Journal of Ambient Intelligence and Humanized Computing (2022) 13:3433–3456
https://doi.org/10.1007/s12652-020-01839-w

ORIGINAL RESEARCH

Stock market prediction using machine learning classifiers and social media, news

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

Received: 31 March 2019 / Accepted: 25 February 2020 / Published online: 14 March 2020

Abstract
Accurate stock market prediction is of great interest to investors; however, stock markets are driven by volatile factors such as microblogs and news that make it hard to predict stock market index based on merely the historical data. The enormous stock market volatility emphasizes the need to effectively assess the role of external factors in stock prediction. Stock markets can be predicted using machine learning algorithms on information contained in social media and financial news, as this data can change investors' behavior. In this paper, we use algorithms on social media and financial news data to discover the impact of this data on stock market prediction accuracy for ten subsequent days. For improving performance and quality of predictions, feature selection and spam tweets reduction are performed on the data sets. Moreover, we perform experiments to find such stock markets that are difficult to predict and those that are more influenced by social media and financial news. We compare results of different algorithms to find a consistent classifier. Finally, for achieving maximum prediction accuracy, deep learning is used and some classifiers are ensembled. Our experimental results show that highest prediction accuracies of 80.53% and 75.16% are achieved using social media and financial news, respectively. We also show that New York and Red Hat stock markets are hard to predict, New York and IBM stocks are more influenced by social media, while London and Microsoft stocks by financial news. Random forest classifier is found to be consistent and highest accuracy of 83.22% is achieved by its ensemble.

Keywords Deep learning · Feature selection · Hybrid algorithm · Natural language processing · Predictive modeling · Sentiment analysis · Stock market prediction

1 Introduction

The stock market is a vital component of a country's economy. It is one of the largest opportunities for investment by companies and investors. A company can gain a considerable amount of money by expanding its business through an Initial Public Offerings. It is a good time for an investor to purchase new stocks and gain extra profits from dividends offered in the company's bonus program for

✉ Mustansar Ali Ghazanfar
mghazanfar@uel.ac.uk

Wasiat Khan
wasiat.khan@gmail.com

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

Amin Karami
a.karami@uel.ac.uk

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

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

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

shareholders. As a trader, an investor can also trade stocks in the stock market.

Stock traders need to predict trends in stock market behavior for correct decision making to either sell or hold the stock they possess or buy other stocks. To gain profits, stock traders need to buy those stocks whose prices are expected to increase in near future and sell those stocks whose prices are expected to decrease. If stock traders predict trends in stock prices correctly, they can realize significant profits. Therefore, prediction of future stock market trends is very important for decision making by stock traders. However, stock markets are volatile (Bastianin and Manera 2018) and therefore challenging to predict, and external factors, like social media and daily financial news, affect stock prices at once in a positive or negative manner. These factors must be considered for accurate stock market prediction.

Investment in the stock market is risky, but when approached in a disciplined manner, it is one of the most efficient ways to enjoy substantial profits. Investors evaluate the performance of a company before deciding to purchase its stock to avoid buying risky stocks. This evaluation includes analysis of the company's performance on the social media and financial news websites. However, such a huge amount of social media and financial news data cannot be completely assessed by investors. Therefore, an automated decision support system is necessary for investors, as this system will evaluate stock trends automatically using such large amounts of data. This automated system can be built using machine learning algorithms. Finding those algorithms that are more effective in predicting stock market trend using external data, like financial news and social media data, is very important. As accurate stock prediction based on external factors will increase investors' profits, so machine learning researchers have taken a keen interest in this field.

Previous research on stock prediction used historical (Hegazy et al. 2014; Shen et al. 2012; Chen et al. 2018; Yetis et al. 2014; Ou and Wang 2009), social media (Urolagin 2017; Chakraborty et al. 2017; Khatri and Srivastava 2016; Yan et al. 2016; Zhou et al. 2016), or news (Dang et al. 2018; Vargas et al. 2018; Chen et al. 2017a, b; Li et al. 2014a, b, c) data to predict the stock market using machine learning algorithms. Different predictive systems have been proposed that use one or the other type of data. These systems provide useful information to investors to make investment decisions for buying or selling a stock. But using one kind of data may not give increased prediction accuracy for a stock market.

Historical data has been used in a technical analysis approach in which mathematics is applied to analyze data for finding future stock market trends (Dang et al. 2018). Researchers used different machine learning techniques, such as deep learning (Li et al. 2014a, b, c) and regression analysis (Jeon et al. 2018), on stock historical price data, but it is important to include external factors because

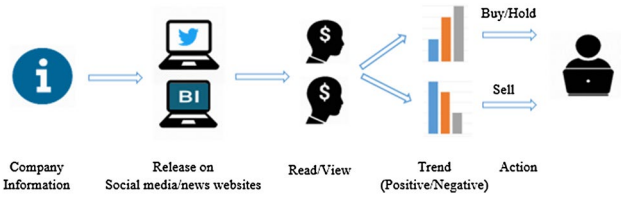

Fig. 1 A general plot that illustrates how social media and financial news affect stock market trends

unexpected events expressed on social media and financial news can also affect stock prices.

Social media is relatively a new form of content on the Internet. One of its important properties is the timely availability of new information and fast interaction among its users. Such interaction can be regarded as a measure of users' attention towards a large number of topics including stock market. But social media alone does not affect the behavior of stock traders and thus, stock markets.

Individuals who are looking to invest in stock markets are often unaware of the behavior of stock markets. As a result, they don't know which shares to purchase and which to sell to maximize their profits. These investors know that stock market growth depends on related news. Therefore, they need accurate and timely information about stock exchange listings, so their trading decisions can be made with timely and accurate information. Since this information can be acquired from financial news websites, most of these websites have evolved to become valuable sources of information that assist traders. Still, investors' expectations based on financial news alone as a trading strategy may not be enough (Brown and Cliff 2004).

In existing predictive systems, researchers used social media posts or news data along with stock market data for stock prediction. As far as we know, no predictive system has used both types of data for stock prediction. Using one kind of data may not give maximum prediction accuracy. Both data can change and affect traders' decisions, so both sources of data, i.e., social media and financial news, should be considered when developing a predictive system for stock markets. Considering both sources for prediction will increase the accuracy of the proposed prediction system. Therefore, using social media and financial news would result in more effective stock trend predictions. A general illustration of how news and social media affect trends in stock market is shown in Fig. 1.

The social media and financial news websites considered as external sources of data for our proposed machine learning model will provide raw text data in the form of tweets and news headlines. This raw data is not understandable by the machine learning algorithms. The data need to be preprocessed. In the fundamental analysis approach, natural

language processing (NLP) is used to analyze social media and financial news data to find positive, neutral, or negative sentiments based on the contents of the documents. Then machine learning algorithms can be used to learn the association between sentiments of text documents and stock market trend movement.

An efficient predictive system is of great significance for stock traders. Traders want such algorithms that can use large data efficiently. The stock data for trend prediction contains a mixture of features from textual data and stock price data in which some features are more relevant while the others are not for making predictions. We don't know which features to remove and which features to select. Therefore, feature selection needs to be performed first on the final data sets.

A quality predictive system that produces quality results is of great value for stock traders. Traders want predictive systems that are accurate and can detect spam data. Due to the increased use of social media, spammers have also started targeting social media. Spammers use multiple twitter accounts to post duplicate tweets for promoting their services and products (Sedhai and Sun 2018). They are very active in spreading spam messages on social media, so removal of spam tweets is necessary from the social media dataset. A predictive algorithm that incorporates historical, social media, and news data into predictions and that performs efficiently after feature selection and spam tweets reduction will be especially useful.

Stock markets usually behave differently from each other. Some behave differently due to stock volatility; prediction of such stock markets is difficult. Identification of such stock markets is also beneficial for stock traders in making trading decisions.

Stock traders are interested in stock markets that are common and well known among other traders. The most common stock markets are discussed more frequently on social media and news platforms, where stock traders look to learn more about these stock markets. Stock traders seek such stock markets that investors are interested in and discuss on social media and financial news platforms. Therefore, the identification of such stock markets is also important for stock traders.

In recent years, deep neural networks have gained numerous successes in different fields such as speech recognition (Noda et al. 2015) and computer vision (He et al. 2016). The concept of deep learning can be used in stock prediction too due to its efficiency on data sets of large sizes (Li et al. 2014a, b, c).

In machine learning, combining classifiers is a popular approach and has proven superior in performance compared to using single classifiers (Tsai et al. 2011). Different classifiers can be combined using ensemble methods in machine learning, thereby improving prediction accuracy of the individual classifiers.

Based on the analysis given above, we suggest a novel machine learning approach for stock market prediction by analyzing social media posts and financial news as external factors. For this research, we select Twitter as the source of social media data due to its conciseness (Tayal and Komaragiri 2009). Financial news headlines are collected from Business Insider[1] and historical data is gathered from Yahoo Finance.[2]

The main contributions of this research are as follows.

- Proposing a combination of financial news and social media data for predicting stock market trends.
- Proposing feature selection from final data sets to improve prediction performance.
- Proposing spam tweets reduction for improving algorithms prediction performance.
- Proposing such algorithm that gives consistent results.
- Proposing stock markets that are difficult to predict.
- Proposing stock markets that are more influenced by social media and financial news.
- Proposing deep learning for predicting stock markets.
- Proposing a hybrid algorithm for stock market prediction.

The remaining part of this text is organized as follows. Existing work in stock trend prediction is given in Sect. 2. In Sect. 3, we explain research methodology and describe each step with detail. In Sect. 4, we present implementation details of the proposed system. Section 5 gives the experimental results and discussion. In Sect. 6, we present conclusion and suggestions for future work in this field.

2 Related work

2.1 Sentiment analysis

In the last decade, sentiment analysis has gained importance because of the availability of huge amount of textual data on the social media and news platforms. This textual data can be mined for finding opinions of users for different application areas. For sentiment analysis of this huge volume of textual data, data mining and machine learning carry great importance. Therefore, machine learning researchers have carried out research on mining opinions of users of these platforms.

Tweets can be classified into different classes based on their contents. Yuan (2016) explored rule, lexicon, and

[1] https://www.businessinsider.com.

[2] http://www.finance.yahoo.com.

machine learning-based sentiment categorization methods. For the lexicon-based techniques, feature scoring and word count approaches were tried. For the machine learning-based technique, support vector machine (SVM), maximum entropy (ME), and Naïve Bayes (NB) were used. Part-of-Speech linguistic annotations and Bag of Words (BoW) with N-Gram features were compared. They found that BoW was an effective and simple feature that achieved optimum performance. The linguistic features also showed better performance. A survey on the categorization of Twitter data was performed by Lakshmi et al. (2017) using NB. They examined the information contained in tweets and concluded that this information is very structured and heterogeneous, and can be classified into positive, neutral, or negative classes.

Sentiment analysis of user opinions can be performed for different application areas. For example, Joshi and Tekchandani (2016) performed comparative analysis of SVM, ME, and NB machine learning algorithms to classify movie review data from Twitter using unigram features, bigram features, and a combination of unigram and bigram features. They found that performance of SVM was better than the other classifiers. Qasem et al. (2015) evaluated logistic regression (LR) and neural networks on two weighting schemes namely, unigram term frequency (TF) and bigram TF inverse document frequency (TF-IDF) on tweets related to technology stocks, i.e., Facebook, Google, Twitter, and Tesla. From the experimental results, they concluded that in terms of overall accuracy, classifiers gave the same results. However, empirical experiments showed that unigram TF-IDF outperformed TF.

Similar to social media, news is also an important external factor that conveys important events related to stocks and affect stock markets, so machine learning researchers also performed sentiment analysis of news data. Dang and Duong (2016) performed sentiment analysis of news and classified news into upward, neutral, and downward classes. They used SVM on stock price data and news related to VN30 Index companies to find correlation between stock prices and financial news. They found that a correlation exists between stock prices and news. Tirea and Negru (2015) used automated text categorization to extract stock related information from news ontology that affected stock behavior. They found that there was a relationship between news and stock price behavior. They used news data related to Bucharest Stock Exchange companies. The news data was crawled using Google custom search.

In addition to news websites, breaking news are also posted on social media platforms like Twitter. Alostad and Davulcu (2015) used breaking news on Twitter and financial news from the NASDAQ website for 30 companies listed on DJI to predict hourly direction of stocks of these companies. They proved that information contained in news articles lead to significant increase in hourly directional prediction accuracies for the stocks mentioned in the articles.

2.2 Stock market prediction

The following subsections describe the usage of various kind of data for stock market prediction in literature.

2.2.1 Price data

Stock market researchers have employed various machine learning techniques for mining historical, social media, and news data to develop prediction models. Before social networking and financial news platforms were so common, stock price data was usually used for predicting stock market. For example, a machine learning model was proposed by Hegazy et al. (2014) for price forecasting of S&P 500. The model used technical indicators and price data of various stock markets listed on S&P 500. The proposed model integrated least square SVM (LS-SVM) and particle swarm optimization (PSO) algorithms. The PSO algorithm was used for optimizing LS-SVM by selecting the best combination of parameters to predict the daily stock prices. A new algorithm was proposed for prediction by Shen et al. (2012) that used temporal relationships among worldwide stock markets and different financial products to forecast stock market coming day trends using SVM. Stock price data was used as input parameters to SVM. Chen et al. (2018) used CSI 300 stock price data from the Chinese stock market for comparing price prediction of traditional neural networks with deep learning and found that prediction performance of deep learning was better than that of traditional neural networks. Yetis et al. (2014) used NASDAQ stock market index price data in feed forward artificial neural network (FFANN) for predicting stock value and found that ANN showed good predictive performance for NASDAQ. Ten machine learning algorithms were used by Ou and Wang (2009) on stock price data to predict index price movement in the Hong Kong stock market. SVM and LS-SVM were found to demonstrate good predictive performance than the other forecasting models.

2.2.2 Social media data

Machine learning researchers have examined opinions of investors available on social media platforms to inform stock market forecasting. These platforms contain substantial information about companies and the products and services they offer. Urolagin (2017) explored the relationship of stock prices from Yahoo Finance and social media text sentiments of a company. He used NB and SVM classifiers for performing sentiment classification. For this classification, N-gram feature vectors were formed from the most significant words

in tweets. Furthermore, the relationship pattern between the number of positive or negative tweets and stock prices was explored. He found that an association existed between tweet features, like number of positive, neutral, negative tweets, and total number of tweets, to forecast the stock market status using SVM.

Opinions on social media platforms illustrate the sentimental state of many users on these platforms. These users express their opinions for a company or its products in tweets or comments. Extraction of sentiments from these tweets or comments is used to detect the user view for a particular company or product. Chakraborty et al. (2017) collected tweets that contained keywords 'AAPL', 'stock market', and 'stocktwits'. Tweets that contained the keyword 'AAPL' were used to predict Apple Inc. stock index, whereas tweets that contained keywords 'stock market' and 'stocktwits' were used to predict stock market movement in the United States. They used SVM for sentiment classification while a boosted regression tree model was used to predict next day stock difference. Khatri and Srivastava (2016) collected tweets and comments from the Stock Twits[3] website, while the market data was extracted from Yahoo Finance for Facebook, Apple, Google, Oracle, and Microsoft stock markets. The tweets and comments were classified into four categories: up, down, happy, and rejected. This polarity index and stock data were used as input in ANN to predict stock closing prices.

Previous research has shown that stock market price and the public mood expressed in tweets are related to some extent. For example, Yan et al. (2016) proposed a model called Chinese Profile of Mode States (C-POMS) for analyzing sentiments in microblog feeds. Then, Granger causality test was performed, which showed that there was a relationship between C-POMS analysis and stock market price series. Probabilistic neural network (PNN) and SVM were used for making predictions. Experimental results showed that SVM was better than PNN for predicting stock market movements.

Financial analysts convey their analyses in tweets for some stocks. Data mining algorithms and NLP techniques can be employed to discover such analyses from tweets. For example, Zhou et al. (2016) performed correlation analysis on ten million stock relevant tweets from Sina Weibo.[4] They found that five features of China's stock market namely, opening, closing, intra-day lowest index, intra-day highest index, and trading volume can be predicted from different emotions expressed in tweets such as fear, joy, unhappiness, and revulsion. Their model outperformed baseline models in predicting these features of China's stock market. They used K-means discretization for predicting these features.

Different events that are related to stock markets are posted on social media platforms and can affect stock market returns. These events provide meaningful labels for events classification such as losses or gains. Makrehchi et al. (2013) suggested a new technique to estimate sentiments that were based on events related to stock markets. An efficient classifier was built to judge sentiments of tweets so the information could be used in building an efficient strategy for trading.

2.2.3 Financial news data

Online news is an interesting data that can be mined and analyzed to acquire helpful information for stock market prediction. These news contents can be classified into general, political, and financial news, for example.

The desire of any trader is to predict market behavior to inform decisions on if and when to buy or sell stock market shares to maximize profit. Such forecasting is difficult because the behavior of the stock market is always changing, affected by numerous external factors such as political situations (Khan et al. 2019), global economy, and investor expectations. Therefore, some research works exist on the usage of financial news for predicting stock markets. For example, Vargas et al. (2018) used technical indicators from stock prices of Chevron Corporation and financial news titles from Reuters[5] to predict daily directional movement of the stock market using deep learning models. Results showed a positive impact of the hybrid input of news data and technical indicators on the directional movement of the stock. Dang et al. (2018) proposed a novel framework called Stock2Vec and two-stream gated recurrent unit (TGRU) network for predicting stock prices directions of S&P 500 using Reuters financial news and Bloomberg[6] and Harvard IV-4 sentiment dictionary. They showed that the TGRU network performed better than the baseline models and that Stock2Vec was very efficient using financial data sets. Chen et al. (2017a, b) used GRU and recurrent neural networks (RNN) on stock price data from the CSI 300 Index and news from Sina Weibo to predict volatility of the Chinese stock market. Their proposed model outperformed the baseline methods and showed good prediction performance. Li et al. (2014a, b, c) explored the impact of news on stock market price returns of Hong Kong's stock exchange. They evaluated stock prediction accuracy of the proposed model and compared stock performance at different market levels like sector, stock, and index. They found that news sentiment analysis improved

[3] https://www.stocktwits.com.

[4] https://www.weibo.com.

[5] https://www.reuters.com.

[6] https://www.bloomberg.com.

prediction accuracy. They also found that sentiment analysis models outperformed the BoW model at sector, stock, and index levels.

2.2.4 Social media and financial news data

Both news and social media are external factors that can affect stock trends. The combination of social media and financial news for predicting stock markets is very scarce in the literature. The effect of this hybrid data can be checked for predicting stock markets. Usmani et al. (2016) used different machine learning techniques on social media and different types of news feeds related to Karachi Stock Exchange (KSE) to predict market performance. They found that KSE-100 index performance can be forecasted using machine learning algorithms. Attigeri et al. (2015) used financial news and social media data to predict future stock values using LR. They found that social media and financial news are correlated in predicting stock markets. Li et al. (2014a, b, c) found relationship between financial indicators, such as social media and news, and future stock prices of Shenzhen and Shanghai Stock Exchanges using support vector regression. They found that company-related information in news articles can affect trading activities, and social media sentiments can influence traders' decision making.

2.3 Feature selection

Different techniques have been used by researchers for feature selection from data sets with a large number of features. Nowadays, machine learning researchers are recognizing the importance of feature selection for analyzing data because data with high dimension not only affect the learning models but also increase computational time and are considered as information poor (Cao et al. 2016; Cheng et al. 2013). Moreover, due to the large number of features, we face the *curse of dimensionality* which states that in space of high dimension, data turn out to be sparse (Cheng et al. 2013; Hastie et al. 2009).

To solve the problems that arise from high dimensional data, researchers use two approaches: feature extraction and selection. In the first approach, new feature space with low dimensionality is created while in the second approach, redundant and irrelevant features are removed and a small subset of more relevant features is selected.

Feature selection has been done increasingly with swarm intelligence (SI) algorithms (Blum and Li 2008; Hassanien and Emary 2016). The reason is that the technique is popular for solving various optimization problems and finding optimal features is definitely this kind of problem. SI algorithms were very popular in recent years (Blum and Li 2008), and nowadays, its two popular algorithms are PSO (Eberhart and Kennedy 1995) and ant colony optimization (ACO) (Dorigo 1992).

For comparing SI algorithms for feature selection, Brezocnik et al. (2018) conducted a comprehensive review of the algorithms. They explained different application areas, techniques, methods, and their settings for various aspects of feature selection. They analyzed that SI algorithms were mostly used on small datasets with up to 150 features. They also found the fact that SI algorithms were mostly used in Bio-Medical Engineering (60.53%) and Computer Engineering (28.95%) application areas.

In the field of Bio-Medical Engineering, for example, gravitational cuckoo search algorithm (Yang and Deb 2009) was used by Jayaraman and Sultana (2019) for feature selection in heart disease classification. In Computer Engineering, artificial fish swarm algorithm (Li 2003) was used by Wang and Dai (2013), bat algorithm (Yang 2010) by Enache et al. (2015), and grey wolf optimizer (Mirjalili et al. 2014) by Seth and Chandra (2016) for feature selection in intrusion detection problems. For feature selection in image steganalysis, Mohammadi and Abadeh (2014) used IFAB novel feature selection method that was based on artificial bee colony algorithm (Karaboga 2005) and Chhikara et al. (2018) used firefly algorithm (Yang 2008). Some additional problems that were addressed in this field are diagnosis of fault in complex structure with bacterial foraging optimization (Passino 2002) algorithm (Wang et al. 2016), improved face recognition with adaptive binary PSO (BPSO) (Sattiraju et al. 2013), identification of malicious web domains using BPSO (Hu et al. 2016), and classification of web pages with ACO (Saraç and Özel 2014). From the comparison of various SI algorithms for feature selection, Brezocnik et al. (2018) concluded that there is no single SI algorithm that is most effective for feature selection.

Some researchers also used hybrid approaches for feature selection by combining individual techniques. For example, Ibrahim et al. (2019) combined slap swarm algorithm (Mirjalili et al. 2017) with PSO and found an enhancement in performance and accuracy. Similarly, Moslehi and Haeri (2019) also proposed a hybrid approach by combining genetic algorithms and PSO and found that the proposed approach was capable to obtain accurate classification. Zhong and Enke (2016) compared three techniques for feature selection namely, kernel-based principal component analysis (PCA), fuzzy robust PCA, and PCA in order to reduce 60 economic and financial features in the data to forecast S&P 500 Index. They found that PCA gave slightly higher accuracy performance than the other two techniques.

2.4 Spam tweets reduction

Spam contents on social media has also attracted researchers due to the problems created by these spam contents. Much

research has been done on spam tweets detection for different purposes. For example, a semi-supervised spam detection framework for spam tweets detection was proposed by Sedhai and Sun (2018). RF, LR, and NB classifiers were trained and tested on HSpam14 data set (Sedhai and Sun 2015) to classify tweets into ham and spam in real time. Chen et al. (2017a, b) created a dictionary of blacklist words to improve detection performance of their model for detecting spam and low-quality contents in real time on Twitter. RF was used as a classifier for classifying tweets.

Similar to spam tweets, spam profiles are also a source of unwanted advertisements and spam tweets and thus a security threat for users of social media platforms. Al-Zoubi and Faris (2017) analyzed twitter spam profiles to extract features using information gain and ReliefF. The extracted features were then used to classify and detect spam profiles of Twitter users by applying the decision tree (DT), multilayer perceptron (MLP), K nearest neighbor (KNN), and NB classifiers.

2.5 Stock market volatility

Stock markets are volatile; therefore, their behavior cannot be predicted accurately. External factors, such as financial crises, can influence stock volatility. Researchers have employed different techniques to identify stock market volatility. For example, Kumar and Patil (2015) employed time series and machine learning techniques to forecast the volatility of the S&P 500 index. The stock market standard deviation was considered to forecast volatility with high accuracy.

Generalized autoregressive conditional heteroscedasticity (GARCH) family models have been proved to predict stock volatility with maximum accuracy. For example, Omer and Halim (2015) used three models from the GARCH family and found that the exponential GARCH model outperformed the other models for forecasting volatility of the Malaysian stock market. Wang et al. (2014) proposed a learning-based multi-kernel extreme learning machine to increase performance of volatility prediction using stock historical data and news.

2.6 Deep learning for stock prediction

In the current decade, deep learning has achieved greater importance in various fields. Researchers have used it in various fields. It can be used in stock trend prediction as well due to its effectiveness on large data sets. For example, Dang et al. (2018) proposed TGRU network for predicting stock price directions of S&P 500 and found that the proposed model outperformed the baseline models with an overall accuracy of 66.32%. Khare et al. (2017) used RNN and FFANN to predict short term future prices of New York Stock Exchange (NYSE) using technical analysis. They found FFANN's performance superior in predicting stock short term prices. Long short-term memory neural network was used by Li et al. (2017) on investor sentiments and stock price data to forecast index values of CSI300 in China's stock market. NB was used to extract these investors' sentiments from posts related to various forums.

2.7 Hybrid algorithm

Combining classifiers has demonstrated good performance compared to individual classifiers. Researchers have used different ensemble methods in different fields to enhance individual classifier's accuracies. For example, Todorovski and Džeroski (2003) proposed a technique for combining classifiers called meta DTs (MDTs). Ordinary DTs, KNN, and NB algorithms were combined using stacking and voting ensemble methods. They compared these methods and showed that stacking performed best compared to voting. Džeroski and Ženko (2004) evaluated different ensemble techniques for constructing ensembles of heterogeneous classifiers namely, J4.8, IBK, and NB, with stacking. They proved that the ensemble performed better compared to selecting the best classifier from the ensemble by cross validation.

Ensemble methods can be used to enhance prediction of stock returns in stock markets. Tsai et al. (2011) used bagging and majority voting ensemble methods to improve prediction performances of MLP, LR, and classification and regression tree classifiers. They showed that classifier ensembles performed best compared to individual classifiers in terms of prediction accuracy. Kim et al. (2003) used genetic algorithms and majority voting techniques for combining multiple neural network classifiers to predict customer purchasing behavior and showed that their method outperformed than that of the individual classifiers. Their results further showed no significant difference in ensemble methods. Sun and Li (2012) used weighted majority voting to combine several SVM classifiers for predicting financial distress. Their experimental results showed that ensemble of SVM performed better than the individual SVM classifier.

Ensemble generation is a well-known approach for increasing the accuracy of decision making by classifiers. As a rule, majority voting is the model usually applied for decision making by the classifiers. Hajdu et al. (2013) created an ensemble-based system using majority voting for detecting optic disc in retinal images. Liu et al. (2018) implemented a positive and unlabeled learning algorithm using multiple classifiers and three ensemble methods, which were based on weighted average, majority, and weighted vote combination rules. From results, they concluded that ensemble methods based on weighted vote and weighted average outperformed individual classifiers.

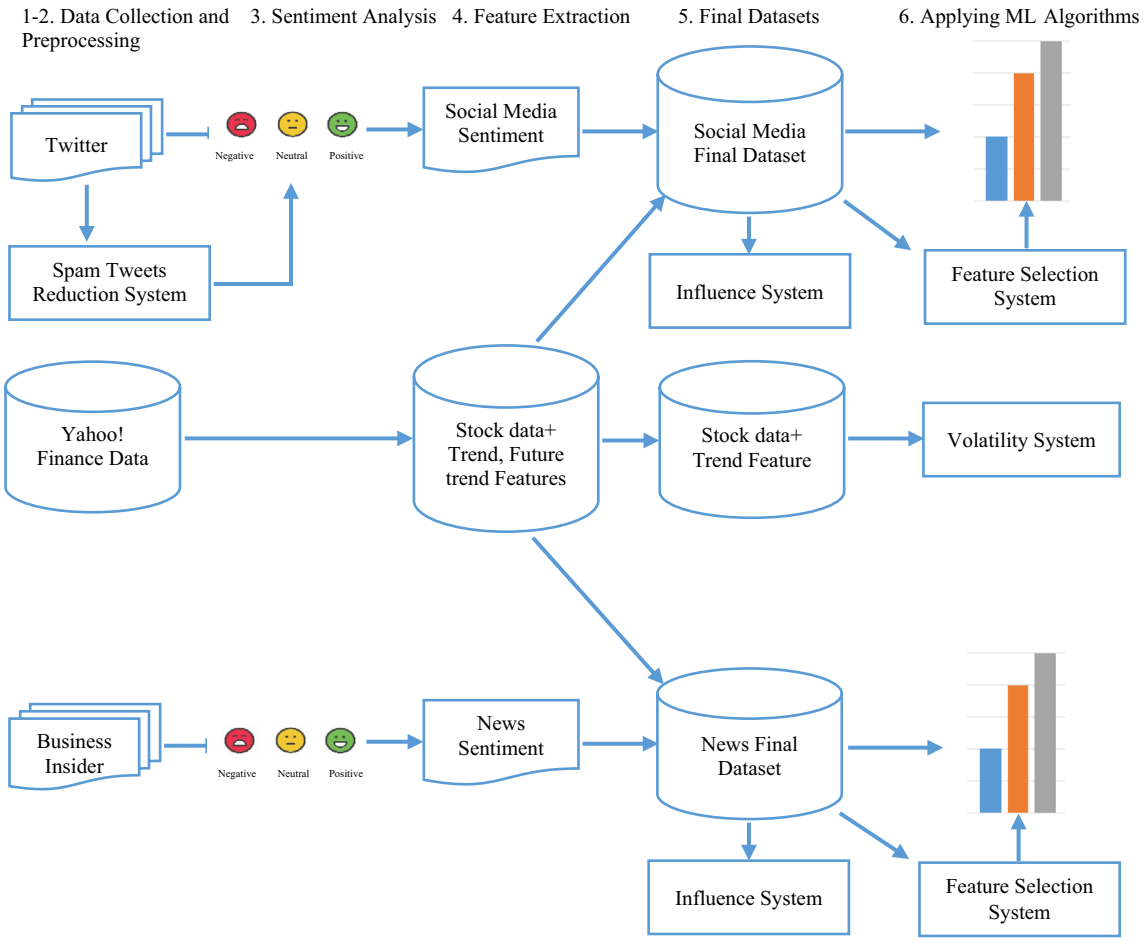

Fig. 2 Flow chart of the steps in our proposed framework for stock market forecasting using financial news and social media

3 Methodology

This section describes the individual steps performed in our proposed framework for stock prediction. Our proposed framework includes six basic steps and sub systems. The steps and the sub systems are shown in Fig. 2. The steps are described in subsequent subsections while the sub systems are given in Sect. 4.

3.1 Data collection

This subsection describes the data collection process, sources of the collected data, and structure of the collected data. The stock markets that are selected as case study for this research and their tweets and news counts are given in Table 1. In this table, the stock exchanges show the overall stock markets while the other stocks are stock markets of individual companies. Note that the stock market terminology will be used interchangeably to refer to the stock market of individual companies and overall markets.

Stock market, social media, and news data of the selected stock markets and S&P 500 index price data are gathered for 2 years from July 1, 2016 to June 30, 2018.

3.1.1 Social media data

Twitter is selected as the source of social media data due to its conciseness (Tayal and Komaragiri 2009). To download desired tweets, Twitter API has been implemented in Python.[7] The Python application gets as input parameters the start date, until date, output file path, and a search query which uses cashtags containing ticker symbols of the selected stock market preceded by $ sign like '$NYSE', '$HPQ', etc. We use cashtags as search query for downloading stock market related tweets because cashtags have been found to be useful for the analysis of financial information and providing new insights into stocks and companies (Hentschel and Alonso 2014). All tweets of the selected

[7] https://github.com/Jefferson-Henrique/GetOldTweets-python.

Table 1 Stock markets symbols and tweets, news count summary

No.	Stock market	Ticker symbol	Country/stock exchange	Tweets count	News count
1	Karachi Stock Exchange	KSE	Pakistan	34	0
2	London Stock Exchange	LSE	United Kingdom	2535	53
3	New York Stock Exchange	NYSE	United States	12,538	0
4	HP Inc.	HPQ	NYSE	27,432	554
5	International Business Machines Corporation	IBM	NYSE	364,601	1700
6	Microsoft Corporation	MSFT	NASDAQ	168,901	3316
7	Oracle Corporation	ORCL	NYSE	51,328	799
8	Red Hat, Inc.	RHT	NYSE	18,120	212
9	Twitter, Inc.	TWTR	NYSE	380,472	2367
10	Motorola Solutions, Inc.	MSI	NYSE	6444	284
11	Nokia Corporation	NOK	NYSE	23,441	301

stock markets are downloaded in raw form between start date and end date in.csv file format. The downloaded raw tweets file has three features—*Source*, *TweetText*, and *Date*—which indicate the source or user of the tweet, text of the tweet, and the date on which the tweet was posted, respectively.

3.1.2 Yahoo Finance stock data

Stock historical price data is available on Yahoo Finance. Price data of the selected stock markets are collected from Yahoo Finance for the selected time period in.csv file format. The downloaded data files have seven features—*Date*, *Open*, *High*, *Low*, *Close*, *Volume*, and *Adjusted Close*—which on a specific date show the stock traded date, stock open price, stock maximum trading price, stock lowest trading price, stock closing price, number of shares traded, and closing price of a stock when dividends are paid to investors, respectively.

3.1.3 Financial news data

The third important data for this research is financial news related to stock markets. Researchers have used different sources like Reuters (Qasem et al. 2015; Vargas et al. 2018), Sina Weibo (Chen et al. 2017a, b), and FINET[8] (Li et al. 2014a, b, c; Liu et al. 2018). For this research, we have selected Business Insider (Todorovski and Džeroski 2003) for collecting financial news because it contains a collection of stock market related news from the world famous news websites such as Reuters, Financial Times, etc. We have crawled news headlines of the financial news of selected stock markets in.csv file format for the selected time period using a web crawler implemented in JSOUP Java library.[9]

The raw news files have four features—*Source*, *Link*, *NewsText*, and *Date*—which indicate news source, URL link, news headline, and publication date, respectively. The use of short news headlines rather than news article bodies makes this step unique, which enables the usage of short text for news sentiment analysis. Moreover, the collected news headlines are general financial news related to stock markets, so there is no need for further filtration.

3.2 Preprocessing

3.2.1 Social media and financial news data

The downloaded tweets and news headlines are in raw form and need to be preprocessed before applying machine learning algorithms. The following steps are performed to convert the tweets and news headlines into an appropriate form for the machine learning algorithms.

1. Tweets and news documents are converted into word tokens.
2. HTML and other tags like author tag (@) and cashtags are removed. These tags need to be removed because they carry no useful information for machine learning algorithms in finding sentiments.
3. URLs are removed.
4. Stop words are also eliminated. These are such words which are frequently existed in tweets and news headlines (for example, is, the, are, an, etc.) and carry no valuable information for classifiers.
5. Words are converted into the same stems; this process is called stemming.
6. Duplicate tweets are removed.

[8] http://www.finet.hk/mainsite/index.htm.

[9] https://jsoup.org.

3.2.2 Yahoo Finance stock data

Adjusted close attribute is removed from the downloaded stock price data as this attribute has no role in our stock prediction model.

3.3 Sentiment analysis

Sentiment analysis of the processed tweets and financial news is performed using Stanford sentimental analysis package of Stanford NLP (Socher et al. 2013). Most of the sentiment analysis systems work by looking at words in isolation, giving positive or negative points for positive or negative words, respectively, and then summing up these points. This method ignores the order of words, and therefore, valuable information is lost. In contrast, the Stanford NLP approach, which is based on the deep learning model, makes representation of complete sentences, based on the sentence structure. This approach uses the Sentiment Treebank, which includes sentiment labels for 215,154 phrases in the parse trees composed of 11,855 sentences and gives new challenges for the compositionality of the sentiment. To address this, the recursive neural tensor network has been introduced. The model has been trained on the new treebank, and it performs best of all previous methods on a number of metrics. The prediction accuracy of sentiment labels for all phrases has reached 80.7%, which is a 9.7% improvement over the bag of features baselines. It is the only model that can accurately capture the effect of conjunctions and negations and their scope at various tree levels for both positive and negative phrases.

According to the Stanford NLP approach, more negative tweet or news has a sentiment of 0, negative tweet or news has a sentiment of 1, neutral tweet or news has a sentiment of 2, positive tweet or news has a sentiment of 3, and more positive tweet or news has a sentiment of 4. This can be expressed as.

$$Sentiment\ score = \begin{cases} 0\ if\ tweet\ or\ news\ is\ more\ negative \\ 1\ if\ tweet\ or\ news\ is\ negative \\ 2\ if\ tweet\ or\ news\ is\ neutral \\ 3\ if\ tweet\ or\ news\ is\ positive \\ 4\ if\ tweet\ or\ news\ is\ more\ positive \end{cases} \tag{1}$$

The overall sentiment of news or tweets posted on a specific date is the aggregated sentiment of individual news or tweets for that particular date. If the overall sentiment count on a specific date is higher, then that will mean that the sentiment positivity is higher on that date. The sentiment analysis step results in sentiment features in processed social media and news data files.

3.4 Feature extraction

3.4.1 From Yahoo Finance stock data

For stock trend prediction, two features namely, *Trend* and *Future Trend*, are extracted from the existing features in stock data files. These features have nominal values of positive, neutral, or negative. The value of the *Trend* feature can be found by subtracting the stock open price from the close price on a certain date. The criteria for selecting these values is given in the following equation.

$$Trend_d = \begin{cases} Positive & if\ P_c - P_o > 0 \\ Neutral & if\ P_c - P_o = 0 \\ Negative & if\ P_c - P_o < 0 \end{cases} \tag{2}$$

where $Trend_d$ is the trend, P_c is the closing price, and P_o is the stock opening price on a certain date, respectively.

The *Future Trend* feature is the attribute that will be predicted. It is the difference between a stock's current day closing price and closing price after n days. If the difference is positive, it means that the trend will be positive after n days. If the difference is zero, the future trend will be neutral, and finally, if the difference between the two is negative, that stock future trend will shift downward after n days. Future trend after n days can be determined by the following equation.

$$Future\ Trend_n = \begin{cases} Positive & if\ P_{tc} - P_{nc} > 0 \\ Neutral & if\ P_{tc} - P_{nc} = 0 \\ Negative & if\ P_{tc} - P_{nc} < 0 \end{cases} \tag{3}$$

where P_{tc} is stock today closing price and P_{nc} is stock closing price after n days.

We have selected the value of n as 10, which means that we will identify the stock's future trend up to 10 days. In other words, we will find the impact of news and social media for 10 days predictions in future.

3.4.2 From news and social media data

Two features namely, *News Sentiment* and *Social Sentiment* are created in news and social media data files, whose values are the aggregated sentiments of individual news or tweets posted on a particular date.

3.5 Final data sets

In this subsection, we discuss the process of creating the final data sets which we will use for stock prediction.

Table 2 A view of the final dataset for stock market forecasting using social media data of HPQ

Date	Open	High	Low	Close	Volume	Trend	Social sentiment	Future trend
7/1/2016	12.55	12.76	12.51	12.73	11.865149	Positive	0	Negative
7/5/2016	12.62	12.71	12.26	12.36	11.520288	Negative	33	Positive
7/6/2016	12.26	12.62	12.04	12.6	11.743982	Positive	41	Positive
7/7/2016	12.66	12.88	12.59	12.85	11.976999	Positive	20	Positive
7/8/2016	13.05	13.16	12.95	13.08	12.191372	Positive	20	Positive
7/11/2016	13.17	13.3	13.03	13.16	12.265938	Negative	7	Positive
7/12/2016	13.47	13.83	13.42	13.72	12.787891	Positive	26	Positive
7/13/2016	13.73	13.85	13.63	13.81	12.871778	Positive	50	Negative
7/14/2016	13.89	13.94	13.75	13.8	12.862456	Negative	39	Positive
7/15/2016	13.89	13.93	13.77	13.85	12.909061	Negative	48	Positive

Table 3 Selected machine learning algorithms with optimal parameter values

No.	Algorithm	Abbreviation	Optimal parameter values
1	Gaussian Naïve Bayes	GNB	NA
2	Multinomial Naïve Bayes	MNB	alpha:0.2
3	Support Vector Machine	SVM	kernel: rbf, C: 0.5
4	Logistic Regression	LR	NA
5	Multilayer Perceptron	MLP	alpha: 0.0001, activation: tanh, solver: adam, learning_rate: constant, hidden_layer_sizes:(5)
6	K Nearest Neighbor	KNN	n_neighbors: 3
7	Classification and Regression Tree	CART	max_features: log2, min_samples_split: 13, random_state: 123, min_samples_leaf: 1
8	Linear Discriminant Analysis	LDA	shrinkage: None, solver: lsqr
9	AdaBoost	AB	n_estimators: 100, learning_rate: 0.1
10	Gradient Boosting Classifier	GBM	n_ estimators: 250
11	Random Forest Classifier	RF	n_jobs: -1, min_samples_leaf: 1, n_estimators: 20, random_state: 123, criterion: gini, min_samples_split: 5
12	Extra Tree	ET	n_jobs: -1, min_samples_leaf: 1, n_estimators: 20, random_state: 123, criterion: entropy, min_samples_split: 6

3.5.1 For stock market forecasting using social media

The final data set for this subsystem is created by adding the *Social Sentiment* feature into the stock data file. A tabular view of the data set is shown in Table 2.

3.5.2 For stock market forecasting using financial news

Similarly, the final data set for this subsystem is created by adding the *News Sentiment* feature into the stock data file. This data set is similar to the social media data set shown in Table 2.

3.5.3 For stock market forecasting using social media and financial news

The final data set for this subsystem is created by adding the *Social Sentiment* and *News Sentiment* features into the stock data file. This data set has two sentiment attributes, one from the social media and the other from the news data.

3.5.4 For stock volatility

For creating the final data set for stock volatility, only the *Date* and *Close* features in the stock price data files are

retained and all the other features are discarded. This data set and that of the S&P 500 are then used for finding volatility.

3.5.5 For spam reduction system

The data set for the spam reduction system comprises a training data set of 380 spam and ham tweets.

3.6 Applying machine learning classifiers

In this research study, 12 machine learning classifiers are selected and compared in terms of their prediction performance. These classifiers are first trained and then tested on the final data sets to identify future stock market trends. The final data sets for prediction systems are split into 70% training data (350 samples) and 30% testing data (150 samples) before applying machine learning algorithms. For training and testing the algorithms, we develop prediction models using *Scikit-learn* (Pedregosa et al. 2011), which is a Python library for machine learning. The algorithms used for stock prediction systems in this research are listed in Table 3.

3.6.1 Standardizing the final data sets

Before applying classifiers, the final data sets are standardized using *StandardScalar* class of *Scikit-learn*. Standardization is useful for transforming attributes with Gaussian distribution because some algorithms, like GNB and LDA, assume Gaussian distribution of the input data. After transformation, the attributes have a mean of 0 and a standard deviation of 1.

During standardization of data sets, data may be leaked from the training to the testing data set. To overcome this problem, a strong test harness having a strong separation of training and testing is needed. This requires preparation of data, in which the knowledge of the whole training data set may be leaked to the algorithm. To avoid this, *Pipeline* utility of *Scikit-learn* is used. *Pipelines* transform the data in a linear sequence that can be bound together, resulting in such a modeling process that can be evaluated. It ensures that standardization is constrained to each fold of the cross validation (CV) process which helps in preventing leakage of data in the test harness. Its main goal is to ensure that the whole process of pipelining is constrained to the available training data for evaluation.

3.6.2 Classifiers performance evaluation

Performance of the selected classifiers is evaluated using different evaluation metrics. Since our problem is a multi-class classification problem and the distribution of classes is not uniform, therefore we have used accuracy primary classification metric and three within-class classification metrics namely, precision, recall, and F-measure. Accuracy is a classification metric for evaluating classifiers and can be expressed as.

$$Accuracy = \frac{Number\ of\ correct\ predictions}{Total\ number\ of\ predictions} \tag{4}$$

When the class distribution in the dataset is not uniform, then accuracy may not be a good metric for evaluating classifier performance. Therefore, for evaluating classification performance, a confusion matrix is used and its precision, recall, and F-measure are found.

Precision is the skill of the model to classify samples accurately and can be calculated as follows.

$$Precision = \frac{TP}{TP + FP} \tag{5}$$

where TP is the true positive rate and FP is the false positive rate of the algorithm.

Recall shows the skill of the model to classify the maximum possible samples, and is represented by the following equation.

$$Recall = \frac{TP}{TP + FN} \tag{6}$$

where FN is the false negative rate of the algorithm.

F-measure describes both precision and recall and can be represented as follows.

$$\mathrm{F-measure} = 2 \times \frac{Precision \times Recall}{Precision + Recall} \tag{7}$$

3.6.3 Proposed method for model's validation

Prediction models can be validated using different methods like substitution, holdout, and CV which include k-fold, leave-one-out, and leave-more-out CVs (Chou and Lin 2012; Kuhn and Johnson 2013).

Some tuning parameters that are used by the algorithms for optimizing classification performance are explored in terms of variance using tenfold CV. The number of folds selected are 10 which is recommended (Kohavi 1995) when algorithms performance is compared (Thu et al. 2011; Chou and Lin 2012). Therefore, in the current work, the tenfold CV method is used in the development of each predictive model for all possible configurations of parameters. In this method, data of the training set is divided randomly into ten subsets. Nine subsets are used as new training set for developing all prediction models, while the hold out set is used to find predictive performances of the fitted models. This procedure is repeated ten times on various training datasets until every instance of the subset is used just one time for testing. Then the CV overall accuracy estimation is

measured by taking average of the ten individual accuracies. This method is used for avoiding overfitting and selection of best parameters for the prediction models. The testing data set is not used in the model development, but used for testing the predictive performance of the concerned model after completion. Prediction models are developed using best parameters combination and training data sets and then the final models are applied to testing data sets.

3.6.4 Parameters optimization

Machine learning classifiers use one or more tuning parameters to avoid either underfitting or overfitting. The *fit* method from the *Scikit-learn* GridSearchCV class creates a grid of tuned classification algorithms, and allows a consistent environment to train each machine learning algorithm and tune their parameters. When the optimal values for parameters are found, then the complete training data set is used for building the final model. To select optimal values for the tuned parameters based on the training data set, tenfold CV is used and the testing data set is completely removed during the CV process. The tuned parameter values considered optimal achieve the overall highest classification accuracy during the CV process. The parameters and their values in the last column of Table 3 are the optimal parameter values that are used in the respective classifier class of the *Scikit-learn*.

4 Proposed system

The proposed system considers each aspect of the data and the system itself for achieving accurate predictions. So, the proposed system for stock prediction is divided into eight subsystems. This section describes these subsystems.

4.1 Stock prediction system

4.1.1 Using social media

Extensive experiments are performed on the final data sets using machine learning algorithms to predict future stock market trends of the selected stock markets for the next 10 days. The selected algorithms are first trained and then tested using tenfold CV. We use tenfold CV to ensure that each instance is used equally for training and testing, while reducing the variance. Parameter tuning for the algorithms is also performed to ensure the selection of optimal parameter values for getting maximum prediction accuracy.

4.1.2 Using financial news

The selected classifiers are used on the final data sets of the selected stock markets for the subsequent 10 days to predict future trends. These classifiers are first trained and then tested using tenfold CV on the final data sets. Parameters tuning is also performed for the news-based prediction system to get maximum prediction accuracy.

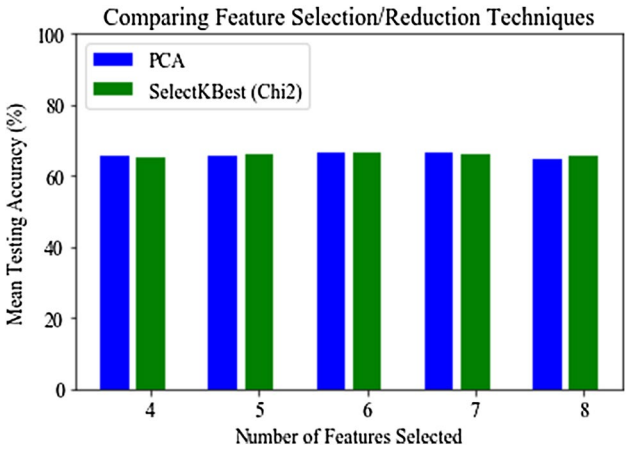

Fig. 3 Mean testing accuracy comparison of PCA and SelectKBest techniques on different values of *K*

4.1.3 Using social media and financial news

Prediction of stock markets is performed for 10 days using the chosen machine learning classifiers over the final data sets that have news and social media sentiments as external features. The classifiers are first trained and then tested using the tenfold CV on the final data sets and future predictions are performed.

4.2 Dimensionality reduction/feature selection

To develop an efficient prediction model, dimensionality reduction or feature selection is performed in this research work. The *feature_selection* and *decomposition* modules of *Scikit-learn* are used for feature selection to improve accuracy or performance of the algorithms on large data sets. The modules support a large number of classes for feature selection. We use SelectKBest Chi2 class from the *feature_selection* module and PCA class from the *decomposition* module for feature selection in this research for different *K* values, where *K* is the number of features to select in SelectKBest and the number of components to keep in PCA. Figure 3 shows mean testing accuracy for both methods using different *K* values.

4.2.1 SelectKBest (Chi2)

SelectKBest is a univariate feature selection method that selects best *K* features and discards the remaining features. This selection of features is based on univariate statistical

tests. Chi2 computes Chi squared stats between each positive feature and the classes. The statistics measure dependency among the features. Those features that are expected to be independent of classes and are not relevant for classification are removed.

4.2.2 PCA

PCA is a linear dimensionality reduction technique that uses singular value decomposition of the data for projecting the data to a lower dimensional space. It is used to overcome redundant features/components in a data set. In PCA, we find components that explain maximum variance; these components are used to retain maximum possible information.

4.3 Spam tweets reduction

To obtain quality prediction results, spam reduction is performed in this research work. For spam tweets reduction, we train and test the MNB classifier for classifying raw tweets data set into spam and ham tweets and note the accuracy, precision, recall, and F-measure. The MNB classifier is used for spam/ham tweets classification because of its accuracy on text classification (Afzal and Mehmood 2016). The trained model which gives a testing accuracy of 81.74% is then used to classify the raw tweets into spam and ham tweets. The percentage of spam and ham tweets is found for each stock market. After classification, spam tweets are removed and basic steps of preprocessing and sentiment analysis are performed on ham tweets to create final data sets. Machine learning algorithms are then trained and tested on the final data sets, and their performance is noted for comparison purposes.

4.4 Identification of a consistent classifier

It is of great significance to identify a classifier that gives consistent results in all scenarios, i.e., in stock market prediction using social media and news data, in dimensionality reduction, and in spam tweets reduction. Therefore, we compare prediction accuracies of all classifiers in these cases and identify a classifier that gives consistent results in all the cases for stock market prediction.

4.5 Identification of stock markets that are difficult to predict

Stock volatility can be used to find the prediction difficulty of stock markets. The higher the volatility, the riskier the stock. There are different methods for finding stock market volatility. We use three methods for finding volatility of the selected stocks and compare their results.

4.5.1 Using variance and standard deviation

Volatility is usually measured using variance and standard deviation. Standard deviation can be calculated by taking the square root of the variance. The variance is calculated using these steps: (1) find mean of the data set, (2) find the differences between the mean and each data value, (3) take squares of each deviation, (4) add these squared deviations, and (5) divide the sum of squared deviations by the number of data values. These steps result in different standard deviations for different stocks. The higher the standard deviation of a stock, the higher the volatility, deeming the stock as difficult to predict.

4.5.2 Using beta

The second general technique for the performance evaluation of stock markets is based on the beta values of the stocks. The beta value reports the stock's movement or behavior with respect to some index, for example, S&P 500 index. It is calculated by using linear regression on stock data points.

We use the selected stocks closing prices and the closing level of S&P 500 for the selected time period to calculate the beta values for the selected stock markets. To calculate the beta value, first, the daily percentage change is calculated for stocks using the equation.

$$C\Delta_s = (C_{ts} - C_{ys})/C_{ys} \times 100 \tag{8}$$

where $C\Delta_s$ is the daily percentage change, C_{ts} is today's closing price, and C_{ys} is yesterday's closing price of the stock.

Similarly, daily percentage change for the S&P 500 index is calculated using the same equation.

$$C\Delta_i = (C_{ti} - C_{yi})/C_{yi} \times 100 \tag{9}$$

where $C\Delta_i$ is the daily percentage change, C_{ti} is today's closing price, and C_{yi} is yesterday's closing price of the index.

In the second step, the movement of stock and index relative to each other is compared by covariance formula. The result of the covariance is then divided by the variance of the index. The formula for beta calculation is illustrated by the following equation.

$$\beta = COVARIANCE\left(C\Delta_s, C\Delta_i\right)/VAR\left(C\Delta_i\right) \tag{10}$$

This results in different beta values for different stocks. Stocks that have a beta value greater than 1 are more volatile, while stocks that have a beta value less than 1 are more stable (Gidofalvi and Elkan 2001).

4.5.3 Using fluctuations in stocks closing prices

The third method that we use to determine stock volatility is to find fluctuations in closing prices of stocks graphically. The closing prices of the selected stocks are plotted using *Matplotlib* python plotting library (Hunter 2007) to show fluctuations in closing prices. Stocks with more fluctuations in their closing prices will be more volatile and hence difficult to predict.

4.6 Identification of stock markets that are more influenced by social media and news

In this research, the sentiment-based approach is used to find those stock markets from the selected stocks that are more influenced by social media and news. Sentiments of tweets and news are found using the Stanford NLP approach for sentiment analysis. Stocks with more positive sentiments are considered to be more influenced by social media or news.

4.7 Application of deep learning in stock prediction

As neural networks perform best, we also use neural networks specifically MLP over stock data for prediction. We introduce deep learning by adding hidden layers in MLP. Hidden layers are increased by 1 at a time and prediction accuracies are noted to find the optimal number of hidden layers for this research problem.

4.8 Hybrid algorithm

Different ensemble methods are used for combining predictions of individual classifiers for getting superior performance. In this research, individual predictions of best classifiers are combined using the *voting* ensemble method to get the highest prediction accuracy. The *voting* ensemble method is selected because there are minor differences among the voting and other ensemble methods, and the former has shown best performance in terms of prediction accuracy (Kim et al. 2003).

We combine predictions of RF, ET, and GBM classifiers using the *majority voting* ensemble method. These are the classifiers that performed best in our prediction models. Python *VotingClassifier* class is used to implement *majority voting*. In this technique, the predicted class value for a specific sample is the class value that shows the majority of the class values predicted by each of the individual classifiers. For example, if the prediction for a given sample in our problem is RF → Positive, ET → Positive, GBM → Negative,

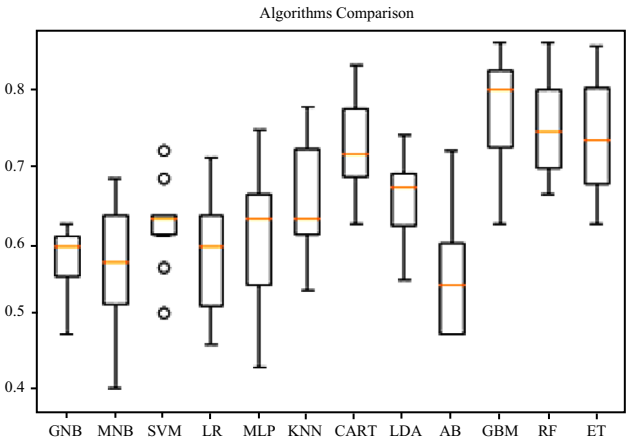

Fig. 4 Box plot distribution of the natural training dataset of HPQ, comparing accuracies of algorithms

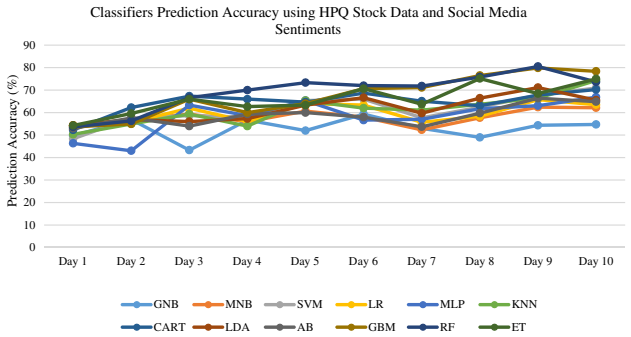

Fig. 5 Prediction accuracies of algorithms on the HPQ stock data and social sentiments for 10 days

the *VotingClassifier* will classify the sample as "Positive" based on the majority class value.

5 Experimental results and discussion

Like the proposed system, results are also divided into eight subsections. This section describes results of these sub-systems.

5.1 Stock prediction system

In this subsection, prediction results of the proposed algorithms on the HPQ stock market data set are presented.

5.1.1 Using social media

1. Results from the tenfold CV

Table 4 Classification performance of the algorithms on the testing data set of HPQ in terms of precision, recall, and F-measure

Classes	Metrics	Algorithms											
		GNB	MNB	SVM	LR	MLP	KNN	CART	LDA	AB	GBM	RF	ET
Positive	Precision (%)	66.00	62.00	67.00	69.00	66.00	77.00	72.00	72.00	65.00	83.00	81.00	75.00
	Recall (%)	54.00	100.00	96.00	82.00	84.00	67.00	77.00	86.00	98.00	85.00	89.00	73.00
	F-measure (%)	60.00	77.00	79.00	75.00	74.00	72.00	75.00	79.00	78.00	84.00	85.00	74.00
Neutral	Precision (%)	0.00	NA	NA	NA	NA	00.00	NA	NA	NA	NA	NA	NA
	Recall (%)	0.00	NA	NA	NA	NA	0.00	NA	NA	NA	NA	NA	NA
	F-measure (%)	0.00	NA	NA	NA	NA	0.00	NA	NA	NA	NA	NA	NA
Negative	Precision (%)	47.00	100.00	76 .00	57.00	53.00	57.00	59.00	68.00	82.00	75.00	79.00	58.00
	Recall (%)	54.00	2.00	23.00	40.00	30.00	65.00	53.00	47.00	16.00	72.00	67.00	61.00
	F-measure (%)	50.00	3.00	35.00	47.00	38.00	61.00	56.00	56.00	26.00	73.00	72.00	60.00

Performance comparison of the selected classification algorithms on the tenfold CV is shown in Fig. 4 over the HPQ data set.

The box plot shows performance of each algorithm on the tenfold CV in terms of variance and average accuracy on the training data set. The central line in each box shows the median, while the edges of the boxes show the 25th and 75th percentiles. The whiskers are extended to the most extreme data points that are not considered as outliers, while the outliers are shown individually using the solid circles. Figure 4 summarizes the overall accuracy measure of 12 algorithms before standardization and parameters tuning. Average accuracy is achieved in the 40–88% range across the 12 algorithms.

Clearly, the GBM classifier achieves the highest average accuracy of 76.50%, followed by ET and CART with average accuracies of 73.61 and 73.27%, respectively. RF shows relatively low performance with an average accuracy of 72.13%, followed by LDA, KNN, MLP, and SVM with average accuracies of 65.54, 64.88, 62.58, and 61.70%, respectively. LR, GNB, and MNB behave worse, and AB shows the lowest average accuracy of 55.85%.

2. Results from the independent testing data set

The accuracy of the testing data set is in the range 43.04–80.53%. Figure 5 shows the performance of different algorithms over the subsequent 10 days. Clearly, the RF classifier shows the highest accuracy of 80.53% on day 9, followed by the GBM classifier with an accuracy of 79.86%. MLP has the lowest accuracy of 43.04% on the 2nd day after the date on which the trade was executed.

Results for stock market future trends prediction using sentiments from social media show that the maximum prediction accuracy is attained on day 9 by RF followed by GBM on day 10. Without using the social media sentiment feature, an accuracy of 75.16% is achieved on day 9 by RF which shows a 5.37% decrease in prediction accuracy without using social media sentiments. The graph shows that the effect of sentiments on prediction accuracy is increasing gradually from day 1 to day 9.

It is clear from the results of training and testing data sets that GBM performs best on both data sets, while RF performs best on the testing data set. The improvement in the RF performance on the testing data set may be due to data set *standardization* and *parameters tuning*.

For comparing accuracy of the classifiers in classifying among the three future trend classes (positive, neutral, and negative), confusion matrices are created and three accuracy metrics: recall, precision, and F-measure, are found for each class for two prediction models. We show precision, recall, and F-measure for prediction models using social media and news only. Table 4 displays values for these metrics for the social media-based prediction model.

For the positive future trend class, maximum precision is acquired with GBM (83.00%), while highest recall is achieved with MNB (100.0%). However, the overall performance of RF is best (precision, recall, and F-measure are 81.00, 89.00, and 85.00%, respectively). The value of

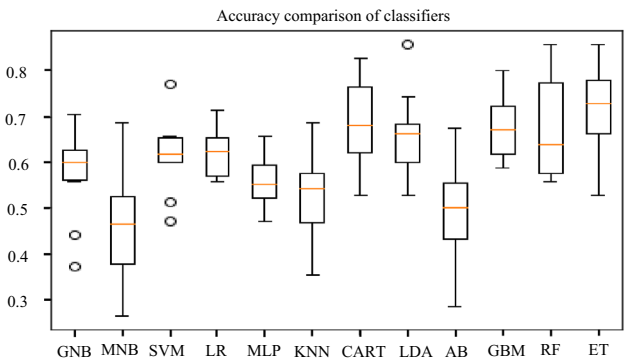

Fig. 6 Box plot distribution of the natural news training dataset, comparing accuracies of algorithms

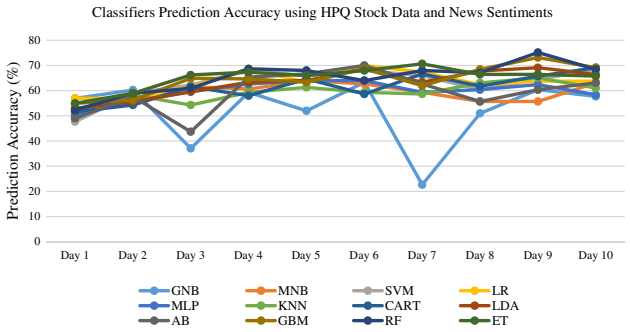

Fig. 7 Prediction accuracies of algorithms on the HPQ stock data and news sentiments for 10 days

precision is greater than 62.00% for all the algorithms, while the value of recall is greater than 67.00% except for the GNB (54.00%). For the negative future trend class, maximum precision is achieved by MNB (100.0%) but it shows very poor recall (2.00%) and an F-measure of (3.00%). Therefore, the algorithms show relatively low recall (2.00–72.00%) and F-measure (3.00–73.00%) for the negative future trend class. GBM shows relatively good performance in classifying the negative future trend class with precision, recall, and F-measure of 75.00, 72.00, and 73.00%, respectively. From Table 4, it is evident that the critical issue is the undesirable classification rate of the algorithms for the neutral future trend class. GNB and KNN show 0.00% precision, recall, and F-measure, while all the other algorithms give no precision, recall, and F-measure. Particularly, from the confusion matrix, it is evident that the most notable error source is the misclassification of the neutral future trend class into positive or negative future trend classes. The reason may be the small number of samples in this class. Some classifiers outperformed the others to a great extent although the data set is not balanced. Clearly, RF and GBM are both efficient in showing highest accuracy for the positive and negative future trend classes regardless of the imbalance data set.

5.1.2 Using financial news

1. Results from the tenfold CV

Performance comparison of the selected classification algorithms on the tenfold CV is shown in Fig. 6 for the HPQ data set. The box plot shows performance of each algorithm on the tenfold CV in terms of variance and average accuracy on the training data set. The figure summarizes the overall accuracy measures of 12 algorithms before *standardization* and *parameters tuning*. Average accuracy is achieved in the 25–82% range across the 12 algorithms.

On average, the RF classifier shows the highest average accuracy of 73.71%, followed by CART and ET classifiers with average accuracies of 70.14 and 69.05%, respectively. GBM shows relatively low performance with an average accuracy of 66.72%, followed by LDA, LR, and SVM classifiers with average accuracies of 65.86, 62.39, and 61.52%, respectively. GNB, MLP, KNN, and AB classifiers behave worse, and MNB has the lowest average accuracy of 46.74%.

2. Results from the independent testing data set

The results for stock predictions using financial news show that accuracy over the testing data set falls in the range of 22.66–75.16%. Figure 7 shows the performance of different classifiers over 10 subsequent days. Clearly, the RF classifier achieves the highest accuracy of 75.16% on day 9, followed by GBM with an accuracy of 73.15%. GNB has the lowest accuracy of 22.66% on day 7. The results also show that the maximum accuracy is reached on day 9 followed by day 8. Without using the news sentiment feature, an accuracy of 69.79% is achieved on day 9 by the RF, which shows a 5.37% decrease in prediction accuracy without using the news sentiments. This decrease in accuracy is the same as in the case of using the social media sentiment. This may be because of using the same technique for sentiment analysis of news and social media.

Table 5 Classification performance of the algorithms on the testing dataset of HPQ in terms of precision, recall, and F-measure

Classes	Metrics	Algorithms											
		GNB	MNB	SVM	LR	MLP	KNN	CART	LDA	AB	GBM	RF	ET
Positive	Precision (%)	64.00	56.00	61.00	61.00	60.00	69.00	70.00	66.00	58.00	74.00	75.00	68.00
	Recall (%)	65.00	100.00	99.00	98.00	95.00	70.00	67.00	90.00	100.00	82.00	86.00	77.00
	F-measure (%)	65.00	72.00	76.00	75.00	74.00	69.00	69.00	77.00	74.00	78.00	80.00	72.00
Neutral	Precision (%)	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	Recall (%)	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	F-measure (%)	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Negative	Precision (%)	57.00	0.00	93.00	88.00	78.00	60.00	61.00	78.00	100.00	72.00	76.00	64.00
	Recall (%)	55.00	0.00	22.00	22.00	22.00	60.00	65.00	43.00	11.00	63.00	63.00	54.00
	F-measure (%)	56.00	0.00	35.00	35.00	34.00	60.00	63.00	55.00	19.00	67.00	69.00	58.00

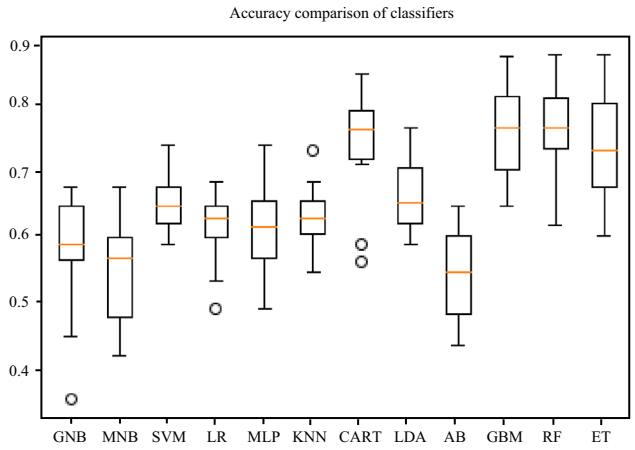

Fig. 8 Box plot distribution of the natural news and social media training dataset comparing accuracies of algorithms

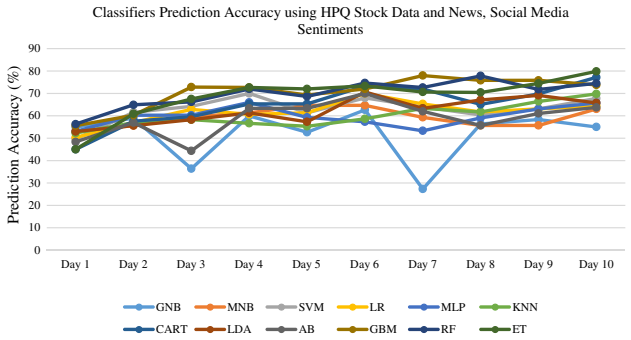

Fig. 9 Prediction accuracies of algorithms on the HPQ stock data and news, social media sentiments for 10 days

Table 5 shows precision, recall, and F-measure for the news-based prediction model. For the positive future trend class, the maximum precision is attained by the RF classifier (75.00%), while maximum recall is obtained by MNB (100.0%) and AB (100.00%). However, overall performance of the RF classifier is best with precision, recall, and F-measure of 75.00, 86.00, and 80.00%, respectively. The overall precision is above 56.00%, while the recall is above 77.00%, except for GNB (65.00%) and CART (67.00%).

For the negative future trend class, maximum precision is obtained with AB (100.0%), but it shows very poor recall (11.00%) and F-measure (19.00%). MNB shows the lowest precision, recall, and F-measure (0.0%), while highest recall achieved is 65.00% by CART. Therefore, the algorithms show relatively low recall (0.00–65.00%) and F-measure (0.00–69.00%) for the negative future trend class. Performance of the RF classifier is relatively good for the negative future trend class with precision, recall, and F-measure of 76.00, 63.00, and 69.00%, respectively.

Table 5 shows that the critical issue in the news-based prediction system is also the unsatisfactory classification performance of the algorithms for the neutral future trend class. All classifiers show no precision, recall, and F-measure for this class. In particular, it is proved from the confusion matrix that the most notable error source is the misclassification of neutral future trend class into positive or negative future trend classes. The reason in this case may also be the small number of samples in this class. Some classifiers show best performance compared to the other classifiers to a great extent for this classification problem. Clearly, RF and GBM are both efficient in showing highest accuracies for the positive and negative future trend classes, regardless of the imbalance data set.

From Tables 4 and 5, it is evident that the RF classifier also showed good performance in terms of precision, recall, and F-measure on the testing data sets, so it may be recommended for stock market prediction.

5.1.3 Using social media and financial news

1. Results from the tenfold CV

Performance comparisons of the selected classification algorithms on the tenfold CV are shown in Fig. 8 over the HPQ data set. The box plot shows performance of each algorithm on the tenfold CV in terms of variance and average accuracy on the training data set. The figure summarizes the overall accuracy measures of 12 algorithms before *standardization* and *parameters tuning*. Average accuracy is attained in the range of 42–85% across the 12 algorithms.

On average, the RF classifier shows the maximum average accuracy of 77.10%, followed by GBM, CART, and ET classifiers with average accuracies of 76.84, 74.20, and 73.99%, respectively. LDA shows relatively low performance with an average accuracy of 66.45%, followed by SVM, KNN, and MLP classifiers with average accuracies of 65.28, 63.31, and 62.10%, respectively. LR, GNB, and MNB classifiers behave worse, and AB has the lowest average accuracy of 53.78%.

2. Results from the independent testing data set

The results for stock prediction using news and social media sentiments show that accuracy over the testing data set fall into the range of 27.33–79.86%. Figure 9 shows performance of different classifiers for 10 days on the HPQ testing data set. The ET classifier achieves the highest accuracy of 79.86% on day 10, followed by GBM with accuracy of 78.0%. GNB has the lowest accuracy of 27.33% on day 7. Results also show that the maximum accuracy is reached on day 10.

The highest prediction performance of the RF algorithm is decreased, but overall prediction accuracies of most of the

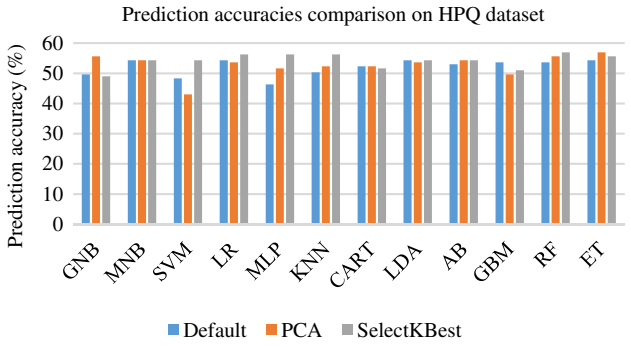

Fig. 10 Accuracies comparison before and after applying PCA and SelectKBest feature selection techniques on the HPQ stock dataset

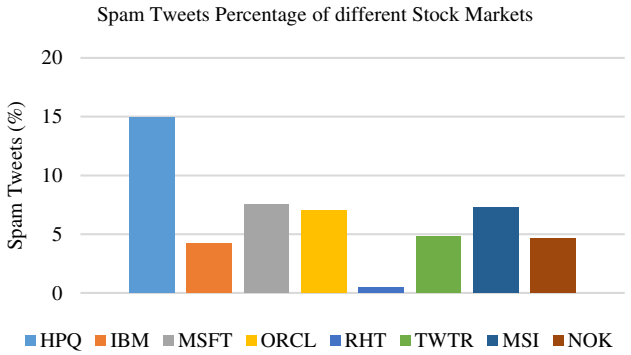

Fig. 11 Percentage of spam tweets for selected stock markets

algorithms are increased after day 3. Overall accuracy of the GNB is low, and a significant decrease in accuracy of GNB (27.33%) is observed on day 7.

The results of three prediction models discussed in previous subsections show the lowest performance of GNB and MNB on both the training and testing data sets. The lowest accuracy of GNB may be due to its assumption of *Gaussian distribution* of data because it performs best on data with *Gaussian distribution*. Similarly, MNB performs best on classification with discrete features, while our dataset is a mixture of nominal and discrete features, and therefore, causes a decrease in performance of the MNB.

5.2 Dimensionality reduction/feature selection

From results analysis, we find that accuracies of most of the classifiers (GNB, SVM, LR, MLP, KNN, AB, RF, and ET) improve by one or both of the feature selection methods, while GBM shows a decrease in accuracy, and MNB, CART, and LDA show no change in accuracy after dimensionality reduction/feature selection. The number of features/components on which maximum accuracy is achieved by both techniques are 6 out of 8 features in this problem as shown in Fig. 3. The best technique between SelectKBest and PCA for improving prediction accuracy is SelectKBest as illustrated in Fig. 10. From the results, it can be concluded that classifiers performance can be improved or the same performance can be achieved using a subset of features.

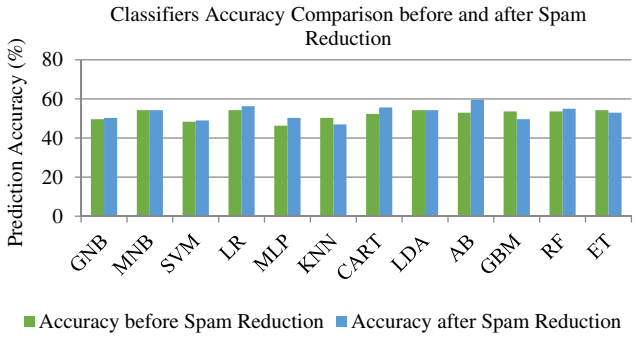

Fig. 12 Classifiers accuracy comparisons before and after spam reduction

5.3 Spam tweets reduction

The percentage split of spam and ham tweets show that the HPQ stock market is more influenced by spammers followed by MSFT stock, as shown in Fig. 11. About 14.97% of the HPQ tweets and 7.54% of the MSFT tweets are found to be spam. RHT is the stock that is least affected by spammers (0.49% spam tweets). Similarly, among the overall stock markets, LSE is found to be more influenced by spammers (spam tweets are 14.0%) while KSE is not affected at all. The reason for KSE may be that it is such a stock market that is not discussed very often as is evident from its tweets and news count given in Table 1.

After spam tweets reduction, prediction accuracies of most of the classifiers (GNB, SVM, LR, MLP, CART, AB, and RF) are improved, which indicate their robustness. Highest improvement in accuracy after spam reduction is shown by AB (6.62%). Similarly, accuracies of some of the classifiers (KNN, GBM, and ET) are decreased, while MNB and LDA show no change in their performance after spam reduction. Classifiers accuracy comparisons are given in Fig. 12 before and after spam reduction.

Similarly, accuracies of some of the classifiers (KNN, GBM, and ET) are decreased, while MNB and LDA showed no change in their performance after spam reduction. Classifiers accuracy comparisons are given in Fig. 12 before and after spam reduction.

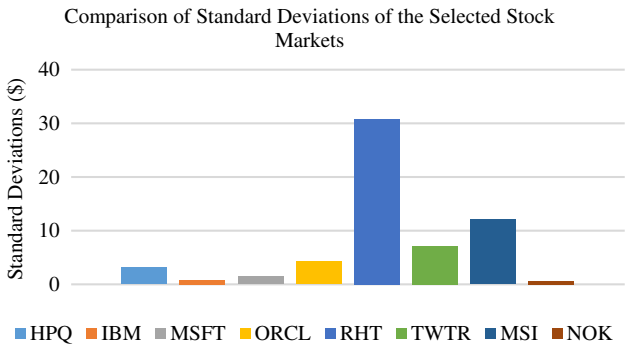

Fig. 13 Comparison of *standard deviations* for the selected stocks

Table 6 Beta values of the selected stock markets

Stock market	Beta value
HPQ	1.34
IBM	0.91
MSFT	1.33
ORCL	0.97
RHT	1.15
TWTR	0.88
MSI	0.79
NOK	0.97

5.4 Identification of a consistent classifier

From the results analysis of some subsystems discussed in previous subsections, we conclude that the best classifier that gives consistent results is RF for the following reasons.

- It gives highest prediction accuracy (80.53%) in stock prediction using social media.
- It shows highest prediction accuracy (75.16%) in stock prediction using financial news.
- Its prediction accuracy improves after feature selection for both SelectKBest and PCA techniques by 3.31 and 1.98%, respectively.
- Its prediction accuracy improves after spam reduction by 1.32%.
- It shows best performance in terms of classification accuracy and precision, recall, and F-measure.
- It outperforms on both training and testing data sets.

Best performance of the RF may be because our problem is a multiclass problem and RF is suitable for multiclass problems. The second reason may be that our data set contains a mixture of numerical (*Open*, *High*, *Low*, *Close*, *Sentiment*) and categorical (*Trend*, *Future Trend*) features, and RF works well with such types of data sets. Finally, RF is an ensemble learning method for classification problems, which can be used for boosting the accuracy.

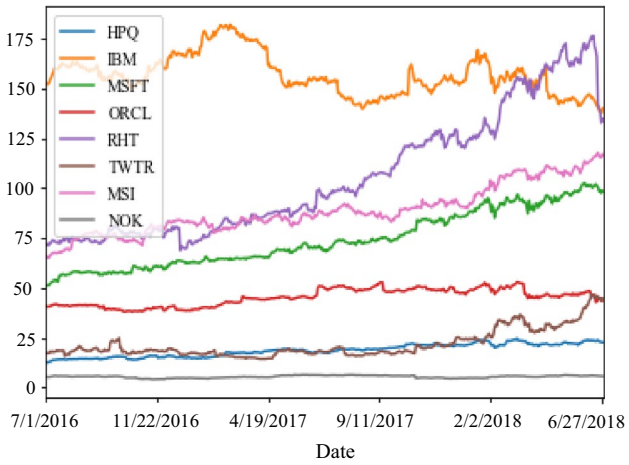

Fig. 14 Fluctuations in closing prices of selected stocks

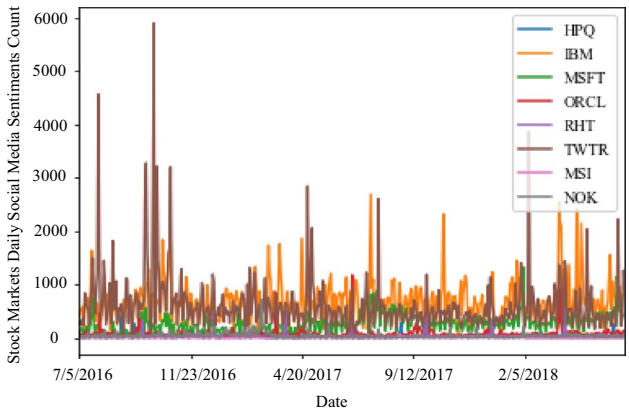

Fig. 15 Social media influence on selected stocks

Due to the consistent results of the RF classifier, the algorithm may be recommended for predicting trends in stock markets.

5.5 Identification of stock markets that are difficult to predict

5.5.1 Using variance and standard deviation

Standard deviations for selected stocks have been plotted in Fig. 13. The maximum standard deviation is that of RHT (30.78), which shows that RHT is a more volatile stock, and therefore difficult to predict. The second higher *standard deviation* is that of MSI (12.09), while NOK stock shows the lowest *standard deviation* of 0.60. Similarly, among the overall stock markets, NYSE is found to be hard to predict.

5.5.2 Using beta

Equation (10) gives β value for each stock as given in Table 6. Using this method, the β value of RHT (1.15) is

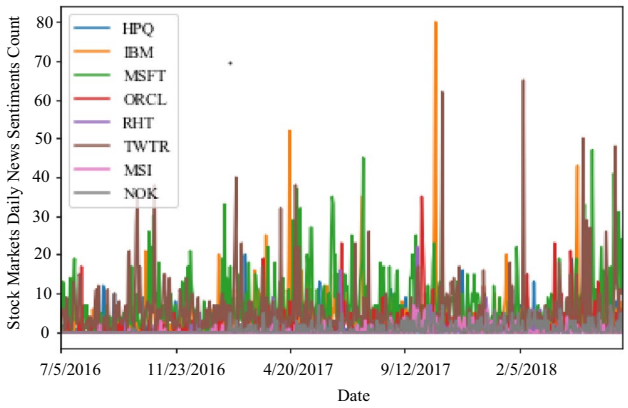

Fig. 16 News influence on selected stocks

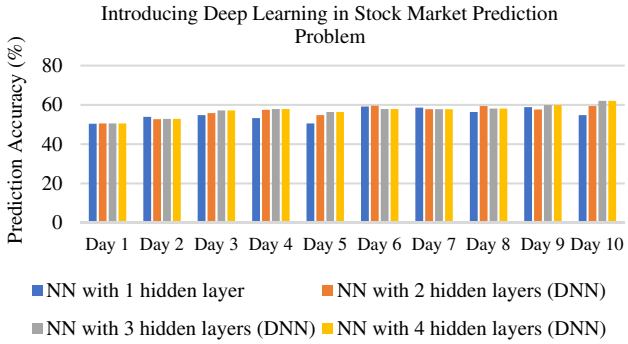

Fig. 17 Accuracy comparison on different hidden layers of the neural network

greater than 1, which shows that RHT stock market is more volatile and therefore difficult to predict. Similarly, HPQ and MSFT stocks have β values of 1.34 and 1.33 respectively, and therefore also volatile stocks and difficult to predict. Likewise, among the overall stock markets, NYSE is found to have β value of 1.01 and therefore hard to predict.

5.5.3 Using fluctuations in stock closing prices

Figure 14 graphically shows maximum fluctuation in closing price of RHT. According to this method too, RHT stock market is deemed difficult to predict. Similarly, IBM and MSI stock markets exhibit maximum fluctuation and therefore are difficult to predict. The lowest fluctuation is shown by NOK stock, followed by HPQ. Among the overall stock markets, NYSE have shown maximum fluctuation in closing price and is therefore difficult to predict.

Results analysis of all the methods for stock volatility prediction shows that β method for finding volatility exhibit some conflicting results. For example, according to β method, HPQ is volatile stock, while it shows lowest fluctuation. It shows the fact that β is seldom used for identifying volatile stock markets.

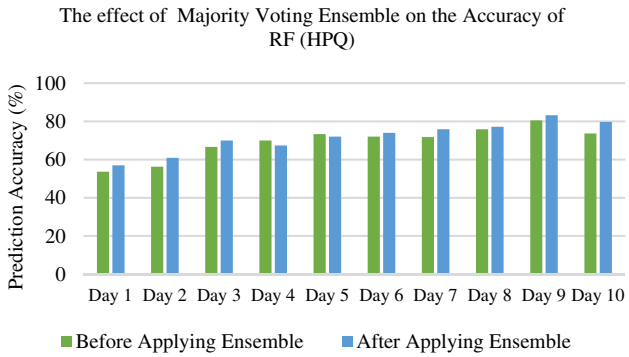

Fig. 18 The effect of *VotingClassifier* ensemble on the accuracy of RF

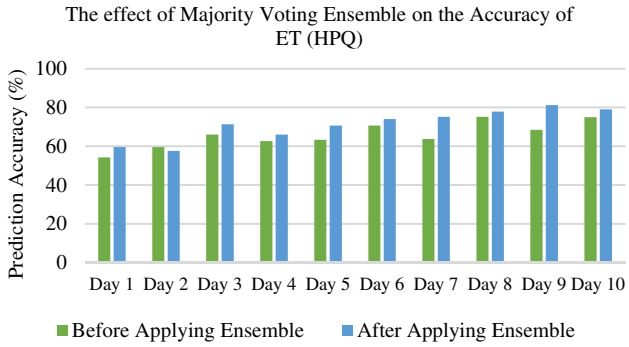

Fig. 19 The effect of *VotingClassifier* ensemble on accuracy of ET

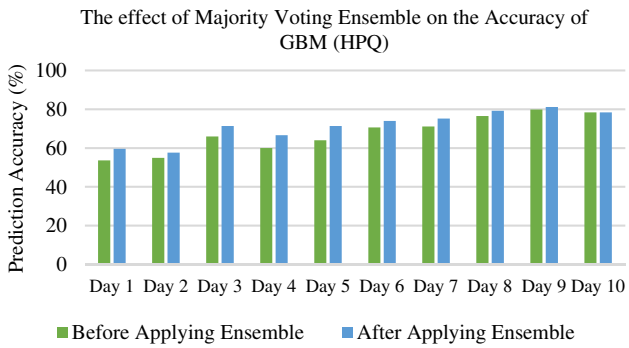

Fig. 20 The effect of *VotingClassifier* ensemble on accuracy of GBM

5.6 Identification of stock markets that are more influenced by social media and news

Our experimental results show that IBM stock market is more influenced by social media, followed by TWTR, while MSI and ORCL stocks are least influenced, as can be seen from the spikes shown in Fig. 15. Similarly, among the overall stock markets, NYSE is found to be more influenced by social media.

Experimental results on news sentiments show that MSFT stock is more influenced by news, followed by TWTR while RHT, MSI, and NOK stocks are least influenced by news as can be seen from the spikes shown in Fig. 16. Similarly, LSE is more influenced by news among the overall stock markets.

From the analysis of both results, we can conclude that TWTR stock market is influenced by both social media and news.

5.7 Application of deep learning in stock prediction

By applying deep learning, prediction accuracy of the neural network (MLP) increases up to 7.3%. Prediction accuracy gradually increases by increasing the number of hidden layers up to 3 as shown in Fig. 17. The neural network accuracy decreases only on day 2 and day 7 after increasing hidden layers. Its performance does not improve by using 4 hidden layers, which shows that the optimal number of hidden layers for this problem are 3.

5.8 Hybrid algorithm

When individual predictions of RF, ET, and GBM classifiers were combined, their prediction accuracies improved. Prediction accuracies of RF are plotted before and after applying ensembles over the HPQ social media final data set as shown in Fig. 18. The highest prediction accuracy of RF classifier increases from 80.53 to 83.22% on day 9 after applying the voting ensemble method.

Similarly, the highest prediction accuracy of the ET classifier improves from 75.16 to 81.2% on day 9 (Fig. 19), which is a significant increase in accuracy. Prediction accuracy of ET decreases only on day 2.

Lastly, maximum prediction accuracy of GBM improves from 79.86 to 81.2% on day 9, as shown in Fig. 20.

From results analysis, it can be concluded that ensemble methods enhance classifier prediction accuracy and can be used in any field, including stock prediction, for boosting accuracies of individual classifiers.

6 Conclusion and future work

This research presents a framework for stock market future trends prediction using news and social media as external factors. We examined the effect of social media and financial news on stock prediction for 10 days in future. By including sentiment attributes, we found that the social media has more influence in stock prediction on day 9, while financial news show its greater effects on day 9 and then on day 8. We also concluded that by combining sentiments of social media and financial news, the highest accuracy decreased but the overall accuracies of most of the classifiers increased after day 3. We presented different aspects of the data and the algorithms used for prediction. More specifically, we examined the effect of feature selection and spam tweets reduction on prediction performance of the algorithms and found that there is a positive effect of feature selection and spam tweets reduction on the performance of most of the classifiers. Moreover, we examined each aspect of the selected classifiers and found that RF gives consistent results in all the cases and therefore it is recommended for stock trends prediction. Selected stocks behavior was also examined using different techniques, and it was found that NYSE and RHT are more volatile stocks and therefore difficult to predict. Similarly, HPQ, MSFT, and IBM stocks were also found to be volatile and difficult to predict. The effect of social media and news is also explored, and it is proposed that NYSE, IBM, and TWTR stocks are more influenced by social media, while LSE and MSFT stocks are found to be more influenced by news. Similarly, TWTR stock is found to be influenced by both news and social media. The application of deep learning in stock prediction showed improvement in neural network performance in terms of prediction accuracy. Lastly, the ensemble of predictions of individual classifiers using voting ensemble method showed an improvement in the performance of individual classifiers in terms of prediction accuracy.

For future study, the use of a more systematic technique for determining stock relevant keywords for searching social media and news will result in obtaining more quality results for stock market prediction. Another possible direction for future study is to use other social media data, such as Google + and Facebook, and to compare their effects on the stock market prediction.

References

Afzal H, Mehmood K (2016) Spam filtering of bi-lingual tweets using machine learning. In: IEEE 18th international conference on ICACT, pp 710–714

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

Al-Zoubi A, Faris H (2017) Spam profile detection in social networks based on public features. In: IEEE 8th international conference ICICS, pp 130–135

Attigeri GV, MM MP, Pai RM, Nayak A (2015) Stock market prediction: a big data approach. In: IEEE region 10 conference on TENCON, pp 1–5

Bastianin A, Manera M (2018) How does stock market volatility react to oil price shocks? Mach Dyn 22(3):666–682

Blum C, Li X (2008) Swarm intelligence in optimization. In: Dorigo M (ed) Swarm intelligence. Springer, Berlin, pp 43–85

Brezočnik L, Fister I, Podgorelec V (2018) Swarm intelligence algorithms for feature selection: a review. Appl Sci 8(9):1521

Brown GW, Cliff MT (2004) Investor sentiment and the near-term stock market. J Empir Financ 11(1):1–27

Cao J, Cui H, Shi H, Jiao L (2016) Big data: a parallel particle swarm optimization-back-propagation neural network algorithm based on MapReduce. PLoS ONE 11(6):e0157551

Chakraborty P, Pria US, Rony M, Majumdar MA (2017) Predicting stock movement using sentiment analysis of Twitter feed. In: IEEE 6th international conference ICIEV-ISCMHT, pp 1–6

Chen W, Yeo CK, Lau CT, Lee BS (2017a) A study on real-time low-quality content detection on Twitter from the users' perspective. PLoS ONE 12(8):e0182487

Chen W, Zhang Y, Yeo CK, Lau CT, Lee BS (2017b) Stock market prediction using neural network through news on online social networks. In: IEEE international ISC2, pp 1–6

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

Cheng S, Shi Y, Qin Q, Bai R (2013) Swarm intelligence in big data analytics. In: International conference on intelligent data engineering and automated learning. Springer, Berlin, pp 417–426

Chhikara RR, Sharma P, Singh L (2018) An improved dynamic discrete firefly algorithm for blind image steganalysis. Int J Mach Learn Cybern 9(5):821–835

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

Dang M, Duong D (2016) Improvement methods for stock market prediction using financial news articles. In: IEEE 3rd national foundation for science and technology development conference on information and computer science (NICS), pp 125–129

Dang LM, Sadeghi-Niaraki A, Huynh HD, Min K, Moon H (2018) Deep learning approach for short-term stock trends prediction based on two-stream gated recurrent unit network. IEEE Access 6:55392–55404

Dorigo M (1992) Learning and natural algorithms. Ph.D. Thesis, Politecnico di Milano, Milano, Italy

Džeroski S, Ženko B (2004) Is combining classifiers with stacking better than selecting the best one? J Mach Learn 54(3):255–273

Eberhart R, Kennedy J (1995) Particle swarm optimization. In: Proceedings of the IEEE international conference on neural networks, pp 1942–1948

Enache AC, Sgarciu V, Petrescu-Niţă A (2015) Intelligent feature selection method rooted in Binary Bat Algorithm for intrusion detection. In: 2015 IEEE 10th Jubilee international symposium on applied computational intelligence and informatics. IEEE, pp 517–521

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

Hajdu A, Hajdu L, Jonas A, Kovacs L, Toman H (2013) Generalizing the majority voting scheme to spatially constrained voting. IEEE Trans Image Proc 22(11):4182–4194

Hassanien AE, Emary E (2016) Swarm intelligence: principles, advances, and applications. CRC Press, Boca Raton

Hastie T, Tibshirani R, Friedman J (2009) The elements of statistical learning: data mining, inference, and prediction, 2nd edn. Springer, New York, p 745

He K, Zhang X, Ren S, Sun J (2016) Deep residual learning for image recognition. In: Proceedings of IEEE conference on CVPR 2016, pp 770–778

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

Hentschel M, Alonso O (2014) Follow the money: a study of cashtags on Twitter. First Monday 19(8). https://doi.org/10.5210/fm.v19i8.5385

Hu Z, Chiong R, Pranata I, Susilo W, Bao Y (2016) Identifying malicious web domains using machine learning techniques with online credibility and performance data. In: 2016 IEEE congress on evolutionary computation (CEC). IEEE, pp 5186–5194

Hunter JD (2007) Matplotlib: a 2D graphics environment. Comput Sci Eng 9(3):90–95. https://doi.org/10.1109/MCSE.2007.55

Ibrahim RA, Ewees AA, Oliva D, Elaziz MA, Lu S (2019) Improved salp swarm algorithm based on particle swarm optimization for feature selection. J Ambient Intell Humaniz Comput. https://doi.org/10.1007/s12652-018-1031-9

Jayaraman V, Sultana HP (2019) Artificial gravitational cuckoo search algorithm along with particle bee optimized associative memory neural network for feature selection in heart disease classification. J Ambient Intell Humaniz Comput. https://doi.org/10.1007/s12652-019-01193-6

Jeon S, Hong B, Chang V (2018) Pattern graph tracking-based stock price prediction using big data. J Future Gener Comput Syst. https://doi.org/10.1016/j.future.2017.02.010

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

Karaboga D (2005) An idea based on honey bee swarm for numerical optimization. Technical report-tr06, Erciyes University, Engineering Faculty, Computer Engineering Department, vol 200, pp 1–10

Khan W, Malik U, Ghazanfar MA, Azam MA, Alyoubi KH, Alfakeeh AS (2019) Predicting stock market trends using machine learning algorithms via public sentiment and political situation analysis. Soft Comput. https://doi.org/10.1007/s00500-019-04347-y

Khare K, Darekar O, Gupta P, Attar VZ (2017) Short term stock price prediction using deep learning. In: 2nd IEEE international conference RTEICT, pp 482–486

Khatri SK, Srivastava A (2016) Using sentimental analysis in prediction of stock market investment. In: IEEE 5th international conference ICRITO, pp 566–569

Kim E, Kim W, Lee Y (2003) Combination of multiple classifiers for the customer's purchase behavior prediction. J Decis Support Syst 34(2):167–175

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

Kuhn M, Johnson K (2013) Applied predictive modeling. Springer, New York

Kumar PH, Patil SB (2015) Volatility forecasting using machine learning and time series techniques. IJRCCE 3(9):8284–8292

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

Li X (2003) A new intelligent optimization-artificial fish swarm algorithm. Ph.D. Thesis, Zhejiang University, Hangzhou, China

Li Q, Wang T, Li P, Liu L, Gong Q, Chen Y (2014a) The effect of news and public mood on stock movements. J Inf Sci 278:826–840. https://doi.org/10.1016/j.ins.2014.03.096

Li X, Huang X, Deng X, Zhu S (2014b) Enhancing quantitative intraday stock return prediction by integrating both market news and stock prices information. J Neuro Comput 142:228–238

Li X, Xie H, Chen L, Wang J, Deng X (2014c) News impact on stock price return via sentiment analysis. J Knowl-Based Syst 69:14–23. https://doi.org/10.1016/j.knosys.2014.04.022

Li J, Bu H, Wu J (2017) Sentiment-aware stock market prediction: a deep learning method. In: IEEE international conference ICSSSM, pp 1–6

Liu R, Li W, Liu X, Lu X, Li T, Guo Q (2018) An ensemble of classifiers based on positive and unlabeled data in one-class remote sensing classification. IEEE J Sel Top Appl Earth Obs Remote Sens 11(2):572–584

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

Mirjalili S, Mirjalili SM, Lewis A (2014) Grey wolf optimizer. J Adv Eng Softw 69:46–61

Mirjalili S, Gandomi AH, Mirjalili SZ, Saremi S, Faris H, Mirjalili SM (2017) Salp Swarm Algorithm: a bio-inspired optimizer for engineering design problems. Adv Eng Softw 114:163–191

Mohammadi FG, Abadeh MS (2014) Image steganalysis using a bee colony based feature selection algorithm. J Eng Appl Artif Intell 31:35–43

Moslehi F, Haeri A (2019) A novel hybrid wrapper–filter approach based on genetic algorithm, particle swarm optimization for feature subset selection. J Ambient Intell Humaniz Comput. https://doi.org/10.1007/s12652-019-01364-5

Noda K, Yamaguchi Y, Nakadai K, Okuno HG, Ogata T (2015) Audio-visual speech recognition using deep learning. J Appl Intell 42(4):722–737

Omer NAB, Halim FA (2015) Modelling volatility of Malaysian stock market using garch models. In: IEEE international symposium iSMSC, pp 447–452

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

Passino KM (2002) Biomimicry of bacterial foraging for distributed optimization and control. IEEE Control Syst Mag 22:52–67

Pedregosa et al (2011) Scikit-learn: machine learning in Python. JMLR 12:2825–2830

Qasem M, Thulasiram R, Thulasiram P (2015) Twitter sentiment classification using machine learning techniques for stock markets. In: IEEE international conference on ICACCI, Kochi, India, pp 834–840

Saraç E, Özel SA (2014) An ant colony optimization based feature selection for web page classification. Sci World J 2014:649260. https://doi.org/10.1155/2014/649260

Sattiraju M, Manikantan K, Ramachandran S (2013) Adaptive BPSO based feature selection and skin detection based background removal for enhanced face recognition. In: 2013 4th national conference on computer vision, pattern recognition, image processing and graphics (NCVPRIPG). IEEE, pp 1–4

Sedhai S, Sun A (2015) HSpam14: a collection of 14 million tweets for hashtag-oriented spam research. In: 38th ACM conference on SIGIR, pp 223–232

Sedhai S, Sun A (2018) Semi-supervised spam detection in Twitter stream. IEEE Trans Comput Soc Syst 5(1):169–175

Seth JK, Chandra S (2016) Intrusion detection based on key feature selection using binary GWO. In: 2016 3rd international conference on computing for sustainable global development (INDIACom). IEEE, pp 3735–3740

Shen S, Jiang H, Zhang T (2012) Stock market forecasting using machine learning algorithms. Department of Electrical Engineering, Stanford University, Stanford, pp 1–5

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 2013 conference on empirical methods in natural language processing, pp 1631–1642

Sun J, Li H (2012) Financial distress prediction using support vector machines: ensemble vs. individual. J Appl Soft Comput 12(8):2254–2265

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

Thu HLT, Marrero-Ponce Y, Cansañola-Martin GM, Cardoso GC, Chávez MC, Garcia MM, Morell C, Torrens F, Abad C (2011) A comparative study of nonlinear machine learning for the "in silico" depiction of tyrosinase inhibitory activity from molecular structure. Mol Inform 30(6–7):527–537

Tirea M, Negru V (2015) Text mining news system-quantifying certain phenomena effect on the stock market behavior. In: IEEE 17th international symposium on SYNASC, pp 391–398

Todorovski L, Džeroski S (2003) Combining classifiers with meta decision trees. J Mach Learn 50(3):223–249

Tsai CF, Lin YC, Yen DC, Chen YM (2011) Predicting stock returns by classifier ensembles. J Appl Soft Comput. https://doi.org/10.1016/j.asoc.2010.10.001

Urolagin S (2017) Text mining of tweet for sentiment classification and association with stock prices. In: IEEE ICCA, pp 384–388

Usmani M, Adil SH, Raza K, Ali SA (2016) Stock market prediction using machine learning techniques. In: IEEE 3rd international conference on ICCOINS, pp 322–327

Vargas MR, dos Anjos CEM, Bichara GLG, Evsukoff AG (2018) Deep learning for stock market prediction using technical indicators and financial news articles. In: IEEE international joint conference IJCNN, pp 1–8

Wang G, Dai D (2013) Network intrusion detection based on the improved artificial fish swarm algorithm. J Comput 8(11):2990–2996

Wang F, Zhao Z, Li X, Yu F, Zhang H (2014) Stock volatility prediction using multi-kernel learning based extreme learning machine. In: IEEE joint conference IJCNN, pp 3078–3085

Wang H, Jing X, Niu B (2016) Bacterial-inspired feature selection algorithm and its application in fault diagnosis of complex structures. In: 2016 IEEE congress on evolutionary computation (CEC). IEEE, pp 3809–3816

Yan D, Zhou G, Zhao X, Tian Y, Yang F (2016) Predicting stock using microblog moods. J China Commun 13(8):244–257

Yang X-S (2008) Firefly algorithm. In: Nature-inspired metaheuristic algorithms. Luniver Press, Beckington, pp 128

Yang XS (2010) A new metaheuristic bat-inspired algorithm. In Nature inspired cooperative strategies for optimization (NICSO 2010) Springer, Berlin, pp 65–74

Yang XS, Deb S (2009) Cuckoo search via Lévy flights. In: 2009 world congress on nature & biologically inspired computing (NaBIC). IEEE, pp 210–214

Yetis Y, Kaplan H, Jamshidi M (2014) Stock market prediction by using artificial neural network. In: IEEE WAC, pp 718–722

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

Zhong X, Enke D (2016) Forecasting daily stock market return using dimensionality reduction. Exp Syst Appl 67:126–139. https://doi.org/10.1016/j.eswa.2016.09.027

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