Chapter

Recent Advances in Stock Market Prediction Using Text Mining: A Survey

Faten Subhi Alzazah and Xiaochun Cheng

Abstract

Market prediction offers great profit avenues and is a fundamental stimulus for most researchers in this area. To predict the market, most researchers use either technical or fundamental analysis. Technical analysis focuses on analyzing the direction of prices to predict future prices, while fundamental analysis depends on analyzing unstructured textual information like financial news and earning reports. More and more valuable market information has now become publicly available online. This draws a picture of the significance of text mining strategies to extract significant information to analyze market behavior. While many papers reviewed the prediction techniques based on technical analysis methods, the papers that concentrate on the use of text mining methods were scarce. In contrast to the other current review articles that concentrate on discussing many methods used for forecasting the stock market, this study aims to compare many machine learning (ML) and deep learning (DL) methods used for sentiment analysis to find which method could be more effective in prediction and for which types and amount of data. The study also clarifies the recent research findings and its potential future directions by giving a detailed analysis of the textual data processing and future research opportunity for each reviewed study.

Keywords: machine learning, deep learning, natural language processing, sentiment analysis, stock market prediction

1. Introduction

Stock market prediction aims to determine the future movement of the stock value of a financial exchange. The accurate prediction of share price movement will lead to more profit investors can make. Predicting how the stock market will move is one of the most challenging issues due to many factors that involved in the stock prediction, such as interest rates, politics, and economic growth that make the stock market volatile and very hard to predict accurately. The prediction of shares offers huge chances for profit and is a major motivation for research in this area; knowledge of stock movements by a fraction of a second can lead to high profits [1]. Since stock investment is a major financial market activity, a lack of accurate knowledge and detailed information would lead to an inevitable loss of investment. The prediction of the stock market is a difficult task as market movements are always subject to uncertainties [2]. Stock market prediction methods are divided into two

main categories: technical and fundamental analysis. Technical analysis focuses on analyzing historical stock prices to predict future stock values (i.e. it focuses on the direction of prices). On the other hand, fundamental analysis relies mostly on analyzing unstructured textual information like financial news and earning reports. Many researchers believe that technical analysis approaches can predict the stock market movement [3–5]. In general, these researches did not get high prediction results as they depend heavily on structured data neglecting an important source of information that is the online financial news and social media sentiments. These days more and more critical information about the stock market has become available on the Web. Examples include BBC, Bloomberg, and Yahoo Finance. It is hard to manually extract useful information out of these resources. This draws a picture of the significance of text mining techniques to automatically extract meaningful information for analyzing the stock market. In this research, the most crucial past literature was reviewed, and a major contribution was made to the subject of using text mining and NLP for market prediction.

We revealed the finding of the selected studies to show the significantly improved performance of stock market forecasting via many machine learning methods. This study also clarifies the recent innovation researches and its potential future contribution. Comparisons and analyses of different researches are made on the financial domain of market prediction that can help to establish potential opportunities for future work. In this research, we also focused on the promising results accomplished by machine learning methods for analyzing the stock market using text mining and natural language processing (NLP) techniques.

In contrast to the other current survey articles that concentrate on summarizing many methods used for forecasting the stock market, we aim to compare many machine learning (ML) and deep learning (DL) methods used for sentiment analysis task of social media and financial news articles to find which method could be more effective in prediction. **Figure 1** represents the reviewed study framework. The rest of this work is organized as follows. Section 2 provides a review of background concepts that are needed to be known before the detailed analysis of the literature. Section 3 illustrates the relationship between stock market prediction and text mining. Section 4 includes a review of the machine learning main methods used for stock market prediction based on textual resources. Section 5 explained the least frequently used algorithm for stock prediction based on text mining. Section 6 describes the reviewed work text sources and period and number of collected items. Section 7 contains the reviewed works finding, limitation, the measurement used, and future work. Finally, Section 8 concludes this paper.

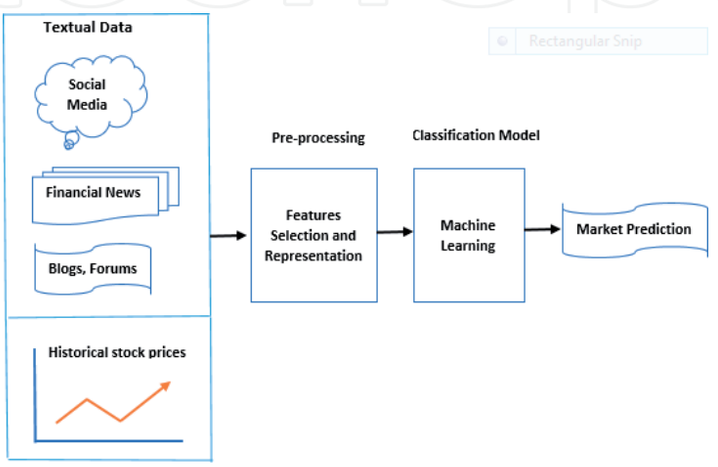

Figure 1.
The reviewed study framework.

2. A review of background concepts

Our work defined the following concepts as important to understand this research topic.

2.1 Sentiment analysis

Sentiment analysis uses text mining, natural language processing, and computational techniques to automatically extract sentiments from a text [6]. It aims to classify the polarity of a given text at the sentence level or class level, whether it reflects a positive, negative, or neutral view [7]. In stock market prediction task, two important sources of the text are used either social media mainly using Twitter data or online financial news article.

2.1.1 Twitter sentiment

Twitter is a significant source of data, and many researchers have examined its relationship with stock market movements [8]. While each tweet is restricted to 140 characters, it is believed that the information can accurately reflect public mood [9].

2.1.2 Online financial news sentiment

Financial news articles are perceived to be a more consistent and reliable source of information. Many researchers suggested that the financial news articles have a strong relationship with stock market fluctuation; therefore, analyzing financial news reports can help in predicting the stock market movements [10]. In [11], the author used a unified latent space model to examine the relationship between stock prices and news article releases. The result indicates a good return accuracy, which proves that news article analysis has an important impact on stock market movement.

2.2 Textual data preprocessing

Textual data need to be prepared before used by the machine learning algorithm for sentiment analysis task using these methods.

2.2.1 Feature extraction

Feature extraction or sometimes called attribute selection aim to select features, attributes, or piece of text that is more relevant to the prediction task. Many methods have been used for feature selection. The commonly used feature selection procedure for document or sentence classification task is the bag-of-words (BOW) approach, which was recently used for market prediction by many authors [12–14]. In the mentioned model, each word in a text or document will be treated as a feature neglecting the grammar or word order and only preserving the abundance. The second most popular method used recently for the feature selection process is Word2vec [12]. In this technique, the aim is to learn word embedding using a two-layer neural network. The input to that neural network is a text, and the output is a group of vectors (i.e. the input is a corpus and the output is a vector of words).

Another important feature selection method is the latent Dirichlet allocation (LDA) technique used recently for market prediction in [13]. In the LDA model, the text is viewed as probabilistic collections of terms or words, and the collections are then treated as selected features. Other researches [12, 14] used a Skip-Gram model

that aims to predict the context word (surrounding words) for a given target word. However, feature selection is a crucial step in the textual data preprocessing, and many other strategies may also be used for text analysis.

2.2.2 Feature representation

After feature selection, every feature must be illustrated by a numeric value so that it can be analyzed by machine learning techniques. The most common technique of feature representation is a binary representation (BR), which is a number system that uses two values such as 0 and 1 exclusively to represent the information. This technique has been exploited for market prediction researches by many authors [15–17]. The second most popular method used in text mining for financial application is the term frequency-inverse document frequency (TF-IDF), which is a numeric value that represents the significance of a word for a document or corpus that is used recently by many authors [12, 18]. Other feature representation methods can also be used successfully in text preprocessing, and we will discuss those with more details in the following sections.

3. The relationship between stock market prediction and text mining

Many papers study the relationship between stock price movements and the market sentiments, and the most relevant studies will be discussed in this section.

Ref. [19] examined the ability to use sentiment polarity (positive and negative) and sentiment emotions selected from financial news or tweets to predict the market movements. For sentiment analysis, they have collected a large dataset of the top 25 historical financial news headlines in addition to a large set of financial tweets collected from Twitter. Furthermore, they collected stock historical price data for many S&P 500 companies and used the close price as an indicator of the stock movements. For evaluation, they used the Granger causality test [20] that is a statistical test technique commonly used to reveal causality in time series data and explore if one-time series data can predict the other. For sentiment analysis, the authors examined two machine learning methods SVM and LSTM. The experiment result illustrated that in some cases sentiment emotions contribute to Granger-cause stock price fluctuation, but the finding was not inclusive and must be examined for each case. Also, it has been revealed that for some stocks, adding sentiment emotions to the machine learning market prediction model will increase the prediction accuracy. Comparing the two machine learning methods, SVM achieved better and more balanced results, and that's because the size of the dataset is quite small to be sufficiently used with SVM.

Another paper [21] examined the efficiency of using sentiment analysis of microblogging sites to forecast the stock price returns, volatility, and trading volume. The extracted intraday data from the two sources of information, Twitter and StockTwits, were collected for 2 years. For the evaluation, the authors used five famous stocks, namely, Amazon, Apple, Goldman Sachs, Google, and IBM. Prices were represented every 2 min, and the sentiment data were collected for the same period span of each trading day. To find the links between stock price outcomes and tweet sentiment, they applied Granger causality analysis. The experiments indicate that there is a causal link between Twitter sentiments and stock market returns, volatility, and volume. Among all five stocks, market volatility and volume seem to be more predictable than market direction or return.

In [22], the author exploited a multiplex network approach to study the correlation between market movements and social media sentiments. The proposed

model merges information from two sources of data: Twitter posts and market price data. The authors selected 100 of the biggest capitalized companies of the S&P 500 index for a 5-year period from May 2012 to August 2017. In their model, they suggested that financial network correlation was established by the integration of the two techniques. The first one suggests that two stocks tend to be associated if they share joint neighbors. The other techniques suggest that two connected stocks usually remain connected in the future. The findings demonstrated that a multiplex network approach incorporating information from both social media and financial data can be used to forecast a causal relationship framework with high accuracy.

The authors in [23] investigated the ability of economic news to predict Taiwan stock market returns. The proposed model used text mining techniques throughout many steps. Firstly, they converted the textual news into numerical values. Secondly, they append the resulting numerical variable to regression models with macroeconomic attributes to examine the role of news articles in predicting stock price returns. The model also defines specific keywords and calculates the number of positive, negative, and neutral words in each news text and then converts them into three news attributes, which are then fed to the regression model. The experiments find that adding news articles was able to reduce the root mean square error (RMSE) that proves that the economic news has crucial impacts on market returns. The experiments also indicate that negative news has more influence on the stock market returns than positive news articles.

The study proposed in [24] aims to analyze whether tweet messages could be used to predict future trends of stocks for particular companies listed on the Dow Jones stock market, focusing on 12 companies related to 3 distinct and crucial economic branches in technology, services, and health care. The authors gathered the company's market data and Twitter posts for a 70-day period for analysis. The companies of each category were chosen based on the volume of messages that mention the company names on the StockTwits website. The study illustrates that some of the proposed ad hoc forecasting models well predict the next day direction of the stock movements for some companies with 82% of success and there is no unified method to be used with all cases. The results also indicate that more volume of a tweet will yield better prediction results. Moreover, the study proved the robust correlation between tweet's posts and the trend movements for some companies.

Overall, past studies indicate that there is a strong relationship between market movements and information published in news and social media. The information on social media contributes to enhancing the prediction models with all of the discussed papers. The evaluation of event sentiment may affect the market returns further and boost the outcome of forecasting.

4. Machine learning for market prediction

Recently, many research studies used machine learning via text mining innovation methods to successfully predict the stock market changes, and the most significant ones are going to be discussed in this section.

4.1 Support vector machines

Support vector machines (SVMs) are a supervised machine learning model used extensively in classification and regression tasks. SVM is a hyperplane that divides a collection of documents into two or more classes with a maximum margin [25].

SVM was first applied to the text classification task by Joachims [26]. In his approach, the author used a limited vocabulary as the feature collection by using

a list of the most occurred words and discard of uncommon words from the feature set. Utilizing 12,902 documents from the Reuters-21578 document group and 20,000 medical summaries, the author compared the effectiveness of many machine learning techniques such as SVM and Naive Bayes (NB). For both document groups, the experiments demonstrated that the SVM achieve better classification result compared to NB classifier.

For stock market prediction, many research papers used the SVM for text classification and sentiment analysis. Combining both textual information and historical stock prices for stock market prediction [27] research applied the SVM to forecast the Chinese stock direction and stock prices between the years 2008 and 2015. For text mining, the authors formed a stop word and sentiment dictionary based on a specific domain. In the study, there were two kinds of input. The first one includes 2,302,692 news items, whereas the other contains only stock data of the largest 20 Chinese stocks based on trading volume. Support vector regression (SVR) is used to predict stock price, and support vector classification (SVC) is exploited to predict stock direction. The result indicates that both audience numbers and news quality have a crucial impact on the stock market. Moreover, for SVC, the direction accuracy was 59.1734%, which illustrates better progress than other works. The result also indicates that news articles have an important effect on the stock market fluctuations.

Another research [28] introduced a stock market prediction framework. For sentiment analysis, the researchers used two financial sentiment dictionary, namely, the Harvard IV-4 sentiment dictionary (HVD) and Loughran and McDonald (LMD) [29] financial dictionary. The dataset consists of 5 years of historical Hong Kong Stock Exchange prices and financial news collected from January 2003 to March 2008. For text classification SVM was used for training. Experiments indicate that the techniques with sentiment analysis perform better than a bag-of-words model in accuracy measures. It also revealed the small difference between the two models LMD and HVD. For LMD the accuracy was 0.5527, whereas HVD accuracy was 0.5460, which indicates that the two dictionaries can be used effectively for the market prediction task.

Another paper [30] developed a model to predict three stock price directions with 1-day, 2-day and 3-day lag. The dataset contains financial news of SZ002424 stock from September of 2012 to March of 2017. In order to analyze the structure of news and get the hiding information inside the contents, the authors proposed a semantic and structural kernel (S&S kernel). The kernel was based on SVM and uses medical industry news for evaluation. Experiments find that the proposed kernel can reach up to 73% accuracy when predicting the price trend with 2-day lag, which proves that content structure hidden in daily financial news can predict the stock market movements. The result also reveals that financial news has an important influence on stock movements that typically last for 2–3 days.

In the work of [31], the authors used a lexicon-based approach to predict the stock market based on Twitter user feelings. The authors used historical stock data in addition to Twitter messages to predict DJIA and S&P 500 indices movements. Twitter data were obtained to train support vector machine and neural networks (NN) for 7 days. The dataset was created by adding a normalized set of tweets that contains 8 categories of emotions in about 755 million tweets. The collected tweets were downloaded from the period of February 13, 2013, to September 29, 2013. For sentiment analysis, a dictionary approach has been created manually by an expert in the field. The best average accuracy was obtained by using the SVM algorithm to forecast the DJIA indicator with an accuracy equal to 64.10%. However, using NN to predict S&P 500 achieves only a 62.03% in accuracy measure, which proves that

SVM performs better than the NN algorithm for market prediction. Moreover, the results achieved by the model indicate that it is possible to increase the prediction accuracy using human sentiment analysis and a lexicon-based approach.

In the paper of [15], the authors proposed a model with the user interface to predict the market movement for 1 day ahead. The proposed model consists of historical stock prices, technical indicators, Wikipedia company pages, and Google news. The model employs three machine learning methods to compare and select from, namely, ANN, SVM and decision tree (DT). The model concentrates on forecasting the AAPL (Apple NASDAQ) stock movement for a period from May 1, 2012, to June 1, 2015. For the APPL prediction case study, the authors used SVM recursive feature elimination (RFE) to choose the most important features. RFE is applied via backward choosing of predictors relying on feature importance ranking. Combining many data sources, the financial expert system achieves 85% accuracy in prediction. The result indicates that incorporating data from multiple sources will improve the efficiency of market prediction.

In [32], the author introduced a method to predict the stock movement for 1 day ahead. The proposed technique used a manually labeled corpus. The dataset contains 16 randomly selected stocks that are commonly discussed by StockTwits users collected from the period of March 13, 2012, to May 25, 2012. The collected tweets were about 100,000 posts. For text analysis, the model used SVM to analyze sentiment in StockTwits. The results prove the outstanding performance of SVM for sentiment classification tasks with accuracy that can reach up to 74.3%, whereas the overall accuracy for predicting the market up and down change based on the suggested model was 58.9%.

From the findings recorded in **Table 1**, it can be noted that SVM efficiency surpasses the effectiveness of approaches that used neural network models as we discussed earlier.

4.2 Deep learning

A deep learning concept is derived from machine learning methods that utilize many layers of data processing for the extraction of features, patterns, and classification. Recently, deep learning techniques are launched to sentiment analysis tasks, and they are considered effective in most cases [33].

In [34] the authors investigated whether deep learning methods can be modified to improve the accuracy of StockTwits sentiment analysis. Several neural network variants such as LSTM, doc2vec, and CNN were examined to discover stock market sentiments posted on StockTwits. The results prove that the convolutional neural network is one of the best deep learning methods for predicting authors' sentiment in the StockTwits dataset. Many other types of research discussed the successful use of deep learning for sentiment analysis and natural language processing tasks. On the survey research in [35], some of the different methods used in sentiment analysis tasks are compared. The main result showed the excellent performance of deep learning methods for sentiment analysis, in particular, CNN and LSTM methods.

Another paper [36] proposed a method to predict the French stock market based on sentiment and subjectivity analysis of Twitter data. The author applied a simple feedforward neural network to analyze tweets and predict CAC40 index movements for the next day. The Twitter collected data for the period of February 27, 2013, to June 16, 2013, was about 25,930 tweets. In addition to Twitter data, Martin also used historical stock market prices for the CAC40 index and other stocks. The results yield a direction accuracy of 80%, which indicates that using a neural network can be used successfully to predict the stock market movements.

Reference	Data type	Methods	Feature selection of textual data	Feature representation	Measure used	Results
Porshnev et al. [31]	Twitter, historical stock data of DJIA and S&P 500	SVM, NN, and sentiment dictionary	Emotion lexicon	Sentiment score of 8 scales	Accuracy	SVM ACC = 64.10%. NN ACC = 62.03%
Li et al. [28]	Five years historical Hong Kong Stock Exchange prices and financial news	SVM and two financial sentiment dictionary HVD and LMD	Polarity asymmetry of the news	Sentiment score	Accuracy	LMD ACC = 0.5527, HVD ACC = 0.5460
Xu and Keelj [32]	StockTwits	Manually labeled corpus and SVM	Unigram, bigram, line length, and punctuation	Sentiment score	Accuracy	ACC 58.9%
Weng et al. [15]	Historical stock prices, technical indicator, Wikipedia company pages, and Google news Apple NASDAQ stock	ANN, SVM, and DT	RFE	Binary	Accuracy for including all data sources	Approximately 85%
Xie and Jiang [27]	Financial news, 20 Chinese stock prices	SVM and specific sentiment dictionary	BOW	Sentiment score from −5 to +5, where −5 represents the most negative impact, +5 represents the most positive impact, and 0 represents for stop word	Accuracy	ACC 59.1734%
Long et al. [30]	Financial news of SZ002424 stock	SVM and S&S kernel	BOW	Keyword frequency	Accuracy	ACC = 73% with 2-day lag

Table 1.
Support vector machine for stock market prediction based on text mining studies.

4.2.1 Artificial neural networks

Artificial neural networks are a subset of deep learning technology that falls within the large artificial intelligence domain, and it mimics the human brain and its nervous system work. The simplest form of artificial neural networks is a feedforward neural network where the data go through the different input nodes until they reach the output node using only one direction, which is obtained by using a categorizing activation function.

In [37], the authors proposed a market investment recommendation system to predict intraday stock returns. The authors tested many prediction methods to find the best resulting algorithm. The dataset includes 72 S&P 500 companies for evaluation. Using both historical market data with financial news, the authors implemented the modeling technique many times to select the best model. For the first time, they have applied a feedforward neural network algorithm. For the second time, they used a stepwise logistic regression (SLR). For the third time, they implemented the decision trees with a genetic algorithm (GA) proposed by [38]. The best result was obtained by using the neural network prediction technique, which indicates that the NN algorithm is profitable for any initial investment. The result also confirms that combining market data with financial news can predict the market movement with better accuracy.

In [39] the producers predicted the stock market movements based on sentiment analysis of comments and tweets extracted from Twitter and StockTwits famous social media sites. User comments are classified into four different categories, which are up, down, happy, and rejected. The market data of the popular companies like Apple, Microsoft, Oracle, Google, and Facebook was collected from the period of January 1, 2015, to February 22, 2016. Both market data and polarity data were fed to an artificial neural network to predict the movements of the stock. The best prediction result was obtained for Apple Company with MSE equal to 0.14.

In [40], the proposal adopted a two-layer RNN-GRU technique to forecast the Chinese stock market movements. The model exploited sentiment analysis of Sina Weibo (a very popular Chinese social network) news and posts. The authors constructed their sentiment dictionary using user posts on the website. The authors also collected stock prices of the Shanghai Shenzhen 300 Stock Index (HS300) to use as an input to the recurrent neural network (RNN) model with gated recurrent units (GRU). The experiments revealed that the news and posts on Sina Weibo can predict the market movements with MAE equal 0.625 and with MAPE equal to 9.38.

In [13] the authors proposed a multi-source multiple instance (M-MI) model to predict the stock market index movements. In the proposed frameworks, the authors collected data from multiple resources, namely, quantitative data of Shanghai Composite Index historical prices for each trading day, financial news data to extract events, and social media data taken from Xueqiu (a famous trader social network in China to explore user sentiments user posts). Then, the analyzed sentiments, events, and the stock historical data are given as input to the M-MI model to make the prediction. For event extraction, the authors used HanLP (the popular method used for text parsing to grab the syntax of a sentence). Event extracted is used to feed the Restricted Boltzmann Machines (RBMs), which is a creative theoretical artificial neural network. In the model, the authors also examined the importance of specific sources to the index movements by giving them specific weights. The proposed framework prediction accuracy was about 60%, which reveals many findings. Firstly, the integration of features from multiple resources can make a more effective prediction. Secondly, both news events and

market historical data have a more important effect on stock movements than social media sentiments. Thirdly, both news events and quantitative data have larger impacts on stock fluctuations than using sentiments alone.

Recently [41] applied a technique to forecast the stock directions. The authors used sentiment analysis of news headlines in addition to historical market data of Apple stock to predict the market trend. Hive ecosystem was used to preprocess the data, and the naive Bayes classifier was utilized to calculate the sentiment scores. With two inputs from news headlines sentiment score and historical numeric market data, the multilevel perception artificial neural network (ANN) is applied to forecast the stock movements. In the training procedure, the authors used back-propagation, and in the output layer, they used the identity function. Moreover, the model tested two different periods for training the data; in the first method, they trained a 3-year data period, and the second method trained a year data period. The result represents an accuracy of 91% in the first methods, while 98% accuracy was achieved in the second method, which indicates that stock price forecasting is more efficient for a shorter time.

More recently [42] predicted future market trends by using both market historical prices and financial news article sentiments as input to the neural network. The authors collected historical prices of the 20 biggest companies listed in the NASDAQ100 index to predict the fluctuations of the stock for the portfolio that consists of 20 firms historical stock prices, with a periodicity of 15 min, obtained from Google Finance API. For new article analysis, two approaches of feature selection have adopted the dictionary of Loughran and McDonald (2011) (L&Mc) and affective space [43]. The Loughran and McDonald dictionary is commonly used for market prediction and consists of many critical words for the classification task that represents negative, positive, and uncertain sentiments that can be found commonly in financial news, whereas affective space (AS) dictionary is a vector space dictionary that depends on the similarity and relationships between words as natural language processing methods. For dimensionality reduction, the affective space mapped each term to a 100-dimensional vector that allows concepts to be grouped based on their semantics and relations.

The proposed model with Loughran and McDonald's dictionary confirms to be more effective, resulting in an annualized return of 85.2%, while the use of affective space feature dictionary as an input to the neural network model proved to be more effective in obtaining high accuracy results. **Table 2** summarizes the studies that used NN extensively for market prediction techniques.

4.2.2 Recurrent neural network

Recurrent neural network is an important variant of artificial neural network that starts as normal with front direction but preserves the relevant data that may need to be utilized later. In other words, every node will act as a memory cell that remembers some information it had in the earlier step.

A well-known variant of RNN model is long short-term memory (LSTM), which was proposed by Hochreiter and Schmidhuber in 1997 [44]; it is a standard recurring neural network that solves the exploding gradient problem. LSTM can depict the long dependencies in a sequence by adopting a memory unit and a gate mechanism to determine how information stored in the memory cell can be used and updated [45]. Each LSTM is a set of cells or system modules that catch and store streams of data. The cells represent a transport line that carries data from the past and collects them for the present module from one module to another. Through the use of certain gates in each cell, data can be disposed of, filtered, or added for the next cells [46].

Reference	Data type	Methods	Feature selection of textual data	Feature representation	Measure used	Results
Martin [36]	CAC40 index data, Twitter	NN	Tokens	Average sentiment score	Direction accuracy	80%.
Geva and Zahavi [37]	72 companies in the S&P 500 index data, financial news	NN, SLR, and DT with GA	BOW	Calibrated sentiment scores and binary indicator	Return over the initial investment 200 k	NN: 8.57% SLR: −0.20% GA: 0.16%
Khatri and Srivastava [39]	Twitter and StockTwits, index data of Apple (APPL), Microsoft (MSFT), Oracle (ORCL), Google (GOOG), and Facebook (FB)	ANN	Predefined words	Sentiment score between 0 and 1	MSE	AAPL: 0.14 MSFT: 0.18 ORCL: 0.22 FB: 0.28 GOOG: 0.27
Zhang et al. [13]	Shanghai Composite Index, financial news data, and social media from Xueqiu	(RBMs) and ANN	For event extraction: HanLP, Sentence2vec For sentiments: latent Dirichlet allocation	Two polarities: positive or negative	Prediction accuracy	60%
Zhang et al. [13]	HS300 Index and Sina Weibo news and posts	Own sentiment dictionary and two-layer RNN-GRU	Positive and negative keywords	Probability value for fall or rise	MAE, MAPE, and RMSE	0.625, 9.381, and 0.803
Picasso et al. [42]	20 companies in NASDAQ-100 index and financial news articles	NN with L&MC and NN with affective space	L&Mc dictionary and AS dictionary	Counts of negative, positive, uncertainty, superfluous, and other words of the dictionary found in news and number of news in the slot	Accuracy	NN AS 68% NN L&MC 60%
Shastri et al. [41]	Apple stock and news headlines	Hive ecosystem, NB , and multilevel perception artificial neural network (ANN)	Unique positive and negative words	Sentiment score	MAPE, trend prediction accuracy of 1-year period	8.21 98%

Table 2.
The main study that used NN extensively for market prediction based on text mining.

In the paper of [47], the proposal adopted a method to predict the stock market movements based on the bidirectional gated recurrent unit (BGRU), which is considered a variant of LSTM. The model used financial news that comes from Reuters and Bloomberg websites and historical stock prices to predict the market fluctuation with a better result. The S&P stock prices and news data were collected in the period of 2006–2013. Also, the model examined the method performance on the individual stock that comes from different sectors, namely, Google Inc., Walmart, and Boeing. In the proposed method, the authors used the word embedding model introduced by [48] to select the most efficient features from the collected financial news. In word embedding model, the words were encoded as vectors in a high-dimensional space, and then the analogy between words in meaning is interpreted to closeness in the vector space. The proposed model achieved accuracy equal to 59.98% in the S&P 500, whereas individual stock prediction accuracy was more than 65%. The authors also examined the performance of many LSTM variants like standard LSTM, GRU, and BGRU. The finding shows that BGRU obtained the best results compared to other LSTM variants.

However, conventional LSTM is unable to detect what is the most crucial part of the sentence for the sentiment categorization task. Therefore, [49] proposed a design mechanism capable of detecting the crucial part of the sentence related to a specific aspect and explained the architecture of attention-based LSTM in detail.

To predict the stock market directional movements, [50] proposed an Attention-based LSTM model (AT-LSTM) to predict the movements of Standard & Poor's 500 index and individual companies' stock price using financial news titles. The attention techniques were divided into two classes. The first class of attention assigns there weight to the news that contains positive sentiments to the stock market such as "raise," "growth," etc. While the second class of attention assigns there weight to the news that mentions the major companies in the S&P 500 such as "Microsoft" and "Google." Therefore, the attention model is trained continuously to assign more attention to the relevant news based on its content. The proposed method achieved more than 66% accuracy, and the company WALMART obtained a max accuracy of 72.06%. The results prove that attention mechanisms can achieve good results for market prediction in specific cases.

In [51] proposal support decision system based on deep neural networks and transfer learning was applied. To enhance the prediction accuracy, the authors pretrain the networks on a different corpus. The main aim of the study was to recommend the best deep learning techniques in terms of market prediction. The system provides its corpus with a length of 139.1 million words. The authors trained the deep neural networks by using the Adaptive Moment Estimation Algorithm (Adam), which can effectively solve sparse gradient problems. Then the use of transfer learning aims to initialize the weights of parameters with values that might be close to the optimized ones. In order to account for unbalanced classes in their dataset, they have used classification balanced accuracy that can be defined as the arithmetic mean of sensitivity and specificity. They also predicted the direction of nominal returns. The result proves that LSTM models surpass all traditional machine learning models based on the bag-of-words technique, specifically when they used transfer learning to pretrain word embeddings.

Recently [12] examined the effect of financial news articles on stock trend fluctuation either rise or fall. The financial new articles related to the Taiwan 50 Index were collected from Google. For textual data analysis and NLP tasks, the authors used their lexicon and then exploited the LSTM to make the final prediction. The use of LSTM features was joint with historical data and adjusted in each step. The results prove that individual stock prediction using the study polarity lexicon was better than the benchmark model. Moreover, the proposed model reaches an

accuracy of 76.32, 80.00, and 77.42% for each of the following stocks TSMC, Hon Hai, and Formosa Petrochemical, respectively, which reveals the effectiveness of the LSTM model in market prediction based on text analysis.

Another study proposed in [52] examined the effectiveness of using the LSTM technique to predict market movements, using market data and textual resources as input to the model. The authors analyzed user sentiments from forum texts about the CSI300 index using the naive Bayes algorithm and then using LSTM, which contains a merged layer, a ReLU layer, and a softmax layer to combine the investor sentiment taken from forum posts with the historical market. The fall or rise trend prediction accuracy achieved was 87.86%, outperforming other commonly used machine learning methods such as SVM algorithm by at least 6%, which highly indicates that LSTM can achieve a better result in prediction when using larger datasets. **Table 3** summarizes the recent studies that used RNN networks for stock market prediction based on text analysis.

4.2.3 Convolutional neural network (CNN)

Convolutional neural network used for natural language processing was first explained by Collobert and Weston in [53]. A typical convolutional neural network is composed of multiple convolutional layers at the bottom of a classifier. Conventional inputs for text processing are characters, phrases, paragraphs, or documents that are converted into a matrix representation. Each row of the matrix represents a token, which is typically a word or character [54].

In [16], framework proposal for stock market prediction based on long-term events and short-term events extracted from financial news articles about the S&P 500 index was applied. The collected financial news articles come from October

Reference	Data type	Methods	Feature selection of textual data	Feature representation of textual data	Measure used	Results
Li et al. [52]	CSI300 index data, investors forum posts	LSTM, NB	Manually labeled sentiments by experts	Pos, neg, and neutral	Direction accuracy	87.86%
Huynh et al. [47]	S&P 500 index, financial news	BGRU	Word embedding	Real valued vectors	Prediction accuracy	59.98%
Kraus and Feuerriegel [51]	German ad hoc announcements	Transfer learning with RNN and LSTM	Word embedding (em)	Polarity score	Direction accuracy of nominal return	RNN 0.552 LSTM 0.576 LSTM-em 0.578
Liu [50]	S&P 500 index, financial news titles	AT-LSTM	Own word embedding trained with Skip-Gram	Word embedding and character-composition vector	Direction accuracy	More than 66% for each stock
Chen et al. [12]	Financial news articles, Taiwan 50 Index	LSTM and polarity lexicon	Word2vec and Skip-Gram	TF-IDF and polarity score	Accuracy	Up to 80% for Hon Hai stock

Table 3.
Recent studies that concentrate on RNN variants for market prediction based on text analysis.

2006 to November 2013, which was released initially by Ding et al. [55]. The long-term events represent events over the past month, while the short-term events represent events on the last day of the stock price fluctuate. The proposed frameworks train the extracted events using a neural tensor network and then a convolutional neural network to predict both the short-term and the long-term impact of extracted events on stock price fluctuations. The proposed framework examined two different ways for representing the input to CNN. The first method (WB-CNN) used word embedding as input and convolutional neural networks for prediction. The second method (EB-CNN) used event embedding as input and convolutional neural networks for prediction. The experiments achieve accuracy of 61.73% for WB-CNN, while the EB-CNN method achieved an accuracy equal to 65.08%, which illustrates that the proposed model is more effective in stock market prediction than other models that predicted the S&P 500 index based only on stock historical data analysis. The model also proves that CNN can extract the longer-term influence of financial news events than traditional feedforward neural networks.

In [17], writers proposed a model to predict the intraday stock market directional movements of the S&P index using financial news title and financial time series market data as input. The paper compared two commonly used deep learning methods, which are RNN and CNN algorithms using many text representation methods. The RNN method used in the paper was the LSTM model. The proposed model examined many types of text representation as an input to the CNN prediction model. The (W-CNN) represents a word embedding as input and a CNN as a forecast model. The (S-CNN) represents sentence embedding input and CNN forecast model. The (W-RCNN) word embedding input and RCNN forecast model. The (S-RCNN) represents sentence embedding input and RCNN forecast model. The (WI-RCNN) shows word embedding and historical time series input and RCNN prediction model. The (SI-RCNN) illustrates sentence embedding and historical time series data input and RCNN prediction model. Experiments on each of the previous models revealed that CNN is more effective than RNN on capturing

Reference	Data type	Methods	Feature selection of textual data	Feature representation of textual data	Measure used	Results
Ding et al. [16]	Financial news articles	The train uses a neural tensor network, WB-CNN, EB-CNN	Word embedding WB, event embedding EB	Binary	Accuracy	WB-CNN 61.73 EB-CNN 65.08%
Vargas et al. [17]	Financial News articles, S&P index market data	Many text representation	Word embedding, sentence embedding	Binary	Accuracy	W-CNN 57.22% S-CNN 60.96% W-RCNN 60.22% S-RCNN 61.49% WI-RCNN 61.29% S1-RCNN 62.03%

Table 4.
CNN use for stock market prediction based on text mining results.

semantic from new financial, and RNN is more efficient in capturing the context information for the stock market prediction. Moreover, the results prove that the sentence embedding for text representation is more effective than the word embedding. **Table 4** summarizes the studies that used CNN for stock market prediction based on sentiment analysis and NLP.

5. Other machine learning methods

Many other machine learning methods were used successfully and less frequently for market predation applications based on text mining. Summaries of these studies are illustrated in **Table 5**. In the study of [18], a method was proposed to predict the stock trend movements of three NASDAQ companies, namely, Yahoo Inc., Microsoft Company, and Facebook Inc. (FB Inc). The model used financial news sentiment analysis with historical stock data to predict the market with higher accuracy. The task is accomplished with two steps: Firstly, they used naive Bayes classifier to classify news sentiment into two classes, positive or negative. Secondly, to forecast the stock trend fall or raise, they used k-Nearest Neighbor algorithm (K-NN) (a clear algorithm that saves all possible instances of data and categorizes the new data based on a scale of closeness and is often used to classify a new data based on the current classification of its neighbors). The results show that the accuracies of sentiment analysis of news only can go up to 63%, while combining news sentiments with historical stock prices can achieve trend prediction accuracy up to 89.80%, which proves that adding historical stock prices to the classification model will be able to improve the prediction performance.

In the work of [56], the authors suggest a method to predict the daily up and down price fluctuation of four tech companies of NASDAQ stock, which are Apple (AAPL), Google (GOOG), Microsoft (MSFT), and Amazon (AMZN). The model analyze Twitter user messages in addition to three previous days of the stock price movement. The model constructs a named-entity recognition (NER) approach to identify and remove the noise of Twitter data. A decision tree approach was used to build the classification model. The proposed model achieved the highest accuracy of 82.93% in predicting the daily up and down changes of Apple Company, which indicates that using named-entity recognition method for noise removal of Twitter data can improve the accuracy results.

The research in [8] proposed a method to predict the stock market movements based on two feature extraction methods, using a novel aspect-based sentiment model to improve the prediction performance. The first methods tempt to excerpt hidden topics and sentiments together and use them for the prediction, while the aspect-based sentiment methods treat every message as a list of topics and correlative sentiment values. To build the prediction model, the authors used SVM with the linear kernel and collected data of 18 stocks for a period of 1 year from July 2012 to July 2013. Exploiting the aspect-based sentiment feature method obtained the best result with 54.41% average accuracy. The proposed model also proves to be 3.03% more effective than using the human sentiment method for stock movement prediction.

In [61] proposal a method to forecast the Indonesian stock movements based on Twitter sentiment analysis was introduced. Naive Bayes and random forest algorithm was used to find the user sentiments of the 13 most popular companies in Indonesia. The linear regression technique was used to build the prediction model. The highest accuracy was achieved by the categorization model using the random forest algorithm with 60.39% accuracy, whereas naive Bayes classifier was able to classify tweet data with 56.50% accuracy. For the price movement's prediction, the

Reference	Data type	Methods	Feature selection of textual data	Feature representation of textual data	Measure used	Results
Vu et al. [56]	Twitter user messages, AAPL, GOOG, MSFT, and AMZN indices data	Decision tree approach, NER	Predefined bullish-bearish anchor words	The real number for daily Neg_Pos and Bullish_Bearish	Daily prediction accuracy	AAPL 82.93% GOOG 80.49% MSFT 75.61% AMZN 75.00%
Moniz and de Jong [57]	News stories for 598 global companies	Ensemble tree, LDA	LDA	Binary	F1- measure	0.508
Bing et al. [58]	30 NASDAQ and New York stock indices and Twitter	Association rule	Sentiment word list	TF-IDF, vector space model, which is an arithmetic model to represent text as vectors	Average accuracy	76.12%
Li et al. [59]	HSI 23 stocks indices and financial news	Multiple kernel learning	Word list extreme positive, positive, neutral, negative, and extreme negative	TF-IDF, vector space model	RMSE	0.139 for 30 m
Shynkevich et al. [60]	Five stock from the S&P 500 index, SS, and SIS news items	Multiple kernel learning	BOW	TF-IDF	Highest accuracy	81.63% for WLP stock with six kernels
Nguyen et al. [8]	Social media message board and 18 stocks index data	SVM	POS tagging Stanford CoreNLP for aspect-based sentiment	Average sentiment score or values	Prediction Accuracy	Aspect-based model 54.41%
Cakra and Trisedya [61]	Twitter data and many companies in Indonesia indices data	NB and RF and linear regression	Sentiment lexicon and sentiment shifters	Positive, negative, and neutral	Prediction accuracy	NB 67.37% RF 66.34%

Reference	Data type	Methods	Feature selection of textual data	Feature representation of textual data	Measure used	Results
Ghanavati et al. [62]	Hong Kong market index and financial news articles and summaries	Loughran and McDonald dictionary, ML, and metric learning-based methods	Tokenization using OpenNLP tools	Sentiment value vectors	The average error rate	The average error rate of ML for large cape stock 0.15 The average error rate of ML for small cape stock 0.20
Khedr and Yaseen [18]	Financial news, index data of 3 NASDAQ companies	K-NN and NB	TF-IDF and N-gram	Values for pos, neg, and equal	Trend prediction accuracy	89.80%
Gálvez and Gravano [63]	Twelve stocks of the MERVAL Index and online message boards	Combining LSA with ridge regression	Latent semantic analysis (LSA)	Numbers for each special token	Maximum accuracy when using technical indicators and topics from the online message board	Up to 0.750
Liu and Wang [15]	China Security Index 300 (CSI300) and the Standard & Poor's 500 (S&P500). News reports and numerical data	LSTM and many textual representations	News embedding	Numerical Vectors	Accuracy	NBAd raises the accuracy of 2.32% and 1.35% higher than the best baseline models of the dataset
Maqsood et al. [64]	Many USA, Hong Kong, Turkey, and Pakistan company indices. Twitter	Event sentiment, linear regression (LG), support vector regression (SVR), and deep learning	A comprehensive dictionary with their own generated word list	Sentiment value that is calculated for each day separately	Average root mean square error (RMSE)	For each country using LG, SVR, and DL, respectively. US 4.35, 1.33, and 1.65 Hong Kong 0.90, 0.31, and 0.35 Turkey 0.27, 0.11, and 0.11 Pakistan 0.70, 0.34, and 0.33

Table 5.
Summaries of machine learning methods that were used successfully and less frequently for market prediction based on text mining.

proposed models can predict the upcoming price fluctuation of either rise or fall with the accuracy of 67.37% achieved by the naive Bayes algorithm and 66.34% obtained by using Random Forest classifier.

Other research [62] introduced a stock market prediction service framework that allows users to choose different data sources and machine learning techniques. The authors gathered all news summaries and historical prices of all the stocks for a 1-year period. Using the Hong Kong market stock dataset for evaluation, they found that metric learning-based methods can improve the prediction results. The study also shows that adding news to the historical prices for stock market prediction will be more useful on large and popular stocks.

Recently [14] applied a numerical-based attention (NBA) method for multiple sources of stock market prediction. News headlines and numerical data combined to predict the stock prices. For evaluation, the authors collected news headlines and numerical data from two sources: the China Security Index 300 (CSI300) and the Standard & Poor's 500 (S&P500). They used NBAa-NBAd to denote different variations of the models with different textual representations. In these three datasets, the proposed structure accomplishes the best outcomes. Especially, NBAd raises the accuracy of 2.32 and 1.35% higher than the best baseline models on S&P500 and CSI300.

More recently, [64] investigated the effect of the most important event from 2012 to 2016 into the stock exchange prediction of four selected countries, which are the USA, Hong Kong, Turkey, and Pakistan. The events are then categorized into local and global events for each country according to their economic effects on the country stocks. Twitter data were gathered to find the sentiment for each one of these events. The model used a total of eight events for all countries. For classification, the authors investigated linear regression, support vector regression, and deep learning model for market prediction. The results revealed that linear regression achieves the worst prediction results compared to the other two methods used in their analysis, while the support vector regression achieves the best results. Event sentiment illustrates noted development in the forecasting results. For example, the US election 2012 event achieves the best prediction results in all methods, which indicates that a local event that appears in the USA has a very great effect on stock market future forecasting.

In [63], the authors predicted the Argentinian stock market by using online message boards with topic discovery methods in addition to daily historical stock prices. The authors exploited Latent Semantic Analysis (LSA) approach that finds the latent topics in the text. The experiments are trained with multiple combinations of features selected from online texts. The results show that the most predictive features are derived from the texts that contain the most relevant semantic content. Moreover, the experiments illustrate that combining LSA with ridge regression was able to identify the structure of the texts that later improves the prediction performance of the model.

In [57], the authors proposed a model that aims to find the influence of negative terms represented by the financial media on investor behavior. The proposed model relays on the counting of negative words from the dictionary and word counting methods to extract contextual information. The model also used a Latent Dirichlet allocation model to derive the financial media statements of negative influence. The model combines the two inputs in an ensemble tree to categorize the effect of financial media news on stock market fluctuation. The results indicate that there is a strong relationship between negative effect derived from financial media news and a company stock market fluctuation.

In the same year, authors in [58] suggested algorithm predicts 30 NASDAQ and New York stock exchange companies' movements. The algorithm used NLP

methods to categorize Twitter messages. Then the authors applied association rules to find interesting rules and associations between the stock movements and the Twitter messages. The collected tweets were about 15 million Twitter messages. The big data then stored it in MongoDB, which is an open-source database used to save and process the huge data. The suggested method has explained the relationships hidden in social media as a graph with several layers, with the top layer, intermediate layer, and the bottom layer attributes to show the relations. The proposed method has increased the dimensionality of whole variables that would measure the hidden and embedded data among the Twitter messages. The results indicate the outstanding performance of using tweet message sentiment to predict the stock market movements 3 days later.

In [60], the researchers exploited the multiple kernel learning method to integrate data from the stock special (SS) and subindustry special (SIS) news items effectively to predict future market movements. Multiple kernel learning (MKL) applies many different kernels to learn from various sections of data. Pairs of Gaussian, linear, and polynomial kernels were used to compare each model performance. For evaluation, the authors used five stocks from the S&P 500 index that belongs to managed healthcare subindustry. The results indicate that using Gaussian, linear, and polynomial kernels jointly in MKL achieves higher prediction results. The results also indicate that exploiting two types of news increases prediction accuracy in comparison with models that used only a single news source.

The study in [59] combined information on historical stock prices with financial market news to enhance the market forecasting accuracy of intraday trading status. For evaluation the model used the Hong Kong Stock Exchange (HKEx) tick prices; more specifically the authors used 23 stocks in Hang Seng Index10 (HSI) intraday prices in the year 2001. Multi-kernel support vector regression (MKSVR) was used with two subkernels: one for the news items and the other kernel for the stock historical prices. The results indicate that MKSVR outperforms other benchmark models that exploited only one source of information.

The evaluation measurements vary in all of the reviewed works; some of the researches calculate accuracy, F-measure, or recall and precision with accuracy being the most commonly used. However, other researchers calculated the error in prediction using mean absolute percent error (MAPE), mean squared error (MSE), or root mean square error (RMSE).The variances in using different evaluation measurements and exploratory data make an accurate comparison between different models difficult to achieve.

6. The reviewed work text source and period and number of collected items

The textual data input comes from different several sources, and the period and the numbers of collected data are varied, and all are illustrated in **Table 6**.

The majority of writers have analyzed primary news websites like the Reuters and Bloomberg [16, 17, 37, 47, 50], Dow Jones [57], and Yahoo Finance [8, 18]. Most authors use financial news because it is associated with less noise compared to the general news. They either select the news text or the news headline as input to their machine learning model. Recently news titles and headlines are specifically extracted and are regarded to be more clear, concise, and associated with less noise [14, 16, 17, 50]. Other authors have examined less formal sources of news information such as Google News [12, 15]. Other researchers collect their textual information merely from social media websites especially Twitter to analyze the public user sentiments to predict the market more effectively [39, 56, 61, 64].

Reference	Text type and source	Period	Number of collected items
Vu et al. [56]	Twitter user messages	April 1, 2011 to May 31, 2011	5,001,460 daily tweets
Porshnev et al. [31]	Twitter	April 132,013, to September 29, 2013	About 755 million tweets
Martin [36]	Twitter data.	February 27, 2013, to June 16, 2013	About 25,930 tweets
Li et al. [28]	23 stocks in Hang Seng Index10 (HSI) intraday prices and financial news from the website Caihua, http://www.finet.hk/	Intraday prices of the year 2001	28,885 pieces of news
Xu and Keelj [32]	StockTwits	March 13, 2012, to May 25, 2012	100,000 tweets
Bing et al. [58]	Twitter messages	October 2011 to March 2012	15 million Twitter messages
Li et al. [28]	Financial news articles from FINET (a main financial news seller in Hong Kong)	January 2003 to March 2008	Not mentioned
Geva and Zahavi [37]	Financial news from Reuters 3000 Extra Service	September 15, 2006, to August 31, 2007	51,263 news items
Moniz and de Jong [57]	News source is a corpus extracted from Dow Jones Newswires (DJNW). News articles are collected from financial blogs, online newspapers, financial magazines, and many online websites	January 1, 2009, to December 31, 2013	The corpus consists of 35,678 daily news stories
Ding et al. [16]	Financial news titles from Reuters and Bloomberg	October 2006 to November 2013	442,933 for training 110,733 for development 110,733 for testing
Cakra and Trisedya [61]	Twitter data	April 14, 2015 to April 30, 2015	Not mentioned
Nguyen et al. [8]	Texts in a message board from Yahoo Finance Message Board	July 2012 to July 2013	The different numbers of messages for each stock that follows between 89 and 11,220 in maximum
Shynkevich et al. [60]	News of 5 stock from the S&P 500 index that belongs to managed healthcare sub from LexisNexis database	September 1, 2009, to September 1, 2014	More than 400 news articles
Ghanavati et al. [62]	News summaries(source not mentioned)	June 1, 2014, and June 1, 2015	Not mentioned

Reference	Text type and source	Period	Number of collected items
Khatri and Srivastava [39]	Twitter and StockTwits	January 1, 2015, to February 22, 2016	Not mentioned
Gálvez and Gravano [63]	Message board texts from the webpage http://foro.ravaonline.com.	June 1, 2010, and July 31, 2015	More than 20,000 posts
Weng et al. [15]	Wikipedia company pages and Google news	May 1, 2012, to June 1, 2015	Not mentioned
Chen et al. [40]	Sina Weibo news and posts	January 1, 2015, to March 8, 2017	Not mentioned
Li et al. [52]	Forum posts from guba.eastmoney.com	January 1, 2009, to October 31, 2014	More than 18 million posts
Kraus and Feuerriegel [51]	German ad hoc announcements from www.dgap.de	2010–2013	10,895 observations
Huynh et al. [47]	Financial news from Reuters and Bloomberg websites	2006–2013	5816 news for training 2904 news for testing
Khedr and Yaseen [18]	News data from different resources, Google finance, Reuters, wall street journal, marketwatch.com, zacks.com, Yahoo Finance, and ecomomics.com, nasdaq.com	Not mentioned	Not mentioned
Vargas et al. [17]	Financial news title from Reuters and Bloomberg	October 2006 to November 2013	13,149 for training 1976 for development 2046 for testing
Liu [50]	Financial news titles collected from Reuters and Bloomberg	2006–2013	445,262 for training 55,658 for development 55,658 for testing
Zhang et al. [13]	Financial news articles from financial news websites in China and Xueqiu social media posts	2015–2016	38,727 news in 2015 and 39,465 news in 2016 6,163,056 posts for 2015 and 2016
Xie and Jiang [27]	Financial news of Wallstreetcn, Stockstar, China news, and many other resources	2008 and 2015	2,302,692 news items
Long et al. [30]	Financial news from ifeng.com financial channel in China	September 2012 to March 2017	18 news per day at maximum
Shastri et al. [41]	News headlines from http://www.nasdaq.com/	2013–2016	Not mentioned
Picasso et al. [42]	News articles from intrinio.com API	July 3, 2017 to June 14, 2018	Not mentioned

Reference	Text type and source	Period	Number of collected items
Chen et al. [12]	News articles from Google	January 4, 2016, to December 29, 2017	130,000 articles
Liu and Wang [14]	News headlines from five famous financial news websites in china	January 1, 2016, to December 31, 2016	780,920 financial news headlines
Maqsood et al. [64]	Twitter data	2000–2018	11.42 million tweets

Table 6.
Summaries of the reviewed work text source, period, and number of collected items.

Also, as **Table 6** illustrated, the data were collected in a variety of periods; some few papers collected data in several months, while others extracted data within a maximum of 7-year period, which resulted in more sufficient data and better results in prediction.

However, it can be noted that the insufficiency of highly structured datasets containing text data of markets prevents researchers from accumulating their analysis and assessment efforts with others. Another problem is the imbalanced dataset that has been used by many researchers, which is discriminating the accuracy of prediction. In future, potential researchers are encouraged to locate new datasets for market forecasting based on text mining analysis.

Market predictive text mining could become much more advanced by concentrating on a particular source of text, such as a specific social media website or the new news source from specialized financial news websites. As mentioned in Section 3 of this research, there is a strong relationship between the behavioral economics and the market fluctuations; due to this fact focusing on behavioral economics studies and its impact on market movements will be of great research opportunity in the future.

7. The reviewed work findings, limitations, and future work

Developments in sentiment analysis approaches and deep learning have enabled the development of stock market prediction systems to turn future web content, tweets and financial, and news contents into investment decision systems. Online text mining processes are evolving and have been intensively investigated using machine learning advancements, and this trend will continue to achieve progression especially for market prediction.

Many researchers believe that analyzing only the historical prices of the stock market will be able to predict the stock market movement [3–5]. However, other researchers combine both textual information with historical prices of stock to predict the stock market movements [8, 13, 15, 47, 62]. The previous studies' major limitation is that they depend heavily on either structured data (historical stock prices) or unstructured data (news articles or social media). However, for the researchers that used both structured and unstructured data, the major limitation for most of them is that they combined either news articles or social media with past stock prices to predict the stock movements and they neglect the critical impact of combining social media and financial news information's with time series market data to improve the forecasting results.

Reference	Finding	Limitation and future work	Year
Vu et al. [56]	Using the named-entity recognition method for noise removal of Twitter data improves the accuracy results.	Increase the collected tweet data and the collection period and expand the number of companies.	2012
Porshnev et al. [31]	It is possible to increase the prediction accuracy using human sentiment analysis and a lexicon-based approach.	They need to expand the training period to achieve better outcomes. Use more effective sentiment analysis method to increase the prediction accuracy.	2013
Martin [36]	Twitter sentiment analysis using the neural network can be used to predict the stock market movements.	Adding a different source of information such as financial news articles will be able to improve the prediction performance more.	2013
Li et al. [28]	The sentiment analysis model performs better than a bag-of-words model inaccuracy measures. There was a small difference between using the two models, LMD and HVD.	We need to automatically expand the HVD and LMD dictionaries without affecting the accuracy of the dictionary.	2014
Xu and Keelj [32]	The result shows the outstanding performance of SVM for the sentiment classification task.	Expand the data analysis period. Use a more effective expanded lexicon. Exploit the user profile features.	2014
Geva and Zahavi [37]	NN algorithm is profitable for any initial investment. Combining market data with financial news can predict the market movement with better accuracy.	Study the effect of using other prediction models, and investigate the impact of using different textual data processing.	2014
Moniz and de Jong [57]	There is a strong relationship between negative affect derived from financial media news and a company stock market fluctuation.	Adding social media data to the dataset to improve the prediction performance.	2014
Bing et al. [58]	The study algorithm has an outstanding performance in using tweet message sentiment to predict the stock market movements 3 days later.	Needs to add other textual sources for social media data such as Facebook. Adding news items to the dataset.	2014
Li et al. [59]	The results indicate that MKSVR outperforms other benchmark models that used only one source of information.	Adding more sources of textual data. Apply more subkernel using the same textual data. Positive and negative news could be classified by using the use of sentiment analysis to categorize positive and negative news. The use of multiple subkernels for each news in different sentiment classes.	2014

Reference	Finding	Limitation and future work	Year
Ding et al. [16]	CNN can extract the longer-term influence of financial news events than traditional feedforward neural networks.	Adding different textual data sources and improvement in classification algorithm will yield a better result.	2015
Nguyen et al. [8]	The proposed model proves to be more effective than using the human sentiment method for stock movement prediction.	They have to define the number of topics and sentiment beforehand. The model can predict the stock movements either up or down only and can be improved to predict the degree of the movements. Adding different text data sources like financial news.	2015
Cakra and Trisedya [61]	The highest classification accuracy was achieved by using the random forest classification model.	Have to expand the data collection period. Needs to improve the sentiment classification model by adding different features.	2015
Shynkevich et al. [60]	Using of Gaussian, linear, and polynomial kernels jointly in MKL achieves higher prediction results. Exploiting two types of news increases the prediction accuracy in comparison with models that used only a single news item.	Add historical stock prices to the dataset with the news articles to enhance the prediction results.	2015
Khatri and Srivastava [39]	It is better to invest in a company whose sentimental score is high and positive rather than choosing a close price as an indicator of stock movements.	The datasets should be taken for a longer time to achieve better results.	2016
Ghanavati et al. [62]	The metric learning methods can improve the results. Adding news to the historical prices for stock market prediction will be more useful on large and popular stocks.	Needs to add the different sources of textual information like social media.	2016
Weng et al. [15]	Incorporating data from multiple sources will improve the efficiency of market prediction.	The use of different rank values selected from a different data source. Expand the work to include the certainty level of the prediction, which can be achieved by using Bayesian Belief Networks (BBN) or ensemble methods. Try to forecast the actual price instead of the movement. Adding other data sources also will increase the prediction performance.	2017
Chen et al. [12]	News and posts on Sina Weibo can predict the market movements.	The use of more improved machine learning techniques for sentiment analysis such as interdependent Latent Dirichlet allocation (ILDA) will improve the prediction performance.	2017

Reference	Finding	Limitation and future work	Year
Li et al. [52]	There is a strong relationship between investor sentiments and CSI300 prices.	Utilized only naive Bayes algorithm for classification and did not test other classification methods that may achieve better results.	2017
Huynh et al. [47]	BGRU obtained the best results in predicting the market compared to other LSTM variant.	Adding another textual source of information such as social media may enhance the model performance.	2017
Kraus and Feuerriegel [51]	LSTM models surpass all traditional machine learning models based on the bag-of-words technique, specifically when using transfer learning to pretrain word embeddings.	Increasing the number of collected news for a longer time and applying the deep learning model will improve the predictive performance.	2017
Vargas et al. [17]	CNN is more effective than RNN on capturing semantic from financial news. RNN is more effective in capturing the context information for the stock market prediction. Sentence embedding for text representation is more effective than the word embedding.	Exploiting the reinforcement learning models to train the proposed methods on trading simulation may yield better results.	2017
Khedr and Yaseen [18]	Adding historical stock prices to the classification model will be able to improve the prediction performance.	Adding technical analysis and social media sentiment analysis will improve the prediction results.	2017
Gálvez and Gravano [63]	The results indicate that the most predictive features derived from the texts that contain the most relevant semantic content. Moreover, the results prove that combining LSA with ridge regression was able to identify the structure of the texts, which improves the prediction performance of the model.	Adding even sentiment and more text resources such as social media data will improve the results.	2017
Checkley et al. [21]	There is a causal link between Twitter sentiments to stock market returns, volatility, and volume. Among all five stocks, market volatility and volume seem to be more predictable than market direction or return.	The consideration of event sentiment may affect the market return more and improve the forecasting result.	2017
Bujari et al. [24]	Some of the proposed ad hoc forecasting models well predict the next day direction of the stock movements for some particular companies with 82% of success, and there is no unified method to be used with all cases. The more volume of a tweet will yield better prediction results. There is a strong correlation between tweet posts and the trend movements for some companies.	Investigate another source of textual information such as online financial news.	2017

Reference	Finding	Limitation and future work	Year
Zhang et al. [13]	Both news events and market historical data have a more important effect on stock movements than social media sentiments. Both news events and quantitative data have larger impacts to drive stock fluctuations than sentiments.	Increasing the dataset collection period may improve prediction performance.	2018
Liu [50]	Adding news articles was able to predict the individual stock prices with better accuracy compared to predicting the market using time series prices alone.	Predicting price changes at a different time horizon in the future to achieve better performance. The study used the full corpus as input for the prediction model, which may add noise to the data and affect prediction accuracy.	2018
Xie and Jiang [27]	Both audience numbers and news quality have a crucial impact on the stock market.	Have to develop a better sentiment evaluation system.	2019
Long et al. [30]	Content structure hidden in daily financial news can successfully predict the stock market movements. Financial news influence on stock movements lasts for 2–3 days.	Adding the structural information to the prediction model will be able to improve the prediction performance. The use of different models to process news texts may also improve the results.	2019
Shastri et al. [41]	Stock price forecasting is more efficient for a shorter time.	Upgrade the sentiment analysis task by increasing the words that may affect the stock movements more.	2019
Picasso et al. [42]	The model with the LMD dictionary is more effective in annualized return measure, while the use of AS dictionary proved to be more effective in obtaining high accuracy results.	The model could not achieve overwhelming results compared to using news set alone. The use of advanced feature fusion methods will improve the results. Collect more news data for a longer period.	2019
Chen et al. [12]	Individual stock prediction using the study polarity lexicon was better than the benchmark model.	The research did not analyze detailed data; it only has the data that can be achieved by any public users.	2019
Liu and Wang [14]	NBAd structure accomplishes the best outcomes. Market predictions of the stock price at the minute time frame obtain better outcomes than those at day level.	Apply the NBA model in an index or industry-level data.	2019
Mudinas et al. [19]	In some cases, sentiment emotions contribute to Granger-cause stock price fluctuates, but the finding was not inclusive and must be examined for each case. For some stocks, adding sentiment emotions to the machine learning market prediction model will increase the prediction accuracy. SVM achieved better and more balanced results.	Enhancing the sentiment classification model and increasing the number of collected items will yield a better result.	2019

Reference	Finding	Limitation and future work	Year
Souza and Aste [22]	Multiplex network approach incorporating information from both social media and financial data can be used to forecast the causal relationship framework with high accuracy.	Investigate the impact of financial crises by expanding the historical data period. Use different techniques of the financial correlation establishment and apply it to portfolio management mechanisms.	2019
Wu et al. [23]	Adding news articles was able to reduce the RMSE that proves that the economic news has crucial impacts on market returns. The negative news has more influence on the stock market returns than positive news articles.	The research only tested the news texts published in the Knowledge Management Winner newspaper. Future study may include other online news datasets. Apply the proposed model to examine the stocks of smaller companies.	2019
Maqsood et al. [64]	Not all the main events have a crucial impact on stock market movements. More crucial local events affect the performance of the prediction model. Support vector regression gives the best prediction performance	Needs to exploit more than one social media website to produce sentiment analysis for a specific event. The use of financial news may improve the prediction result.	2020

Table 7.
Summaries of the reviewed work findings, limitations, and future work.

Moreover, as **Tables 2–5** indicate, the main trends in recent studies are to utilize deep learning methods instead of conventional machine learning to analyze the stock market textual information in the news or social media due to the advantages of DL that offer overconventional machine learning. DL promises enough amount of data and training time that conventional machine learning methods are unable to handle effectively.

Many recent studies only exploit sentiment analysis of textual data, and they neglect the important influence of historical stock prices, which affect their prediction accuracy results; this suggests that the incorporation of data from multiple sources will improve market prediction effectiveness. The more data fed into the prediction model, the better accuracy can be achieved.

Machine learning models described previously have been discussed to show how SVM and LSTM are highly preferred by investigators because of their high accuracy result in text classification and market prediction, whereas many other machine learning methods like K-nearest neighbors (k-NN), random forest (RF), linear regression, decision tree, artificial neural networks (ANN), etc. illustrate promising results for text mining and sentiment analysis task for market analysis but are least frequently used and need to be further investigated.

However, the reviewed work has some limitations; one of the main limitations is the insufficiency of highly structured datasets containing text data on markets for certain periods that researchers can use to integrate their analysis and assessment efforts; another problem is the imbalanced dataset that has been used by many researchers, which make discriminating result in prediction.

Future work should focus on predicting the movement of the stock market using structured data (past stock prices) along with textual data from different resources like financial news and social media. Moreover, to achieve better results

in predicting the stock market, the text mining procedure should improve feature selection, feature representation, and dimensionality reduction methods.

In general, many techniques will be able to improve the prediction methods such as adding the structural information to the prediction model, expanding the training period, using more effective expanded lexicons, adding different sources of information such as financial news articles, increasing the number of collected news for longer period, applying the deep learning models, upgrading the sentiment analysis task by increasing the words that may affect the stock movements more, using of more improved machine learning techniques for sentiment analysis such as Interdependent Latent Dirichlet allocation (ILDA), adding historical stock prices to the dataset with the news and social media information, and considering of event sentiments analysis as illustrated in **Table** 7.

8. Conclusion

Knowledge of stock movements by a fraction of a second can lead to high profits investors can make which makes stock market studies a major motivation for a researcher. The great advances and success of natural language process and sentiment analysis of online news based on machine learning and deep learning have gained huge popularity recently in the financial domain especially in market prediction models. This survey has discussed the recent current studies on market prediction systems based on text mining techniques with comprehensive clarifying of the model's main limitations and future improvement methods. The survey was undertaken on many major portions such as text preprocessing, machine learning algorithms, evaluation mechanisms, findings, and limitations associated with detailed discussion and explanation of the most successful used techniques. Moreover, this review provides a serious attempt to address the problem of market prediction based on the most recent text mining methods and provide a clear view of the future research direction. Recently, more extensive observations into the financial markets are required in the current dynamic world, since the absence of it can have a detrimental effect on the investments around the globe. It is therefore essential to undertake prediction models based on text mining research as a practical solution that can lead to a much greater degree of confidence in the understanding of market movements and make valuable investments. With the considerable amount of textual data available online, the need to build specialized text mining systems gradually evolves for each field of market analysis.

This study is intended to support other researchers to place the different theories in this research area more easily into practice and become able to make key decisions in the development of future models. The researches mentioned in this paper proved the effectiveness of text mining and sentiment analysis methods in predicting market movements. By comparing many ML methods such as SVM or decision tree and deep learning models like LSTM or CNN, we discussed some of these model's limitations and future work and debated the best result obtained by each one of these models. After all, the proposed survey displayed the need of improving the prediction methods such as adding the structural information, considering of event sentiments analysis, using more effective expanded lexicons, increasing the number of collected news, expanding the training period, applying the deep learning models, adding different sources of information, upgrading the sentiment analysis task by increasing the words that may affect the stock movements more, and using unified benchmark dataset and evaluation measures.

IntechOpen

Author details

Faten Subhi Alzazah* and Xiaochun Cheng
Department of Computer Science, Middlesex University, London, UK

*Address all correspondence to: fatensubhi@gmail.com

IntechOpen

References

[1] Gupta A, Dhingra B. Stock market prediction using hidden Markov models. In: 2012 Students Conference on Engineering and Systems. IEEE; 2012. pp. 1-4

[2] Asadi S, Hadavandi E, Mehmanpazir F, Nakhostin MM. Hybridization of evolutionary Levenberg–Marquardt neural networks and data pre-processing for stock market prediction. Knowledge-Based Systems. 2012;**35**:245-258

[3] Saravanan S, Mala S. Stock market prediction system: A wavelet based approach. Applied Mathematics and Information Sciences. 2018;**12**:579-585. DOI: 10.18576/amis/120312

[4] Chung H, Shin KS. Genetic algorithm-optimized long short-term memory network for stock market prediction. Sustainability. 2018;**10**(10):3765

[5] Long W, Lu Z, Cui L. Deep learning-based feature engineering for stock price movement prediction. Knowledge-Based Systems. 2019;**164**:163-173

[6] Agarwal B, Mittal N, Bansal P, Garg S. Sentiment analysis using common-sense and context information. Computational Intelligence and Neuroscience. 2015;**2015**

[7] Rajput V, Bobde S. Stock market forecasting techniques: Literature survey. International Journal of Computer Science and Mobile Computing. 2016;**5**(6):500-506

[8] Nguyen TH, Shirai K, Velcin J. Sentiment analysis on social media for stock movement prediction. Expert Systems with Applications. 2015;**42**(24):9603-9611

[9] Sun A, Lachanski M, Fabozzi FJ. Trade the tweet: Social media text mining and sparse matrix factorization for stock market prediction. International Review of Financial Analysis. 2016;**48**:272-281

[10] Schumaker RP, Chen H. Textual analysis of stock market prediction using breaking financial news: The AZFin text system. ACM Transactions on Information Systems (TOIS). 2009;**27**(2):1-9

[11] Ming F, Wong F, Liu Z, Chiang M. Stock market prediction from WSJ: Text mining via sparse matrix factorization. In: 2014 IEEE International Conference on Data Mining. IEEE; 2014. pp. 430-439

[12] Chen MY, Liao CH, Hsieh RP. Modeling public mood and emotion: Stock market trend prediction with anticipatory computing approach. Computers in Human Behavior. 2019;**101**:402-408

[13] Zhang X, Qu S, Huang J, Fang B, Yu P. Stock market prediction via multi-source multiple instance learning. IEEE Access. 2018;**6**:50720-50728

[14] Liu G, Wang X. A numerical-based attention method for stock market prediction with dual information. IEEE Access. 2018;**7**:7357-7367

[15] Weng B, Ahmed MA, Megahed FM. Stock market one-day ahead movement prediction using disparate data sources. Expert Systems with Applications. 2017;**79**:153-163

[16] Ding X, Zhang Y, Liu T, Duan J. Deep learning for event-driven stock prediction. In: Twenty-Fourth International Joint Conference on Artificial Intelligence; 2015

[17] Vargas MR, De Lima BS, Evsukoff AG. Deep learning for stock market prediction from financial news

articles. In: 2017 IEEE International Conference on Computational Intelligence and Virtual Environments for Measurement Systems and Applications (CIVEMSA). IEEE; 2017. pp. 60-65

[18] Khedr AE, Yaseen N. Predicting stock market behavior using data mining technique and news sentiment analysis. International Journal of Intelligent Systems and Applications. 2017;**9**(7):22

[19] Mudinas A, Zhang D, Levene M. Market trend prediction using sentiment analysis: Lessons learned and paths forward. 2019. arXiv preprint arXiv:1903.05440

[20] Granger CW. Investigating causal relations by econometric models and cross-spectral methods. Econometrica: Journal of the Econometric Society. 1969;**1**:424-438

[21] Checkley MS, Higón DA, Alles H. The hasty wisdom of the mob: How market sentiment predicts stock market behavior. Expert Systems with Applications. 2017;**77**:256-263

[22] Souza TT, Aste T. Predicting future stock market structure by combining social and financial network information. Physica A: Statistical Mechanics and its Applications. 2019;**535**:122343

[23] Wu GG, Hou TC, Lin JL. Can economic news predict Taiwan stock market returns? Asia Pacific Management Review. 2019;**24**(1):54-59

[24] Bujari A, Furini M, Laina N. On using cashtags to predict companies stock trends. In: 2017 14th IEEE Annual Consumer Communications & Networking Conference (CCNC). IEEE; 2017. pp. 25-28

[25] Dumais S, Platt J, Heckerman D, Sahami M. Inductive learning algorithms and representations for text categorization. In: Proceedings of the Seventh International Conference on Information and Knowledge Management; 1998. pp. 148-155

[26] Joachims T. Text categorization with support vector machines: Learning with many relevant features. In: European conference on machine learning. Berlin/Heidelberg: Springer; 1998. pp. 137-142

[27] Xie Y, Jiang H. Stock market forecasting based on text mining technology: A support vector machine method. 2019. arXiv preprint arXiv:1909.12789

[28] Li X, Xie H, Chen L, Wang J, Deng X. News impact on stock price return via sentiment analysis. Knowledge-Based Systems. 2014;**69**:14-23

[29] Loughran T, McDonald B. When is a liability not a liability? Textual analysis, dictionaries, and 10-Ks. The Journal of Finance. 2011;**66**(1):35-65

[30] Long W, Song L, Tian Y. A new graphic kernel method of stock price trend prediction based on financial news semantic and structural similarity. Expert Systems with Applications. 2019;**118**:411-424

[31] Porshnev A, Redkin I, Shevchenko A. Machine learning in prediction of stock market indicators based on historical data and data from twitter sentiment analysis. In: 2013 IEEE 13th International Conference on Data Mining Workshops. IEEE; 2013. pp. 440-444

[32] Xu F, Keelj V. Collective sentiment mining of microblogs in 24-hour stock price movement prediction. In: 2014 IEEE 16th Conference on Business Informatics, Vol. 2. IEEE; 2014. pp. 60-67

[33] Uysal AK, Murphey YL. Sentiment classification: Feature selection based

approaches versus deep learning. In: 2017 IEEE International Conference on Computer and Information Technology (CIT). IEEE; 2017. pp. 23-30

[34] Sohangir S, Wang D, Pomeranets A, Khoshgoftaar TM. Big data: Deep learning for financial sentiment analysis. Journal of Big Data. 2018;**5**(1):3

[35] Singhal P, Bhattacharyya P. Sentiment Analysis and Deep Learning: A Survey. Bombay: Center for Indian Language Technology, Indian Institute of Technology; 2016

[36] Martin V. Predicting the French stock market using social media analysis. In: 2013 8th International Workshop on Semantic and Social Media Adaptation and Personalization. IEEE; 2013. pp. 3-7

[37] Geva T, Zahavi J. Empirical evaluation of an automated intraday stock recommendation system incorporating both market data and textual news. Decision Support Systems. 2014;**57**:212-223

[38] Goldberg DE. Genetic Algorithms in Search, Optimization, and Machine Learning. Reading, MA: Addison-Wesley; 1989

[39] Khatri SK, Srivastava A. Using sentimental analysis in prediction of stock market investment. In: 2016 5th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions) (ICRITO). IEEE; 2016. pp. 566-569

[40] Chen W, Zhang Y, Yeo CK, Lau CT, Lee BS. Stock market prediction using neural network through news on online social networks. In: 2017 International Smart Cities Conference (ISC2). IEEE; 2017. pp. 1-6

[41] Shastri M, Roy S, Mittal M. Stock price prediction using artificial neural model: An application of big data. EAI Endorsed Transactions on Scalable Information Systems. 2019;**6**(20)

[42] Picasso A, Merello S, Ma Y, Oneto L, Cambria E. Technical analysis and sentiment embeddings for market trend prediction. Expert Systems with Applications. 2019;**135**:60-70

[43] Cambria E, Fu J, Bisio F, Poria S. AffectiveSpace 2: Enabling affective intuition for concept-level sentiment analysis. In: Twenty-Ninth AAAI Conference on Artificial Intelligence; 2015

[44] Hochreiter S, Schmidhuber J. Long short-term memory. Neural Computation. 1997;**9**(8):1735-1780

[45] Rao G, Huang W, Feng Z, Cong Q. LSTM with sentence representations for document-level sentiment classification. Neurocomputing. 2018;**308**:49-57

[46] Siami-Namini S, Namin AS. Forecasting economics and financial time series: ARIMA vs. LSTM. 2018. arXiv preprint arXiv:1803.06386

[47] Huynh HD, Dang LM, Duong D. A new model for stock price movements prediction using deep neural network. In: Proceedings of the Eighth International Symposium on Information and Communication Technology; 2017. pp. 57-62

[48] Mikolov T, Sutskever I, Chen K, Corrado GS, Dean J. Distributed representations of words and phrases and their compositionality. In: Advances in Neural Information Processing Systems; 2013. pp. 3111-3119

[49] Wang Y, Huang M, Zhu X, Zhao L. Attention-based LSTM for aspect-level sentiment classification. In: Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing; 2016. pp. 606-615

[50] Liu H. Leveraging financial news for stock trend prediction with attention-based recurrent neural network. 2018. arXiv preprint arXiv:1811.06173

[51] Kraus M, Feuerriegel S. Decision support from financial disclosures with deep neural networks and transfer learning. Decision Support Systems. 2017;**104**:38-48

[52] Li J, Bu H, Wu J. Sentiment-aware stock market prediction: A deep learning method. In: 2017 International Conference on Service Systems and Service Management. IEEE; 2017. pp. 1-6

[53] Collobert R, Weston J. A unified architecture for natural language processing: Deep neural networks with multitask learning. In: Proceedings of the 25th International Conference on Machine Learning; 2008. pp. 160-167

[54] Ho CC, Baharim KN, Fatan AA, Alias MS. Deep neural networks for text: A review. In: The 6th International Conference on Computer Science and Computational Mathematics. Langkawi, Malaysia; 2017

[55] Ding X, Zhang Y, Liu T, Duan J. Using structured events to predict stock price movement: An empirical investigation. In: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP); 2014. pp. 1415-1425

[56] Vu TT, Chang S, Ha QT, Collier N. An experiment in integrating sentiment features for tech stock prediction in twitter. In: Proceedings of the Workshop on Information Extraction and Entity Analytics on Social Media Data; 2012. pp. 23-38

[57] Moniz A, de Jong F. Classifying the influence of negative affect expressed by the financial media on investor behavior. In: Proceedings of the 5th Information Interaction in Context Symposium; 2014. pp. 275-278

[58] Bing L, Chan KC, Ou C. Public sentiment analysis in Twitter data for prediction of a company's stock price movements. In: 2014 IEEE 11th International Conference on e-Business Engineering. IEEE; 2014. pp. 232-239

[59] Li X, Huang X, Deng X, Zhu S. Enhancing quantitative intra-day stock return prediction by integrating both market news and stock prices information. Neurocomputing. 2014;**142**:228-238

[60] Shynkevich Y, McGinnity TM, Coleman S, Belatreche A. Stock price prediction based on stock-specific and sub-industry-specific news articles. In: 2015 International Joint Conference on Neural Networks (IJCNN). IEEE; 2015. pp. 1-8

[61] Cakra YE, Trisedya BD. Stock price prediction using linear regression based on sentiment analysis. In: 2015 International Conference on Advanced Computer Science and Information Systems (ICACSIS). IEEE; 2015. pp. 147-154

[62] Ghanavati M, Wong RK, Chen F, Wang Y, Fong S. A generic service framework for stock market prediction. In: 2016 IEEE International Conference on Services Computing (SCC). IEEE; 2016. pp. 283-290

[63] Gálvez RH, Gravano A. Assessing the usefulness of online message board mining in automatic stock prediction systems. Journal of Computational Scienc. 2017;**19**:43-56

[64] Maqsood H, Mehmood I, Maqsood M, Yasir M, Afzal S, Aadil F, et al. A local and global event sentiment based efficient stock exchange forecasting using deep learning. International Journal of Information Management. 2020;**50**:432-451