Soft Computing
https://doi.org/10.1007/s00500-019-04347-y

FOCUS

Predicting stock market trends using machine learning algorithms via public sentiment and political situation analysis

Wasiat Khan[1] · Usman Malik[2] · Mustansar Ali Ghazanfar[3] · Muhammad Awais Azam[4] · Khaled H. Alyoubi[5] · Ahmed S. Alfakeeh[5]

Abstract

Stock market trends can be affected by external factors such as public sentiment and political events. The goal of this research is to find whether or not public sentiment and political situation on a given day can affect stock market trends of individual companies or the overall market. For this purpose, the sentiment and situation features are used in a machine learning model to find the effect of public sentiment and political situation on the prediction accuracy of algorithms for 7 days in future. Besides, interdependencies among companies and stock markets are also studied. For the sake of experimentation, stock market historical data are downloaded from Yahoo! Finance and public sentiments are obtained from Twitter. Important political events data of Pakistan are crawled from Wikipedia. The raw text data are then preprocessed, and the sentiment and situation features are generated to create the final data sets. Ten machine learning algorithms are applied to the final data sets to predict the stock market future trend. The experimental results show that the sentiment feature improves the prediction accuracy of machine learning algorithms by 0–3%, and political situation feature improves the prediction accuracy of algorithms by about 20%. Furthermore, the sentiment attribute is most effective on day 7, while the political situation attribute is most effective on day 5. SMO algorithm is found to show the best performance, while ASC and Bagging show poor performance. The interdependency results indicate that stock markets in the same industry show a medium positive correlation with each other.

Keywords Natural language processing · Predictive models · Stock markets · Sentiment analysis

Communicated by M. D. Lytras.

✉ Wasiat Khan
wasiat.khan@gmail.com

Usman Malik
usman.malik@insa-rouen.fr

Mustansar Ali Ghazanfar
eng.musi@gmail.com

Muhammad Awais Azam
awais.azam@uettaxila.edu.pk

Khaled H. Alyoubi
kalyoubi@kau.edu.sa

Ahmed S. Alfakeeh
asalfakeeh@kau.edu.sa

[1] Department of Software Engineering, University of Engineering and Technology, Taxila 47050, Pakistan

[2] LITIS Lab, Normandy University, INSA Rouen, 76000 Rouen, France

[3] School of Architecture, Computing and Engineering, University of East London, London, UK

[4] Department of Computer Engineering, University of Engineering and Technology, Taxila 47050, Pakistan

[5] Faculty of Computing and Information Technology, King Abdulaziz University, Jeddah, Saudi Arabia

Published online: 17 September 2019

1 Introduction

The stock market prediction has always been an intriguing and challenging problem for finance and statistics experts. A basic motivation behind the stock market prediction is to buy stocks that are likely to rise in price in future and sell those stocks which are likely to decline. Traditionally, two approaches are used for stock market prediction. One is fundamental analysis technique that relies on a company's fundamental information, such as annual growth rates, market position, revenues, and expenses (Murphy 1999). The second is the technical analysis approach that focuses on historical stock prices for stock market prediction. Technical analysis practitioners use historical price charts and patterns to make predictions (Turner 2007).

In the past, stock market predictions were usually made by financial experts; however, with the advancements in data acquisition techniques, computer scientists have also embraced the prediction problem. The amount of data available for processing is so large that an entire discipline, known as data mining, has emerged for extracting useful information from data. Data scientists started employing machine learning algorithms to develop prediction models for stock markets, resulting in the development of several stock prediction models. However, none of the models effectively included external factors, such as public sentiment and political situations. This paper aims to use these features as well in machine learning algorithms and predict stock prices based on these attributes.

Stock market prediction is a challenging task, and researchers may also face challenges while developing a predictive system. The main challenge in stock market prediction is that the market is complex and nonlinear, with much of the complexity caused by the correlation between market behaviour and investment psychology (Fan et al. 2009). Stock markets are also volatile, or unstable. This volatility occurs when prices of underlying securities fluctuate. These prices move when the expected value of securities changes due to real events; since these events are unexpected, volatility also occurs unexpectedly. Traders face these kinds of challenges in predicting the future of stock markets.

Machine learning techniques can be used in stock market prediction systems to help financial analysts and traders in their decision-making. These algorithms aim to automatically learn and recognize patterns in large amounts of data. Since these algorithms are self-organizing and self-learning, they can be effective in tackling the task of predicting stock price fluctuations and can help in developing automated trading strategies based on these predictions. Algorithmic trading is also more systematic (Lu 2016). Human investors are very emotional, and emotional factors can easily destroy the trading discipline. Therefore, a machine learning approach is advantageous in developing stock market prediction models.

Soft computing is the idea of building wiser and intelligent systems that can work in a similar way as human beings can do. It has many applications in various fields, including civil engineering, biomedical engineering, etc. The research area is interesting and therefore attracted researchers to study the application of soft computing methods, such as fuzzy systems and artificial neural network (ANN) in various fields.

In recent years, soft computing methods like ANN, neuro-fuzzy models, etc., have been successfully applied in solving complex civil engineering problems that were difficult to solve using traditional methods. This includes the prediction of flood, earthquake, tsunami, and landslide. These problems badly affect the service life of the structure by originating premature deterioration. The prediction of these natural hazards involves a complex process that depends on different environmental and physical parameters that are naturally stochastic. Degradation of reinforced concrete (RC) structures can also be considered a natural hazard which depends on different physical and environmental parameters. Although the existing models are reliable to estimate the service life of RC, they are computationally expensive in reducing computational time to a great extent, thus needing a long completion process (Papadrakakis and Lagaros 2002; Cardoso et al. 2008). To overcome this problem, Dey et al. (2019) used various well-known corrosion models and compared their results with the predicted results of the ANN model to forecast the reliable service life of RC structures. The predicted output of ANN was validated with the intended outputs resulting in good results for predicting the reliable service life of RC structures. The purpose of ANN was to avoid complicated mathematical interpretations so that consistent results for predicting the service life of RC could be obtained. Similarly, researchers (Naderpour and Mirrashid 2019) studied joints in the RC beam–column. The study aimed to find the maximal shear capacity of the joints as an important parameter in the damage of RC structures. For this reason, they used two soft computing methods including group method of data handling (GMDH) and a neuro-fuzzy model called adaptive neuro-fuzzy inference system (ANFIS). Experiment results showed that the proposed methods were capable of finding the shear capacity of RC joints with maximum accuracy. It was also shown that ANFIS gave improved results than GMDH. The predicted results were compared with the results obtained using conventional methods. From the comparison, they concluded that the proposed soft computing methods could be effectively used to find the shear capacity of RC beam–column joints. Naderpour et al. (2019) presented a new computational technique to determine the bond strength of fibre-reinforced polymer strip-to-concrete joints using the ANFIS neuro-fuzzy model. The results obtained using ANFIS method showed improved accuracy in the model and less error in comparison with existing methods.

By studying the application of soft computing methods in solving civil engineering problems, it can be concluded that the methods can also be employed for stock market prediction to achieve good prediction results. In the current work, the authors have used different soft computing methods for stock market prediction.

The existing models for stock prediction either find the mood dimension with the highest effect on stock prices or they simply find correlations between Twitter sentiments and stock prices. The current work focuses on developing a machine learning model independent of the mood state, which predicts stock trends based on the overall sentiment of people towards a particular company. The current approach is entirely novel in that the impact of an analysis of people's sentiments towards individual companies has not yet been studied. Also, the effects of political events on stock markets have been found by Taimur and Khan (2015) using correlation, but no machine learning model has been developed which incorporates a political situation as an attribute and which predicts stock prices on that basis. The current research has done this for the first time. The purpose of this research is to establish evidence in support of or against the notion that public sentiment and political situations, either both or singly, affect the stock market and then, based on this evidence, to develop a machine learning model for stock prediction. The important contribution of this research is the inclusion of an analysis of external factors, such as public sentiment and political situations, in developing a machine learning model to predict stock markets. This paper aims to include these features in machine learning algorithms and predict stock prices using these attributes. To develop this model, the authors crawled tweets from Twitter and political news from Wikipedia[1] and pre-processed the data to discover the sentiment and situation features for stock market prediction.

The rest of the paper is organized as follows. Section 2 reviews the related work on optimization techniques and the effects of social media and political news on stock market prediction. The research methodology is presented in Sect. 3. The process of creating the final data sets used in this research is given in Sect. 4. Section 5 discusses the proposed systems, while the results analysis and comparative study are described in Sect. 6. Conclusions are given in Sect. 7.

2 Related work

The main aspects of the literature are the following:

1. Different optimization techniques have been used to solve complex optimization problems.
2. Researchers have used different machine learning algorithms and different types of data, such as historical, social, and news data for stock market prediction with varying accuracy.
3. It is proved by researchers that Naïve Bayes is the best machine learning algorithm for text classification.
4. It has been found that public emotions and opinions expressed in Twitter posts can be used for stock market forecasting.
5. Prediction accuracy can be improved if more data sources are used.

2.1 Optimization techniques

With the rapid progression of society and economics, optimization problems are further complicated; therefore, traditional optimization techniques are unable to get satisfactory results. Therefore, researchers have started working on improving optimization algorithms for improving performance in various optimization problems like solving complex optimization problems, feature extraction for mechanical fault diagnosis, multi-objective gate assignment in an airport, etc. For example, Deng et al. (2019a, b) recommended a better version of an ant colony optimization (ACO) algorithm called multi-population co-evolution ACO (ICMPACO) to balance the solution diversity and convergence speed and to improve optimization performance to solve gate assignment and travelling salesmen optimization problems. The algorithm was based on co-evolution and multi-population techniques. ICMPACO's optimization performance was compared with that of ACO. From the experiment, they concluded that the proposed optimization algorithm shows better optimization performance while solving the problem of travelling salesmen and solved the problem of gate assignment effectively. The proposed algorithm also took on improved optimization capability and stability. Deng et al. (2017a, b) suggested a genetic adaptive collaborative ACO algorithm called MGACACO by introducing a disordered optimization technique, collaborative multi-population technique, and adaptive control parameters into genetic algorithm (GA) and ACO to solve complex optimization problems. The purpose of the MGACACO was to solve the problems of slow convergence speed in ACO and GA's weak searching ability. The experiment findings showed that the proposed algorithm improved the

[1] https://www.wikipedia.org.

searching ability and convergence speed of GA and ACO algorithms, respectively. Zhao et al. (2016) presented a novel features extraction technique, named EDOMFE, to diagnose a fault in a motor bearing. The technique used multi-scale fuzzy entropy, mode selection, and empirical mode decomposition to accurately identify the fault. SVM was employed to classify fault type and its severity in motor bearings. The technique was compared with existing methods and was found effective for diagnosing a fault in a motor bearing. An innovative and intelligent fault diagnosis technique for motor bearing was also studied by Deng et al. (2019a, b) by introducing fuzzy information entropy, empirical mode decomposition, improved PSO, and LS-SVM algorithms into fault diagnosing process. The purpose of using PSO was to optimize LS-SVM parameters for constructing optimum LS-SVM. This optimal algorithm was then used to categorize faults in a motor bearing. From experiment findings, they concluded that the proposed fault diagnosis technique outperformed the existing methods. An effective and multi-objective optimization algorithm was presented by researchers (Deng et al. 2017a, b) for solving gate assignment problem in airports. They proposed an enhanced adaptive PSO algorithm called DOADAPO that used dynamic fractional calculus and alpha-stable distribution's advantages. The dynamic fractional calculus was used for improving the convergence speed. The purpose of using alpha-stable distribution was to enhance the global searching capability of the algorithm. Then, DOADAPO was employed to construct the proposed optimization algorithm for gate assignment problem for efficient and fast assignment of gates to flights. The effectiveness of the proposed model was then verified by actual flight data. Results revealed that the DOADAPO algorithm improved the convergence speed and searching capability and multi-objective optimization technique enhanced the gate assignment service in an airport. Zhao et al. (2018) proposed a technique for identifying fault severity in rolling bearings. The technique used high-order difference mathematical morphology gradient spectrum entropy (HMGSEDI) for completely analysing fault severity and quantitatively describing the degree of fault damage in rolling bearings. They showed that the HMGSEDI method effectively identified the fault damage degree and provided an innovative way for identifying the degree of fault damage and forecasting of fault in rotating machinery.

We can conclude from studying optimization techniques to solve optimization problems that machine learning algorithms can also be optimized to solve stock prediction problem.

2.2 Stock market prediction using stock market historical data

Researchers have used different machine learning algorithms to mine historical, social media, and news articles data to develop prediction models for stock markets. When social media and online news websites were less popular than they are at present, researchers had to rely on the usage of historical data to predict the stock market. For example, researcher (Mostafa 2010) used generalized regression neural networks and multilayer perceptron (MLP) neural network (NN) architectures over daily closing prices to forecast the movement of closing price in Kuwait stock exchange. He proved that NN can learn the relationship between input and output features for making predictions over the data on which the network is trained. Hegazy et al. (2014) presented a machine learning model for stock market price prediction based on technical indicators and historical data of different stocks. A prediction model was also proposed by Kazem et al. (2013) that used the firefly algorithm and support vector regression (SVR) to forecast the stock market price. Shen et al. (2012) used SVM to forecast the coming day's stock market trend. They used global stock data as input features to SVM. To predict stock prices, researchers (Olaniyi et al. 2011) analysed price data of three banks using regression analysis. They predicted stock prices for the banks using the information contained in daily and weekly activity summaries published by the Nigerian Stock Exchange. Two models were developed and compared by researchers (Kara et al. 2011) to predict the movement direction in the Istanbul Stock Exchange using ANN and SVM on ten technical indicators selected as inputs to the proposed models. Results revealed that the performance of ANN was better than SVM. Ou and Wang (2009) used different machine learning algorithms to predict Hong Kong Stock Market index price movement using historical price data. They showed that SVM and least-square SVM gave better forecasting performance compared to other predictive models. Sadhukhan et al. (2016) used historical stock data of various companies for future stock price prediction using the ANN model. The structure of ANN was studied by Shahbaz et al. (2014) for analysing and processing data, and it was concluded that ANN is a suitable model for forecasting stocks and currency. Egeli et al. (2003) found the best model for predicting the Istanbul Stock Exchange market index values, finding that prediction models based on ANNs were more accurate.

In summary, we conclude that these prediction models used stock market historical data to forecast stock markets.

2.3 Stock market prediction using social media data

Social networking platforms allow their users to interact directly with each other. These platforms have become popular venues for their users to share financial analysis results concerning financial securities. This results in the availability of enormous amounts of social media data that attracted researchers to mine these data for financial purposes. Therefore, research has been conducted on investor's opinions published on social media to predict stock markets. For example, researchers (Lakshmi et al. 2017) analysed Twitter data and found that opinions in tweets are highly heterogeneous and structured and are positive, negative, or neutral. Sentiment classification methodologies, namely lexicon-, rule-, and machine learning-based techniques, concerning Twitter data were explored by a researcher (Yuan 2016). For lexicon-based methods, simple word count and the feature scoring approaches were used. For machine learning-based methods, the study used maximum entropy, Naïve Bayes, and SVM. Two features—the bag-of-words (BOW) model and the *n*-gram and part-of-speech linguistic annotations—were compared. It was found that the BOW feature was simple and effective and achieved the best performance. Rahman and Ali (2016) proposed two techniques, SVM and a sentiment classification algorithm (SCA), to classify tweet sentiment labels accurately. Results revealed that SCA always performed better than SVM.

Stock market traders and financial analysts express their emotions in tweets concerning a specific stock. Machine learning algorithms and natural language processing (NLP) techniques can be used to find such emotions in tweets. There exists a correlation between emotions expressed on Twitter and stock markets, and the correlation can be used to forecast features of a stock market (Zhou et al. 2016). Ahuja et al. (2015) performed sentiment analysis of public opinion data from Twitter to find a relationship between the stock market and public mood, using a self-organizing fuzzy neural network. A model was developed by researchers (Nguyen et al. 2015) to predict the movement of stock prices, using sentiments of company-related topics from social media. Results revealed that an average accuracy of 54.41% was achieved by the model. An improvement of 2.07% was also observed when sentiment analysis data were used in the prediction model.

With the availability of social media websites, investors have access to the most useful and latest information consisting of public emotions that influence investors' decisions. Researchers (Bing et al. 2014) proposed an association rule mining technique to find patterns between stock price movements and public sentiment. The proposed technique was applied to tweets related to different industries. Results showed that the proposed algorithm showed better accuracy for some industries, such as IT and media, and that the public sentiment was more effective on the third day after the trading day. Average accuracy of 66.48% was achieved by the proposed algorithm for the selected industries. A regression model was used by researchers (Oliveira et al. 2013) on the data downloaded from StockTwits[2] to predict three variables related to the market: trading volume, volatility, and returns, and found that stock return could not be predicted using sentiment indicator, while posting volume could be used to predict volatility and trading volume.

From this discussion on the usage of social media data for stock market prediction, we can find that there exists some relationship between opinions expressed on social media and the stock market.

2.4 Stock market prediction using political news articles

News articles that are publicly available online provide an interesting type of data for mining and analysis to extract useful information. The news may be categorized as general news, financial news, and political news. This type of data can also be used for stock market prediction. Like social media, political situations also impact stock price returns. Therefore, researchers have also worked on the analysis of political news for stock market prediction. For example, a textual analysis of Euro-periphery crisis-related news was performed by a researcher (Chouliaras 2015) to find stock market volatility and returns. It was found that increased pessimism results in an increase in volatility and a decrease in stock prices. The importance of text analysis for stock price changes prediction (i.e. up, down, or stay) was investigated by Lee et al. (2014) using RF classifier on financial events reported in 8-K documents and stock prices in order to forecast a company's stock price changes. These authors found that using text improved the prediction accuracy by over 10%, especially for short-term prediction, i.e. 1 or 2 days after the financial event occurred. They also found that the impact of the event persisted up to 5 days.

Researchers have found many political and economic factors that can influence stock markets. For example, the effect of political and catastrophic events on stock market data was studied by researchers (Taimur and Khan 2015). They collected 43 political and four catastrophic events. The political events were divided into favourable and non-favourable political events for the country. The impact of these events was studied for 1, 5, 10, and 15 days event

[2] http://www.stocktwits.com.

window to find a correlation between events and stock prices. They showed that political and favourable political events have impacts on the stock market returns up to 5 days after the event occurred, while unfavourable political events had an abrupt impact. They used a mathematical model for this purpose. Suleman (2012) used the enhanced GARCH model (Bollerslev 1986) to find the impact of political news related to political parties on the volatility and returns of the Karachi Stock Exchange. Political news was categorized into good and bad news. The experiment results showed that good political news had a positive influence on the returns of stock and decreased stock volatility. On the other hand, bad political news decreased the stock returns and increased the volatility of the stock market. The correlation between trading volume and stock returns, based on political events, was studied statistically by researchers (Malik et al. 2009) using the Phillips–Perron unit root test method. They showed that stock returns and trading volume fluctuated in a positive or negative way depending on the intensity of the political event. Market-adjusted techniques, namely ordinary least-square (OLS) regression, and risk-adjusted technique, namely multivariate regression model (MVRM), were used by researchers (Chen et al. 2005) to investigate the effect of political events on the performance of Taiwan stock market. From results obtained using market-adjusted OLS, it was found that political events in Taiwan which were related to political elections, the development of cross-strait relationships, and economic policies were associated with the bad performance of the Taiwan stock market. From MVRM results, they concluded that stock price reaction to political events was somewhat unimportant which implied that, with a few exceptions, these political events are not informative. The influence of political news related to the separation of Quebec from Canada was studied on the return volatility of firms in Quebec by researchers (Beaulieu et al. 2005). They used a GARCH model for this purpose. From experiment results, they showed that volatility of stock return varied with the degree to which a company was exposed to political risks. They also showed that critical political events had more significant effects on stock return volatility compared to favourable political events.

From the analysis of related work, we can see that all the prediction models either find a mood dimension which has the highest effect on stock prices, or they simply find a relationship between stock prices and Twitter sentiment. In this study, the authors have tried to develop a machine learning model independent of the mood state which predicts stock trends based on an analysis of the overall sentiment of people towards a particular company, a novel approach that has not yet been studied. For political news analysis, most of the research studies used conventional models and got different accuracies. The summary of presented related work is given in Table 1. The literature in this table is presented with respect to the research problem, contributions, research variables, and the main insights into our research problem.

In summary, we can conclude from the literature that stock market prediction problem can be optimized by using some innovative techniques and that social media, for example, Twitter posts, as well as political events, influence stock market prices, trends, and returns. Researchers have used different machine learning algorithms and data types, such as Twitter posts, topics discussed on social media, and political news, to forecast stock market trends. The main focus of these studies is on three important research variables: accuracy, performance, and the correlation between stock trends and opinions expressed on social media. The authors can conclude from the previous research that ANN is good for obtaining accurate results; Naïve Bayes is the best algorithm for text classification, and SVM shows good prediction performance. The authors have also concluded from the literature that using news articles as a data source can improve accuracy. The literature helped us to select the best machine learning algorithms, data sources, and research variables for our research design.

3 Research methodology

In this section, the authors discuss the research variables used in this study and the research questions which will be answered.

3.1 Research variables

Variables are simplified portions of a complex problem that a researcher intends to study. They are measurable and can be categorized as dependent or independent. Following are the main research variables of this study:

(1) prediction accuracy using external features (sentiment);
(2) the impact of political situations on stock prices; and
(3) stock market interdependency.

3.2 Research questions

The following are the research questions of the current study, which are answered in Sect. 6.

(1) What is the effect of the sentiment feature on prediction accuracy?

Table 1 Summary of previous research on stock market prediction

No.	References	Research problem	Main contribution	Research variables	Insights into the current research
1	Gidofalvi and Elkan (2001)	Prediction of stock price movements using news articles	Numerical indicators were extracted from financial text to predict stock price behaviour	Relationship between news articles and stock price behaviour	Classification of articles
2	Egeli et al. (2003)	Stock market prediction using ANN	ANN model produces accurate results in stock prediction	Accuracy	ANN is best for getting accurate results in stock forecasting
3	Beaulieu et al. (2005)	Political risk impact on stock return volatility	Political news affects stock return volatility	Stock return volatility	Political news impacts volatility of stock market returns
4	Schumaker and Chen (2009)	Prediction of stock market via financial news	Prediction of stock market via financial news	Prediction accuracy	Increasing data sources can enhance accuracy
5	Tang et al. (2009)	Prediction of stock market using time series data and news articles	Prediction of stock market using time series data and news articles	Performance	Increasing data sources can enhance performance
6	Tayal and Komaragiri (2009)	Comparing blogging and microblogging impacts on market performance	Microblogging platforms (Twitter) have a higher predictive accuracy than traditional blogs	Blogging and microblogging impacts on market performance	Twitter has the highest predictive accuracy, owing to its conciseness
7	Malik et al. (2009)	Political events' impact on stock returns and trading volume	Stock return and trading volume fluctuate due to the impact of political events on stock market price	Trading volume, stock return	Political events' impact on stock market returns
8	Go et al. (2009)	Sentiment classification of Twitter using distant supervision	Introduced pre-processing steps for text mining	Accuracy	Text pre-processing steps
9	Ou and Wang (2009)	Predicting index movement of stock market using ten data mining techniques	Found SVM and LS_SVM as best techniques for stock prediction	Predictive performance	SVM and LS_SVM show good performance in stock prediction
10	Pak and Paroubek (2010)	Using Twitter corpus for opinion mining and sentiment analysis	Naïve Bayes shows good performance in text mining	Sentiments	Naïve Bayes is good classifier for text classification
11	Suleman (2012)	Stock market reaction to good and bad political news	Found effect of political news on stock market volatility and returns	News impact on stock returns and volatility	Impact of news on stock volatility and returns
12	Makrehchi et al. (2013)	Prediction of stock market through sentiment analysis of events related to market	Training data set was built from market-related events for the supervised learning of event sentiments	Market performance	Market-related events data can be used for stock prediction
13	Hagenau et al. (2013)	Stock price prediction using financial news	The approach allowed selection of semantically relevant features and reduced over-fitting problem; also used market feedback in feature selection	Classification accuracy	Accuracy can be improved by using a good feature selection method, for example selection of semantically relevant features
14	Li et al. (2014a, b)	The effect of public mood and news on stock movements	Investigated quantitatively the effect of media on stock markets	Media impact on stock market	News and social media can affect stock market movement
15	Li et al. (2014a, b)	The impact of news on stock price returns through sentiment analysis	News sentiment analysis helped in improving prediction accuracy at different stock levels	Impact of news on stock price return, prediction accuracy	Accuracy can be improved by using news sentiments

Table 1 (continued)

No.	References	Research problem	Main contribution	Research variables	Insights into the current research
16	Bing et al. (2014)	Stock price movement prediction using patterns in public sentiments and real stock data	Prediction of stock price movement using patterns between sentiments and real stock data	Stock price movement	Patterns between public sentiments and stock data can be used to find price movement in stocks
17	Ahuja et al. (2015)	Relation between public mood and stock market	Prediction of stock market movement using public opinions	Relation between public moods and stock market	Understanding the relationship between public opinions expressed in tweets and stock market movement
18	Taimur and Khan (2015)	Political and catastrophic events' impact on stock market returns	Political events have impact on stock market up to 5 days after the event	Correlation between events and stock prices	A correlation exists between political events and stock market
19	Nguyen et al. (2015)	Analysis of public sentiments on social media for prediction of stock movement	Proposed a novel feature called "topic-sentiment" for improving prediction performance of stock market	Stock market performance	Stock market can be predicted based on specific topics
20	Zhou et al. (2016)	Prediction of stock market by using emotions from tweets	Correlation between emotions expressed in tweets and stock market	Correlation between emotions expressed in tweets and stock market	Understanding the correlation between emotions expressed in tweets and stock market prediction
21	Zhao et al. (2016)	Diagnose fault in motor bearing	Proposed EDOMFE	Fault type and severity	Prediction problem can be optimized
22	Deng et al. (2017a, b)	Improve optimization performance	Proposed MGACACO	Convergence speed	Machine learning techniques can be optimized
23	Deng et al. (2017a, b)	Solving gate assignment problem	Proposed DOADAPO	Convergence speed, searching capability	Machine learning techniques can be optimized
24	Zhao et al. (2018)	Identify fault severity in motor bearing	Proposed HMGSEDI	Fault severity	Innovative technique can be proposed to solve a problem
25	Deng et al. (2019a, b)	Improve optimization performance	Proposed ICMPACO	Optimization performance	Machine learning techniques can be optimized
26	Deng et al. (2019a, b)	To categorize faults in motor bearing	Proposed innovative and intelligent fault diagnosis technique	Classification performance	Machine learning algorithms can be optimized

(2) On what day after prediction day, the highest prediction accuracy is achieved?
(3) What is the effect of political situations on stock prices?
(4) How different stocks are interdependent?

To answer the first two questions, the authors will perform a sentiment analysis of the social media data and predictions will be made for 7 days after the day on which the trade is executed. Prediction accuracies will be compared to find the day on which the sentiment feature is most effective. For answering question 3, political events occurred in Pakistan will be analysed to find their impact on the stock market of Pakistan. The relationship among stock trends of Pakistan, London, and New York stock markets will be found to answer the last research question of this study. Figure 1 shows the block diagram of the overall flow of events in the proposed methodology for the proposed system.

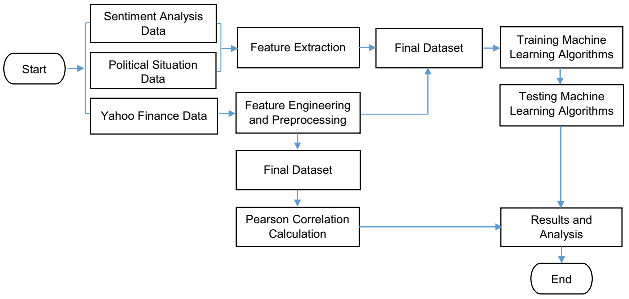

Start
Sentiment Analysis Data
Political Situation Data
Yahoo Finance Data
Feature Extraction
Final Dataset
Training Machine Learning Algorithms
Testing Machine Learning Algorithms
Feature Engineering and Preprocessing
Final Dataset
Pearson Correlation Calculation
Results and Analysis
End

Fig. 1 Block diagram for the overall flow of events in the proposed system for stock trend prediction using social media and political situations and for stock interdependency system

Table 2 Selected stock markets and the training data information

Data set type and stock name	Time period	
	From	To
Yahoo! Finance data[a] or stock market historical data (Google, Microsoft, Apple, Karachi Stock Exchange, London Stock Exchange, NASDAQ, New York Stock Exchange, Facebook, LinkedIn, Twitter)	1 July 2016	30 June 2017
Sentiment analysis data[b] (Google, Microsoft, Apple)	1 July 2016	30 June 2017
Political situation analysis data[c] (Karachi Stock Exchange)	28 May 1998	30 June 2017

[a] https://finance.yahoo.com/quote/MSFT?p=MSFT&.tsrc=fin-srch

[b] https://developer.twitter.com/en/docs.html

[c] https://en.wikipedia.org/wiki/Timeline_of_Pakistani_history_(1947–present)

4 Data sets

The complete and final data sets for this research problem are not available. Therefore, we perform some steps to create the final data sets for our proposed three subsystems. The stock historical raw data will be needed for all selected companies and stock markets, while sentiment analysis data will be needed for Google, Microsoft, and Apple only. Political situation data will be needed for Karachi Stock Exchange only as we will analyse the impact of political events of Pakistan on the Karachi Stock Exchange. The historical and sentiment analysis data will be downloaded from 1 July 2016 to 30 June 2017, while stock historical and political situations data of Pakistan will be downloaded from 28 May 1998 to 30 June 2017. Table 2 shows the data set types and time period for collecting data for different stock markets and individual companies.

4.1 Stock market historical data sets

These data sets will be used to create final data sets for sentiment analysis, political situation analysis, and stock market interdependency systems for the selected stock markets and individual companies.

4.1.1 Data collection process

Stock market data of different stock markets and individual organizations can easily be fetched via Web services like Yahoo! Finance[3] and Google Finance.[4] In this paper, the authors use Yahoo! Finance as their primary source of stock market data. The reason for selecting it as a data source of historical data is that all stock markets historical information is available from the date when the company joined the market to the current date. Another reason for selecting Yahoo! Finance as a source of stock market historical data is that data can be downloaded using the stock market ticker symbol. Also, most of the researchers (e.g. Ou and Wang 2009; Hegazy et al. 2014; Lee et al. 2014; Nguyen et al. 2015; Taimur and Khan 2015) used Yahoo! Finance as a data source of stock market historical data for stock market prediction problems.

Researchers can fetch data for research purposes for different stock markets and individual companies between two inclusive dates. Figure 2 shows the fetching process of downloading stock historical data from Yahoo! Finance for Apple Inc. The data of a company can be downloaded using the ticker symbol or stock market symbol of that company. To download historical data, historical data option and start date and end date are selected to download the data for a specific time period. Data can be fetched at a daily, weekly, or monthly basis. For the current research, we download daily stock historical price data. The downloaded stock data file has seven attributes: Date, Open, High, Low, Close, Adjusting Close, and Volume.

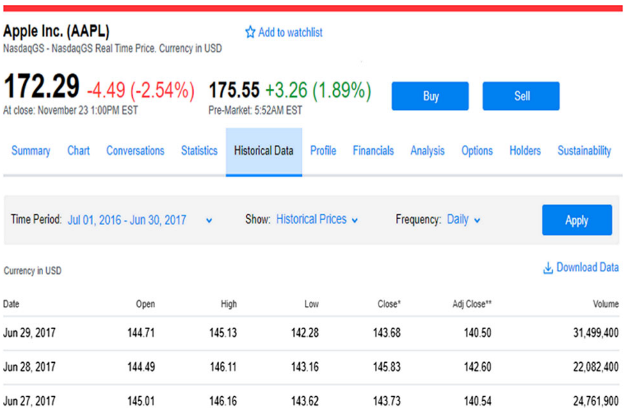
4.1.2 Feature engineering

The Yahoo! Finance stock historical data are pre-processed by removing attributes not required for stock price prediction. The Adjusting Close attribute is removed as it is unnecessary, and two more attributes, Trend and Future Trend, are created in the stock historical data file since they are necessary for the proposed model of stock prediction. The Trend is the difference between the opening and closing price of a stock on a specific date. Future Trend is the stock trend after n number of days from the date on which the trade is executed.

The downloaded raw stock historical data for sentiment analysis contain 251 stock data records of selected stock markets. For political situation analysis, the downloaded

[3] https://www.finance.yahoo.com.

[4] https://www.google.com/finance.



Apple Inc. (AAPL)
NasdaqGS - NasdaqGS Real Time Price. Currency in USD

172.29 -4.49 (-2.54%)
At close: November 23 1:00PM EST

175.55 +3.26 (1.89%)
Pre-Market: 5:52AM EST

Time Period: Jul 01, 2016 - Jun 30, 2017
Show: Historical Prices
Frequency: Daily

Currency in USD

Date	Open	High	Low	Close*	Adj Close**	Volume
Jun 29, 2017	144.71	145.13	142.28	143.68	140.50	31,499,400
Jun 28, 2017	144.49	146.11	143.16	145.83	142.60	22,082,400
Jun 27, 2017	145.01	146.16	143.62	143.73	140.54	24,761,900

Fig. 2 Fetching stock historical data from Yahoo! Finance between two dates for Apple Inc

raw stock historical data for the Karachi Stock Exchange (KSE) comprise 4718 stock data records. Attributes of the stock historical data set are described in Table 3.

4.2 Sentiment analysis data set

4.2.1 Data collection process

The best method for concisely extracting sentiments of the general public is via microblogging websites; owing to its conciseness and popularity (Tayal and Komaragiri 2009), Twitter is chosen as the data source for social media data. Another reason for choosing Twitter as a data source for social media data is that most of the researchers (e.g. Tayal and Komaragiri 2009; Ahuja et al. 2015; Yuan 2016; Rahman and Ali 2016; Zhou et al. 2016; Lakshmi et al. 2017) used Twitter as the data source of social media data for stock market prediction. One more reason for selecting Twitter as the data source of social media data is because of Twitter posts' succinctness, high volume, and real-time characteristics, all of which contribute to the predictive power of its data set (Oh et al. 2011).

To fetch the desired data from Twitter, Twitter API is implemented in Python. The following search criteria are used to collect tweets of some stock market:

(1) Search by keyword;
(2) Search between two dates;
(3) Search by user or source; and
(4) Any of the above in conjunction.

Tweets can be scraped 1 day before the stock quote or *n* number of days before the quote; for the sake of simplicity, tweets are scraped 1 day before the stock prices. For sentiment analysis, all tweets are scraped on a given day which contains the keywords "#GOOG", "#MSFT", and "#AAPL". These are stock market symbols preceded by hashtag (also called ticker symbols) for Google, Microsoft, and Apple Inc. stocks, respectively. The downloaded sentiment analysis data files have four attributes: Source, Text, Date, and Sentiment. Note that the Sentiment attribute has no values at this stage.

4.2.2 Feature engineering

Techniques for extracting features find specific information in textual documents (Billsus and Pazzani 2000). This information is then used in model building and in training machine learning classifiers. The following steps are carried out to extract features from the raw tweets.

(1) Pre-processing

Tweets usually consist of strings of words. Machine learning classifiers cannot be used to classify these strings; they must be converted into some proper format acceptable for the classifiers. The following steps are carried out to convert these strings into a proper format for machine learning classifiers.

Step 1 Tweets are converted into tokens.

Step 2 HTML (hypertext markup language) and other tags, such as author tags (@) and hashtags (#), are removed from tweets. These tags need to be removed, as

Table 3 Attributes of the stock market historical data set

Feature	Format	Description
Date	Date	Date value of the day when the stock was traded
Open	Integer	Stock opening price on a specific date
High	Integer	Contains the highest selling price of a stock on a specific date
Low	Integer	Contains the lowest selling price of a stock on a specific date
Close	Integer	Contains closing price of a stock on a specific date
Volume	Integer	The number of shares traded on a specific date
Trend	Nominal	Difference between closing and opening price on a specific date
Future Trend	Nominal	Attribute whose values will be predicted. It is the difference of the closing price for today and the closing price after "*n*" number of days

they carry no relevant information for classifying text documents and can confuse text classifiers.

Step 3 URLs are removed, as URLs also confuse classifiers.

Step 4 Stop words are removed. These are words that frequently occur in sentences (for example, a, an, is, are, the, etc.) but carry no important information for text classifiers. In machine learning, it is a general practice to remove these words before using the data in a learning algorithm.

Step 5 Stemming is performed which means that the words which are based on the same stem are all changed to that one stem. Inflection information and cases are removed from the words.

(2) Indexing

Text documents are usually represented by vectors of *N*-weighted index words. A technique for this representation of text documents, called vector space model, is used for this purpose. In this technique, a text document is represented by word vectors. A group of text documents can be represented with a word-by-document matrix, in which every record represents the presence of a word in a text document.

(3) Bag-of-words representation

The indexing step results in a BOW representation of documents, whose tokens are characterized as features and their corresponding weights as feature values. In this step, weights are assigned to words, based on the presence of words in tweets, using the term frequency approach. Words with relatively higher weights are regarded as more important features compared to words with less weights.

4.2.3 Sentiment analysis

Sentiment analysis of the raw tweets is performed using Jeffrey Breen (JB) (2011) approach and the Stanford Sentiment Analysis (Socher et al. 2013) package of the Stanford NLP (SNLP).[5] JB's approach uses a dictionary or lexicon of positive and negative words to find the total count of positive and negative words in a tweet. The sentiment of a tweet is then the difference between the number of positive words and negative words. On a particular date, the final sentiment of all tweets is the aggregated sentiment of individual tweets. The higher the sentiment count on a particular day, the higher the positivity of the sentiment on that day. The second approach used to perform sentiment analysis is the SNLP approach. This is a learning-based sentiment analysis library that is known to have yielded accurate results in the range of 50–70%, depending upon the document corpus used.

For each date in stock market historical data set file, a corresponding sentiment analysis file is generated. Each file contains four attributes; the authors generate sentiment analysis files for Apple (20,700), Google (4215), and Microsoft (26,178) stocks. The values in brackets show the number of downloaded tweets. The structure of a sentiment analysis data file on 5 July 2016 for Apple Inc. is shown in Table 4. We can see in the Text column of the table that tweets are in pre-processed form. In this table, the sentiment attribute may have different values depending on the positivity or negativity of the tweet words. For example, the sentiment value of 0 means that the sentiment of the particular tweet is neutral, the sentiment value of 1 means that the sentiment of the particular tweet is positive, while the sentiment value of − 1 means that the sentiment of the tweet is negative. Attributes of the sentiment analysis data set are briefly described in Table 5.

4.2.4 Creating the final data set

After performing sentiment analysis of the social media data of Apple, Google, and Microsoft, the index of Twitter sentiments on a specific date is added as an attribute to the stock market historical data sets of these companies between Trend and Future Trend attributes (Table 3). This results in a total of 42 final data sets for each stock market for sentiment analysis: 21 for the JB approach and 21 for the SNLP approach. These are the final data sets that will be used to train the machine learning algorithms. In these datasets, Sentiment is the aggregated sentiment of all tweets posted on a specific date. For example, in Table 6, the Sentiment value 7 shows aggregated sentiment of all tweets posted on 1 July 2016 for Apple Inc. The aggregated value of the sentiment attribute on a particular date may be positive, negative, or zero, depending on the sentiment value of individual tweets. The table shows the structure of the final data set for the sentiment analysis system for Apple Inc.

4.3 Political situation analysis data set

4.3.1 Data collection process

The political situation analysis data set is created by collecting the most significant political events occurred in Pakistan between 28 May 1998 and 30 June 2017. This collection is done manually. Approximately 98 of the most significant political events occurred within this period are crawled from Wikipedia. The reason for selecting Wikipedia as a data source of political situations is that it has historical information of any country.

[5] http://www.nlp.stanford.edu/.

Table 4 Structure of the sentiment analysis data set for Apple Inc

Source	Text	Date	Sentiment
Fredjoneswest	Apple nasdaq aapl stock apple declare latest emoji will	Tue Jul 05 23:27:00 PKT 2016	0
Financeprnews	Apple ipad ban beyonce dublin concert croke park	Tue Jul 05 23:18:00 PKT 2016	0
Wlstcom	Dividend growth trump market apple aapl adp afl bax bdx	Tue Jul 05 23:17:00 PKT 2016	1
Instavest	Apple health app user sign organ donor	Tue Jul 05 23:04:00 PKT 2016	0
Fredjoneswest	Apple bring u organ donation registration iphone fall	Tue Jul 05 22:46:00 PKT 2016	− 1
Fredjoneswest	Apple seed second beta maco sierra developer	Tue Jul 05 22:24:00 PKT 2016	0

Table 5 Attributes of the sentiment analysis data set

Feature	Format	Description
Source	Text	Twitter ID of the user or organization which tweeted the tweet
Text	Text	Text of the tweet
Date	Date	The date and time of the tweet with respect to Pakistan Standard Time
Sentiment	Integer	The degree of positive or negative sentiment of an individual tweet

Table 6 Structure of the final data set of Apple Inc. for sentiment analysis-based machine learning system (JB approach)

Date	Open	High	Low	Close	Volume	Trend	Sentiment	Future trend
1/7/2016	95.489998	96.470001	95.330002	95.889999	93.720169	Positive	7	Negative
5/7/2016	95.389999	95.400002	94.459999	94.989998	92.840523	Negative	1	Positive
6/7/2016	94.599998	95.660004	94.370003	95.529999	93.368317	Positive	3	Positive
7/7/2016	95.699997	96.5	95.620003	95.940002	93.769043	Positive	− 10	Positive
8/7/2016	96.489998	96.889999	96.050003	96.68	94.492294	Positive	− 1	Positive
11/7/2016	96.75	97.650002	96.730003	96.980003	94.7855	Positive	1	Positive

4.3.2 Classification of events

Political events are classified into two categories: positive and negative. Events which brought political stability in the country and had positive impacts for the country are classified as positive, whereas events which lead to political instability are classified as negative. This classification is performed manually since little research has been done in the domain of automatic political situation analysis based on machine learning. The political situation analysis data set has three attributes: Date, News, and Situation. These attributes are briefly explained in Table 7. Here, the Situation attribute will be used to create the final data set for stock prediction using political situation analysis. Note that the text of news is in pre-processed form.

Attributes of the political situation analysis data set are described in Table 8.

4.3.3 Creating the final data set

After performing classification of events, the Situation attribute having classes of the events is added as Situation attribute between Trend and Future Trend in the stock market historical data set (Table 3), resulting in the final data set for stock prediction using political situation analysis. A snapshot of this final data set is shown in Table 9. The table shows the structure of the final data set for the political situation analysis system for the KSE.

4.4 Stock market interdependency data set

4.4.1 Data collection process

For the stock data interdependency system, the stock historical data sets from Yahoo! Finance for the companies and stock markets whose stock trend interdependency needed to be measured are collected for the same date ranges. The range of dates should be the same for all data

Table 7 Structure of the political situation analysis data set for KSE

Date	News	Situation
28/5/1998	Pakistan conduct nuclear test chagai hill balochistan	Positive
26/7/1999	Kargil war end pakistan india	Positive
12/10/1999	Nawaz sharif oust power house arrest attempt sack general pervez Musharraf	Negative
6/4/2000	Nawaz sharif sentence life imprisonment charge hijack terrorism	Negative
12/5/2000	Supreme court validate October coup grant general pervez musharraf executive legislative authority three year	Positive
21/6/2001	General pervez musharraf assume office president remain chief army staff	Positive

Table 8 Attributes of the political situations analysis data set

Feature	Format	Description
Date	Date	The date on which the political event occurred
News	Text	Text of the political event
Situation	Nominal	Category of the news based on its contents

sets whose stock trend interdependency is to be measured. For the stock interdependency system, stock market historical data sets (Table 3) of the selected stock markets and individual companies are used.

4.4.2 Feature engineering

Stock market historical data sets of the selected stock markets and companies are pre-processed by removing all attributes except the Trend attribute. As noted before, the Trend feature is the numeric difference between the Open and Close attributes of the Yahoo! Finance raw stock data file.

4.4.3 Creating the final data set

All the data sets having trend attribute are combined on the basis of common dates. Table 10 shows a view of the final data set for the stock markets interdependency system. The trends may be positive, negative, or even zero depending on the values of opening and closing price of the stocks. This data set will be used to find interdependency among stock market trends using the Pearson correlation. The abbreviations used are stock symbols of the stock markets.

5 Proposed system

The proposed system can be divided into three subsystems: (a) sentiment analysis, (b) political situation analysis, and (c) stock interdependency analysis. The following subsections describe these subsystems.

5.1 Sentiment analysis-based machine learning system

For this subsystem, the selection of raw data source, collection process, feature engineering, sentiment analysis, and creation of the final data sets are discussed in Sect. 4.2. Training, evaluation, and testing of the machine learning algorithms are given in the following subsections. Figure 3 provides a block diagram of the steps involved in the working of the machine learning system based on sentiment analysis for stock market trends prediction.

Table 9 Structure of the final data set for KSE for political situation analysis-based machine learning system

Date	Open	High	Low	Close	Volume	Trend	Situation	Future trend
28/5/1998	1102.969971	1106.709961	1040.189941	1040.189941	1040.189941	Negative	Positive	Negative
26/7/1999	1200.880005	1225.609985	1197.819946	1219.709961	1219.709961	Positive	Positive	Positive
12/10/1999	1249.790039	1257.939941	1243.97998	1256.939941	1256.939941	Positive	Negative	Negative
6/4/2000	1944.160034	1963.319946	1917.339966	1955.430054	1955.430054	Positive	Negative	Negative
12/5/2000	1723.670044	1727.930054	1651.550049	1672.459961	1672.459961	Negative	Positive	Negative
21/6/2001	1370.209961	1372.670044	1347.050049	1351.160034	1351.160034	Negative	Positive	Positive

Table 10 A view of the final data set for interdependency system

No.	Karachi (KSE)	London (LSE)	NASDAQ (NDAQ)	New York (NYA)
1	− 65.19043	NA	15.570068	79.93018
2	247.16016	− 4.605	− 10.189941	− 21.81982
3	113.64941	17.041	25.000000	17.85986
4	181.75000	2.303	5.179932	− 39.12988
5	37.00000	17.962	− 3.429932	− 20.58008
6	− 86.05957	− 23.489	13.729981	− 14.22022

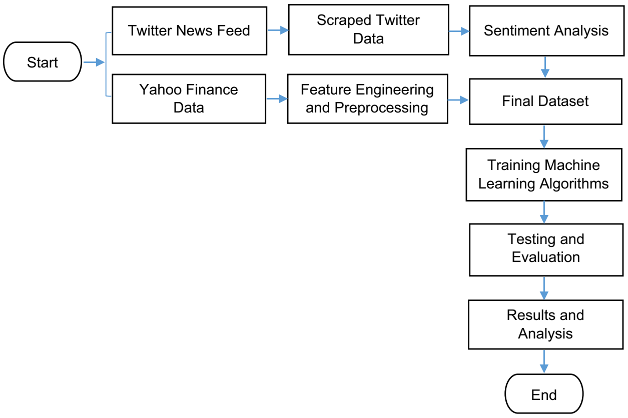


Fig. 3 Block diagram for the sentiment analysis-based machine learning system

5.1.1 Training the algorithms

The final data sets for sentiment analysis system are used to train machine learning algorithms. To do this, the authors use the WEKA machine learning library (Frank et al. 2016). Since the authors are predicting the future trend, which could be positive, negative, or neutral, that is, a nominal attribute, the authors use classification-based algorithms. The WEKA algorithms used are Naïve Bayes (NB), sequential minimal optimization (SMO), k-nearest neighbour or IBK, locally weighted learning (LWL), attribute selected classifier (ASC), PART (partial C4.5 decision tree), multilayer perceptron (MLP), J48, RF, Bagging, and decision table (DT). These algorithms automatically extract parameter values from data sets that are used to train the algorithms.

5.1.2 Performance evaluation of the algorithms

The performance of machine learning algorithms can be evaluated using different error metrics. The choice of an appropriate error metrics is very important for evaluating and comparing the performance of different algorithms to choose the best one. This selection of appropriate error metric depends on the type of the model, e.g. classification, regression, etc., and distribution of the output variable, e.g. uniform distribution. There are different error metrics used to evaluate the performance of machine learning models. Classification problems are the most common machine learning problems, and there is a myriad of metrics used to evaluate predictions for these problems. For example, area under the curve (AUC), classification accuracy, and confusion matrix are used to evaluate classification models. AUC error metric is used for evaluating binary class classification problems. Classification accuracy is a common evaluation metric for classification problems when the class distribution is uniform. When the class distribution is not uniform, a confusion matrix is used to evaluate the model. Similarly, mean squared error (MSE), root-mean-squared error (RMSE), mean absolute error (MAE), etc., are used to evaluate regression models.

Since the current problem is a multi-class classification problem and the class distribution is not uniform (due to a very small number of values for the neutral class), the authors have selected confusion matrix and its accuracy, precision, recall, and *F*-measure metrics for evaluating the selected machine learning algorithms. For the social media sentiment analysis system, the confusion matrix is a 3 × 3 matrix in which the values along the columns show the actual trends, the values along the rows show predicted trends, while the values along the diagonal represent correctly predicted trends. The values in the matrix are represented by some special terminologies. For example, true positive (TP) is used for correct positive trend, true neutral (TNT) for correct neutral, and true negative (TNG) for correct negative trend. Similarly, positive trend classified as neutral or negative (false positive) is represented by $\mathrm{FP}_{\mathrm{nt}}$ and $\mathrm{FP}_{\mathrm{ng}}$, respectively, neutral trend classified as positive or negative (false neutral) is represented by $\mathrm{FNT}_{\mathrm{p}}$ and $\mathrm{FNT}_{\mathrm{ng}}$, respectively, while negative trend classified as positive or neutral (false negative) is represented by $\mathrm{FNG}_{\mathrm{p}}$ and $\mathrm{FNG}_{\mathrm{nt}}$, respectively. From these values, accuracy can be calculated by the following formula:

$$\text{Accuracy} = \frac{\text{The number of correct predictions made}}{\text{The total number of predictions}}$$

Or using the terminologies, we can write accuracy formula as:

$$\text{Accuracy} = \frac{\text{TP} + \text{TNT} + \text{TNG}}{\text{TP} + \text{TNT} + \text{TNG} + \text{FPnt} + \text{FPng} + \text{FNTp} + \text{FNTng} + \text{FNGp} + \text{FNGnt}} \quad (1)$$

This will give performance of the classifier in terms of accuracy.

The confusion matrix also gives some other measures like precision, recall, *F*-measure, etc., for prediction problems in which the distribution of the output variable is not uniform. Precision shows the ability of the algorithm to classify records accurately and can be found by the following formula for the positive trend class:

$$\text{Precision}_p = \frac{\text{TP}}{\text{TP} + \text{FP}_{\text{nt}} + \text{FP}_{\text{ng}}} \quad (2)$$

Recall shows the ability of the algorithm to classify as many records as possible and is given by the following formula for the positive trend class:

$$\text{Recall}_p = \frac{\text{TP}}{\text{TP} + \text{FNT}_p + \text{FNG}_p} \quad (3)$$

F-measure is a description of both precision and recall and can be calculated by the following equation for the positive trend class:

$$F - \text{measure}_p = 2 \times \left(\text{Precision}_p \times \text{Recall}_p\right) / \left(\text{Precision}_p + \text{Recall}_p\right) \quad (4)$$

5.1.3 Testing the algorithms

There are different methods to validate a model in machine learning like holdout, substitution, and cross-validation (CV). The CV method comprises leave-one-out CV, *k*-fold CV, and leave-more-out CV (Chou and Lin 2012). WEKA provides different test options like percentage split, CV, etc. The authors use tenfold CV for model validation. 10 is the number of folds that was recommended by the researcher (Kohavi 1995). The authors use *k*-fold CV because it is widely used to check whether a model is overfit or not. Another reason is that it can be used with any modelling technique. It is also the most likely scenario where the authors can employ different evaluation metrics. Therefore, the selected machine learning algorithms are tested on the final data sets using tenfold CV. Accuracies of the algorithms are reported for analysis and comparison purposes.

5.2 Political situation analysis-based machine learning system

For this subsystem, the selection of political events data source, classification of events, and creation of the final data sets are discussed in Sects. 4.2 and 4.3. Training, evaluation, and testing of the machine learning algorithms are given in the following subsections. Figure 4 shows a block diagram of the political situation analysis-based stock market prediction system.

5.2.1 Training the algorithms

The process of training machine learning algorithms for the political analysis system is similar to that of the sentiment analysis. Since the authors are predicting the future trend, which could be positive, negative or neutral, that is, a nominal attribute, the authors use classification-based algorithms. The selected algorithms are trained on the final data sets for the political situation analysis system.

5.2.2 Performance evaluation of the algorithms

Since the problem of the political situation analysis is also a multi-class classification problem, accuracy, precision, recall, and *F*-measure metrics of the confusion matrix have also been used to evaluate the performance of the algorithms on the final data sets for the political situation analysis system. The confusion matrix will be a 3 × 3 matrix for this system.

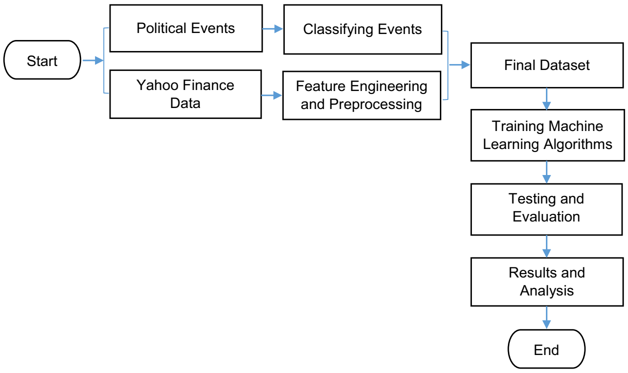


Fig. 4 Block diagram for the political situation analysis-based machine learning system

5.2.3 Testing the algorithms

The selected algorithms are tested on the final data sets for stock prediction using tenfold CV. Accuracies of the algorithms are noted for analysis and comparison purposes.

5.3 System for stock markets interdependency

The third and final problem of this study is to find interdependency among stock market trends of individual companies as well as overall stock markets. The purpose of studying this dependency is to analyse whether or not the stock trend of one company affects that of another. Pearson correlation coefficient is a common method for calculating the correlation coefficient. Correlation denotes the relationship between two variables and how they change with each other. A correlation of 1 means a strong correlation, − 1 shows a weak correlation, while 0 means no correlation.

For calculating correlation coefficients, a simple application is created in *R* programming language which takes the final data set as input, calculates its Pearson product correlation, and returns the interdependency results in graphical form. The basic flow of this system is presented in Fig. 5.

The data collection process, feature engineering, and creation of the final data set are discussed in Sect. 4.4. The correlation calculation step of the system is briefly explained in the following subsection.

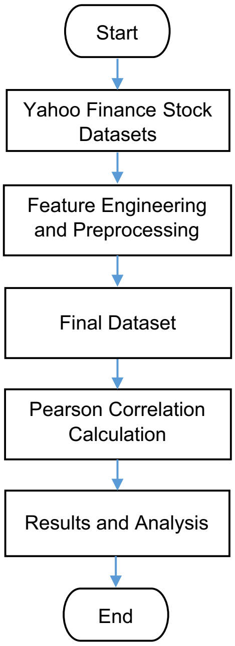

Start
Yahoo Finance Stock Datasets
Feature Engineering and Preprocessing
Final Dataset
Pearson Correlation Calculation
Results and Analysis
End

Fig. 5 Block diagram for stock markets interdependency system

5.3.1 Calculating Pearson correlation

The simplest way of studying interdependency between different variables is by calculating the Pearson correlation among the variables. The *R* programming language (R team 2018) contains a library called "Psych" (Revelle 2018), which contains a method called "pairs.panels". This method takes a data set as input and, in a graphical form, returns the interdependency among trend attributes of the data set. The graphs are kept for analysis and comparison purposes.

6 Results analysis and discussion

Results for the proposed subsystems are presented in this section. A comparative study is also given at the end of this section.

6.1 Results for stock prediction using sentiment analysis

To examine the influence of sentiment feature on stock market prediction, the selected machine learning algorithms are applied on Google, Microsoft, and Apple stock markets final data sets. Predictions are made for 7 days following the trading day for both of the sentiment analysis approaches. All the algorithms are trained and tested using the tenfold CV. Prediction accuracies of the selected machine learning algorithms are found and plotted for comparison purpose. Figure 6 shows the prediction accuracies of algorithms for the Microsoft stock market using the JB approach.

From the figure, it is evident that the highest prediction accuracy is achieved by NB, SMO, IBK, LWL, PART, J48, RF, and DT algorithms on the 7th day, while ASC and Bagging show maximum accuracy on day 6. Since most of

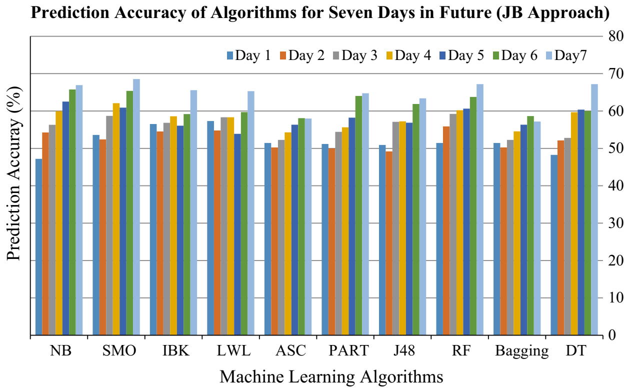


Fig. 6 Graphical representation of prediction accuracy of algorithms on Microsoft stock market final data sets (JB Approach)

the algorithms show maximum accuracy on the 7th day, it can be concluded that the sentiment attribute is more effective for stock market prediction on the 7th day after the trading day. The next higher accuracy is exhibited on day 6 by NB, SMO, LWL, PART, J48, and RF algorithms. A maximum accuracy of 68.56% is achieved by SMO and can be considered the best algorithm. The second-best accuracy of 67.2% is achieved by RF and DT and therefore can be considered second-best algorithms. The lowest accuracy is shown by most of the algorithms (SMO, IBK, ASC, PART, J48, and Bagging) on the 2nd day after the trading day. It means that the sentiment attribute is least effective on the 2nd day from the day on which the trade is executed. Furthermore, accuracies of the algorithms increase gradually from day 2 to day 7, which shows that the impact of sentiment attribute is less effective on day 2 and then gradually increases and becomes maximum on day 7. On average, ASC and Bagging show poor performance by achieving maximum accuracies of 58.1% and 58.64%, respectively, on day 6.

Since the authors want to compare the performance of the selected algorithms using sentiment attribute to assess the impact of social media on stock market prediction, the authors also found prediction accuracy of the best algorithms on stock historical data sets of Apple Inc., Google, and Microsoft without the sentiment attribute. A decrement in the range of 0–3% is observed in the prediction accuracy of best algorithms when no sentiment attribute is used. This is shown in Table 11 for the JB approach. For example, on the 7th day, SMO shows the highest prediction accuracy of 68.56% with the sentiment attribute and 67.75% without including sentiment attribute in Microsoft stock final data set, a 0.85% decrease in accuracy. Similarly, the accuracy of LWL decreases by 1.35% and that of RF by 1.28%. A similar drop in prediction accuracy of the second-best algorithms except DT is also observed when no sentiment attribute is used. Only DT shows no change in its accuracy on both types of the data sets.

This means that the sentiment attribute has some effects on the prediction accuracy of algorithms. A comparison of accuracies of the best algorithms with and without sentiment attribute is shown in Fig. 7. From the figure, it is evident that all of the best algorithms show improved performance over the sentiment analysis final data sets of Apple, Google, and Microsoft.

Table 11 Summary of prediction accuracy of best and second-best algorithms on Apple, Google, and Microsoft final data sets with JB approach with and without sentiment attribute

Stock	Best algorithm/without sentiment	Second-best algorithm/ without sentiment
Apple	LWL (67.20/65.85)	IBK (67.27/65.58)
Google	RF (63.98/62.7)	DT (63.34/63.34)
Microsoft	SMO (68.56/67.75)	RF (67.20/66.66)

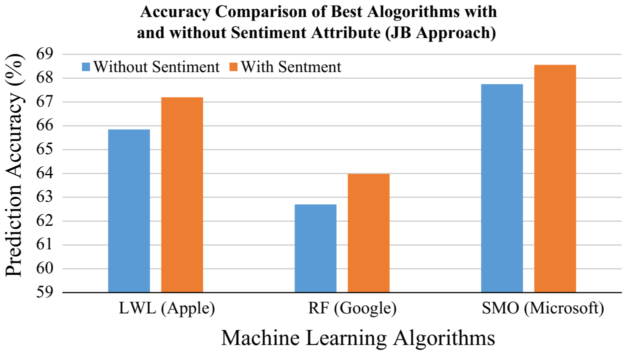


Fig. 7 Prediction accuracy comparison of the best algorithms on selected stock markets with and without the sentiment attribute

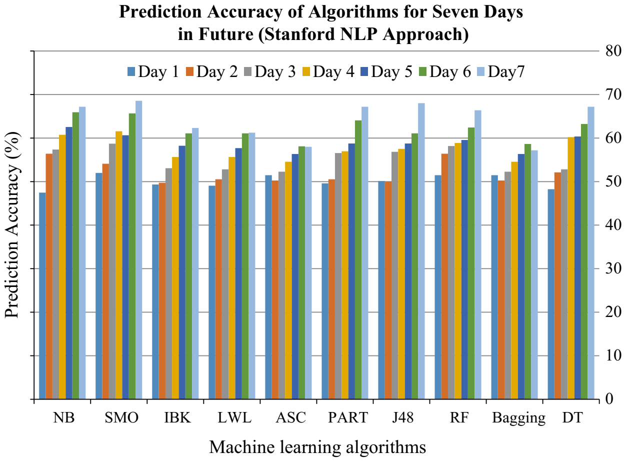


Fig. 8 Graphical representation of prediction accuracy of algorithms on Microsoft stock market final data sets (SNLP Approach)

Prediction accuracies of the algorithms on stock market final data sets for sentiment analysis using the SNLP approach are also found and plotted for comparison purpose. Figure 8 shows the prediction accuracies of algorithms for the Microsoft stock market.

From the figure, it is evident that the highest prediction accuracy is achieved on day 7 by NB, SMO, IBK, LWL, PART, J48, RF, and DT algorithms, while ASC and Bagging show the highest accuracy on day 6. Since most of the algorithms show maximum accuracy on the 7th day, it can also be concluded from the results of the SNLP approach that the sentiment attribute is more effective on the 7th day for stock market prediction. The next higher accuracy is exhibited on day 6 by NB, SMO, IBK, LWL, PART, J48, RF, and DT algorithms.

A maximum accuracy of 68.56% is achieved by SMO and thus can be considered the best algorithm for stock market prediction using the SNLP sentiment analysis approach. The second-best accuracy of 68.02% is achieved by J48. The lowest accuracy is shown by all algorithms except ASC and Bagging on day 1 after the prediction day. Furthermore, accuracies of the algorithms increase gradually from day 2 to day 7, which show that the impact of sentiment attribute is less effective on day 2 and then gradually increases and becomes maximum on day 7th. On average, ASC and Bagging also showed poor performance in this approach.

The authors also found the prediction accuracy of the best algorithms on stock historical data sets of Apple Inc., Google, and Microsoft without the sentiment attribute. A decrement of 2% is observed in the prediction accuracy of the best algorithms when no sentiment attribute is used. Accuracy of only LWL best algorithm decreased on Apple Inc. stock final data set with sentiment attribute. Table 12 summarizes the highest accuracy of the best and second-best algorithms for all stock markets using the SNLP approach. These results are quite similar to those obtained using the JB approach.

While comparing the accuracy of the second-best algorithms with and without sentiment attribute, no change is observed in the performance of LWL and DT algorithms.

This means that the sentiment attribute has some effects on the prediction accuracy of algorithms using the SNLP approach as well. A comparison of accuracies of best algorithms with and without sentiment attribute is shown in Fig. 9 using the SNLP approach. From the figure, it is evident that algorithms show improved performance over the sentiment analysis final data sets of Google and Microsoft using the SNLP approach. However, a decrement in accuracy is shown by LWL on the Apple stock final data set with sentiment attribute.

From the results obtained using the SNLP approach, the authors can conclude that the sentiment improves the efficiency of the algorithms by approximately 2%. Using the SNLP approach, the difference in prediction accuracy with and without sentiment attribute can be observed. From this difference, the authors can again conclude that the prediction accuracy of the algorithms improves with the sentiment attribute.

Table 12 Summary of prediction accuracy of best and second-best algorithms on Apple, Google, and Microsoft final data sets using SNLP approach with and without the sentiment attribute

Stock	Best algorithm/without sentiment	Second-best algorithm/ without sentiment
Apple	LWL (65.85/66.12)	LWL (63.24/63.24)
Google	RF (63.34/63.02)	DT (63.34/63.34)
Microsoft	SMO (68.56/67.75)	J48 (68.02/66.12)

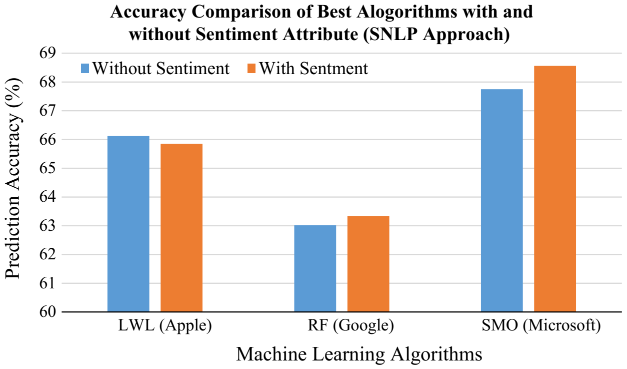


Fig. 9 Prediction accuracy comparison of best algorithms on selected stock markets with and without the sentiment attribute

From the experiment results of both sentiment analysis approaches, the authors can conclude that similar results can be achieved using any sentiment analysis approach.

To quantify the performance of the selected algorithms in differentiating among three trend classes, a confusion matrix is created. Three measures, namely precision, recall, and *F*-measure, are found for each trend class as shown in Table 13. The table obviously demonstrates that precision, recall, and *F*-measure values of each algorithm give a large deviation in results (precision = 0–75.0%, recall = 0–100.0%, *F*-measure = 0–74.5%). For the positive trend class, the highest precision (66.9%) is achieved by IBK, while the highest recall (100%) is shown by Bagging, but overall best performance is shown by the LWL classifier. (Precision, recall, and *F*-measure are 59.6, 99.3, and 74.5%, respectively.) Precision is above 59.0% for all classifiers. From the table, it is clear that the critical problem is the poor classification rate of the neutral trend class because, for the neutral trend class, all the classifiers show zero precision, recall, and *F*-measure. This may be due to the small number of records (one out of 251 for Microsoft) in this class. For the negative trend class, the highest precision (75.0%), recall (50.5%), and *F*-measure (50.0%) are achieved by LWL, NB, and IBK, respectively, while Bagging shows poor performance. (Precision, recall, and *F*-measure are 0%.) Classifiers show relatively low precision (0–75.0%), recall (0–50.5%), and *F*-measure (0–50.0%) for the negative trend class. This may be due to a relatively small number of records in the negative class (99 out of 251 for Microsoft). From the comparison of the results, it can be concluded that LWL shows relatively good performance in classifying both positive and negative trend classes.

From the above results analysis and discussion on the impact of public sentiment attribute for stock market prediction, the authors can deduce the following research findings:

Table 13 Classifiers classification performance in terms of precision, recall, and *F*-measure over the final data set of Microsoft

Class	Metrics	Classifiers									
		NB	SMO	IBK	LWL	ASC	PART	J48	RF	Bagging	DT
Positive	Precision (%)	63.4	63.3	**66.9**	59.6	61.4	65.9	59.2	61.8	59.0	59.7
	Recall (%)	59.0	74.3	75.7	99.3	89.6	75.0	86.8	87.5	**100.0**	81.3
	F-measure (%)	61.2	68.4	71.0	**74.5**	72.9	70.1	70.4	72.4	74.2	68.8
Neutral	Precision (%)	0	0	0	0	0	0	0	0	0	0
	Recall (%)	0	0	0	0	0	0	0	0	0	0
	F-measure (%)	0	0	0	0	0	0	0	0	0	0
Negative	Precision (%)	45.5	50.7	55.6	**75.0**	55.9	53.8	42.4	55.0	0	43.8
	Recall (%)	**50.5**	38.4	45.5	3.0	19.2	43.4	14.1	22.2	0	21.2
	F-measure (%)	47.8	43.7	**50.0**	5.8	28.6	48.0	21.2	31.7	0	28.6

The bold numbers show the highest precision, recall, or *F*-measure for the positive, neutral, or negative classes in the sentiment analysis system

(1) The sentiments of people towards a particular company have some impact on stock prices. Using sentiment attribute improves the prediction accuracy of machine learning algorithms by 0–3% for stock market prediction.
(2) Stock market trends can be predicted with the highest accuracy on day 7 from the day of the initial prediction.
(3) The sentiment of people is least effective on days 1 and 2 from the day of the initial prediction.
(4) Both sentiment analysis approaches produce similar prediction results.
(5) ASC and Bagging show poor performance for both of the sentiment analysis approaches.
(6) SMO shows the best performance for both of the sentiment analysis approaches and thus can be considered the best algorithm for stock prediction using sentiment analysis.
(7) LWL shows relatively good performance in classifying positive and negative trend classes.
(8) There is no single algorithm that can predict stock trends of all companies with the highest accuracy.

6.2 Results for stock prediction using political situation analysis

To analyse the impact of political situation feature on the stock market prediction of KSE, the selected machine learning algorithms are applied on a total of seven political analysis final data sets of the KSE stock market. The impact of a total of 98 political events is studied on the KSE stock market. Prediction accuracies are found for 7 days following the day on which the event occurred. All the algorithms are trained and tested using the tenfold CV. Accuracies of the selected machine learning algorithms are found and plotted for comparison purposes. Table 14 shows the prediction accuracies of the algorithms for 7 days. The accuracies are plotted graphically in Fig. 10.

From the figure, it is evident that most of the algorithms (SMO, IBK, J48, and RF) show good performance on the 7th day, but the highest accuracy of 75.38% is shown by MLP and DT algorithms on day 5. Since the highest accuracy is achieved on day 5, it can be concluded that the political situation is more effective for stock market prediction on the 5th day. MLP and DT algorithms can be considered best due to their highest accuracy of 75.38%. The lowest accuracy is shown by most of the algorithms (SMO, IBK, LWL, MLP, J48, and DT) on 1st day after the day on which the event occurred. It means that the political situation is least effective on the 1st day from the day on which the event occurred. On average, ASC and Bagging algorithms show poor performance by achieving maximum accuracies of 63.07% on day 3 and 66.15% on day 4, respectively.

The authors also found the prediction accuracy of the best algorithms on stock historical final data sets of KSE without the political situation attribute to compare and analyse the performance of the algorithms. A decrement of about 20% is observed in the prediction accuracy of the best algorithms when no political situation attribute is used. For example, the accuracy of NB decreases by 1.54%, SMO by 9.23%, IBK by 20%, MLP by 9.23%, and DT by 21.54% which are significant drops in accuracies of these algorithms. This means that the political situation attribute has a significant effect on the prediction accuracy of algorithms. A comparison of accuracies of best algorithms with and without the political situation attribute is shown in Fig. 11. From the figure, it is evident that all the algorithms show improved performance over the political situation analysis final data set of KSE.

To quantify the performance of the selected algorithms in differentiating among three trend classes, a confusion

Table 14 Prediction accuracy of algorithms on KSE stock data (political situation analysis approach)

Classifiers	Day 1	Day 2	Day 3	Day 4	Day 5	Day 6	Day 7
NB	63.07	61.53	67.69	66.15	**69.23/67.69**	61.53	64.61
SMO	50	56.92	69.23	72.30	70.76	67.69	**73.84/64.61**
IBK ($K = 5$)	56.92	66.15	66.15	61.53	60	60	**70.76/50.76**
LWL	50.76	55.38	70.76	61.53	67.69	60	61.38
ASC	60	50.76	63.07	60	56.92	56.92	58.46
MLP	50.76	58.46	58.46	73.84	**75.38/66.15**	69.23	70.76
J48	55.38	58.46	64.61	58.46	72.30	67.69	74.38
RF	60	61	63.07	66.15	58.46	58.46	69.23
Bagging	60	53.84	63.07	66.15	61.53	61.53	58.46
DT	53.84	60	60	64.61	**75.38/53.84**	69.23	70.76

The bold numbers show accuracy of top 5 best algorithms with and without political situation feature

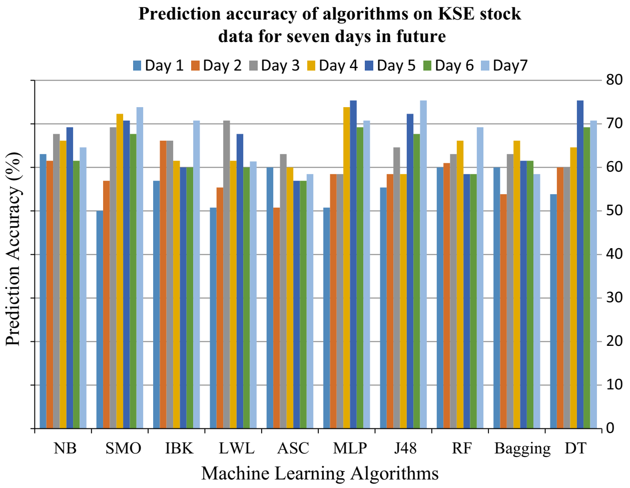


Fig. 10 Graphical representation of accuracy of algorithms on KSE stock market final data sets

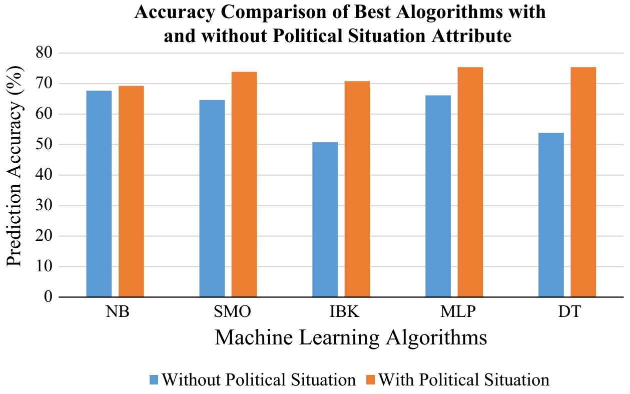


Fig. 11 Prediction accuracy comparison of best algorithms on KSE stock market with and without political situation attribute

matrix is also created for the stock prediction system using political situation analysis. Three measures, namely precision, recall, and *F*-measure, are found for each trend class as shown in Table 15. The table obviously demonstrates that precision, recall, and *F*-measure of each algorithm for the positive and negative trend classes give a large deviation in results (precision = 0–67.6%, recall = 0–100.0%, *F*-measure = 0–76.5%). For the positive trend class, the highest precision (67.6%) is achieved by IBK, while the highest recall (100%) and *F*-measure (76.5%) are shown by ASC, RF, Bagging, and DT, respectively. Overall best performance is shown by ASC, RF, Bagging, and DT classifiers. (Precision, recall, and *F*-measure are 61.9, 100.0, and 76.5%, respectively.) These algorithms show the same performance for classifying positive trend class. Precision is above 55.8%, while recall is above 61.5% for all classifiers for the positive trend class. From the table, it clear that the critical issue is the unavailability of precision, recall, and *F*-measure for the neutral trend class. It is represented by NA in the table. This may be due to the lack of records in this class. For the negative trend class, the highest precision, recall, and *F*-measure (46.2, 50.0, and 48.0%, respectively) are achieved by the IBK classifier. All classifiers except NB and IBK show poor performance for the negative trend class by achieving precision, recall, and *F*-measure of 0%. This is another critical issue that may be due to one-third of records in this class. Classifiers show relatively low precision (0–46.2%), recall (0–50.0%), and *F*-measure (0–48.0%) for the negative trend class. This may also be due to one-third records in this class. From the comparison of the results, it can be concluded that IBK shows relatively good performance in classifying the positive and negative trend classes.

From the above results analysis and discussion on the impact of political situation attribute for stock market prediction of KSE, the authors can deduce the following research findings:

(1) Two algorithms, MLP and DT, exhibited the highest accuracy of 75.38%.

Table 15 Classifiers classification performance in terms of precision, recall, and *F*-measure over the final data set of KSE

Class	Metrics	Classifiers									
		NB	SMO	IBK	LWL	ASC	MLP	J48	RF	Bagging	DT
Positive	Precision (%)	55.8	61.3	**67.6**	61.3	61.9	59.3	60.0	61.9	61.9	61.9
	Recall (%)	61.5	97.4	64.1	97.4	**100.0**	89.7	92.3	**100.0**	**100.0**	**100.0**
	F-measure (%)	58.5	75.2	65.8	75.2	**76.5**	71.4	72.7	**76.5**	**76.5**	**76.5**
Neutral	Precision (%)	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
	Recall (%)	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
	F-measure (%)	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
Negative	Precision (%)	25.0	0	**46.2**	0	0	0	0	0	0	0
	Recall (%)	20.8	0	**50.0**	0	0	0	0	0	0	0
	F-measure (%)	22.7	0	**48.0**	0	0	0	0	0	0	0

The bold numbers show the highest precision, recall, or *F*-measure for the positive, neutral, or negative classes in the political situation analysis system

(2) This highest prediction accuracy is achieved on day 5. This shows that if the impact of political events is included as a feature in stock market historical data, best predictions can be made on day 5. In other words, the impact of a political situation is more effective on the 5th day from the date on which the event occurred.

(3) The lowest accuracy is achieved by most of the classifiers on day 1; it means that political events are least effective on day 1 for stock market prediction.

(4) It can be inferred that political situations have a strong effect (about 20%) on stock prices of the overall stock market of KSE.

(5) IBK has shown relatively good performance in classifying positive and negative trend classes.

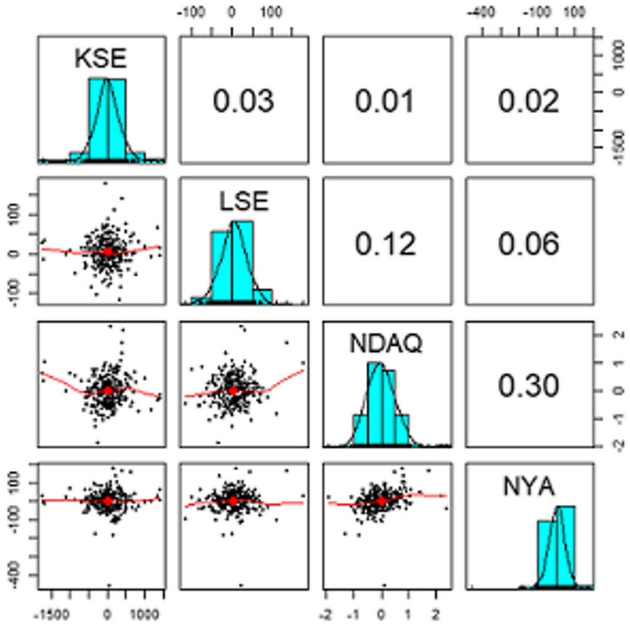


Fig. 12 Scatter plot matrix showing pairwise relationship in lower panel, corresponding Pearson correlation coefficients in the upper panel, while frequency distribution on the diagonal for pairwise relationship between stock market trends

6.3 Results for study of interdependency between stock trends of companies and stock markets

Interdependency among stock markets and individual companies are described in a scatter plot matrix. The output of the *R* Language's "pairs.panels" method is a scatter plot matrix as shown in Fig. 12 for the selected stock markets. There are three panels of this matrix. The lower panel shows the pairwise relationship between stock market trends. The corresponding Pearson correlation coefficients are given in the upper panel, while the diagonal demonstrates marginal frequency distribution for each stock market trend parameter. The marginal frequencies show that NASDAQ trends have Gaussian distribution, while trends of the other stock markets have near Gaussian distribution. The correlation coefficient value of 0.30 in the upper panel shows that trends parameters of NASDAQ and New York stock markets are more correlated, and therefore, these stock markets are more dependent on each other. Similarly, the smallest correlation coefficients of KSE with LSE, NASDAQ, and New York stock markets demonstrate that KSE is least dependent on these stock markets. This may be due to the fact that KSE is situated in a different region (Asia), while LSE is situated in Europe and NASDAQ and New York stock markets in North America and are therefore more dependent on each other compared to KSE stock market.

The correlation or interdependency results of stocks of Twitter, Facebook, and LinkedIn are also shown in a scatter plot matrix as shown in Fig. 13. The lower panel of the matrix shows the pairwise relationship between stock

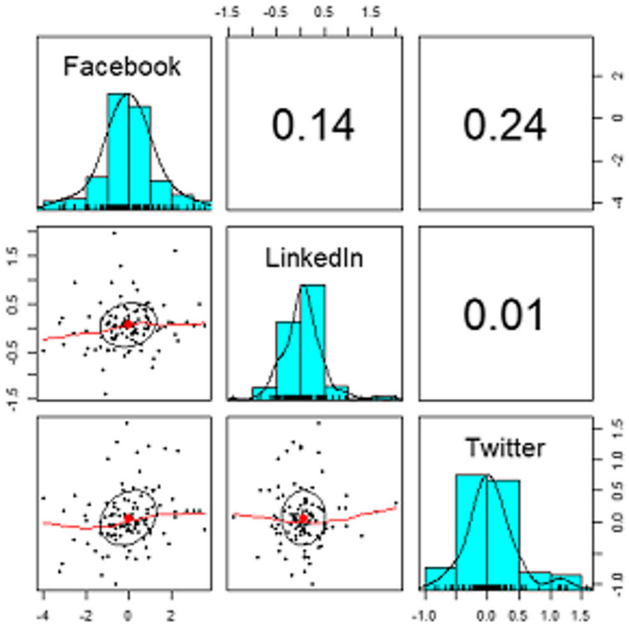


Fig. 13 Scatter plot matrix showing pairwise relationship in the lower panel, corresponding Pearson correlation coefficients in the upper panel, while frequency distribution on the diagonal for the pairwise relationship between stock market trends of social media companies

market trends. The corresponding Pearson correlation coefficients are given in the upper panel, while the diagonal demonstrates marginal frequency distribution for each stock market trend parameter. The marginal frequencies show that all stock markets have Gaussian distribution. The correlation coefficient value of 0.24 in the upper panel shows that the trend parameters of Facebook and Twitter stock markets are more correlated. Therefore, these stock markets are more dependent on each other. Similarly, the correlation coefficient value of 0.01 demonstrates that LinkedIn and Twitter are least correlated and therefore least dependent on each other.

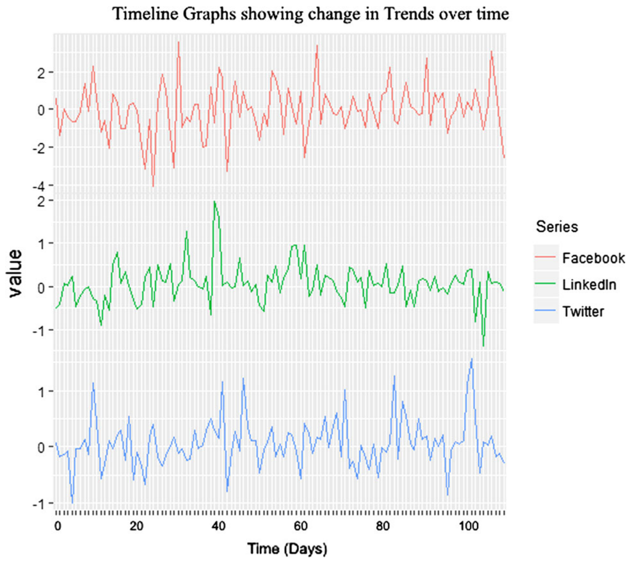


Fig. 14 Timeline graphs showing the change in the trend of social media platforms over time

The timeline graphs are also given for these three social media companies as shown in Fig. 14. The graphs show a change in trend over time.

The spikes in Fig. 14 indicate when the stock market trends fall and when they rise. It can be seen that at a certain time, the spikes of Facebook and Twitter trends rise and fall at the same time, pointing to their interdependency, as compared to LinkedIn. It is possible that there might be some other variable which affects the stock prices of these companies, but that is not our concern. Our concern is to determine whether some relation exists among stock prices of companies belonging to one category.

Similar interdependency results are also found for stock markets belonging to other industries. From these results, it can be concluded that stock markets belonging to the same industry show a positive correlation with each other.

6.4 Comparison with existing methods

Many research studies have been conducted for stock market prediction. To investigate the accuracy of the proposed machine learning model and other such models, a comparative study is presented in this section. This comparison is performed with respect to the accuracy, and the data and methodology used.

By comparing our proposed model for stock prediction using social media, we found that our model performed best in terms of accuracy by achieving the highest accuracy of 68.56% by SMO. The previous models achieved different accuracies which were less than the maximum accuracy achieved by our model. For example, Zhou et al. (2016) model achieved 64.15% accuracy, Nguyen et al. (2015) model showed an average accuracy of 54.41%, while Bing et al. (2014) model achieved an average accuracy of 66.48% on the third day after the day on which the trade occurred. Furthermore, our proposed model improved the prediction accuracy from 0 to 3%, while Nguyen et al. (2015) model showed 2.07% better accuracy which also shows that our model is consistent with the previous research. From these accuracies, we can conclude that our proposed model performed best in terms of prediction accuracy. As opposed to our proposed model which uses overall sentiment of the public towards a particular stock market or company, the existing models used online emotions (Zhou et al. 2016; Ahuja et al. 2015), modes of specific topics of companies (Bollen et al. 2011; Nguyen et al. 2015), and social, political, cultural, natural, and economic events (Bollen et al. 2011). In terms of methodology, the existing models used maximum of up to two machine learning algorithms, e.g. Zhou et al. (2016)

used K-means discretization and SVM, Ahuja et al. (2015) used self-organizing fuzzy neural network, Nguyen et al. (2015) used only SVM, while Bing et al. (2014) applied association rule mining algorithm for stock market prediction. The previous models used only accuracy error metric for evaluating the algorithms, while in the current study, in addition to accuracy error metric, selected algorithms have also been evaluated using confusion matrix.

By comparing our proposed model for stock prediction using political situations analysis, we found that our model performed best in terms of improving accuracy by about 20% compared to existing models. For example, Lee et al. (2014) observed a 10% increase in accuracy when RF classifier was applied to financial events data. As opposed to our proposed model which analyses the impact of political events on stock market forecasting using machine learning algorithms, most of these models though used political events, but they did not use machine learning algorithms. They used different methods, e.g. mathematical models (Taimur and Khan 2015), univariate asymmetric EGARCH model (Suleman 2012), statistical methods (Malik et al. 2009), market and risk-adjusted techniques (Chen et al. 2005), and GARCH (Beaulieu et al. 2005). In terms of methodology, the existing models used no machine learning algorithms. Only Lee et al. (2014) used RF on financial events data and achieved a 10% increase in accuracy which also shows that our model is consistent with previous research for stock market prediction. These models used only accuracy error metric for evaluating the algorithms, while in the current study, in addition to accuracy error metric, selected algorithms have also been evaluated using confusion matrix.

From the comparison, we can conclude that our proposed models for stock market prediction using social media and political events performed best and are unique in terms of the data and methodology used.

7 Conclusion

In this paper, the authors examined the impact of public sentiment and political situations on the prediction accuracy of machine learning algorithms for different stock markets. The authors also found interdependencies among different stocks. By including the sentiment attribute, the authors found that the impact of public sentiment on a particular company is minimal. There are two possible reasons for this: Either sentiment has no impact or the sentiment analysis techniques which are currently existing cannot provide much information about the correlation between sentiment and a company's stock. The authors found that the impact ranged from 0 to 3%. The sentiment attribute also enhanced the prediction accuracy of machine learning algorithms by approximately 2%. Also, stock market trends can be forecasted with the highest accuracy on the 7th day from the day on which the trade is executed.

A very important conclusion from this research is the impact of a country's political events on its stock market; this impact ranges from 10 to 20%. The authors tested their algorithms on 98 political events, and the improvement in prediction accuracy was astonishing. Two algorithms, MLP and DT, shared the highest accuracy of 75.38%. The highest prediction accuracies showed that the impact of a political situation is most effective on the 5th day from the date of its occurrence.

In terms of applied research, the JB and the SNLP approaches produced similar prediction results. There was no single algorithm that could predict the stock prices of all the companies with the highest accuracy using sentiment analysis. However, in political situation analysis, the authors found that including political situations improved the prediction accuracy of MLP and DT by approximately 20%.

In terms of accuracy, the current research improved the accuracy of the Schumaker and Chen method (Schumaker and Chen 2009), which was originally 57.1% for different representations of stock market data. The proposed algorithm predicted results with an accuracy of up to 68%, and that, too, after seven working days. In the political context, the current paper took advantage of the correlation found by Taimur and Khan (2015) and developed a practical machine learning model which yielded an accuracy of 75.38%, something never achieved before. Finally, the authors also proved that positive correlations exist between the trends of companies that belong to a specific industry. Furthermore, a comparative study between the current model and the previous models has been conducted. From the comparison, we investigated that the proposed model performed better than the existing models.

The developed model can be recommended for investors and financial analysts. The prediction model can help them foresee the future behaviour of stocks and decide which stocks to buy and which to sell. The model can also be used in decision-making systems. For example, in a clinical decision support system, the model can be used to classify the medical records of patients into critical and noncritical disease classes. The model can also be used in business and management to identify negative and positive trends and for better allocation of business resources.

For further research, the following improvements are suggested:

(1) The sentiment lexicons and sentiment analysis techniques that currently exist cannot predict stock markets with much accuracy. There is a need for a stock-specific lexicon that can help retrieve more

relevant information from texts in a more efficient manner. Rule-based sentiment analysis with a domain-specific lexicon would result in increased quality and accuracy of results.

(2) A machine learning model should be developed which can automatically calculate a political situation index for a particular day by reading news articles from different sources on that day. This is a complete information retrieval research study in itself.

(3) In this study, the authors only found correlations between stocks of different companies. The authors did not develop a machine learning model on the basis of correlation, as the authors did for political situation analysis. A machine learning model should be built for the correlation analysis of different stocks.

(4) A more advanced real-time prediction model can be developed which can produce prediction results more accurately.

Acknowledgements The authors would like to thank all the reviewers for their constructive comments.

Compliance with ethical standards

Conflict of interest Authors declare that he/she has no conflict of interest.

Ethical approval This article does not contain any studies with human participants or animals performed by any of the authors.

References

Ahuja R, Rastogi H, Choudhuri A, Garg B (2015) Stock market forecast using sentiment analysis. In: IEEE 2nd international conference on computers for sustainable global development, pp 1008–1010

Beaulieu MC, Cosset JC, Essaddam N (2005) The impact of political risk on the volatility of stock returns: the case of Canada. J Int Bus Stud 36(6):701–718

Billsus D, Pazzani MJ (2000) User modelling for adaptive news access. User Modell User Adapt Interact 10(2–3):147–180

Bing L, Chan KC, Ou C (2014) Public sentiment analysis in Twitter data for prediction of a company's stock price movements. In: 2014 IEEE 11th international conference on e-Business engineering (ICEBE), pp 232–239

Bollen J, Mao H, Pepe A (2011) Modeling public mood and emotion: twitter sentiment and socio-economic phenomena. In: 5th international AAAI conference on weblogs and social media

Bollerslev T (1986) Generalized autoregressive conditional heteroscedasticity. J Econ China 31(3):307–327

Cardoso B, Almeida R, Dias M, Coelho G (2008) Structural reliability analysis using Monte Carlo simulation and neural networks. Adv Eng Soft 39(6):505–513

Chen DH, Bin FS, Chen CD (2005) The impacts of political events on foreign institutional investors and stock returns: emerging market evidence from Taiwan. Int J Bus 10(2)

Chou J, Lin C (2012) Predicting disputes in public-private partnership projects: classification and ensemble models. J Comput Civil Eng 27(1):51–60

Chouliaras A (2015) High frequency newswire textual sentiment analysis: evidence from international stock markets during the European financial crisis. Available at SSRN 2572597

Deng W, Zhao H, Zou L, Li G, Yang X, Wu D (2017a) A novel collaborative optimization algorithm in solving complex optimization problems. J Soft Comput 21(15):4387–4398. https://doi.org/10.1007/s00500-016-2071-8

Deng W, Zhao H, Yang X, Xiong J, Sun M, Li B (2017b) Study on an improved adaptive PSO algorithm for solving multi-objective gate assignment. Appl Soft Comput 59:288–302. https://doi.org/10.1016/j.asoc.2017.06.004

Deng W, Xu J, Zhao H (2019a) An improved ant colony optimization algorithm based on hybrid strategies for scheduling problem. IEEE Access 7:20281–20292. https://doi.org/10.1109/ACCESS.2019.2897580

Deng W, Yao R, Zhao H, Yang X, Li G (2019b) A novel intelligent diagnosis method using optimal LS-SVM with improved PSO algorithm. J Soft Comput 23(7):2445–2462. https://doi.org/10.1007/s00500-017-2940-9

Dey A, Miyani G, Sil A (2019) Application of artificial neural network (ANN) for estimating reliable service life of reinforced concrete (RC) structure bookkeeping factors responsible for deterioration mechanism. Soft Comput. https://doi.org/10.1007/s00500-019-04042-y

Egeli B., Badur B, Ozturan M, Badur B (2003) Stock market prediction using artificial neural networks, In: Proceedings of the 3rd Hawaii international conference on business, pp 1–8

Fan Y, Ying SJ, Wang BH, Wei YM (2009) The effect of investor psychology on the complexity of stock market: an analysis based on cellular automaton model. Comput Ind Eng 56(1):63–69. https://doi.org/10.1016/j.cie.2008.03.015

Frank E, Hall MA, Witten IH (2016) The WEKA workbench. Online appendix for data mining: practical machine learning tools and techniques, 4th edn. Morgan Kaufmann, Burlington

Gidofalvi G, Elkan C (2001) Using news articles to predict stock price movements. Department of Computer Science and Engineering, University of California, San Diego

Go A, Bhayani R, Huang L (2009) Twitter sentiment classification using distant supervision. CS224 N project report, Stanford 1(1)

Hagenau M, Liebmann M, Neumann D (2013) Automated news reading: stock price prediction based on financial news using context-capturing features. Decis Support Syst 55(3):685–697

Hegazy O, Soliman OS, Salam MA (2014) A machine learning model for stock market prediction. Int J Comput Sci Telecom 4(12):16–23

Jeffrey B (2011) Twitter text mining. https://www.jeffreybreen.wordpress.com/2011/07/04/twitter-text-mining-r-slides/. Accessed 23 June 2018

Kara Y, Boyacioglu MA, Baykan OK (2011) Predicting direction of stock price index movement using artificial neural networks and support vector machines: the sample of the Istanbul stock exchange. Exp Syst Appl 38(5):5311–5319

Kazem A, Sharifi E, Hussain FK, Saberi M, Hussain OK (2013) Support vector regression with chaos-based firefly algorithm for stock market price forecasting. Appl Soft Comput 13(2):947–958

Kohavi R (1995) A study of cross-validation and bootstrap for accuracy estimation and model selection. Ijcai 14(2):1137–1145

Lakshmi V, Harika K, Bavishya H, Sri Harsha C (2017) Sentiment analysis of twitter data. Int Res J Eng Technol 4(2):2224–2227

Lee H, Surdeanu M, MacCartney B, Jurafsky D (2014) On the importance of text analysis for stock price prediction. LREC 2014:1170–1175

Li Q, Wang T, Li P, Liu L, Gong Q, Chen Y (2014a) The effect of news and public mood on stock movements. Inf Sci 278:826–840

Li X, Xie H, Chen L, Wang J, Deng X (2014b) News impact on stock price return via sentiment analysis. Knowl Based Syst 69:14–23

Lu N (2016) A machine learning approach to automated trading. M.S. thesis. Department of Computer Science, Boston College, Boston, USA

Makrehchi M, Shah S, Liao W (2013) Stock prediction using event-based sentiment analysis. In: 2013 IEEE/WIC/ACM international joint conference on WI and IAT 1, pp 337–342

Malik S, Hussain S, Ahmed S (2009) Impact of political event on trading volume and stock returns: the case of KSE. Int Rev Bus Res Papers 5(4):354–364

Mostafa MM (2010) Forecasting stock exchange movements using neural networks: empirical evidence from Kuwait. Exp Syst Appl 37(9):6302–6309

Murphy JJ (1999) Technical analysis of the financial markets: a comprehensive guide to trading methods and applications. Penguin, London

Naderpour H, Mirrashid M (2019) Shear failure capacity prediction of concrete beam–column joints in terms of ANFIS and GMDH. Pract Period Struct Des Const 24(2):04019006. https://doi.org/10.1061/(ASCE)SC.1943-5576.0000417

Naderpour H, Mirrashid M, Nagai K (2019) An innovative approach for bond strength modelling in FRP strip-to-concrete joints using adaptive neuro–fuzzy inference system. Eng Comput. https://doi.org/10.1007/s00366-019-00751-y

Nguyen TH, Shirai K, Velcin J (2015) Sentiment analysis on social media for stock movement prediction. Exp Syst Appl 42(24):9603–9611

Oh C, Chong, Sheng O (2011) Investigating predictive power of stock micro blog sentiment in forecasting future stock price directional movement. In: ICIS, Shanghai, China

Olaniyi SAS, Adewole KS, Jimoh RG (2011) Stock trend prediction using regression analysis–a data mining approach. ARPN J Syst Softw 1(4):154–157

Oliveira N, Cortez P, Areal N (2013) On the predictability of stock market behavior using stocktwits sentiment and posting volume. Portuguese conference on artificial intelligence. Springer, Berlin, pp 355–365

Ou P, Wang H (2009) Prediction of stock market index movement by ten data mining techniques. Mod Appl Sci 3(12):28

Pak A, Paroubek P (2010) Twitter as a corpus for sentiment analysis and opinion mining. LREc 10(2010):1320–1326

Papadrakakis M, Lagaros D (2002) Reliability-based structural optimization using neural networks and Monte Carlo simulation. Comput Methods Appl Mech Eng 191(32):3491–3507

R Core Team (2018) R: a language and environment for statistical computing. R foundation for statistical computing, Vienna, Austria. https://www.R-project.org/ Accessed 7 May 2018

Rahman A, Ali A (2016) Sentiment analysis on Twitter data. B.S. thesis. Department of Computer Science and Engineering, East West University, Dhaka, Bangladesh

Revelle, W (2018) psych: procedures for personality and psychological research, Northwestern University, Evanston, Illinois, USA. https://CRAN.R-project.org/package=psychversion=1.8.4. Accessed 10 May 2018

Sadhukhan S, Dhadekar M, Bhonar S (2016) Stock market prediction using artificial neural networks. Imp J Interdiscip Res 2(5)

Schumaker RP, Chen H (2009) Textual analysis of stock market prediction using breaking financial news: the AZFin text system. ACM Trans Inf Syst (TOIS) 27(2):12

Shahbaz P, Ahmad B, Reza EA, Jalal JM (2014) Stock market forecasting using artificial neural networks. Eur Online J Nat Soc Sci 2(3):2404–2411

Shen S, Jiang H, Zhang T (2012) Stock market forecasting using machine learning algorithms. Department of Electrical Engineering, Stanford University, Stanford

Socher R, Perelygin A, Wu J, Chuang J, Manning CD, Ng A, Potts C (2013) Recursive deep models for semantic compositionality over a sentiment Treebank. In: Proceedings of the 2013 conference on empirical methods in natural language processing, pp 1631–1642

Suleman MT (2012) Stock market reaction to good and bad political news. Asian J Finance Acc 4(1):299–312

Taimur M, Khan S (2015) Impact of political and catastrophic events on stock returns. VFAST Trans Edu Soc Sci 6(1):21–32

Tang X, Yang C, Zhou J (2009) Stock price forecasting by combining news mining and time series analysis. IEEE/WIC/ACM Int Jt Conf Web Intel Intel Agent Technol 1:279–282

Tayal D, Komaragiri S (2009) Comparative analysis of the impact of blogging and micro-blogging on market performance. Int J Comput Sci Eng 1(3):176–182

Turner T, (2007) A beginner's guide to day trading online. 2nd edn, Adams Media

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

Zhao H, Sun M, Deng W, Yang X (2016) A new feature extraction method based on EEMD and multi-scale fuzzy entropy for motor bearing. Entropy 19(1):14. https://doi.org/10.3390/e19010014

Zhao H, Yao R, Xu L, Yuan Y, Li G, Deng W (2018) Study on a novel fault damage degree identification method using high-order differential mathematical morphology gradient spectrum entropy. Entropy 20(9):682. https://doi.org/10.3390/e20090682

Zhou Z, Zhao J, Xu K (2016) Can online emotions predict the stock market in China?. In: International conference on web information systems engineering, pp 328–342