International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 07 Issue: 03 | March - 2023 Impact Factor: 7.185 ISSN: 2582-3930

Building a Stock Price Prediction Modelusing Random Forest Regression and Sentimental Analysis

Yashmita A[1], Dr. Kavitha D[2],

PG Student[1],

Associate Professor-Finance[2],

PSG Institute of Management, Coimbatore

ABSTRACT

Stock market performance can be influenced by stock prices in either a positive ornegative way because it is a very unpredictable area. The public's opinions and feelings may be affected differently by various occurrences, which could impact the direction in which stock price fluctuates. To increase predictability of stock market indicators, this study examines the potential usage of sentiment data from Twitter users. In this study, relationship between public perceptions about acompany stated in tweets and changes in stock prices are investigated. The resultsof the study demonstrates strong association between changes in stock prices and the opinions shared by the general public in tweets. In this study, a system is createdthat collects tweets, analyses and assesses if tweets has positive or negativesentiment using a model called Random Forest Classifier. The output for the algorithm is then checked with the actual change in stock price the next day to ascertain the prediction's error. The goal of this research is to forecast future market behavior by performing sentiment analysis on a sample of tweets from the past fewdays. The results suggest that Random Forest model is suitable for stock price prediction along with sentimental analysis it generates the accuracy of 84.86% which is accessed by with three widely used measures for the regressor— Mean Squared Error , Root Mean Square Error , Mean Absolute Error and along with sentimental analysis by generating classification report.

Keywords: Stock price prediction, Volatility , Time series, Random Forest Classifier, Sentimental Analysis, Natural Language Tool Kit, VADER**,** Bootstrap method, Back testing, hyper parameter tuning, snscrape, Lemmatization and tokenization.

1. INTRODUCTION

The Stock market movement is important factor that plays a major role in country'sgrowth. Due to vast amount of data that stocks create each day, it is challenging for an individual to study all historical and current price of stock data to predict the movement of price of stock in future. It is influenced by economic, social, andpolitical factors. For opinion mining, there were no established procedures in the past. Back then, people would consult their friends and relatives before making a decision. Similar to this, organizations continued to use manual questionnaires to gather information before making judgments about output.

Surveys were used by businesses to collect feedback from a range of customers, which aided in decision making. Enormous rise in e-commerce, social media content over the past few years, however, has fundamentally altered how decisionsare formed. Users can now write product reviews online on a huge range of websites and applications. Social media platforms, blogs, and discussion forums are the primary venues for exchanging views and comments. Therefore, it has become essential to effectively analyze the data in order to support and aid investors in making wise trading decisions before to making actual investments. With the advent of social media's creation has produced a wealth of knowledge onpublic opinion. Social media is becoming into the perfect platform for expressingpublic opinion on any issue and has a significant influence on broad public opinion.The area of analyzing the sentiment of Twitter is taken into consideration. The emotion of the price of stock classifies tweets as positive, negative, or neutral usingmachine learning and the Random Forest regression model technique. A stock market's potential future trend is ascertained using this sentiment research on classification. There is a greater likelihood that price of stock rise further if sentiment is positive observed and vice versa if negative sentiment is observed

1.1 SOCIAL MEDIA PLATFORM

Over the past few years, there has been a tremendous rise in generation of user data. The growth of Internet-based applications has led to a sharp rise in social media users. This application is growing as a powerful instrument for communication and a source of vast amounts of data regarding public opinion. It is becoming a perfect platform for expressing public mood on any issue and has asignificant influence on public's opinion. Twitter is a platform that gives user the option to read, follow, and comment on the ideas of other users or to immediatelyshare their own opinions.

1.2 SENTIMENTAL ANALYSIS

Sentiment analysis, commonly referred to as opinion mining, gathers and examinesopinions about a range of goods and services using cutting-edge technology and algorithms. To improve the forecast, it is done on the basis of opinions of varioususers of Twitter and scores given for sentiment and pricing are also used. The goalis to identify patterns that are consistent with this link and make use of them to forecast the stock price trend in future. Tweets can give an accurate

representationof the general consensus, which can mirror market views. The tweets in this paper'ssentiment analysis of Twitter data are categorized based on emotions.

1.3 MODEL SET UP FOR MACHINE LEARNING

Modern technology, like as machine learning models, are created and developed in order to provide trading advice. This model must forecast the closing price for example, by using the information available today, the model may forecast the closing price of tomorrow. Using which one can act to buy the stock if the model indicates that its price will increase, and vice versa. To find trends and understandthe industry at the time, machine learning on historical data is used to automate thetrading process. This study illustrates the methodology for incorporating machinelearning into stock forecasting for the price of Microsoft Stock (MSFT). Python and open-source libraries can be used to develop it. The effectiveness of the modeland the precision of the predicted values must be evaluated using error analysis.

1.4 RANDOM FOREST REGRESSION MODEL

This is a model which forecasts, predicts the low and high prices for the stock movement. In light of the anticipated values, decisions will be made whether to sell or hold the stock. Study includes collection of data, processing the data, and developing the reading algorithm. The applied classifier and methodology is the random forest model. This sentiment research on classification is used to the likelypredict the trend of a stock market as it is most adaptable and user-friendly algorithms that provides decent prediction accuracy. Usually, this is done to for classification.

The stock price exhibits complexity and unpredictability due to the stock market'svariability, and uncertainty. There is issue with calculating the stock price will continue to be a concern if better algorithm is not there. Stock market forecasting is vital because the opinions of thousands of individuals frequently determine the direction of the stock market. All of these things has influence on business earnings, which in turn have an impact on investor mood. So there is a need to examine the Machine learning model and sentimental analysis to forecast stock behaviour and predict to reduce investment risk. This study will serve as a useful tool to aid novice traders in making wiser decisions. This model, is the most fittingmodel in case of larger dataset and in terms of accuracy. Since media serves as a platform to read customer reviews, make adjustments as needed, Twitter is used as the data source, retrieving the tweets via the Twitter API, training various models of algorithms.

1.5 BACKGROUND OF STUDY

80% of respondents in a survey felt Machine Learning (ML) method that uses is currently accessible stock data that can accurately predict many features of a specific stock including values of the index value, opening price, closing price, etc.It facilitates quicker decision-making for traders and investors. Some techniques, such SVM, can't be used since stock data isn't linear. So non – linear data using models like RF and LSTM neural networks are analyzed and suggested they are comparatively advantageous. Real stocks are chosen from the stock market, modelling analysis is

carried out to forecast stock prices, and the prediction outcomes of several models are then compared using the RMSE. The RF approachwas used to evaluate its effectiveness against logistic regression. Compared to logistic regression, random forest has a higher mean absolute score. The nonlineartechniques have been successful in picking stocks that can outperform the marketand predicting stock values. For long-term prediction, RF model attains accuracy in the region of 85-95%. When using the Bag-of-words technique, The Stock Market Prediction that combines ML techniques with Sentimental analysis (SA) experimented with using financial data, StockTwits, and Twitter. The sentiment ofthe microblogging data was examined using two SA tools, TextBlob and VADER.

1.6 RESEARCH GAP

It is necessary to take the public mood in consideration to determine stock price accurately and successfully. Due to market volatility, the accuracy of the previousstudy's prediction was unable to anticipate opening price of stock on the followingday. To predict stock prices, data from sentiment analysis of stock-related information will be combined with numerical values associated with past stock values. It may be possible to fix poor performance and existing shortcomings in some tests by improving machine learning techniques. Testing using larger data sets will aid in improving the accuracy of our predictions and to strengthen data validation. Additionally, the model is tested with various performance metrics. Byutilizing both pieces of information, a more effective stock recommendation system is created.

2. LITERATURE REVIEW

When forecasting stock prices, the stockbrokers predominantly rely on fundamental and technical analysis or time series analysis. Machine learning stockmarket predictions are made using the Python programming language. (Reddy, 2018) suggested that Machine Learning (ML) method using the stock market datais accessible that can make accurate prediction. In this case, the study forecasts thevalues of stock values for small and big firms' capitalization using a Support Vector Machine which is one of the Machine Learning model. They collected numerous international financial markets serves as the basis for the SVM algorithm's operations. Moreover, over fitting is not a problem with SVM. Analyzing all of these variables and influencing factors manually would prone to mistakes. Hence he suggests that machine learning is being used to predict many features of specific index including estimated values of index value, closing priceand opening price etc. This will facilitate quicker and better decision-making for traders and investors. (Deshmukh et al., 2019) Study comprehends the variety of information available in the financial market and pinpoint the factors that affect stock prices while taking into account the numerous industrial, macroeconomic, and market indicators. From an empirical point of view, estimating direction of price of stock changes is crucial to creation successful trading methods. According to the results of the survey given to qualified respondents, 80% of the respondentsfeel that the models and approaches they employ are adequate. Findings met the study's goal of accurately predicting the closing price as a result, applying ML to forecast stock behaviour and trends has proven to be a workable alternative to traditional methods, providing investors with exclusive information on market activity in addition to the

potential for usage in conjunction with other methods. (Soni et al., 2022) developed comprehending long-term markets or for projecting the open price of stock by combining use of statistics and machine learning algorithms. Many economists attempted to predict the value of stock in the early days. People later discovered creation of models (statistical) that is beneficial, which includes the time series model as it is simple and the forecasts better resultswhich is efficient. This is used over certain amount of period. However, some machine learning techniques, such support vector machines, can't be used since Stock data isn't linear. (Guo 2022) analyzed the non – linear data using models likeRF and LSTM neural networks and suggested they are comparatively advantageous. Real stocks are chosen from the stock market, modelling analysis iscarried out to forecast stock prices, and the prediction outcomes of several algorithms then compared using the error metrics.

Most of the Stock traders in conduct of making effective decisions they need information at their disposal fast as possible. So there is a concern to make researchefforts the prediction model which provides best accuracy and forecast rate with low error. By providing supporting information, such as the direction of stock prices moving forward, (Obthong et al., 2019) investigated the machine learning algorithms and tactics utilized to improve stock price prediction accuracy. Few algorithms and methodologies are used to predict price of stock depending on accuracy and prediction rate. Along with past data few other information's has influence on price of stock that include news about politics, economy, and the sentiment on social media. Therefore, a high-efficient prediction can be made by combining technical and fundamental analysis. (Shen et al., 2018) conducted a study to predict the stock index movement of using ML algorithms, new predictionalgorithm that examines and shows the relationship between stock markets and financial products to predict the movement of stock trend. In order to estimate daily Forex return (Tristan Fletcher, 2021) discovered that discriminative approaches perform well when complex financial information is included. Along with that thestudy demonstrates that machine learning methods may theoretically be utilized toproduce accurate currency predictions by forecasting the movement of EURUSD between 5 and 200 seconds in the future (up, down, or stay within the bid-ask spread) with an accuracy ranging from 90% to 53%, respectively. The effect of thestock market volatility on the individual stock prices is another difficulty in predicting stock prices. (Erhan Beyaz , 2019) analyzed different inputs—technical,fundamental, and combined—affected machine learning-based stock price predictions while also taking stock market conditions into account. Although ML techniques have been successfully utilized to anticipate stock prices, researchers tend to favour technical indications over fundamental ones because the latter are more difficult. (Lokesh et al., 2018) used technology's rapid advancements, whichhave led to higher processing speed, more storage capacity, and better algorithms, whose approach can inform potential investors about the risks involved and the predicted rise or fall in value of their investments. The study used the dataset of stock market for performing the model (training data), create analysis of sentimenton twitter data, and assessment of risk. The results show that, upon user input, therisk factor generated using Twitter data provides the risk percentage range as wellas the growth per stock in dollar value instantly and with an acceptable degree of accuracy. (Rouf et al., 2021) Developed the overall conceptual framework to explain methodology of algorithms based - systems for prediction of stock marketvolatility. Results from the 2011–2021 era were thoroughly examined after being obtained from online databases including Scopus and the ACM digital library. A thorough analysis was also performed to assess the trend of importance. (Strader et al., 2017) suggested that there is a clear connection between ML techniques andthe stock price prediction. While

assessing the total index (stock) is expected to rise and decline, support vector machines perform well. To determine which stocksto include in a portfolio or to identify higher quality system inputs, genetic algorithms apply an evolutionary problem-solving strategy by applying in findingsfrom Asian stock markets where the data from periods when markets are rising orfalling could be used to evaluate the systems' performance in various market circumstances. (Subhadra Kompella et.al 2019) made stock market predictions using the stock price and the headline as input using sentiment analysis to determine the polarity score and determine the kind of content that has an impact on the stock, either positively or negatively. The collected scores are utilised to calculate the stock prices. Then the exponential moving average method was performed to accuratelyaccess stock's impact. Finally, the random forest approach was used to evaluate itseffectiveness against logistic regression. Compared to logistic regression, randomforest has a higher mean absolute score. Logistic regression is inferior to RandomForest in terms of mean squared score. It is concluded that stock market predictionbased on sentiment research, the random forest algorithm is far more effective thanlogistic regression. Machine learning was applied by (Subba Rao et al., 2019) to forecast the stock market Sensex's behaviour tracking. They are successful to estimate price of stock and characterize the activity of trading securities using algorithms / techniques including random forest regressors, linear regression, decision trees, additional tree regressors, support vector regression. This study estimated the price of the stock and compared the results of each algorithm, an algorithm with high accuracy is ultimately deemed to be a best stock price prediction model. The study concludes that each of these algorithms is capable ofmaking an accurate forecast of the stock price, RF Regressor stand out as the top two. (Abdullah Bin Omar et al., 2022) forecasted the stock index accurately over three time periods: the entire Covid -19 period. According to precision of the model (test) the error metrics determine that the proposed prediction models like autoregressive random forest and autoregressive deep neural network are the betterprediction model for price estimate for the entire period of forecast, which was measured during Covid-19 period. The nonlinear techniques have been successfulin picking the firm's stock that outperforms the stock market and predicting stock values. In the context of the stock selection method (Zheng Tan et al., 2019) Implements the resilience of the random forest (RF) model to the Chinese stock market.

Investors used to look to technical analysts to anticipate stock prices because it was difficult to do so with a high degree of accuracy or certainty. Depending on the stock's starting, closing, and closing prices as well as its worth. (Daori et al., 2022) compared the findings and select the one model that performs the best. Thefindings of this study shows RF method has the good accuracy, it earned a 94.12%accuracy ratio because model was able to increase the number of repeats in training, delivering a higher precision.

(Antonopoulou et al., 2022) proposed the random forest algorithm to predict the four main Greek systemic banks prices. A group of financial metrics based on intraday data used are: Stock prices e of a certain Greek systemic bank. The findings from the study suggested that the factors employed here are essential for forecasting the closing prices of the stocks of systemic institutions. These offer a more accurate forecast of the bank series closing price the next day, where randomforest has also been employed successfully. In all four instances, the forecasted values from the RF-based forecasting models mostly follow the same patterns as the actual values and can successfully forecast Greek bank sector returns. (Abraham et al., 2022) Put forward that a prediction model could determine whether a specific stock had an uptrend. Random Forest classifier is used as to establish the connection between stock indexes and a

single stock's trend. Experiments revealed that forecasting model performs better than the dummy forecast, with precision rate 80%. Investment in clean energy companies is becoming more popular, and main factors driving are environmentally conscious consumers, energy security climate change, divesting from fossil fuels, and technological advancement. (Sadorsky et al., 2019) predicted the stock price direction of clean energy exchange traded funds using machine learning's randomforest technique. The stock price direction predictions made by RF are more accurate than logit models. Random forests techniques yield accuracy rates of between 85% and 90% for a 20-day forecast horizon, while logit models yield accuracy rates of between 55% and 60%. (Sadia et al., 2019) provided a more practical approach to accurately forecast stock movement. The data framework ofstock values the ncludes past year data and preprocessing of the raw data will be amajor emphasis. Therefore Random forest (RF) model was found as most effectivealgorithm that can be used for forecasting the stock price based on number of useful data points which can be reviled from past data once after evaluating the precision of the various models. Due to the algorithm's extensive training on historical data brokers and investors will find it efficient when trading in stock market. In comparison to previously used machine learning models, the experiment shows how the algorithm can estimate the price of a stock with greateraccuracy. Market makers, researchers, policymakers, and economists have all recently shown an increased interest in market forecasting. (Abirami et al., 2022),proposed enhanced supervised learning systems for stock price prediction. The conditions that must be met for random forest to function effectively include: Features must contain some real signal in order for models created using them to outperform guesswork; there should be little correlation between the predictions produced by each tree. Random forest method comes with numerous decision trees (independent) that cooperate and identified the best model for forecasting stock market estimation (Vijayarani et al., 2020). This study presents and evaluates a more useful approach to more accurately forecast stock development. After pre- processing the data, SVM model and random forest model are applied using dataset and the results are discussed. The random forest algorithm was judged to be the best plausible algorithm for forecasting the stock price using the historical data after accessing various models. Since the algorithm was chosen after being evaluated on a sample of data and was developed using a large collection of historical data, it will be a useful resource for financial experts looking to invest money in the stock market. (Ghahramani et al., 2022) Describe a novel approach for simulating candlestick patterns in stock market using Adaboost and RF model.Data preparation, the creation of new features, and data modification are first contributions to the preprocessing section. The second contribution is the introduction of a new approach to forecasting, known as dataset ensembling to forecast daily prices. The models are trained, adjusted, and evaluated on three-yeardaily Bitcoin prices, demonstrating the viability and correctness of the method. (Han, 2019) demonstrates the deep learning and machine learning by investigatingvarious strategies, such as Random Forests (RF) and Long Short-Term Memory. This study aims to identify potential methods for attempting to predict outcomes and using such methods to support human decision-making. These findings can becombined with human experience to make financial judgments by incorporating emotive qualities into the Random Forests model's training set. Their methodology examined the most recent linked articles to determine if popular media outlets form favourable or unfavourable options. In including sentimental as feature with a highweight in the RF model, the accuracy significantly improves. (Elagamy et al., 2018) Shows how the Random Forest algorithm and text mining can be used to create a novel technique for classifying relevant news and detecting key indications. The findings of this study proves that RF

may perform better than other classifier models and achieve high precision, resulting in a three to eight classincrease in classifying using critical indicators. Additionally, it has shown that Random Forest can perform better and obtain the highest level of accuracy when classifying news items using bigram characteristics. Support Vector Machine, LSTM, and Random Forest techniques utilized in predictions. Random Forest is agroup-supervised learning technique with a high accuracy factor for classificationissues. (Hota et al., 2021) compared their results and performance and has provided ANN and RF models for price prediction based on past data and compute the futureprice. The study concludes that although neural networks perform better across, they are only useful for audio, video, and visual datasets. In contrast, models like Random Forest can effectively handle large dataset with ease of computation and implementation, reducing the time and effort required to work with sophisticated Neural Network structures. Prediction is difficult due to the global stock markets' inherent volatility. To ensure low investment risk, market risk, which is closely connected with forecasting errors, needs to be reduced. (Khaidem et al., 2016) proposed a unique method in this study for reducing the risk associated with investing in the stock market by predicting the returns of a stock. Numerous algorithms, including SVM, ANN, etc., have been investigated for resilience in stock market prediction. The study used random forest classifier which yielded better outcomes and shown to be extremely reliable in forecasting the movement of stock price direction the by calculating several metrics, including accuracy, precision and specificity, robustness. For long-term prediction, RF model attains accuracy in the region of 85-95% for all the datasets evaluated, including AAPL and Samsung.

(Darapaneni et al., 2020) Attempted to forecast the movement of price of stock with use of past and the sentiment data. In the study few models are compared, thefirst model, the LSTM, used historical prices as the independent variable. For the second part's Random Forest Model, Analysis of sentiment is captured with Analyzer (Intensity) serves as the primary indicator. Some macro parameters, including the prices of gold, oil, the US dollar rate of exchange, and Government Securities, used to access the model to increase the precision score. The main goalof this paper was to build trading techniques that could aid in the practical use of the models created. In order to determine whether the values predicted by the models were appropriate, examination of RMSE values was performed. This resulted the Mean Absolute Percentage Error (MAPE) varies between 1.36% and 1.81% when using LSTM, while it varies between 1.25% and 1.76% when using Random Forest Model for Sentiment Analysis. Most of the case the accepted theory for prediction of price of stock is Efficient Market Hypothesis. (Joshi et al.,2020) Performed research on sentiment categorization to evaluate the trend of stock price through categorical data such as news articles (Financial) by automating the sentiment recognition process, which allows to determine the overall news polarity based on the words used in the news as they conveysentiment about the current market. If the news is excellent, there is a higher predictability that price of stock will rise because of positive impact on the market.

Additionally, if media's response is bad, price of stock may go downward as aresult. When creating the train set and initially labelling the news, this paperemployed a polarity detection method; dictionary-based method. For the purposeof removing stop terms, lexicon was used that contains stop words related tofinance. When their findings were compared, Random Forest performed admirablyacross performs with an accuracy range of 88% - 92%. SVM

accuracy is around86%. The performance of the naive Bayes algorithm is about 83%. (Bharathi et al.,2017) described an approach that integrates sentiments of common people throughnews feeds and sensex data to forecast stock market behaviour. In this study thedata is gathered from news feeds (RSS) which is combined with stock investmentdata. As a result of which rates are estimated using this model that is trainedexperimentally investigating in the Exchange's stock market (Amman Stock)prices and news feeds are collected from the company. The analysis revealed anincrease in accuracy prediction of 14.43%. (Gondaliya, et al., 2021) Explored theclassification precision score of algorithm/ methods utilized in natural languageprocessing for analysis and stock market forecasting in India based on data during Covid-19 outbreak. Analysis of sentiment was conducted in order to determinebest algorithm. When utilizing Bag-of-words technique, and few algorithmsproduced results that are good than other models prediction. In this study (Ching-Ru Ku et al., 2021) Forecasted stock prices using a variety of parameters. Thefundamental analysis depends on PTT forum conversations and news articles whilethe technical analysis is based on stock past transaction data. The LSTM which isgood at analysing time series data, is applied to forecast the stock price with stockpast transaction information and sentiments of text using processing tool BERT. (Mittal et al., 2022) Investigated the relationship between "sentiment of public andmarket" using algorithms and analysis of sentiment using Twitter data, andforecasted stock market movements based on the forecast and DJIA values fromthe past. Utilizing Self Organizing Fuzzy Neural Networks (SOFNN) and DJIAvalues the study acquired 75.56% accuracy. (Singh, et al., 2022) conducted anemotional analysis to anticipate the prices of Apple stock and the DJIA usingtweets that that has terms "stock market," "StockTwits," and "AAPL." by utilizingSVM sentiment of tweets is identified for every trade day. Twitter APIs are usedto gain the fundamental sentiment analysis. The price (closing) between one day and day after is the same, and the average mood ratings in the data that is trained depends on a day's worth of tweets. As a result, this model can find how much thestock market will rise or fall the following day based on the emotional impact of today's tweets. (Koukaras et al., 2022) study focused on Stock Market Prediction that combines ML techniques with Sentimental analysis (SA) experimented with using financial data, StockTwits, and Twitter. The sentiment of the microbloggingdata was examined using two SA tools, VADER and TextBlob. After preprocessing the data, seven machine learning (ML) methods, including SVM, KNN, and RF etc used to evaluate performance of algorithm in four scenarios. TheAUC and F-score were used - two metrics to assess the model. However, RF, withan F-score of 69.6%, seemed to generate the most accurate predictions since it correctly predicted that the stock prices. (Kolasani et al., 2020) Demonstrated the working of Twitter (tweets) that plays a role in forecasting price of stock using theSentiment, they used 140 trained model dataset to train the algorithms. SVM algorithm was found as better model (0.83 accuracy) in the analysis of sentiment to forecast sentiment of tweet for everyday the market was open. For purpose of predicting the closing values between the AAPL and DJIA prices, two models likeMultilayer Perceptron Neural and Boosted Regression Trees are evaluated. The study demonstrates that when it comes to stock price prediction, Regression outperform conventional models by a wide margin. (Kumar et al., 2020) Developed a system that forecast stock price movement using sentiment analysis on tweets gathered from the Twitter API in combination with stock closing values, obtaining financial news using web scraping and developing a web interface to retrieve the predicted information using real-time data and built the system to trainand update the model using stock data APIs or online scraping stock data.(Heidenet.al., 2021) used the VADER sentiment analysis model to process the news gathered from New York

Times and feed the sentiments as a feature into an LSTM-based stock price prediction model along with the historical data of stocks, this research intends to investigate the impact of financial news inside the stock price prediction problem. Experiments shows algorithm perform good when sentiment are better are taken into account. The findings indicate that the model has the potential to accurately forecast stock values within a 50-day time window, which is very relevant for any type of trading strategy. (Wang et al., 2018) examined the connection between stock price and news sentiments to accurately predict the stockprice. It is suggested to use a revolutionary strategy for predicting the stock price that takes into account impact of sentiment. Using a stock price data set and an improvement in performance in terms of lowering the Mean Square Error is enhanced learning-based method's effectiveness is shown to be superior to existinglearning-based approaches. (Manogna R L, 2022) comprehends that how news impacts changes in emerging market stock indices. This study uses NIFTY50 index and the Economic Times newspaper's headlines. Based on the news headlines, the model anticipates when NIFTY50 index rise or fall by using 3 models; Multinomial NB Classifier, the Random Forest Classifier. Highest precision, 0.73, was attained by the Random Forest Classifier. The investors may receive more decision support from this model. The study concludes that traders utilize sentiment-based trade as a criteria towards their strategy which probably rise chance of profit. In addition to advancements and applications of AI in the financial sector, (Soni et al., 2022) provided an overview of several sentiment andstock price prediction models like LSTM, RF etc. A significant contribution to theadvancement of science and technology is artificial intelligence. Using sentimentsML algorithms and ANN are possible that can increase accuracy of stock market forecast and get more precise results. Individual behaviour can be influenced by the general mood. (Andy, 2022) examined the connection between general sentiment and stock market trends. The stock price of SPY is then estimated using regression and polynomial regression depending on the sentiment of public as expressed in tweets. The outcomes demonstrate the potential of analysis of sentiment using LSTM for predicting the stock price only when more tweets are gathered and examined to confirm link in-between sentiments using LSTM and stock price prediction. The model with the least amount of error is the most effective and frequently advised strategy for prediction. (Prajapati et al., 2022) employed three different models and performed sentiment analysis on news that mentioned the company or the stock in order to conduct this study. The study concludes the stock price prediction system that successfully integrated analysis of sentiment for analysis (fundamental) and long short-term memory for analysis (technical) provided good precision. Inter-day forecasting would be achievable with future iterations of the system, and sentimental analysis would be improved to create a fundamental analysis impact factor. (Bhardwaja et al., 2015) identifiedthe challenge for traders in forecasting stock market, evaluating the Sensex and Nifty forecast numbers. To make it possible, the study shows analysis of sentimentfor the stock market using source Nifty and Sensex live server at different intervals that may be used to forecast the status of the stock by choosing Python programming. It assists traders in knowing which stocks to purchase and will alsocontribute to the preservation of the share market's economic equilibrium. (Oliveira et al., 2016) Explained that the volatility, returns, and trading volume ofv indexes and portfolios are three stock market characteristics that can be used to forecast data from microblogging sites. The strategy makes use of measures of emotion and interest taken from microblogs. The study concluded that the emotionon Twitter and the volume of tweets had an influenced the ability inn forecasting the returns of the portfolios (S&P 500 index), with lower market capitalization, and specific industries. These findings demonstrate the value of microblogging data for

financial expert systems, allowing for the prediction of stock market activity and offering an advantageous replacement for current survey measures. (Gupta et al., 2018) Utilized the NYTimes API to extract news from the New YorkTimes, yahoo finance to get historical prices, the VADER extracts the sentiments and regression is used to forecast stock movement. The findings conclude that using historical prices to analyse social data will produce outcomes with an 80% accuracy rate. In order to forecast stock market movement, the study takes both the significant factor, or news, and historical pricing into consideration. (Lee et al.,2019) focused on the value of text analysis in predicting stock prices by introducing a system that predicts how stock prices of firms will fluctuate in reaction to financial events stated in 8-K filings (UP, DOWN, HOLD).The findings show that adding language improves prediction accuracy by over 10% (relatively) compared to a robust baseline that includes a number of financially- related characteristics. The next day following the financial event is when this influence is most significant, although it can last for up to five days. The study developed a dictionary which is utilized to examine the significance of text analytics for price of stock and by using this dataset, the study demonstrates that including textual data is crucial, particularly in the short term. Based solely on pastprice analysis, it might not be feasible to predict the stock market with any degree of accuracy. (Kumar et al., 2020) conducted sentiment analysis based on the opinions of various Twitter users in order to improve the prediction by adding sentiment scores and pricing. The study reveals a novel sentiment analysis technique using the Twitter API and the Tensor Flow platform to obtain tweets anddetermined the sentiment scores for the tweets by comparing the dates. The modelis trained under consideration and made stock price predictions. The study concludes that with a larger Twitter dataset, random forest method performs well.(Charash et al., 2013) Observed whether media coverage of the general sentiment of investors can forecast changes in stock prices. Firstly the study gathered information on the phrases that express emotions of traders in newspaper and thenclassified these words to reflect where they fall on different emotions/sentiments which predicts the mood indices for every trading day. The time series is used in this study analyses whether these mood indices could predict the opening price thefollowing day. According to the research, activated happy mood predicted rising NASDAQ prices whereas activated unhappy mood indicated falling NASDAQ prices. The findings reveals that in predicting the stock prices, both the valence and activation levels of the mood are significant. (Patrick et al., 2014) Suggested that, for decision-making and risk management on various time scales, sentiment analysis of financial market news is useful. Positive social media posts and news about a firm would undoubtedly entice investors to buy its stock, which would raise the company's stock. Machine learning can be used to create a prediction model for discovering and analyzing the relationship between tweet content and stock values, and then forecasting future prices. (Mohan et al., 2019) discovered that Sentiment analysis has a high association around the stock trend and release of news since it has been shown that predicting stock prices just based on historicaldata or textual information is insufficient by making number of sentiment analysisstudies including other algorithms. In this study deep learning models are utilizedto collect a data (time series) for S&P500 stocks as well as more than two lakh financial news stories on businesses and analyse it by relating it to relevant articles to increase the precision of stock price. The findings states that financial news andprice of stocks has a correlation.

3. RESEARCH METHODOLOGY

In this study Python libraries are employed for development and testing. To do this, softwares like Anaconda Python, Jupyter Notebook, and libraries like OpenCV, Tensorflow, and Keras are installed. Firstly to create a prediction model using the past data set, the stock prices are downloaded from a source and stored as csv file. It is followed by developing a back-testing engine, this work predicts stock values using Python, Pandas, and Scikit-Learn. Using information from today, this model forecasts the closing price of tomorrow. If the algorithm predictsan increase in price then stock is purchased and vice versa if alogorithm predicts afall in price.

For this study Microsoft (MSFT) stock data is examined whose past data is utilizedto predict the price of the MSFT as follows:

- Download MSFT stock values from Yahoo Finance
- Import Libraries
- Explore the data and build up the dataset to forecast prices based on pastprices.
- Evaluate a machine learning model
- Install a back-testing engine
- Improve the model's precision

Install the required libraries

The data file is imported/ loaded as a pandas data frame using the Python library pandas, which is used to analyse. For purpose of visualization package called Matplotlib is installed. An open-source Python data analysis framework called Scikit-learn provides ML algorithm, preprocessing, evaluation of model and training tools. Additionally, serves sub-library for metrics, RandomizedSearchCV,StandardScaler, Random Forest, and Random forest Classifier, seaborn, MinMax Scaler, Confusion matrix, Classification report, Precision score and train test split.Other packages like word cloud, snscrape, nltk.sentiment VADER (sentimental Intensity Analyzer) are installed for sentimental analysis. A Python open-source package called Yfinance is used to access financial data.

Exploring the Microsoft Stock dataset

For exploring the dataset Python scripts is utilized. Installing the yfinance Pythonpackage in a Jupyter notebook, the data is downloaded from Yahoo Finance for a single MSFT stock from time it began trading until the present. The date the stockwas exchanged is represented by the dataframe's row index. Some dates are missing since stock doesn't trade every day as it doesn't trade on weekends or holidays. The MSFT data is downloaded and loaded in frame and then transformedto comma separate value file.

3.1 VARIABLES USED FOR PREDICTION MODEL

Table 3.1 MSFT – Dataframe

Index Features	Meaning
Date	The stock value date
Open	Price of stock at beginning day of trading
High	The stock price at its highest level.
Low	the stock price was at its lowest level
Close	closing price of stock at end of day
Sentiment Polarity	Describes the type of sentiment (Positive, Negative, Neutral)

Date	Open	High	Low	Close	Volume
2022-12-30 00:00:00-05:00	238.210007	239.960007	236.660004	239.820007	21930800
2023-01-03 00:00:00-05:00	243.080002	245.750000	237.399994	239.580002	25740000
2023-01-04 00:00:00-05:00	232.279999	232.869995	225.960007	229.100006	50623400
2023-01-05 00:00:00-05:00	227.199997	227.550003	221.759995	222.309998	39585600
2023-01-06 00:00:00-05:00	223.000000	225.759995	219.350006	224.929993	43597700
2023-01-09 00:00:00-05:00	226.449997	231.240005	226.410004	227.119995	27369800
2023-01-10 00:00:00-05:00	227.759995	231.309998	227.330002	228.850006	27033900
2023-01-11 00:00:00-05:00	231.289993	235.949997	231.110001	235.770004	28669300
2023-01-12 00:00:00-05:00	235.259995	239.899994	233.559998	238.509995	27269500
2023-01-13 00:00:00-05:00	237.000000	239.369995	234.919998	239.229996	21317700

Fig.3.1 MSFT – Dataframe

Visualize the price of Microsoft stock

To visualize the movement of stock price change overtime, is observed by plottinggraph which provides a different perspective on the data's structure using matplotlib package. For forecasting the stock price the dataset from 2019 to Jan 2023 are considered for the study whose volatility is presented in candlestick chartpresented in (Fig.3.2)

Fig. 3.2 Stock Price Volatility

Preparing the data for the Microsoft Stock Price

To evaluate data for a particular time series and identify trends in that data, pandas rolling function is used. Following that, the target is set up by performingthe following:

- This will start by looking at ('2019-01-01', '2019-01-02'), then ('2019-01-02','2019-01-03'), and so on across the DataFrame using the pandas rolling technique.
- To determine whether the second row is greater than the first, they are compared. As seen below, the Target column now shows whether the priceincreased or decreased on the specified day.
- If Target is 1, the value increased. The price dropped if the Target value was

0. The target is predicted using the columns "High," and "Low", "Open,"Volume," Close".

Table 3.2. Target Variables

Date	Actual_Close	Target	Close	Volume	Open	High	Low
2022-12-30 00:00:00-05:00	239.220825	0.0	240.407837	19770700.0	235.061228	241.315567	235.061228
2023-01-03 00:00:00-05:00	238.981430	0.0	239.220825	21930800.0	237.614847	239.360475	236.068717
2023-01-04 00:00:00-05:00	228.527618	0.0	238.981430	25740000.0	242.472686	245.136013	236.806869
2023-01-05 00:00:00-05:00	221.754562	0.0	228.527618	50623400.0	231.699666	232.288188	225.395464
2023-01-06 00:00:00-05:00	224.368011	1.0	221.754562	39585600.0	226.632344	226.981476	221.205933
2023-01-09 00:00:00-05:00	226.552551	1.0	224.368011	43597700.0	222.442841	225.195940	218.801966
2023-01-10 00:00:00-05:00	228.278229	1.0	226.552551	27369800.0	225.884227	230.662268	225.844334
2023-01-11 00:00:00-05:00	235.180939	1.0	228.278229	27033900.0	227.190941	230.732074	226.762022
2023-01-12 00:00:00-05:00	237.914093	1.0	235.180939	28669300.0	230.712121	235.360482	230.532578
2023-01-13 00:00:00-05:00	238.632294	1.0	237.914093	27269500.0	234.672213	239.300620	232.976463

Feature Selection and Data Standardization

Features refer to the columns in the dataset. The model performance has impacted by machine learning application called the feature selection, these selected characteristics has an influence and help to determine how well the forecast turns out. Ineffective features because the test set to perform poorly overall. Sklearn has modules for feature selector and feature importance that canbe used. The importance of the feature is examined and it assigns a score to each features, and the density (curve) in (Fig. 3.3) shows how much of an impact the variable has on the change in price. Feature that has highest score are the most important, dependable variables are always present. Feature selection can improve data visualisation, reduce overfitting, increase accuracy, and reduce training durations. The dataset is standardised using Sklearn's standard scaler function. To expedite training and improve the model's numerical stability, standardisation has been carried out.

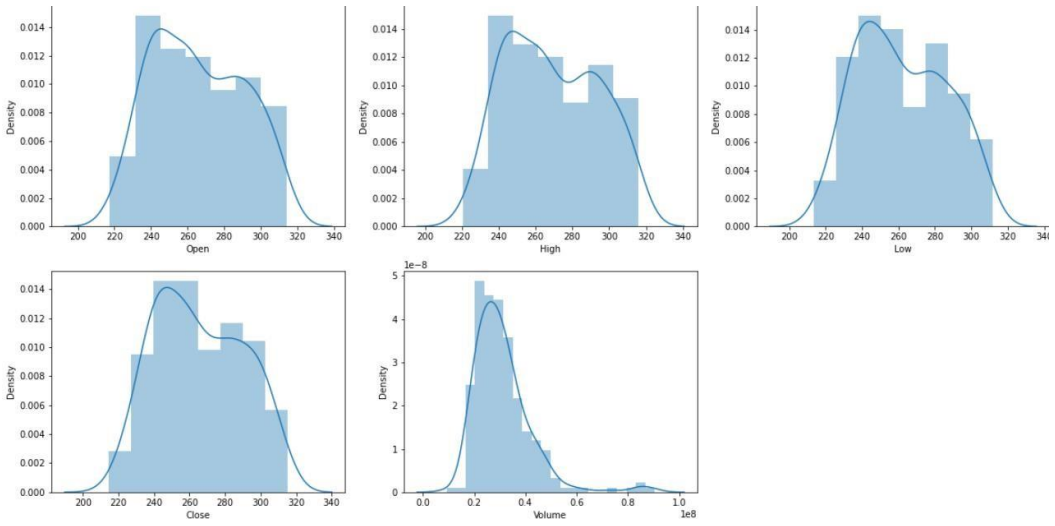

Fig. 3.3. Feature selection
The dataset is sorted for training and testing

It is important to transform the testing data after fitting the scaler on the training data. In doing so, any data leakage throughout the model testing process would be prevented. The fit method will train the model using predictors to predict the Target. The train data set is portion used to develop, build models for prediction.By using the input choices and the raw stock price data to generate the training set's (80%) features, a script for building training datasets is created. They are used for training the model. It is essential that never utilise future data to forecastthe past because it entails working with time series data. Testing dataset (20%) isa portion of the data to predict a future performance of model. It serves as a helpful yardstick for evaluating the model (fig. 3.4). The trained model is evaluated by running it against the expected dataset using the testing set.

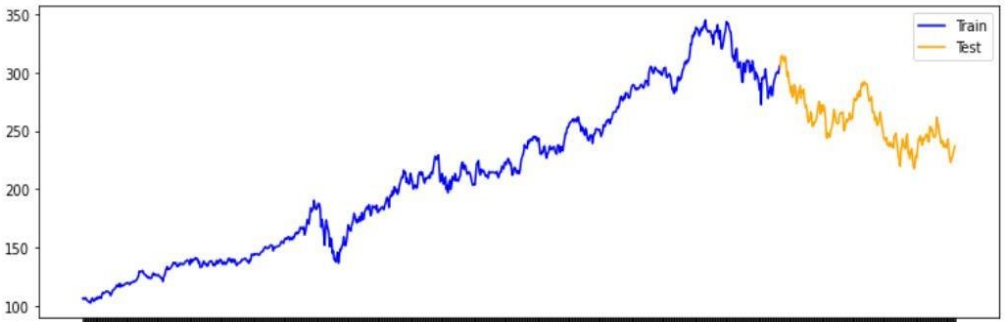

Fig. 3.4 Training and Testing dataset

4. ANALYSIS AND INTERPRETATION

4.1 DEVELOPMENT OF PREDICTION MODEL

Following stage is that machine learning model is developed which determines how precisely the price of stock can be forecasted. The model called random forest which is a classification algorithm which addresses the issue of over fittingin decision trees, is just a collection of several decision trees that are made by picking features that are random in dataset in a random manner. Depending on maximum votes obtained from the decision trees, this model determines a conclusion or forecast.

The predictions are created using the random forest classifier, because it can detect nonlinear correlations in the data. Back testing across the whole price history is required to obtain an accurate error metric. The stock market forecast is made using this method is considered one of approach which is flexible and easy to use, and it provides good prediction accuracy. Typically, this is applied to categorization tasks. They use the same hyper parameters as a decision tree, to predict stock market.

The tool has a tree-like model which is used for decision making (fig. 4.1). It makes a choice depending on potential outcomes, which can include event outcomes, and utility. This algorithm is an example of an algorithm that builds

many decision trees by selecting various observations and features at random, then averaging the results of the various decision trees. On the basis of the inquiries on a label or an attribute, the data is divided into parts. Acquired data set from the public online database of the stock markets from the previous year, using training and testing dataset where supervised learning model's is discovering correlations in data from the training set and reproducing them in thetest set.

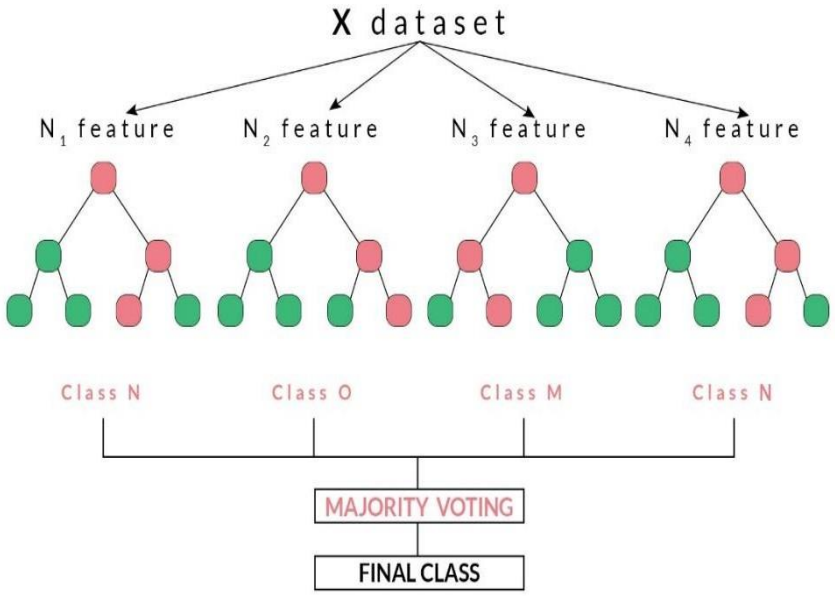

Fig. 4.1 Random Forest Model

Scikit-Random learn's Forest Classifier is installed as the goal is binary (0/1) thisclassification approach where 1 indicates an increase in price, 0 a decrease.
Decision tress (distinct) that the algorithm constructs, expressed as n random, canbe used to initialize this. A decision tree algorithm known as random forest resistant to over fitting than a tree represented in (fig. 4.1). The model is resilient there are more trees. Tress are resilient to over fitting as value is lower.

The framework of the research is represented in the (fig. 4.2) which depicts theworking of the model. This project's goals include gathering, processing, and creating the trading algorithm for forecasting data.

This model predicts the price change using the Random Forest Classifier algorithm which has advantages like avoiding over fitting, can be used for classification and regression, handles missing values, and reduces the time as itcan build large number of decision trees into sub groups, classify accordingly.

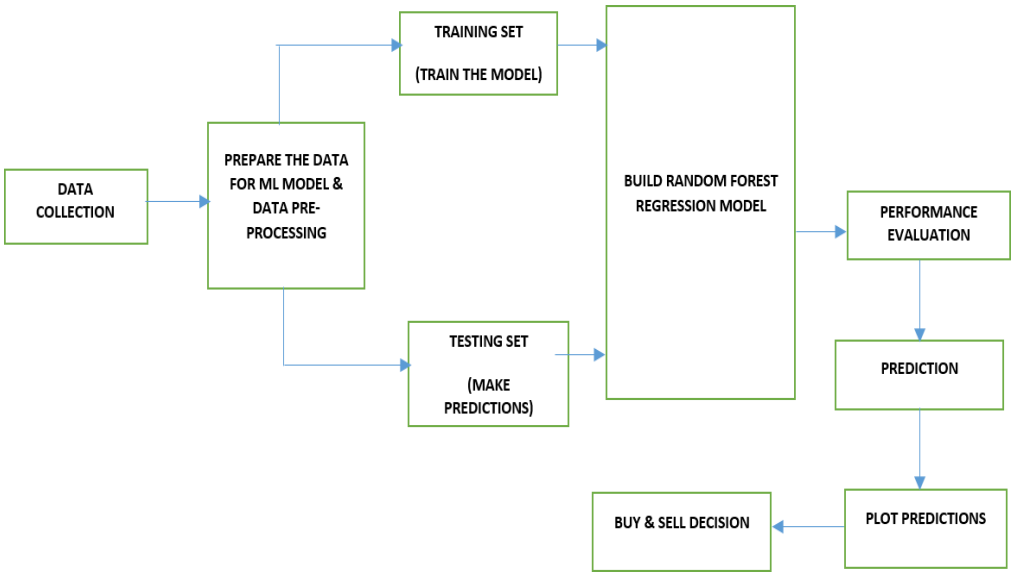

Fig. 4.2 Research Methodology

The ideal parameters taken into account for a random forest are shown (table.4.1)

Table. 4.1 Parameters specifications used to build random model

Parameters	**Criteria**
n_estimators	This parameter defines the tress required while calculating the model's averahe of prediction and highest votes. In this case, n estimators = 500 is chosen because more treesperform better, meaning 500 separate decision trees will be built in model.
max_depth	This parameter indicated by this; in this case, max depth = 10 signifies that all hundreds oftrees are divided at 10 separate nodes.
min_samples_split	When splitting a sample, least number data points are included where its default numberis 2.

min_samples_leaf	Least number of samples included in leaf node is specified. This parameter's default value of 1 is that at least one sample indicated by leaf classifies.
Bootstrap	It is a type of statistical resampling that uses replacement in addition to random samplingof a dataset.
random_state	This parameter generates random set of numbers for the random forest.

Backtesting

Accuracy of model can be increased by using the scikit-learn precision score function is employed. Precision will reveal percentage of days when the price increased as predicted by the algorithm. In order to achieve high precision and reduce risk, this is done. Precision is only 0.4769, that is just 47% of the time didprices increase as anticipated by the model. 47% of the time, the algorithm directionally precise. Back testing can still be used to enhance the model. It will produce a more reliable error estimate. Using data from earlier than the day of prediction, backtesting verifies accuracy. The model will be trained across each all the rows in dataset using the backtesting approach, which loops across the dataset. The predict method is then used to make predictions in order to increase precision. In the backtesting function, separate the training and test data, train a model, then predict proba function to evaluate the prediction that calculates the inaccuracies by combining forecasts with the actual Target.

Assessing the model after back testing

Adding more predictors to the predictions. The projections are roughly 0.537percent more accurate than they were before. To determine how many deals value counts would have generated. Using the technique, there was an almost 54% likelihood that the price increased (fig. 4.3). The Accuracy of the modelbefore and after backtesting process are shown in (table.4.2).

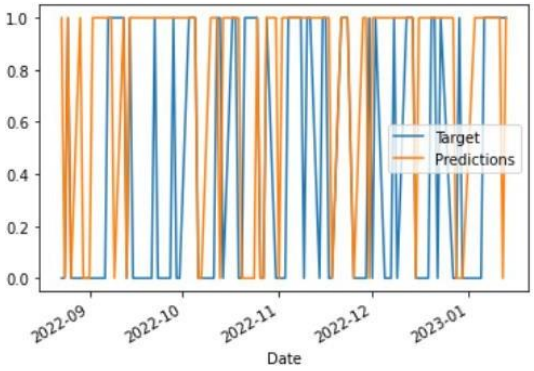

Fig. 4.3 back testing process

Table. 4.2 Accuracy of model

Process	**Accuracy**
Before Back testing Process	47%
After Back testing Process	54%

4.2 SENTIMENTAL ANALYSIS USING MSFT – TWITTER DATASET

The sentiment analysis has recently experienced a boom in popularity. The intentis categorized as either neutral, negative, or positive. Sentiment analysis refers toprocess in examining people's ideas or viewpoints, extracting data from social media, scoring it, and then analyzing it (fig. 4.4)

A lexicon and rule-based sentiment analysis tool called VADER is suited to emotions expressed on twitter. Networking function scraper platforms is calledsnscrape which is used for by scraping things like hashtags, user profiles.

Although Twitter's Developer Account and Application Programming Interface (API) keys are required for obtaining authorized on Twitter in order to scrape tweets, snscrape performs this function without the API keys. When processing text-based records from scraping and sentiment analysis, Natural Language Processing, a vast area of AIis applied. It involves web scraping, which involvesgrabbing a small amount of data from a website.

In this study, source of data is from twitter, and it was determined that text dataneeded for sentiment analysis on Twitter needed to be eliminated. The primary procedures are listed below:

- *Scraping Tweets* - Two methods for scraping Tweets -Tweepy is a python package that offers access to numerous twitter APIs, and Selenium is the browser mimicking tool typically used for testing websites.
- *Identifying Sentiments* - By using sentiment type as the goal variable when training a model, it is possible to distinguish between negative andpositive sentiments on tweets. Sentiment IntensityAnalyzer (SIA) by NLTK can be used for this because it provides superior categorisation.
- *Text Pre-processing* - Since the text extracted from tweets is not sufficiently clean to be utilised for model training, it must first be pre- processed by removing stop words (words used solely for the purpose ofcorrect sentence formulation), links, punctuation, numbers, and special characters. They lack meaningful, comprehensive information, So In order to make text record cleaner, it must be deleted.
- *Lemmatization and tokenization* - In NLP, technique called lemmatization converts sort of word to its base root mode. It is responsible for clubbing words which has similar meaning with root forms. Tokenization is a method used in NLP to separate phrases intoeasier language elements. The first steps in the NLP process are data collection (a sentence) and data reduction into understandable parts.
- *Feature Extraction* - Textual representation must be transformed into numerical features through feature extraction using Bag of Word model(Simple vectorization). Sentiment analysis is carried out using the features that are extracted, and the results are then compared.

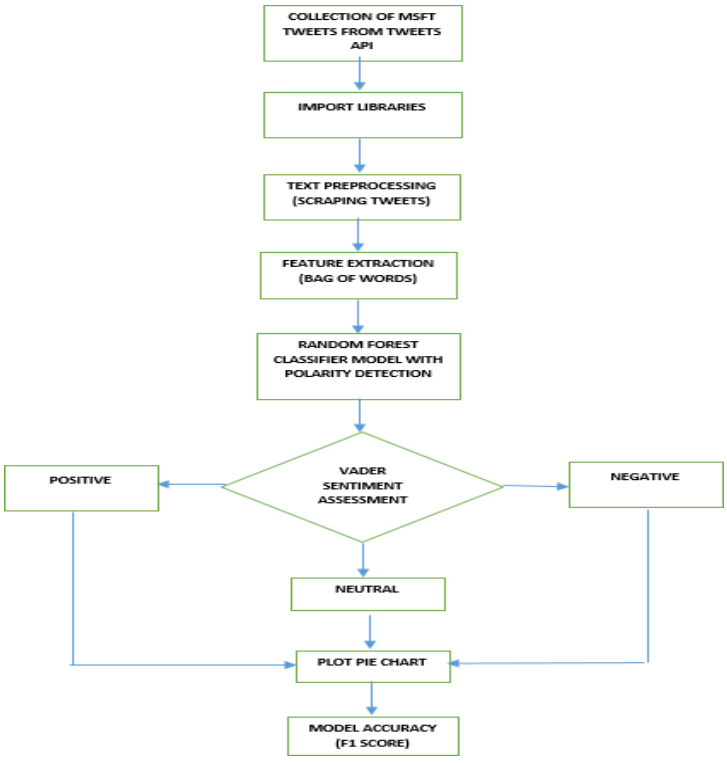

Fig. 4.4 Sentimental Analysis Model

Random classifier Data frame is created which sorts the research variables along with polarity into different sentiment/emotions, by the condition

- The sentiment is positive if polarity is more than score 0.05
- The sentiment is negative if polarity is less than score 0.05
- the sentiment is neutral If polarity is equal to score 0.05

Input is first obtained in the form of a query (fig 4.5), (user input - Number oftweets, Number of days) and analyze is given. It builds the dataframe shown below by appending the tweets list.

	Datetime	Tweet Id	Text	Username
0	23:57:38+00:00	1607888466125393920	The criminal illegal naked shorting is now in ...	SkipperDoodle3
1	23:23:59+00:00	1607880001277288449	if i pick out 4 books tl decided from - will y...	msft_biscuit
2	23:23:12+00:00	1607879802224009218	Long $AMZN at $850 billion cap\nShort 2x $MSFT...	EyeoftheIger
3	23:18:24+00:00	1607878594549751808	Patent DM/078593: HANDSET - TALBOT	MSFTPatentBot
4	23:12:28+00:00	1607877101125865478	How the largest stocks performed today\n\nAppl...	StockMKTNewz
5	23:10:42+00:00	1607876655665582085	People need to begin questioning why Bill Gate...	MM__Stocks
6	23:10:09+00:00	1607876519841665024	$TSLA \nPre-covid high: $64\nToday: $109\n\n$A...	sharon__swing
7	22:38:52+00:00	1607868645392121856	$TSLA \nPre-covid high: $64\nToday: $109\n\n$A...	dr3yec
8	22:18:25+00:00	1607863497550958592	Patent MY135562A: ISSUING A PUBLISHER USE LICE...	MSFTPatentBot
9	22:08:10+00:00	1607860918796562432	$TSLA \nPre-covid high: $64\nToday: $109\n\n$A...	Mr_Derivatives
10	22:05:07+00:00	1607860153642090499	The leaders are all falling hard and fast. $TS...	StayingITMoney
11	22:01:52+00:00	1607859334179856384	$MSFT a break below November lows still needs ...	ElliottForecast
12			snooping is getting boring. need something els...	msft_biscuit
13			MSFT Is down YTD 29% and Perf Year 29%\nAPPLE ...	WayoutFinancial
14			Kowalski Analysis Market on Close\n\n$SPY:-0,...	Kowalski_trader
15			So a common question in DMs keeps coming up. U...	Matt_Stoughton
16			$MSFT Below Avg Volume\nDaily appearances sinc...	dailycandlestix
17			Looks like institutional investors have been t...	PubCoInsight

Fig. 4.5 Twitter dataset

The stock trend is calculated using the stock price and the previous day's stock price, with a value of '0' signifying a price decrease and a value of '1' signifying aprice increase which is represented in (table 4.3).

Table. 4.3 Sentiment Analyzer data frame

	Date	Open	High	Low	Close	Adj Close	Volume	Sentiment	Price Diff	Trend
1003	2023-01-09	226.449997	231.240005	226.410004	227.119995	227.119995	27369800	Positive	2.190002	1
1004	2023-01-10	227.759995	231.309998	227.330002	228.850006	228.850006	27033900	Positive	1.730011	1
1005	2023-01-11	231.289993	235.949997	231.110001	235.770004	235.770004	28669300	Positive	6.919998	1
1006	2023-01-12	235.259995	239.899994	233.559998	238.509995	238.509995	27269500	Positive	2.739991	1
1007	2023-01-13	237.000000	239.369995	234.919998	239.229996	239.229996	21317700	Positive	0.720001	1

The relationship between validation loss and training loss is plotted to show whether the model is over fitting in (fig. 4.6). To display the model's training loss shows how well training data matches and validation loss displays how the model fits new data. Validation dataset has sufficient data to retrieve the model'sgeneralizability because there is a small gap between them, indicating that they behave like datasets from the same distributions.

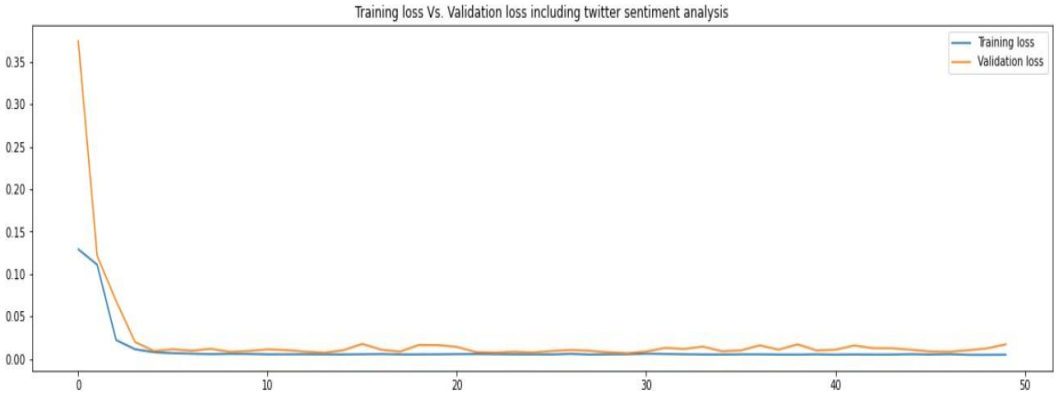

Fig. 4.6 Training and validation loss

Bootstrapping process

To scale both features and target sets, the package MinMaxScaler from sklearn is utilised. Training Testing Data is fitted using the scaler in the Random Forestregressor. Each RF tree is trained using the bootstrap approach, which means that only a portion of the observations are used for training. The chosen portionis known as the bag, and the other samples are called out of bag samples. The results from various trees are pooled once they have all been trained on variousbags. Bootstrap Forest provides many decision trees and, their averages produce a final forecast value. To address over fitting, the primary issue in machine learning, hyper parameter adjustment is used. The RF model's hyper parameters are e utilized which speed up the algorithm or boost its predictive power.

4.3 SENTIMENT ASSESSMENT

After scraping through Twitter for a specific ticker, each message is evaluated and scores are allocated accordingly. Vader can better discern the emotion by cleaning the scraped tweets and run a sentiment analysis on each tweet. A tweet is deemed to be negative if its negative score is higher than its positive score. If atweet receives more positive than negative feedback, it is deemed favorable. If positive equals negative, a tweet is neutral.

Classification Report

	precision	recall	f1-score	support
0	0.56	0.42	0.48	159
1	0.50	0.64	0.56	144
accuracy			0.52	303
macro avg	0.53	0.53	0.52	303
weighted avg	0.53	0.52	0.52	303

Fig.4.7 Classification Report

A classification report, which details how accurately the model predicts positives and negatives, is a metric used to determine if a machine learning program is useful or not. As can be seen from the figures above, the precision for recognizing negatives was 0.56 and for identifying positives it was 0.50. The weighted mean - recall, precision, or F1 score, indicates that ideal outcome lies between 0 and 1. The report's use of a large number of data samples serves as its foundation. There are 303 total data points, of which 159 have negative outcomesand 144 have positive results, as indicated in (fig. 4.7).Using the Natural Language Toolkit's VADER a tool for analysis of sentiment that is rule-based and especially tuned to the sentiments media Utilize a sentiment intensity analyzer to evaluate the emotion. A word cloud representing popular terms and public mood is displayed in (fig. 4.8).

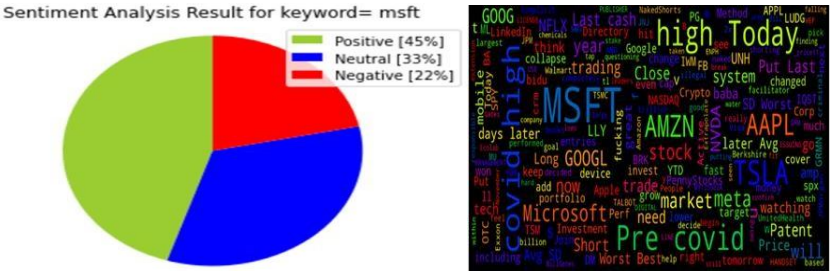

Fig. 4.8 Pie chart and word cloud (sentimental analysis)

5. FINDING OF THE STUDY

Random forest reduces over fitting and choosing set of features and subgroups togenerate smaller trees. Using the Model, the graph which is shown in (fig. 5), depicts the actual and anticipated movement of the MSFT stock during the past 175 days.

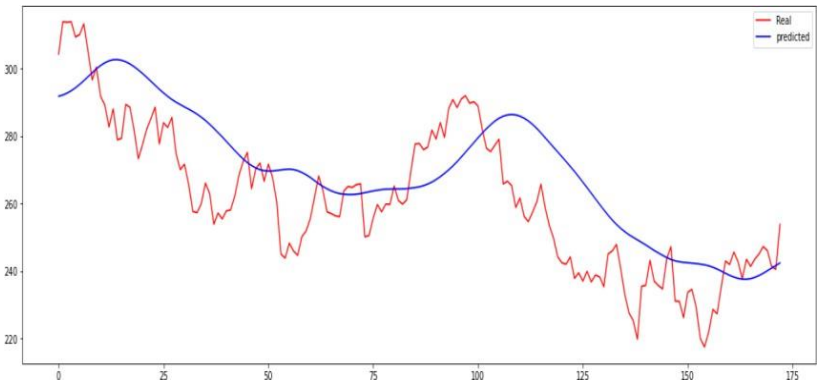

Fig. 5.1 Comparison of Real and Predicted stock movement

5.1 PERFORMANCE MEASURE

- ***Mean Absolute Error (MAE)*** – The actual and projected values differences arecalculated using MEA. MAE number falls, the forecast's accuracy rises. A regression model is evaluated using metrics like MAE, which demonstrate howclosely predictions correspond to the actual data and how big of a variance there is. The lower the MAE, the better a model fits a dataset.

$$\text{MAE} = \underbrace{\frac{1}{n}}_{\text{test set}} \sum_{i=1}^{n} \underbrace{|y_i}_{\text{predicted vaue}} - \underbrace{\hat{y}_i|}_{\text{actual value}}$$

- ***Mean Squared Error (MSE)*** – MSE is function that provides regression linethat resembles data points at closes that is how well the line fits the data

points. In any event, MAE value should be as near to zero as possible.

$$MSE = \frac{1}{n}\Sigma\underbrace{\left(y - \hat{y}\right)}_{\text{The square of the difference between actual and predicted}}^{2}$$

- ***Root mean square error (RMSE)*** **–** This metric quantify deviation in dataset tofit in the regression line. It tells how they are clustered within the line of greatest fit. Model's accuracy increases if RMSE value is less. The RMSE is measuring how the residuals are spread over or how far from line of regression.Better indicator of how effectively the algorithm anticipates the outcome is RMSE.

$$RMSE = \sqrt{\frac{\sum_{i=1}^{N}(Predicted_i - Actual_i)^2}{N}}$$

- ***R squared*** - a measurement of a model's ability to fit a specific data that displays how closely the line of regression and the plot expected and actual value agree the r-squared value are around 0.6 and 1.0, appropriately matches and the model works. The maximum value is 1.0, therefore the better the modelfits the data. The R-Squared statistic in a regression model measures how variance in a dependent value is exhibited by one or more predictor variable.

Table. 5.1 Performance measure before sentimental Analysis

Performance Metric	**Values**
Mean Absolute Error	0.285
Mean Squared Error	0.23669
Root Mean Squared Error	0.4865
(R^2) Score	1.0
Random Tree Regressor	Train set : 80% and Test Set : 20%
Accuracy	52 %.

Table. 5.2 Performance measure after sentimental analysis

Performance Metric	Values
Mean Absolute Error	0.285
Mean Squared Error	0.38
Root Mean Squared Error	0.62
(R^2) Score	0.8944
Random Tree Regressor	Train set : 80% and Test Set : 20%
Accuracy	84.86 %.

Plot of the future days using the predicted values with sentimental analysis Constructing a data frame with the expected values for the upcoming years (fig.5.2). In addition to that the model also helps in decision making process to the sell, buy, and hold prices, (table.5.3) traders attempt maximize their profits by buying at the lowest price, selling at the highest price, and holding price if neither is occurring.

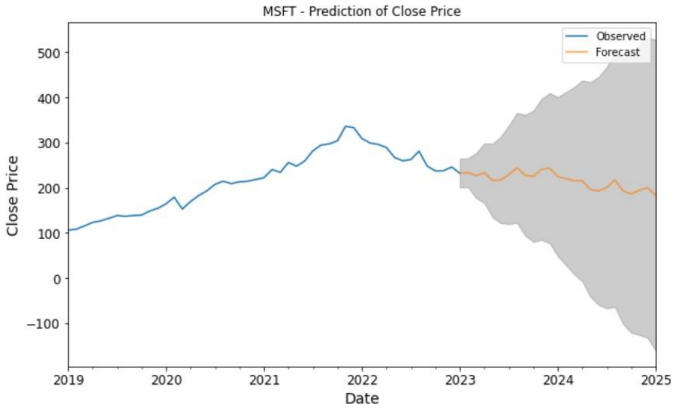

Fig. 5.2 Stock price direction (Forecast)

Table. 5.3 Buy/sell decision

Date	Buy / Sell	Predicted Price
2023-08-15	Buy	235.7
2023-04-12	Sell	477.9

The accuracy of the predicted and actual price of stock is plotted by integratingthe Random Forest Regression and the sentimental analysis using the MSFT twitter dataset. When used with the Twitter dataset, the more precise Random

Forest model offers more accurate time series models for prediction and incorporates real-time public mood monitoring, which can further increase the model's accuracy. The whole data frame, which has been scaled to have 5 columns and 12323 data points for training as shown (fig. 5.3).

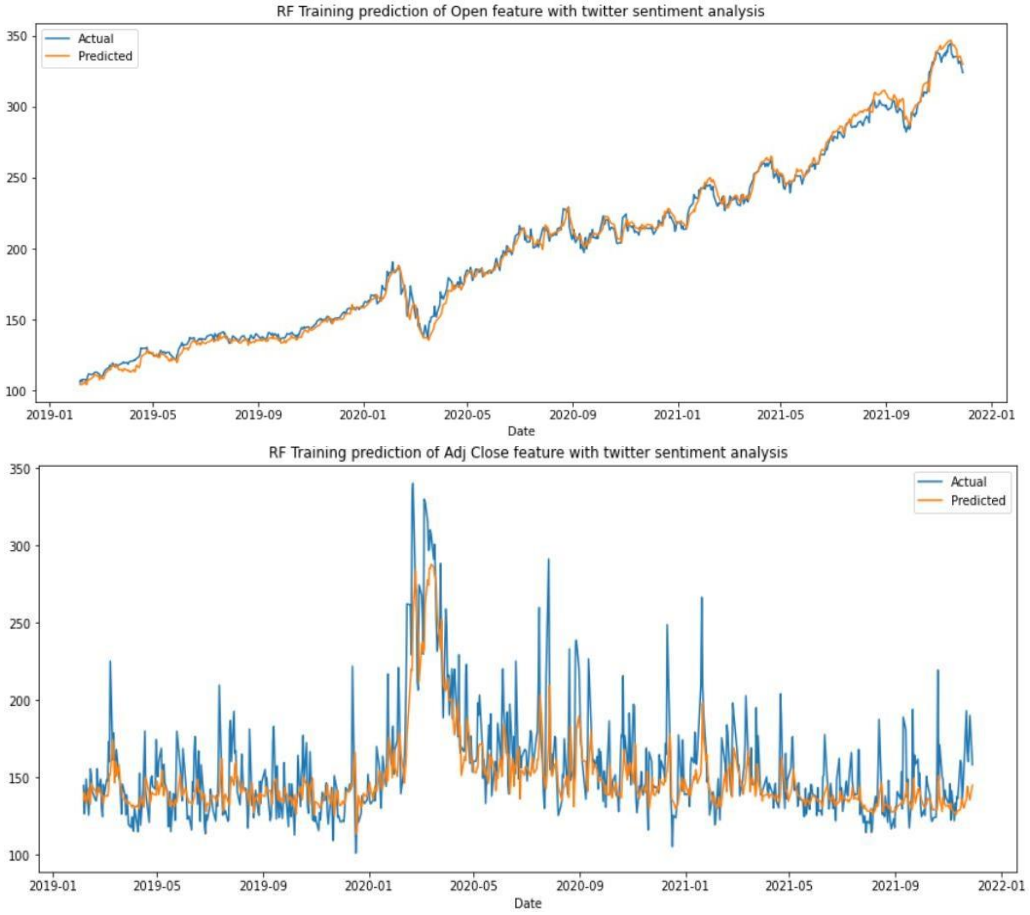

Fig. 5.3 Combining RF and Twitter sentiment analysis for prediction

For accessing the accuracy they are trained and tested for last year 2022 and the real and predicted price are plotted using the Random Forest Regressor. It showsthat the graph is in accordance with the sentimental analysis which is plotted below (fig.5.4) where there was decline in the price during the start of 2023.

Table 5.4 Real vs. predicted price of MSFT stock price

Date	*Real Price*	*Predicted Price*
03-01-22	335.35	318.9835
04-01-22	334.83	318.9835
05-01-22	325.86	318.9835
06-01-22	313.15	318.9835
07-01-22	314.15	318.9835
...	...	...
09-01-23	226.45	232.4847
10-01-23	227.76	232.4847
11-01-23	231.29	232.4847
12-01-23	235.26	232.4847
13-01-23	237	232.4847

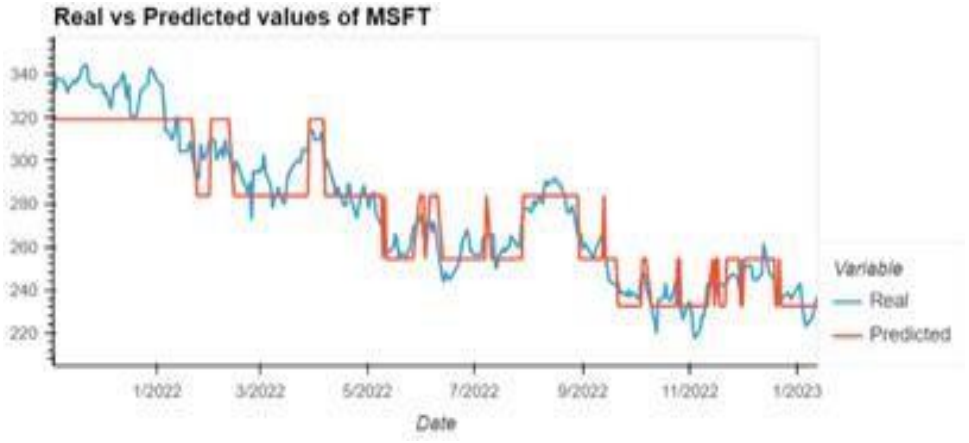

Fig. 5.4 Real vs. predicted price of MSFT stock price

(Fig. 14) depicts the MSFT stock price prediction (Lower and upper close price)from the year 2023 to 2025. This prediction model was built only after the back tesing, bootstrapping process which improved the accuracy of the model from 53% (from previous studies) to 84.89%. (Fig.5.5) indicates the price prediction for upcoming days which was obtained using fit model which reads the dataset, sets the target and regression step and predicts the results. From the figure it is inferred that there will be decline in the MSFT stock price in future days.

	lower Close	upper Close
2023-01-01	200.390844	264.270207
2023-02-01	201.043153	264.922516
2023-03-01	176.667995	276.698384
2023-04-01	166.027869	298.580840
2023-05-01	134.489225	297.337096
2023-06-01	121.323429	312.644346
2023-07-01	119.529402	337.703108
2023-08-01	122.096775	365.650039
2023-09-01	93.889767	361.477717
2023-10-01	79.882983	370.279057
2023-11-01	84.482463	396.569897
2023-12-01	76.878222	409.641468
2024-01-01	48.094021	400.610012
2024-02-01	28.524659	411.719410
2024-03-01	8.189061	422.511079
2024-04-01	-7.749791	437.535859
2024-05-01	-42.125187	433.590519
2024-06-01	-59.910834	445.491640
2024-07-01	-66.999184	467.240120
2024-08-01	-64.246813	497.937682
2024-09-01	-101.973408	487.262970
2024-10-01	-121.548687	493.868519
2024-11-01	-126.083249	514.679553
2024-12-01	-132.928197	532.387772
2025-01-01	-161.355151	527.767175

Fig. 5.5 Price Prediction

6. CONCLUSION

The paper demonstrates significant correlation between firm's price of stock fluctuations, the public sentiment towards that company as represented through tweets on Twitter. The primary objective is to create of a sentiment analyzer that determines type of sentiment in a tweet and the prediction prices in Twitter API.Three categories—positive, negative, and neutral—are used which categorizes tweets. The price of stock would reflect the favourable attitude expressed by the general public on Twitter, which is strongly backed by the outcomes of the prediction made using the RF model. According to tests that assessed the accuracy of the various algorithms, the RF model is the most efficient approach for determining the value of a stock backed by many data points from the historical data. According to this study's analysis oftweets, the stock price of Microsoft, the most well-known IT company, is specifically correlated with or even predicted by public opinion. The findings indicate that shifts in public opinion can have an impact on the market, which suggests that there is a good probability to predict the stock exchange.

This project involves downloading the stock dataset, cleaning and analysing the data, developing a machine learning model, creating a back testing tool, and improving forecast accuracy. The RF model is trained for predicting percentage -a stock price change. Three widely used error metrics is accessed and generates accuracy of 84.86%.

FUTURE SCOPE

As this study is only concerned with determining the price of the MSFT stock, the project's scope can be expanded to include price predictions for other industries or stocks. Further, the graph's more detailed characteristics, such as indications, will give the user knowledge time to buy, sell the stocks using indicators. An indicator for supports and resistance will examine a graph and provide accurate supports and resistance. Another technique to analyse movement of stock price is moving averages. Thus, it may be possible to add numerous indicators at once for a better trading experience. This might help to calculate how much money can be made potentially make trading using this method. The number of predictors should be increased, intraday trading shouldbe included, hourly patterns from the previous day should be incorporated, activity post-close and pre-open should be considered, and trading on other exchanges that open before the NYSE should also be considered to understand the global sentiment. This model can still be enhanced by adding new data sources, including those from sites besides Twitter, so that the sentiment scorescan better aid the system in price prediction.

REFERENCES

[1] Reddy, V. K. S. (2018). Stock market prediction using machine learning. International Research Journal of Engineering and Technology (IRJET), 5(10), 1033-1035.

[2] Deshmukh, Y., Saratkar, D., Hiratkar, H., Dhopte, S., Patankar, S., Jambhulkar, T., & Tiwari, Y. Stock Market Prediction Using Machine Learning. International Journal of Advanced Research in Computer and Communication Engineering Vol. 8, Issue 1, January 2019

[3] Payal Soni, Yogya Tewari, Deepa Krishnan, "Machine Learning Approachesin Stock Price Prediction: A Systematic Review", Journal of Physics: ConferenceSeries, (2022).

[4] Guo, Y. (2022). Stock Price Prediction Using Machine Learning. Electronics 2022, 11, 3414. https://doi.org/10.3390/electronics11203414

[5] Obthong, M., Tantisantiwong, N., Jeamwatthanachai, W., & Wills, G. (2020). A survey on machine learning for stock price prediction: Algorithms and techniques. Conference: 2nd International Conference on Finance, Economics,
Management and IT Business

[6] Shen, S., Jiang, H., & Zhang, T. (2012). Stock market forecasting using machine learning algorithms. Department of Electrical Engineering, Stanford University, Stanford, CA, 1-5.

[7] Ahmad, I., Basheri, M., Iqbal, M.J., Raheem, A., 2018. Performance comparison of support vector machine, random forest, and extreme learning machine for intrusion detection. IEEE Access 6, 33789–33795.

[8] Beyaz, E. (2019). Effective Stock Price Forecasting Using Machine Learning Techniques Whilst Accounting for the State of the Market. The University of Manchester (United Kingdom). Department of Computer Science, Student thesis:
Phd

[9] Lokesh, S., Mitta, S., Sethia, S., Kalli, S. R., & Sudhir, M. (2018). Risk Analysis and Prediction of the Stock Market using Machine Learning and NLP. International Journal of Applied Engineering Research, 13(22), 16036-16041.

[10] Rouf, N., Malik, M. B., Arif, T., Sharma, S., Singh, S., Aich, S., & Kim, H.
C. (2021). Stock market prediction using machine learning techniques: a decade

survey on methodologies, recent developments, and future directions. Electronics,10(21), 2717.

[11] Strader, T. J., Rozycki, J. J., Root, T. H., & Huang, Y. H. J. (2020). Machinelearning stock market prediction studies: review and research directions. Journal of International Technology and Information Management, 28(4), 63-83.

[12] Kompella, S., & Chakravarthy Chilukuri, K. C. C. (2020). Stock market prediction using machine learning methods. International Journal of Computer Engineering and Technology, 10(3), 2019.

[13] Polamuri, S. R., Srinivas, K., & Mohan, A. K. (2019). Stock market prices prediction using random forest and extra tree regression. Int. J. Recent Technol. Eng, 8(1), 1224-1228.

[14] Omar, A. B., Huang, S., Salameh, A. A., Khurram, H., & Fareed, M. (2022).Stock Market Forecasting Using the Random Forest and Deep Neural Network Models Before and During the COVID-19 Period. Frontiers in Environmental Science, 907.

[15] Mohan, S., Mullapudi, S., Sammeta, S., Vijayvergia, P., & Anastasiu, D. C. (2019, April). Stock price prediction using news sentiment analysis. In 2019 IEEEfifth international conference on big data computing service and applications (BigDataService) (pp. 205-208). IEEE.

[16] DAORI H, ALHARTHI M, ALANAZI A, et al. Predicting Stock Prices Using the Random Forest Classifier. Research Square; 2022. DOI: 10.21203/rs.3.rs-2266733/v1.

[17] Antonopoulou, H.; Theodorakopoulos, L.; Halkiopoulos, C.; Mamalougkou,
V. On the Predictability of Greek Systemic Bank Stocks using Machine Learning Techniques.,Preprints2022,2022070462.,https://doi.org/10.20944/preprints2022 07.0462.v1

[18] Abraham, R., Samad, M. E., Bakhach, A. M., El-Chaarani, H., Sardouk, A., Nemar, S. E., & Jaber, D. (2022). Forecasting a stock trend using genetic algorithm and random forest. Journal of Risk and Financial Management, 15(5), 188.

[19] Sadorsky, P. (2021). A random forests approach to predicting clean energy stock prices. Journal of risk and financial management, 14(2), 48.

[20] K. Hiba Sadia, Aditya Sharma, Adarrsh Paul, SarmisthaPadhi, Saurav Sanyal

(2019). Stock Market Prediction Using Machine Learning Algorithms. International Journal of Engineering and Advanced Technology (IJEAT), 08.

[21] Abirami, R. K.Varalakshmi, Maddika. J., Reddy, Kota., R, Chittipi Reddy Akash, (2022) Stock Market Price Prediction Using Random Forest and Support Vector Machine", International Journal of Creative and Research Thoughts(IJCRT), Vol. 10.

[22] Vijayarani, S., Suganya, E., & Jeevitha, T. (2020). Predicting Stock Market Using Machine Learning Algorithms. International Research Journal of Modernization in Engineering Technology and Science.

[23] Ghahramani, M., & Aiolli, F. Price direction prediction in financial markets,using Random Forest and Adaboost. ESANN 2022 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning, 5-7 October 2022, i6doc.com publ., ISBN 978287587084-1.

[24] Nabipour, M.; Nayyeri, P.; Jabani, H.; Mosavi, A.; Salwana, E.; S., S. Deep Learning for Stock Market Prediction. Entropy 2020, 22, 840. https://doi.org/10.3390/e22080840.

[25] Elagamy, M. N., Stanier, C., & Sharp, B. (2018, April). Stock market randomforest-text mining system mining critical indicators of stock market movements. In 2018 2nd international conference on natural language and speech processing (ICNLSP) (pp. 1-8). IEEE.

[26] Hota, L., & Dash, P. (2020). Comparative Analysis of Stock Price Predictionby ANN and RF Model. Computational Intelligence and Machine Learning, 2, 1-9.

[27] Khaidem, L., Saha, S., & Dey, S. R. (2016). Predicting the direction of stockmarket prices using random forest. arXiv preprint arXiv:1605.00003.

[28] Darapaneni, N., Paduri, A. R., Sharma, H., Manjrekar, M., Hindlekar, N., Bhagat, P., & Agarwal, Y. (2022). Stock price prediction using sentiment analysis and deep learning for Indian markets. arXiv preprint arXiv:2204.05783.

[29] Kalyani, J., Bharathi, P., & Jyothi, P. (2016). Stock trend prediction using news sentiment analysis. arXiv preprint arXiv:1607.01958.

[30] Bharathi, S., & Geetha, A. (2017). Sentiment analysis for effective stock market prediction. International Journal of Intelligent Engineering and Systems, 10(3), 146-154.

[31] Gondaliya, C., Patel, A., & Shah, T. (2021). Sentiment analysis and prediction of Indian stock market amid Covid-19 pandemic. In IOP Conference Series: Materials Science and Engineering (Vol. 1020, No. 1, p. 012023). IOP Publishing.

[32] Ko, C. R., & Chang, H. T. (2021). LSTM-based sentiment analysis for stock price forecast. PeerJ Computer Science, 7, e408.

[33] Ganthade, A., Kulkarni, E., Agrawal, A., Garodia, S., & Lahane, P. Stock Market Prediction Based On Twitter Sentiment Analysis. International Journal of Advances in Engineering and Management (IJAEM) Volume 4, Issue 5 May 2022, pp: 1087-1092

[34] Singh, S., & Kaur, A. (2022). Twitter sentiment analysis for stock prediction. Available at SSRN 4157658.

[35] Koukaras, P., Nousi, C., & Tjortjis, C. (2022, May). Stock Market Prediction Using Microblogging Sentiment Analysis and Machine Learning. In Telecom (Vol. 3, No. 2, pp. 358-378). MDPI.

[36] Kolasani, S. V., & Assaf, R. (2020). Predicting stock movement using sentiment analysis of Twitter feed with neural networks. Journal of Data Analysis and Information Processing, 8(4), 309-319.

[37] KUMAR, D. S., UG, P. K. R., & Arun, A. (2020). Real-Time Stock Market Prediction Based On Social Sentiment Analysis Using Machine Learning. Ilkogretim Online, 19(3), 4329-4334.

[38] Heiden, A., & Parpinelli, R. (2021). Applying LSTM for Stock Price Prediction with Sentiment Analysis. In 15th Brazilian Congress of Computational Intelligence (pp. 1-8).

[39] Wang, Z., Ho, S. B., & Lin, Z. (2018, November). Stock market prediction analysis by incorporating social and news opinion and sentiment. In 2018 IEEE International Conference on Data Mining Workshops (ICDMW) (pp. 1375-1380). IEEE.

[40] Manogna R L. Sentiment analysis of financial news for stock price prediction: Empirical evidence from India, 07 November 2022, (Version 1), Research Square, https://doi.org.

[41] Soni, S., Shirvastava, A. K., Motwani, D., & Pradesh, G. Feasibility Study of Stock Market Prediction for Sentiment Analysis using Artificial Intelligence.

MDPI Electronics 2021, 10(21), 2717

[42] Fuller, A. (2022). Predicting Stock Market Indicators through Sentiment Analysis on Twitter (Doctoral dissertation, University of Iowa).

[43] Jishtu., Prajwal., Prajapati, D., (2022) Prediction of the Stock Market Basedon Machine Learning and Sentiment Analysis, IEEE International Conference onData Mining Workshops (ICDMW).

[44] Bhardwaj, A., Narayan, Y., & Dutta, M. (2015). Sentiment analysis for Indianstock market prediction using Sensex and nifty. Procedia computer science, 70, 85-91.

[45] Oliveira, N., Cortez, P., & Areal, N. (2017). The impact of microblogging data for stock market prediction: Using Twitter to predict returns, volatility, trading volume and survey sentiment indices. Expert Systems with applications, 73, 125-144.

[46] Gupta, J., Jain, A., & Bohra, Y. (2018). Sentimental analysis on news data for stock market prediction. Int J Manage Appl Sci, 4(6), 84-86.

[47] Lee, H., Surdeanu, M., MacCartney, B., & Jurafsky, D. (2014, May). On theImportance of Text Analysis for Stock Price Prediction. In LREC (Vol. 2014, pp.1170-1175).

[48] Kumar, K. S. M. V., Kumar, G. R., & Rao, J. N. (2020). Use sentiment analysis to predict future price movement in the stock market. International Journal of Advanced Research in Engineering and Technology, 11, 1123-1130.

[49] Cohen-Charash, Y., Scherbaum, C. A., Kammeyer-Mueller, J. D., & Staw,
B. M. (2013). Mood and the market: can press reports of investors' mood predict stock prices?. PloS one, 8(8), e72031.

[50] Uhr, P., Zenkert, J., & Fathi, M. (2014, October). Sentiment analysis in financial markets A framework to utilize the human ability of word association for analyzing stock market news reports. In 2014 IEEE International Conference on Systems, Man, and Cybernetics (SMC) (pp. 912-917). IEEE.