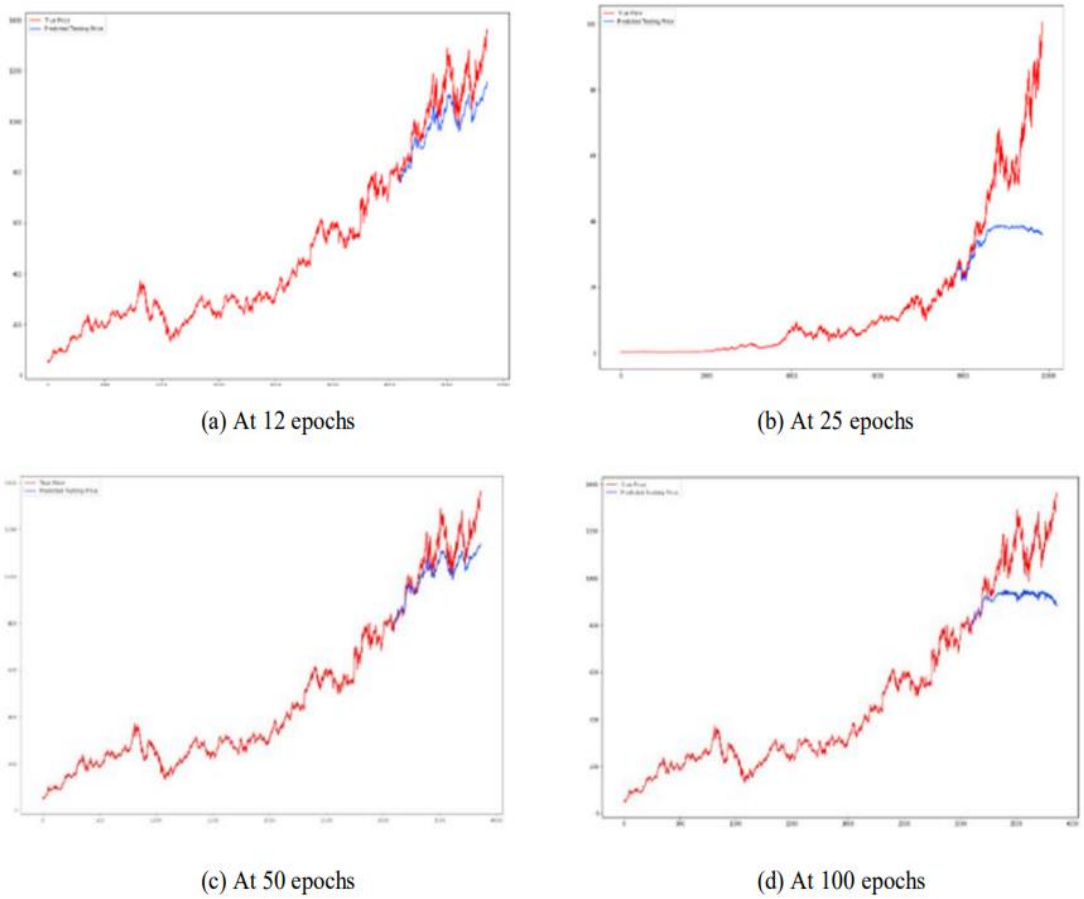


Figure 5: Architecture of Proposed Methodology

Results

After generating the results, it was noted that the dataset was initially less volatile and had lower values. The results of testing the SVM-KNN models are displayed in the figure below, with the red lines representing the actual market price and the blue lines representing the values predicted by our model.

DOI: https://doi.org/10.17762/turcomat.v11i1.13585
Research Article



(a) At 12 epochs (b) At 25 epochs

(c) At 50 epochs (d) At 100 epochs

Figure 6: Generation of Results using the SVM-KNN model

The table below displays the precision of our training and testing over all epochs.

Table 1: Processing time for TCS Stocks

Number of Epochs	Processing Time/sec
12 epochs	264
25 epochs	550
50 epochs	1100
100 epochs	2200

During the training and validation phase, the graphs below indicate a significant reduction in loss and a steady improvement in accuracy.

DOI: https://doi.org/10.17762/turcomat.v11i1.13585

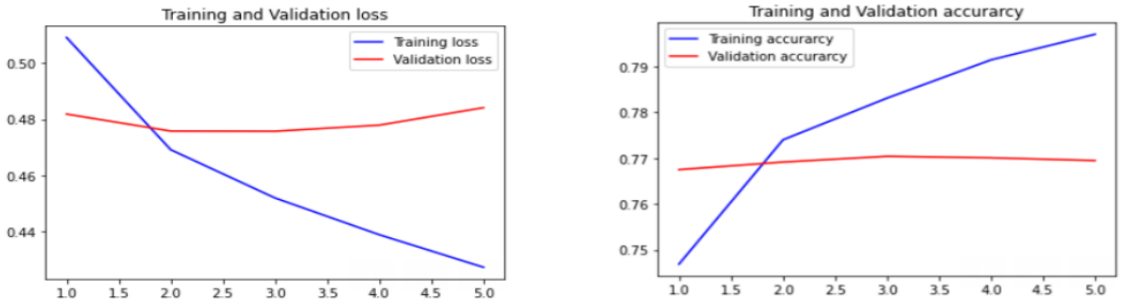


Figure 7: Training and Validation phase of the SVM-KNN model

Conclusions and Future Work

Due to the volatility nature of stocks and other factors, such as financial news on social media platforms, it has always been difficult to predict the price of a company. Thus, the working implementation of the proposed study is based on an SVM-KNN model that predicts TCS Stock prices using sentiments from the Twitter repository. This prediction method tends to collect the Twitter opinions of investors on TCS stocks and the corresponding Yahoo finance stock price. In order to do sentiment analysis, we employed the highest, lowest, and closing stock prices as inputs with the VADER framework. Together with historical stock data, the suggested architecture has been utilised to anticipate stock prices using an algorithm based on machine learning.

In addition to the model, the dataset is also limited and might be expanded to increase the model's overall accuracy. At later stages, the model might also be utilised for day trading, where investors are more likely to be concerned about short-term forecasts. Also, the emotions employed in this study are only categorised as good and negative, which can be expanded to include the recognition of dread, anxiety, and disgust. Thus, this could be the subject of future investigation.

References

[1] Alostad H, Davulcu H (2015) Directional prediction of stock prices using breaking news on Twitter. In: IEEE/WIC/ACM international conference on WI-IAT 1, pp 523–530

[2] Yuan B (2016) Sentiment analysis of Twitter data. M.S. thesis, Department of Computer Science, Rensselaer Polytechnic Institute, New York

[3] Lakshmi V, Harika K, Bavishya H, Harsha CS (2017) Sentiment analysis of twitter data. Int Res J Eng Technology 4(2):2224–2227

[4] Joshi R, Tekchandani R (2016) Comparative analysis of Twitter data using supervised classifiers. In: IEEE international conference ICICT, 3 pp 1–6

[5] Hegazy O, Soliman OS, Salam MA (2014) A machine learning model for stock market prediction. International Journal Computer Science Telecommunication 4(12):16–2

[6] Chen L, Qiao Z, Wang M, Wang C, Du R, Stanley HE (2018) Which artificial intelligence algorithm better predicts the Chinese stock market? IEEE Access 6:48625–48633

[7] A. B. Pawar, M. Jawale, D. Kyatanavar, Fundamentals of sentiment analysis: concepts and methodology, in: Sentiment analysis and ontology engineering, Springer, 2016, pp. 25–48

[8] Combining Review Text Content andReviewer-Item Rating Matrix toPredict Review Rating DJK Nilesh P Sable nternational Journal of Innovative Research in Science, Engineering and Technology

[9] Fischer T, Krauss C (2017) Deep learning with long short-term memory networks for financial market predictions

[10] A Survey on Mapping Bug Reports to Relevant Files: A Ranking Model, AFine Grained Benchmark, A Feature Evaluation PNS Anuja Hemant Shinde nternational Journal of Innovative Research in Computer and Communication Engineering

[11] Hileman, Garrick and Michel Rauchs. 2017. Global cryptocurrency benchmarking study. Technical report, Cambridge Centre for Alternative Finance.