ISSN 2090-3359 (Print)
ISSN 2090-3367 (Online)

Advances in Decision Sciences

Volume 26
Issue 2
June 2022

Published by Asia University, Taiwan

A Comprehensive Review of Stock Price Prediction Using Text Mining

Maede TajMazinani

Department of Finance and Insurance, University of Tehran

maedetaj@ut.ac.ir

Hosein Hassani (corresponding author)

Research institute for energy management and planning, University of Tehran;

Hassani.stat@gmail.com

Reza Raei

Department of Finance and Insurance, University of Tehran

raei@ut.ac.ir

Received: December 31, 2021; First Revision: February 23, 2022;

Last Revision: May 5, 2022; Accepted: May 13, 2022;

Published: May 31, 2022

Abstract

Purpose. In various studies, the sentiment analysis identifies as an essential part of stock price behavior prediction. The availability of news, social media networks, and the rapid development of natural language processing methods resulted in better forecasting performance. However, there is a lack of a comprehensive framework and review paper to address the advantages and challenges of this very timely topic.

Design/methodology/approach. This paper aims to promote the existing literature in this field by focusing on different aspects of previous studies and presenting an explicit picture of their components. We, furthermore, compare each system with the rest and identify their main differentiating factors. This paper summarized and systematized studies that seek to predict stock prices based on text mining and sentiment analysis in a systematic review paper.

Findings. It discussed the developments made during recent years and addressed the existing gap in this field to the research community.

Keywords: Stock price prediction, Sentiment analysis, Text mining, Big data

1 Introduction

The stock price fluctuates between buyers and sellers until a reasonable price is reached. It can be said that the supply and demand rule sets stock prices. Discussing stock market prediction from various points of view is widespread (see, for example, Darsono et al., 2022; Franses, 2018; Kudryavtsev, 2020). With more people gaining access to the markets via mobile apps and other connected devices, more are keen on figuring out whether this market has any predictability. According to various studies, stock market forecasts are arduous (Nazário et al., 2017). Stock price analysis can be divided into two types: technical and fundamental. In technical analysis, prices are considered historically, price trends are predicted using diagrams, and investment decisions are then made. According to the efficient market hypothesis (EMH), news, events, and product releases determine financial markets' movement, influencing stock value (Fama, 1965). According to EMH, stock prices follow a random walk pattern. It is impossible to predict them with more than 50% probability, but several studies have rejected that notion (see, for instance, Ballings et al., 2015; Bollen et al., 2011; Chong et al., 2017; Qian & Rasheed, 2007).

Economists believe that stock prices depend on the macroeconomy, seasonal effects, and political events (Kao et al., 2013). Behavioral finance indicates that decision-making depends on emotions and moods (Diamond, 2008). Hon et al. (2021), and Wong (2020) review behavioral finance and its applications in the stock market in detail. With the rise of technology and developments in information release techniques, we are bombarded by information every second. Social media is a perfect platform to transfer opinions, thoughts, and views about any topic and issue between the public and significantly affects people's ideas and decisions (Hassani et al., 2020, Lee et al., 2022).

The stock market is affected by thousands of news items broadcast daily by news sites. Social networks such as Twitter, LinkedIn, and Facebook have inspired many people's lives as a core source of acquiring information these days. Social media has become an essential tool to spread knowledge about finance, especially the capital market. Predicting stock price based on sentiment analysis has attracted many kinds of research in finance and the natural language processing area to discuss the relationship between stock price behavior and sentiments of investors and news. Applying this approach steers many pieces of research to use such online sources (Geva & Zahavi, 2014; Moat et al., 2013; Nguyen et al., 2015; Weng et al., 2017).

Unstructured data can also be analyzed with fundamental analysis to forecast market trends. These data may be financial reports, formal documents, and online discussions (Nassirtoussi et al., 2014). To perform a fundamental analysis, news and articles about a company are considered as core and vital information to predict future trends (Kalyani et al., 2016). Text mining in stock price prediction refers to the recent ten years. Most of the previous studies are done in English because natural language processing tools are mainly developed for it. However, text mining and analysis has spread to other languages and developing markets in recent years, including Chinese, Arabic, and Persian (Derakhshan & Beigy, 2019), (Liu et al., 2017), (Chen et al., 2018).

Review articles make practitioners familiar with crucial research sources and provide a sound vision of the current situation in the field under study (Nassirtoussi et al., 2014). The research path in stock price prediction has changed during the last few years. They are mainly based on technical analysis in the past as the data was more available and easy to process (Mizuno et al., 1998), (Adebiyi et al., 2012), (Dash & Dash, 2016).

This research intends to give a detailed study of various stock market prediction techniques based on sentiment analysis. It considers several existing news-based stock market prediction techniques to identify the possible relationship between textual information and the stock exchange. This review paper is carried out using the applied methodologies, sentiment sources, and datasets employed. Thus, the study acts as the motivation for the future stock market prediction based on sentiment analysis. Section 2 presents the foundational concepts and theories of stock price prediction and text mining. Section 3 provides a review of the primary studies and Section 4 concludes this work and suggests future research ideas.

2 Technical and Methodological

2.1 Stock price prediction

Stock price prediction literature is primarily based on the efficient market hypothesis (EMH) and random walk theory. According to random walk theory, the stock price trend does not help in predicting the future. However, the EMH states we can better analyze the financial market and a company's stock value by utilizing news, events, and product release (Fama, 1965). Though, several studies refuted this hypothesis and demonstrated that the market is somewhat predictable (Ballings et al., 2015), (Bollen et al., 2011), (Chong et al., 2017). Many researchers have investigated stock exchange patterns and discussed this hypothesis in behavioral economics and finance (Bikas et al., 2013; Fox & Sklar, 2009; John R Nofsinger, 2005; Smith, 2003). According to EMH, market prices reflect the value precisely and answer recent changes and knowledge (Chen et al., 2018). The technical analysis uses historical stock price and volume data to predict future returns on financial assets (Nazário et al., 2017). Technical indicators such as relative strength index, moving averages, difference of exponential moving averages (MACD), and money flow index are among the most known besides the chart patterns that constitute the principles of technical analysis. Several studies apply technical analysis rules with intelligent system techniques (Bisoi & Dash, 2014; Kazem et al., 2013; Wei et al., 2011). We can utilize historical prices and volume data to form a trading system that can help predict the direction of asset prices in the future. Here, studies on stock price prediction based on technical analysis apply different approaches to examine the effectiveness of technical analysis, such as trading systems

(Berutich et al., 2016; Cervelló-Royo et al., 2015; Taylor, 2014), computational techniques (Bisoi & Dash, 2014; da Costa et al., 2015; Zhu et al., 2015), and chart patterns (Friesen et al., 2009; Wang & Chan, 2007; Zapranis & Tsinaslanidis, 2012). In terms of operational tools, studies continue to use many sophisticated machine learning-based models such as neural networks (Sang & Di Pierro, 2019; Sezer et al., 2017; Ticknor, 2013) and genetic algorithms (Ahmadi et al., 2018; Sezer et al., 2017).

Behavioral finance studies the effect of psychological processes on decision-making. This method is based on external factors such as company and market conditions, political and economic factors, textual information in financial news articles, social networks, and even blogs by economic analysts (Vijh et al., 2020). Fundamental analysis is usually based on processing unstructured data. These data can be financial reports, formal documents, or online comments and debates (Nassirtoussi et al., 2014). In fundamental analysis, a company's financial conditions and macroeconomic indicators like EBITDA, P/E, income, return on equity, and the dividend yield is determinant factors for decision-making. Because of this, financial analysts trade stocks with values higher or lower than their intrinsic values. According to EMH, the intrinsic value of a stock equals its current price (Picasso et al., 2019).

2.2 Text mining

"*Text mining refers to obtaining information and patterns implicit, previously unknown, and potentially valuable in big unstructured textual data, such as natural-language texts*"
(Hassani et al., 2020).

Generally, Natural language processing (NLP) describes how humans communicate, evaluate, and analyze text and speech (Brownlee, 2017). It also refers to the extraction of knowledge from unstructured data. The process of text mining includes. Text mining provides text pre-processing, applying techniques, and text analysis (Hagenau et al., 2013). A text representation model shows a text as a vector of features (Hassani et al., 2020). There are various approaches for text vectorization; word Embedding, deep learning-based methods like embedding layers in neural networks, using the pre-trained vectors such as Elmo, BERT (Brownlee, 2017), noun phrases (Schumaker & Chen, 2009), n-gram (Hagenau et al., 2013), topic modelling (Nguyen et al., 2015), sentiment words (Li et al., 2014), and a bag of words (Groth & Muntermann, 2011; Shynkevich et al., 2016) are applied in the literature.

2.2.1 Text pre-processing

A textual dataset may be derived from various sources. Researchers prefer to select financial websites due to less noise and more relevant data. Reuters and Bloomberg are among the most preferred ones (Ding et al., 2015; Peng & Jiang, 2015; Nassirtoussi et al., 2014; Bollen et al., 2011). Investors use different platforms such as Twitter to share their opinions about stocks, investments, or capital markets. Studies on these platforms are done primarily after 2017 (see, for example, Chen et al., 2018; Derakhshan & Beigy, 2019; P. Liu et al., 2018). In recent years, attempts have been made in languages other than English (Alkubaisi et al., 2018; Chen et al., 2018; Derakhshan & Beigy, 2019; Guo et al., 2017; Katayama & Tsuda, 2018; Liu et al., 2018; Liu et al., 2017; Zhang et al., 2018). Here, we also consider them to steer the future works in non-English speaking countries.

Text pre-processing is a method for cleansing and preparing textual content in a particular context. The final purpose is to lessen the textual content to only the necessary phrases for NLP desires. The first step in text pre-processing is noise removal. The noise required to put off textual content typically relies on its source. For instance, access to data can be achieved via the Twitter API or web pages. The second is tokenization. We need access to each word in a string, and it should be broken into smaller components. Breaking text into smaller parts is called tokenization, and the individual pieces are called tokens. We need Stemming and Lemmatization to bluntly remove prefixes and suffixes from a word and replace a single-word token with its root. The latter one is stopping word elimination. Stop words are removed during pre-processing when we do not attend sentence structure. They are usually the most common words that do not provide any considerable information about the statement, including "a," "an," and "the." After pre-processing the text, it is required to convert it into an appropriate form for machine learning algorithms. In this step, word embedding comes in.

Word embedding refers to techniques in natural language processing (NLP) in which words or phrases are converted to vectors of real numbers. The most common word embedding methods are BOW, N-gram TD-IDF, and Word2Vec. The BOW is usually employed in document classification methods, where each word's incident is considered a characteristic for training a classifier (Nassirtoussi et al., 2014; Geva & Zahavi, 2014; Nguyen et al., 2015; Peng & Jiang, 2015; Ruiz et al., 2012; L. Zhang, 2013). In the following, limitations of the BOW technique are given. Semantic meaning: the bare BOW

does not keep in mind the concept of the word in the text. It relinquishes the context which is used. The same word can be utilized in various places according to the context of close words. In some cases, it may be necessary to disregard words based on their relevance to the case because a colossal vector size in a large document can result in much computation.

The Bag-of-words model is an order-less document representation. The n-gram model can store this spatial information. Conceptually, the BOW model can be viewed as a specific state of the n-gram model, with n=1. Term frequency-inverse document frequency displays the importance of a word. A numerical statistic is often used as a scoring element in information retrieval searches and text mining. The TF–IDF is equilibrated by the number of texts containing the word. Zhang (2013) applied this method to choose the best features. Chen et al. (2019) used it to propose a model, and Nguyen et al. (2015) used TF-IDF for feature weighting.

Topic modeling is another important field of NLP besides word embedding. Word embedding is also employed to grab features from text data and map words, sentences, vectors, and numbers. Topic modeling is similar to clustering for numeric data and applied for unsupervised classification of documents. It is important to note that topic modeling is not the same as topic classification. As a supervised learning technique, topic classification involves training a model using manual annotations and predefined topics. After training, the model accurately classifies unseen texts according to their topics. However, grammatical content and order of words are not considered in the model.

2.2.2 Sentiment analysis

Behavioral economics has furnished plenty of evidence that financial decisions are notably driven by employing sentiment (Zhang et al. 2018). A society's aggregate optimism or pessimism can influence investor choices (Nofsinger, 2005). The sentiment is substantially used in data mining or social media analysis since attitudes are critical to investigating human behavior (Chakraborty, 2019). Sentiment analysis corresponds to identifying the view associated with a piece of text (Henrickson, 2019). The goal of sentiment evaluation is to discover opinions, classify the attitude they create, and categorize them as division-sensitive. Sentiment evaluation can be used to review enterprise products, ascertain the highs and lows of stock prices (Yu, 2013), and understand the idea of people reading news and views

expressed via humans in political debates (Chakraborty, 2019). The field of sentiment evaluation focuses on the judgments, responses, and emotions derived from the text. Sentiment analysis is applied mainly on three levels: document level, sentence level, and aspect level (Liu, 2020). Machine learning and lexicon-based approaches are common ways of sorting sentiments (Chakraborty et al., 2018). *"Lexicon is a collection of predefined words where a polarity score is associated with each word" (Sharma et al., 2020)*. In this approach, the classifier detects the polarity using a lexicon, and the quality of classification depends on lexicon size. Dictionaries can be created manually or using seed words (Taboada et al., 2011). A polarity analysis and opinion mining system identify the sentiment of textual content as positive, negative, or neutral by applying natural language processing and text mining (Mostafa, 2013). Positive and negative values are assigned to every positive term (a word or phrase) and negative term. In the simplest case, the sum of all values for the document is added; if the sum is positive, the document is in a positive category, negative is in the negative category, and zero is in the neutral category (Liu, 2020). The most commonly used dictionaries are WordNet, SentiWordNet, and SenticNet.

Polarity analysis is hugely dependent on its fields, such as the stock market, products, or education (Padmaja & Fatima, 2013). The polarity of an input sentence is detected in machine learning by analyzing features from labeled data. Supervised and unsupervised are two branches of the machine learning method (Sharma et al., 2020). In supervised learning, the first stride is training the model using a labeled dataset, and the second is the prediction of a given test dataset. Unsupervised learning is employed once the dependability of tagged data is rugged. It is more straightforward to gather untagged data than labeled data, and the sentence is classified based on keyword lists of each cluster. When analyzing domain-dependent data, the unsupervised approach is easier to use (Sharma et al., 2020). A vital source for text evaluation is news articles. The investor sentiments can be exploited from social media platforms and news sites. Several studies revealed that financial news could affect stock price changes. In Table 1, details of the pre-processing of reviewed works are shown.

Table 1.

Pre-processing

Reference	Text processing and sentiment analysis
Bollen et al.(2011)	Opinion finder and Google profile of mood states are employed to assess the sentiments of tweets.
Ruiz et al. (2012)	Stock tickers and hashtags are used to extract tweets. Tweets are represented as graphs based on users, retweets, URLs, etc. then features are extracted according to the graphs.
Zhang et al. (2013)	A bag of words is employed for tokenization, and stop words are removed. N-gram and SentiStrength lexicon are applied for feature extraction. Pearson's Chi-squared test and TD-IDF are applied to choose the best features.
Geva, Zahavi(2014)	As feature representations, news-item count and bag of words are utilized. Commercial sentiment scoring software by Digital Trowel is applied for sentiment score.
Dickinson and Hu (2015)	Stanford N.L.P. Sentiment Classifier is employed to predict the sentiment of tweets. Both n-gram and Word2Vec techniques are applied for feature representation and pre-process raw text before random forest classifier to evaluate the accuracy of Stanford prediction
Ding et al.(2015)	Events are represented as tuples. Structured events are extracted from text using Open IE technology and dependency parsing. ReVerb to extract the candidate tuples of the event and then parse the sentence with ZPar to extract the subject, object, and predicate.
Nguyen et al.(2015)	Six features are used: price, human sentiment (messages labeled by users), classified sentiments, LDA-based and JST-based features, and Aspect-based sentiments. Words are lemmatized by Stanford CoreNLP, a bag of word technique is used for feature representation, and the feature

	weighting is TF-IDF.
Pagolu et al. (2016)	Tokenization, stop words removal, and regex matching for removing special Characters are done for pre-processing. N-gram and Word2Vec are used for feature representation and sentiment analysis.
Peng et al. (2016)	Word embedding, keep sentences that mention at least one stock name, group them based on the publication date and stock name, labeled positive, negative based on the next day's closing price, using a bag of words, polarity score, category tag for extract features
Liu et al.(2017)	A classical Chinese text segmentation tool called "Jieba" is employed. A sentiment score based on logistic regression, tf--idf, and n-gram is defined to assess the polarity of posts.
Chen et al.(2018)	Technical indicators are calculated based on historical prices. Content features are divided into two categories: sentiment features and LDA features. Sentiment features are extracted using sentiment dictionary: pos and neg keywords based on next day trend.
Nisar and Yeung(2018)	A lexicon-based sentiment classifier, Umigon, is applied to extract sentiment
Liu et al.(2018)	Since social media variables are number-based, any sentiment analysis or text processing method is not applied.
Zhang et al. (2018)	Naïve Bayes is used to infer tweets' sentiments, and a sentiment index is defined based on positive and negative sentiments.
Katayama and Tsuda (2018)	The Japanese evaluation polarity dictionary evaluates the polarity of news.
Das et al.(2018)	Stanford core NLP is employed for sentiment analysis.
Alkubaisi et al.(2018)	Tokenization, and removing stop words are done. The polarity of tweets is done by expert labeling technique
Weng et al.(2018)	PCA is applied for dimensionality reduction.

Chen et al. (2019)	Stock names are used as keywords—the tool jieba and a self-defined keyword dictionary segment the articles by part of speech. Word to Vec is applied, and TF-IDF is used to propose a lexicon.
Broadstock and Zhang (2019)	The Canadian National Research Council (NRC) definitions of tone and sentiment are employed to extract the sentiments of tweets.
Derakhshan & Beigi(2019)	Buy and sell labels are considered human semantics features; aspect-based sentiment, LDA, and LDA-POS are other methods.
Guo et al. (2017)	The sentiment classification model is based on a Chinese sentiment corpus from Renmin university in china. The classification model is Logistic regression, and then a Sentiment index based on the number of positive, negative, and neutral comments is calculated.
Maqsood et al (2020)	Tweets are tokenized into word vectors. Alex Davies' word list and SentiWordNet are applied for sentiment analysis.
Li et al (2020)	News articles are tokenized and converted into word vectors, stop words are excluded, and then word vectors are mapped into sentiment space using four sentiment dictionaries which are SenticNet 5, SentiWordNet 3.0, Vader, and LoughranMcDonald Financial Dictionary 2018
Audrino et al(2020)	The Deep-MLSA technique is employed to classify the polarity of tweets. Raven Pack News Analytics is utilized to measure the relevance score of news.
Xuan Ji et al (2021)	Doc2Vec is used for feature vectors, and a stacked auto-encoder is applied to balance the dimensions.

2.3 Machine learning algorithms

After the text is processed and converted to numbers, machine learning algorithms are applied. Prices (open, close, first, last, low, high) are employed to train the machine learning algorithms for prediction purposes. Predictions focus on either stocks like Apple, AMAZON (Nguyen et al., 2015), Microsoft (Pagolu et al., 2016), or indexes like Taiwan 50 index (Chen et al., 2019), DowJones Index (Audrino et al., 2020). Most of the works are categorical forecasts: up, down, rise, fall, positive, and negative, indicating the reaction of stock prices to events. Only a few pieces of research apply regression and econometric models (see, for example, Audrino et al., 2020; Geva & Zahavi, 2014; Maqsood et al., 2020; Ruiz et al., 2012). Some researchers utilize the trading hours, but others consider the time between trading hours of two days. The period varies from 5 days (Nisar & Yeung, 2018) to 7 years (Ding et al., 2015). Further details are shown in Tables 2 and 3.

2.3.1 Support vector machines (SVM)

Support vector machines (SVM) are used for classification and regression using hyperplanes in a high-dimensional or infinite-dimensional space. SVM constructs a model that devotes new instances to one class or another using a set of classified training examples. A linear support vector machine is a parametric model despite an RBF kernel SVM, and the latter's complexity grows with the size of the training set. Using a linear classifier is not justified if the data is separable linearly. The linear SVM is used in various studies (see, for instance, Derakhshan & Beigy, 2019; Nguyen et al., 2015; Zhang, 2013), and SVM with RBF-kernel is also applied by Zhang et al. (2018). Another typically used execution of SVM is LIBSVM which uses an SMO-type (Sequential Minimal Optimization) algorithm. Pagolu et al. (2016) predict stock price movements using this implementation.

2.3.2 Neural networks

Biological neural networks inspire an Artificial Neural Network (ANN) in the human brain to process information (Geva & Zahavi, 2014). In recent works, deep neural networks (DNNs), with multiple hidden layers between the input and output layers, are used (Ding et al., 2015; Peng & Jiang, 2015). DNNs can model complex nonlinear relationships. A multilayer perceptron (MLP) is composed of three or more layers. It applies a nonlinear

activation function (usually hyperbolic tangent or logistic) to classify not linearly severable data. For instance, multilayer perceptron applications in (NLP) are speech recognition and machine translation. Zhang et al. (2018) indicated that MLP achieves better performance due to effective learning of hidden relations between the features and the price movements. Convolutional neural networks (CNNs or ConvNets) are deep neural networks used for analyzing visual imagery. A convolutional neural network includes one or more convolutional layers and uses a variation of multilayer perceptron. CNNs are deeper networks with fewer parameters and perform very well in image and speech processing. In Ding's work (Ding et al., 2015), the model that considers CNN outperforms other models, including standard neural networks and SVM. A recurrent neural network (RNN) employs sequential data or time-series data. The output of recurrent neural networks depends on the primary elements within the sequence.

Apple's Siri and Google's voice search apply RNNs. Chen et al. (2018) apply RNN-boosting to betterment the prediction accuracy, and the results indicated that the proposed model is better than MLP and SVR. Deep learning uses Long Short-Term Memory (LSTM) networks, a type of recurrent neural network. It can process single data points (like images) and entire data sequences (such as video or speech). As a result, in recent studies, LSTM has been used in stock prediction based on text mining and sentiment analysis (Chen et al., 2019; Li et al., 2020). Chen et al. (2019) applied LSTM to compare the performance of his proposed lexicon and NSTUD in the Hong Kong market. The fuzzy neural network is also a neural network where the inputs and the connection weights are fuzzy numbers and combine the neural network's learning ability with the noise-handling capability of FL (Bollen et al., 2011).

2.3.3 Naïve Bayes

Naïve Bayes is a probabilistic classifier based on the Bayes Theorem. Zhang (2013) applied the naïve Bayes to assess the effect of tweets on stock prices and then compared naïve Bayes with SVM and maximum entropy as the rival method.

2.3.4 Random Forest

The "forest" refers to an ensemble of decision trees, and the output class is the mode of the classes (classification) or mean/average prediction (regression) of the individual trees. Dickinson & Hu (2015) applied random forest to predict the price movement of thirty DJIA

stocks. Random forest has been also used in conjunction with logistic regression and SVM for predicting Microsoft stock prices based on tweets (Pagolu et al., 2016). Random forest regression is used for stocks analysis such as Amazon and Pfizer (Weng et al., 2018).

2.3.5 Regression models

Support Vector Regression (SVR) is used in (Weng et al., 2018) and (Maqsood et al., 2020) to assess for stock market analysis. Ruiz et al. (2012) and Audrino et al. (2020) applied the Autoregressive (AR) model in their works. The Autoregressive model represents a random process applied to describe specific time-varying processes in nature, economics, etc. The autoregressive model determines that the output hinges linearly on its initial values and a stochastic term. Audrino et al. (2020) examine the volatility and value at risk using the heterogeneous autoregressive (HAR) model, one of the most popular models for realized volatility, due to its simplicity and suitable predictive performance. One of the ways to predict the values of one variable from the values of another is through univariate linear regression. The logistic model (or logit model) models the probability of a particular class, such as pass/fail, win/lose, or healthy/sick. Logistic regression employs a logistic function to model a binary dependent variable though more advanced extensions exist (Pagolu et al. (2016; Liu et al., 2018).

Table 1.

Market data

Reference	Market	Index	Time-Frame	Period	Forecast type
Bollen et al. (2011)	stocks	Dow Jones Industrial Average (DJIA)	daily	February 28, 2008, to December 19, 2008	categorical: up-down
Ruiz et al. (2012)	stocks	S&P500 Index	daily	The first half of 2010	regression
Zhang et al. (2013)	stocks	Technology ETF	intraday	March 2013	correlation
Geva, Zahavi (2014)	stocks	S&P500 Index	intraday	September 2006 to August 2013	regression
Dickinson and Hu (2015)	stocks	30 companies of the Dow Jones Industrial Average	daily	November 2014 through March 2015	correlation
Ding et al. (2015)	stocks	S&P500 Index	daily	October 2006 to November 2013	categorical: increase, decrease
Nguyen et al. (2015)	stocks	Eighteen stocks, including apple, amazon, Dell, eBay, and Google.	daily	July 23, 2012, to July 19, 2013	categorical: up, down
Pagolu et al. (2016)	stocks	Microsoft	daily	August 31, 2015, to August 25, 2016	categorical: rise, fall

Peng et al. (2016)	stocks	CRSP database	daily	October 2006 to December 2013	categorical: up-down
Liu et al. (2017)	stocks	Ten different stocks	daily	September 25, 2015, to September 30, 2016	categorical: rise, go down
Chen et al.(2018)	stocks	Shanghai-Shenzhen 300 Stock Index (HS300) and CSI.	daily	January 1, 2015, to February 14, 2017	categorical: up-down
Nisar and Yeung (2018)	stocks	FTSE100	daily	May 4, 2016- May 9, 2016	regression
Liu et al. (2018)	stocks	77 stocks	Daily	2009-2016	regression
Zhang et al. (2018)	stocks	A-share market	daily	November 2014 to May 2015	categorical: up-down
Katayama and Tsuda (2018)	stocks	TOPIX 500	daily	January 1983 until the end of December 2016	regression
Das et al. (2018)	stocks	Google, Microsoft, Apple	daily	May 2005 to June 2017	regression.
Alkubaisi et al. (2018)	stocks	Al-marai Saudi Arabia	daily	18-September-2016 to 25-May-2017	categorical
Weng et al. (2018)	stocks	Nineteen stocks, including Amazon, Pfizer, I.B.M	daily	January 2013 to December 2016	regression
Chen et al. (2019)	stocks	Taiwan 50 Index	daily	2016-2017	categorical: fall, rise

Broadstock and Zhang (2019)	stocks	Exxon Mobil, General Electric (GE), Chesapeake Energy (CHK), Ford Motor Company (F), Disney (DIS), and Walmart	intraday	August 2018	regression
Derakhshan & Beigi (2019)	stocks	18 shares	daily	July 2012 to July 2013	categorical: up-down
Guo et al. (2017)	stocks	Two industries	daily	January 2014 to June 2015	Lead-lag relation
Maqsood et al. (2020)	stocks	Stocks from the US, Turkey, Pakistan, and Hong Kong	daily	2012-2016	regression
Li et al. (2020)	stocks	Three representative stocks in each sector of HIS(the Hang Seng Index)	daily	January 2003 to March 2008	categorical: fall, rise, horizontal
Audrino et al. (2020)	stocks	Dow Jones Industrial Average Index,18 companies from NYSE: Intel, Microsoft, Citigroup, ...	Daily	the beginning of 2012 until the end of 2016	regression
Xuan Ji et al (2021)	Stocks	Meinian Health	Daily	January 2010 to November 2019	regression

Table 3.

Prediction algorithms

Reference	Algorithm Details	News and tech data
Bollen et al.(2011)	Granger causality analysis and Self-organizing Fuzzy Neural Network (SOFNN.)	No
Ruiz et al. (2012)	Autoregressive and twitter-augmented regression	No
Zhang et al.(2013)	Pearson's correlation, Naive Bayes classification, Maximum Entropy classification, Support Vector Machines.	No
Geva, Zahavi(2014)	Neural networks, Genetic algorithm, stepwise logistic regression	Yes
Dickinson and Hu (2015)	Pearson's correlation, Random forest	No
Ding et al.(2015)	Deep convolutional neural network	No
Nguyen et al.(2015)	SVM with linear kernel	
Pagolu et al (2016)	Random forest, Logistic regression, SVM	No
Peng et al. (2016)	Deep neural network	No
Liu et al.(2017)	Recurrent neural network	No

Chen et al.(2018)	Recurrent neural networks	Yes
Nisar and Yeung(2018)	Pearson's correlation and multiple regression	No
Liu et al(2018)	univariate regression	No
Zhang et al (2018)	SVM and MLP	Yes
Katayama and Tsuda (2018)	OLS regression	No
Das et al(2018)	R.N.N.	No
Alkubaisi et al. (2018)	Naïve Bayes Classifiers	No
Weng et al. (2018)	Neural network regression, support vector regression, boosted regression tree, random forest regression	Yes
Chen et al. (2019)	LSTM	No
Broadstock and Zhang (2019)	OLS regression	No
Derakhshan & Beigi (2019)	Neural network	No
Guo et al(2017)	The thermal optimal path (TOP) method	No
Maqsood et al (2020)	Linear regression, support vector regression, and deep learning	No

Li et al. (2020)	LSTM, SVM, MKL	Yes
Audrino et al(2020)	Heterogeneous autoregressive (HAR) model, economic-HAR and sentiment-HAR	No
Xuan Ji et al (2021)	LSTM,ARIMA models, RNN	Yes

3 Applications

We will review previous studies by examining the various components of the process used. Studies in this field have changed since previous reviews (Nassirtoussi et al., 2014). For instance, recent studies focus mainly on Twitter in US markets. In a research study entitled ``Twitter mood predicts the stock market``, Bollen et al. (2011) examined the effect of mood derived from Twitter on the value of the Dow Jones index over time. They use the two opinion finder tools, which measure positive and negative mood, and the Google profile of the mood states, which measure mood from six aspects (composure, alertness, confidence, caution, kindness, and happiness). The results of their research show an accuracy of 86.7% in predicting an increase or decrease in the final price of the Dow Jones Industrial Average. Ruiz et al. (2012) examined the correlation between Twitter activity, stock prices, and trading volume changes. Their research indicated that the correlation with trading volume is stronger than the price. However, using a simulator, they showed that a trading strategy performed better than other basic strategies using this low correlation between news and stock prices. Zhang et al. (2013) examined the effect of different machine learning methods on presenting positive and negative concepts in Twitter posts. In addition, they used the concepts extracted from Twitter to find correlations with stock prices. They also determined what words in the tweets correlated with changes in stock prices. Their research showed that the support vector machine has higher accuracy in forecasting.

Dickinson and Hu (2015) have explained the correlation between emotions and stock price movements using Twitter. Their research used both Word2Vec and n-gram methods with random forest. Their results demonstrate that some companies' stock values, such as Microsoft and Walmart, have a strong positive correlation. In contrast, others, such as Goldman Sachs, have a strong negative correlation. Pagolu et al. (2016) also used Twitter's emotion analysis to predict the Dow Jones index. For this purpose, they used Twitter data and word2vec and Ngram methods to analyze emotions. Their research reveals a correlation between rising and falling stock prices with public sentiment in tweets. In several other works from 2018 to 2020, the positive effect of Twitter data on stock market prediction has been confirmed. For example, Das et al. (2018) examined the impact of Twitter data on the forecast of US market shares using recurrent neural networks to analyze market sentiment

and trends. Their results show that Twitter data positively affects stock market forecasting. Broadstock and Zhang (2019) examine the impact of social media (Twitter) on the daily returns of the US stock market. They looked at Twitter sentiment at both company and financial market levels and studied the returns on a sample of US corporate stocks in 1,5, and 30-minute intervals using CAP-based regression. Their research suggests a positive relationship between news and daily stock returns.

Maqsood et al. (2020) investigated the impact of Twitter on stocks in the United States, Hong Kong, Pakistan, and Turkey using linear regression, support vector regression, and in-depth learning. They used dictionaries from SentiWordNet and AlexDavies to show that domestic and foreign events affect stock market forecasts and that some global events, such as the US election, affect other countries. Nisar and Yeung (2018) also consider a unique event. They collected a sample of 60,000 tweets from 6 days ago, during and after the local elections, and examined its relationship to changes in the London FTSE100 index. Their research findings reveal a correlation between the FTSE100 index and general trends and mood and investor behavior in the short term, although this relationship is not yet statistically significant. Back to the Twitter effect, Audrino et al. (2020) surveyed the impact of emotion and attention metrics on stock market volatility, particularly the Dow Jones Industrial Average and 18 stocks on the New York Stock Exchange. They used heterogeneous autoregressive models and the Deep-MLSA technique to determine the polarity of tweets.

In Saudi Arabia, Alkubaisi et al. (2018) presented a stock market classification model based on Twitter emotion analysis. Tweets related to the shares of the Saudi company El Marai and New Bay networks to achieve a 90% accuracy. Geva and Zahavi (2014) evaluated the effect of adding market data to textual news using data mining methods to predict stock returns in daily trading. They found that considering the two data sources enriches the information available to the prediction model and can reveal common patterns that may not be identified when using each source separately. The highest prediction accuracy is obtained using the nonlinear neural network prediction algorithm. Their research sample included 72 companies from the S&P index and Reuters news. Ding et al. (2015) used the deep convolutional neural network approach and the financial news of Reuters and Bloomberg to predict the S&P 500 index. Their research shows that the prediction accuracy of the proposed method has improved by 6% compared to the previous best methods. Peng et al. (2015) also examined the impact of Reuters and Bloomberg News on predicting US stock market

movements. In their research, they used deep recurrent neural networks and word embedding. The accuracy of forecasting using financial news is significantly improved based on their results.

Nguyen et al. (2015) used Yahoo Finance Message Board as the corporate news source and discussed the effect of emotions and trends of companies' news and headlines on stock price forecasts. Comparing the average prediction accuracy of 18 stocks during a trading year demonstrated that their proposed model performed 2.07% better than considering historical prices alone. In the Chinese market, several researchers discussed the effect of local social networks on the stock market. In this regard, Liu et al. (2017) have predicted stock fluctuations by classifying posts on the East Money Forum, a Chinese stock exchange discussion room, into positive and negative. The results indicate that the accuracy of predicting stock fluctuations is increased by using the posts in discussion rooms. After that, Chen et al. (2018) proposed a model based on the news published on social media Sina Weibo and deep recurrent neural networks. They were able to improve the accuracy of the prediction.

In the same year, Liu et al. (2018), in a study entitled "Financial Market Tweets: The Impact of Media in the Big Data Age", collected daily data from the search engines Baidu 360, the financial platform Hexun and Sina Weibo. Their findings from a survey of 77 companies on the Shanghai-Shenzhen Stock Index confirm that trading volume and turnover ratios positively correlate with media activity. In contrast, the relationship is negatively correlated with stock returns. This relationship is also weaker for social media than other channels. Zhang et al. (2018), based on Xueqiu social network tweets, which are similar to Twitter but for investors, proposed a model to predict the share price movements in the Chinese stock market using a support vector machine and the perceptron network. New Biz networks categorize tweets into positive, negative, and neutral. Their research findings suggest that predictive performance improves by considering the characteristics of emotions. Guo et al. (2017) also consider Xueqiu for their research. Their sentiment classification model was based on a Chinese sentiment corpus from Renmin University in China, and the logistic model was implemented. According to the results, sentiment does not always lead to stock price fluctuations. The effect of comments on the Oriented Fortune website on stock Meinian health by the LSTM model has also been studied (Ji et al., 2021).

In Japan, Katayama and Tsuda (2018) examined the impact of news on companies listed on Japan's Topix 500 Index. The results showed that if positive news is published, the company's stock price will increase. The effect of this case will be more significant when the news is on the home page. The amount of this effect is higher for companies with smaller market value. Li et al. (2020) discussed this issue in the Hong Kong market. They implement four different dictionaries and fully connected neural networks. Their model considers both technical and textual data. Their results showed that the specific finance domain dictionary outperforms the others, and incorporating both prices and sentiments in models improves the models' results.

Weng et al. (2018) propose models based on machine learning methods such as neural networks, backup vector regression, random tree, and textual sources, including Google search engine information, Wikipedia, and technical indicators, to predict stock prices in The US stock market. Their proposed model has improved forecasting accuracy. Derakhshan and Beigi (2019) have examined the effect of shareholder tweets on the price movement of 18 shares in the Iranian market. Their research is one of the few works based on the Persian language. They considered buying and selling labels in SAHAMYAB as an emotional feature and used LDA, LDA-POS methods, and neural networks. The results indicate a 55.33% accuracy for Iranian stocks. In the following, further detail about reviewed works in case of textual data and findings are shown in Tables 4 and 5.

Table 4.

Textual data

Reference	Text Type	Text Source	No. of items	language
Bollen et al. (2011)	tweets	Twitter	9,853,498 tweets	English
Ruiz et al. (2012)	tweets	Twitter	26 million tweets	English
Zhang (2013)	tweets	Twitter	Over 1 million tweets	English
Geva, Zahavi (2014)	news	Reuters 3000	51263 news item	English
Dickinson and Hu (2015)	tweets	Twitter	Over 2 million tweets	English
Ding et al. (2015)	news	Reuters & Bloomberg	More than 10 million events	English
Nguyen et al.(2015)	news	Yahoo Finance Message Board	18 message boards	English
Pagolu et al. (2016)	tweets	Twitter	250000 tweets	English
Peng et al. (2016)	news	Reuters & Bloomberg	553666 articles	English
Liu et al. (2017)	tweets	East Money Forum	96000 posts	Chinese
Chen et al. (2018)	tweets	Sina Weibo	808,283 tweets	Chinese
Nisar and Yeung (2018)	tweets	Twitter	60000 tweets	English

Liu et al. (2018)	news and tweets	The baidu ‹360 ‹Hexun‹ Sina Weibo	Not mentioned	Chinese
Zhang et al. (2018)	tweets	Xueqiu	6.48 million tweets	Chinese
Katayama and Tsuda(2018)	news	Nikkei Telecon	-	Japanese
Das et al (2018)	tweets	Twitter	560,000 tweets	English
Alkubaisi et al. (2018)	tweets	Twitter	3246 tweets	Arabic
Weng et al. (2018)	news	Wikipedia, Google, Financial News	-	English
Chen et al. (2019)	news	China Times Finance, Yahoo Stock Market News, Google Finance News, and China Electronics News	130000 articles	English
Broadstock and Zhang (2019)	tweets	Twitter	-	English
Derakhshan & Beigi (2019)	tweets	SAHAMYAB/Yahoo message board	787,547 comments on yahoo & 21205 comments on Sahamyab	Persian/English
Guo et al. (2017)	tweets	Xueqiu	Not mentioned	Chinese
Maqsood et al. (2020)	tweets	Twitter	11.42 m	English
Li et al. (2020)	news	FINET	Not mentioned	English
Audrino et al. (2020)	tweets, financial news	Twitter, Stock Twits, Raven Pack News Analytics, Google, Wikipedia,	2,524,369 tweets, 1,862,690 Stock Twits tweets	English

Xuan Ji et al (2021)	news	Oriented Fortune website	530813 documents	Chinese

Table 2.

Findings of the reviewed works

Reference	Findings	Trading Strategy
Bollen et al.(2011)	The accuracy of DJIA predictions can be significantly improved by including specific public mood dimensions. The accuracy is 86.7%.	No
Ruiz et al. (2012)	The correlation of micro-blogging is stronger with traded volume than with the stock price.	Yes
Zhang et al.(2013)	Naïve Bayes and SVMs with TD-IDF scoring feature selection outperform other methods, and the accuracy is about 87% in scoring tweets. There was almost no correlation between intraday tweets and stock prices.	No
Geva, Zahavi(2014)	integrating market data with textual data improves the modeling performance, and using more advanced textual data representations further improves predictive accuracy.	Yes
Dickinson and Hu (2015)	There is not a uniform connection between sentiment and price across all companies. The correlation is strongly positive in several companies, particularly consumer-facing corporations	No
Ding et al.(2015)	Compared to the state-of-the-art baseline models, the proposed model can achieve a 6% improvement in S&P500 index prediction and individual stocks.	Yes
Nguyen et al.(2015)	The average accuracy is 54.41 %, but the proposed model can predict the stock price movements with more than	No

	60% accuracy for a few stocks.	
Pagolu et al. (2016)	Sentiment analysis results show that the accuracy of word2vec and n-gram techniques are very close and about 70%.	No
Peng et al. (2016)	Features derived from financial news can significantly improve the prediction accuracy, and the best performance is 56.87%	No
Liu et al.(2017)	Using RNN with sentimental indicators can boost prediction accuracy.	No
Chen et al.(2018)	The average accuracy of Tech+Sent+LDA20 is 65.28% and outperforms the other feature subsets. RNN-Boost is better than a single RNN.	No
Nisar and Yeung(2018)	There is a correlation between public sentiment regarding the local elections and FTSE movements with various time lags.	No
Liu et al.(2018)	Media effect exists in new media channels: trading volume and turnover ratio are positively related to activities in new media, while the stock return is negatively related	No
Zhang et al. (2018)	Involving Xueqiu features achieves better performance in both ACC and AUC metrics. Besides, the MLP model outperforms the SVM model	No
Katayama and Tsuda (2018)	The sentiment of news has a specific influence on stock prices. The influence becomes more remarkable when the article is on the front page and for companies with smaller capitalization.	No
Das et al.(2018)	The study indicates that sentiment analysis of public mood derived from Twitter feeds can eventually forecast individual stock price movements.	No
Alkubaisi et al.(2018)	The best prediction accuracy is 90.38 % with hybrid naïve Bayes classifiers.	No
Weng et al. (2018)	Boosted regression tree (BRT) and the Random Forest Ensemble (RFR) as the best models for predicting the 1-day	Yes

	ahead stock price. Online data contributes considerably to prediction accuracy. However, the importance of those online features reduces or varies remarkably over time.	
Chen et al. (2019)	The proposed lexicon outperforms the NTUSD. Lexicon, and there is a 1.56% gap in prediction accuracy between two lexicons in different stocks.	No
Broadstock and Zhang (2019)	Lagged sentiment terms are significant for all stocks.	No
Derakhshan & Beigi (2019)	The LDA-POS method outperforms the human sentiments method. The average accuracy on both the English and Persian datasets is 56.24% and 55.33%, respectively.	No
Guo et al. (2017)	The investor sentiment cannot always forecast the stock price; it is a momentous measure to present the market performance most time	No
Maqsood et al. (2020)	Both local and global events affect the prediction of the stock market. Some global events, such as the US elections, affect other countries.	No
Li et al (2020)	Applying both sentiment and technical is better than applying them solo. BASED ON BOTH INFORMATION SOURCES, the LSTM outperforms the MKL and the SVM in prediction accuracy and F1 score.	No
Audrino et al(2020)	Sentiment and attention variables are both influential in prediction volatility and can improve the precision of one-day-ahead value-at-risk predictions	No
Xuan Ji et al (2021)	The experimental results showed that the proposed model outperformed baseline models in MAE, RMSE, and R^2.	No

4 Conclusion

Text mining and sentiment analysis have become an integral part of stock price prediction. This paper summarizes the previous studies that focused on predicting the stock market using text mining in recent years to clarify the current situation and upcoming path for researchers. Here, we provide a comprehensive summary of the latest developments in stock price prediction using text mining techniques. The background concepts, including stock price prediction, text mining, and machine learning algorithms, are discussed. The key findings are as follows. The most common evaluation metrics are the F1 score, ACC, AUC, R2, and MSPE. Most accuracies are between 50% to 70%; however, a few cases have higher accuracy.

Some studies focus on the sentiment analysis sector and compare the effect of different lexicons on stock prices (Chen et al., 2019). On the other hand, some others pay more attention to different text processing techniques (Derakhshan & Beigy, 2019; Pagolu et al., 2016; Chen et al., 2018, Li et al., 2020). Sixty-eight per cent of the studies were in English, and Twitter data and the Twitter platform were the most used among others by researchers. Besides support vector machines, different neural network types, including deep learning, recurrent neural networks, and LSTM, are the most frequently used algorithms. Limited research studies applied technical and textual data together.

The previous studies confirm correlations between stock price and investor sentiments. As a result, investors, banks and financial institutions can apply the provided systems to achieve better returns. It is then recommended to focus on designing trading strategies and techniques, including text mining analysis.

References

Adebiyi, A. A., Ayo, C. K., Adebiyi, M., & Otokiti, S. O. (2012). Stock price prediction using neural network with hybridized market indicators. *Journal of Emerging Trends in Computing and Information Sciences, 3*(1).

Ahmadi, E., Jasemi, M., Monplaisir, L., Nabavi, M. A., Mahmoodi, A., & Jam, P. A. (2018). New efficient hybrid candlestick technical analysis model for stock market timing on the basis of the Support Vector Machine and Heuristic Algorithms of Imperialist Competition and Genetic. *Expert Systems with Applications, 94*, 21-31.

Alkubaisi, G. A. A. J., Kamaruddin, S. S., & Husni, H. (2018). Stock Market Classification Model Using Sentiment Analysis on Twitter Based on Hybrid Naive Bayes Classifiers. *Computer and Information Science, 11*(1), 52-64.

Audrino, F., Sigrist, F., & Ballinari, D. (2020). The impact of sentiment and attention measures on stock market volatility. *International Journal of Forecasting, 36*(2), 334-357.

Ballings, M., Van den Poel, D., Hespeels, N., & Gryp, R. (2015). Evaluating multiple classifiers for stock price direction prediction. *Expert Systems with Applications, 42*(20), 7046-7056.

Berutich, J. M., López, F., Luna, F., & Quintana, D. (2016). Robust technical trading strategies using GP for algorithmic portfolio selection. *Expert Systems with Applications, 46*, 307-315.

Bikas, E., Jurevičienė, D., Dubinskas, P., & Novickytė, L. (2013). Behavioural finance: The emergence and development trends. *Procedia-social and behavioral sciences, 82*, 870-876.

Bisoi, R., & Dash, P. K. (2014). A hybrid evolutionary dynamic neural network for stock market trend analysis and prediction using unscented Kalman filter. *Applied Soft Computing, 19*, 41-56.

Bollen, J., Mao, H., & Zeng, X. (2011). Twitter mood predicts the stock market. *Journal of computational science, 2*(1), 1-8.

Broadstock, D. C., & Zhang, D. (2019). Social-media and intraday stock returns: The pricing power of sentiment. *Finance Research Letters*.

Brownlee, J. (2017). *Deep learning for natural language processing: develop deep learning models for your natural language problems*: Machine Learning Mastery.

Cervelló-Royo, R., Guijarro, F., & Michniuk, K. (2015). Stock market trading rule based on pattern recognition and technical analysis: Forecasting the DJIA index with intraday data. *Expert Systems with Applications, 42*(14), 5963-5975.

Chakraborty, K., Bhattacharyya, S., Bag, R., & Hassanien, A. A. (2018). Sentiment analysis on a set of movie reviews using deep learning techniques. *Social Network Analytics: Computational Research Methods and Techniques, 127*.

Chen, M.-Y., Liao, C.-H., & Hsieh, R.-P. (2019). Modeling Public Mood and Emotion: Stock Market Trend Prediction with Anticipatory Computing Approach. *Computers in Human Behavior*.

Chen, W., Yeo, C. K., Lau, C. T., & Lee, B. S. (2018). Leveraging social media news to predict stock index movement using RNN-boost. *Data & Knowledge Engineering, 118*, 14-24.

Chong, E., Han, C., & Park, F. C. (2017). Deep learning networks for stock market analysis and prediction: Methodology, data representations, and case studies. *Expert Systems with Applications, 83*, 187-205.

da Costa, T. R. C. C., Nazário, R. T., Bergo, G. S. Z., Sobreiro, V. A., & Kimura, H. (2015). Trading system based on the use of technical analysis: A computational experiment. *Journal of Behavioral and Experimental Finance, 6*, 42-55.

Darsono, S. N. A. C., Wong, W.-K., Nguyen, T. T. H., Jati, H. F., & Dewanti, D. S. (2022). Good Governance and Sustainable Investment: The Effects of Governance Indicators on Stock Market Returns. *Advances in Decision Sciences, 26*(1), 69-101.

Das, S., Behera, R. K., & Rath, S. K. (2018). Real-time sentiment analysis of Twitter streaming data for stock prediction. *Procedia computer science, 132*, 956-964.

Dash, R., & Dash, P. K. (2016). A hybrid stock trading framework integrating technical analysis with machine learning techniques. *The Journal of Finance and Data Science, 2*(1), 42-57.

Derakhshan, A., & Beigy, H. (2019). Sentiment analysis on stock social media for stock price movement prediction. *Engineering Applications of Artificial Intelligence, 85*, 569-578.

Diamond, Peter A., Behavioral Economics (March 15, 2008). MIT Department of Economics Working Paper No. 08-03, Available at SSRN: https://ssrn.com/abstract=1108588 or http://dx.doi.org/10.2139/ssrn.1108588

Dickinson, B., & Hu, W. (2015). Sentiment analysis of investor opinions on twitter. *Social Networking, 4*(03), 62.

Ding, X., Zhang, Y., Liu, T., & Duan, J. (2015). *Deep learning for event-driven stock prediction.* Paper presented at the Twenty-Fourth International Joint Conference on Artificial Intelligence.

Fama, E. F. (1965). The behavior of stock-market prices. *The journal of Business, 38*(1), 34-105.

Fox, J., & Sklar, A. (2009). *The myth of the rational market: A history of risk, reward, and delusion on Wall Street*: Harper Business New York.

Franses, P. H. (2018). Prediction Intervals for Expert-Adjusted Forecasts. *Advances in Decision Sciences, 22*, 1-12.

Friesen, G. C., Weller, P. A., & Dunham, L. M. (2009). Price trends and patterns in technical analysis: A theoretical and empirical examination. *Journal of Banking & Finance, 33*(6), 1089-1100.

Geva, T., & Zahavi, J. (2014). Empirical evaluation of an automated intraday stock recommendation system incorporating both market data and textual news. *Decision support systems, 57*, 212-223.

Groth, S. S., & Muntermann, J. (2011). An intraday market risk management approach based on textual analysis. *Decision support systems, 50*(4), 680-691.

Guo, K., Sun, Y., & Qian, X. (2017). Can investor sentiment be used to predict the stock price? Dynamic analysis based on China stock market. *Physica A: Statistical Mechanics and its Applications, 469*, 390-396.

Hagenau, M., Liebmann, M., & Neumann, D. (2013). Automated news reading: Stock price prediction based on financial news using context-capturing features. *Decision support systems, 55*(3), 685-697.

Hassani, H., Beneki, C., Unger, S., Mazinani, M. T., & Yeganegi, M. R. (2020). Text Mining in Big Data Analytics. *Big Data and Cognitive Computing, 4*(1), 1.

Hon, T.-Y., Moslehpour, M., & Woo, K.-Y. (2021). Review on behavioral finance with empirical evidence. *Advances in Decision Sciences, 25*(4), 1-30.

Ji, X., Wang, J., & Yan, Z. (2021). A stock price prediction method based on deep learning technology. *International Journal of Crowd Science*.

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

Kao, L.-J., Chiu, C.-C., Lu, C.-J., & Yang, J.-L. (2013). Integration of nonlinear independent component analysis and support vector regression for stock price forecasting. *Neurocomputing, 99*, 534-542.

Katayama, D., & Tsuda, K. (2018). A Method of Measurement of The Impact of Japanese News on Stock Market. *Procedia computer science, 126*, 1336-1343.

Kazem, A., Sharifi, E., Hussain, F. K., Saberi, M., & Hussain, O. K. (2013). Support vector regression with chaos-based firefly algorithm for stock market price forecasting. *Applied Soft Computing, 13*(2), 947-958.

Koyel Chakraborty, S. B., Rajib Bag, Aboul Alla Hassanien. (2019). *social network analytics*: academic press.

Kristian Henrickson, F. R., Francisco Camara Pereira. (2019). *Mobility Patterns, Big Data and Transport Analytics*: Tools and Applications for Modeling.

Kudryavtsev, A. (2020). Immediate And Longer-Term Stock Price Dynamics Following Large Stock Price Changes. *Annals of Financial Economics, 15*(01), 2050002.

Lee, C. S., Cheang, P. Y. S., & Moslehpour, M. (2022). Predictive Analytics in Business Analytics: Decision Tree. Advances in Decision Sciences, 26(1), 1-29.

Li, Q., Wang, T., Li, P., Liu, L., Gong, Q., & Chen, Y. (2014). The effect of news and public mood on stock movements. *Information Sciences, 278*, 826-840.

Li, X., Wu, P., & Wang, W. (2020). Incorporating stock prices and news sentiments for stock market prediction: A case of Hong Kong. *Information Processing & Management*, 102212.

Liu, B. (2020). *Sentiment analysis: Mining opinions, sentiments, and emotions*: Cambridge university press.

Liu, P., Xia, X., & Li, A. (2018). Tweeting the financial market: Media effect in the era of Big Data. *Pacific-Basin Finance Journal, 51*, 267-290.

Liu, Y., Qin, Z., Li, P., & Wan, T. (2017). *Stock volatility prediction using recurrent neural networks with sentiment analysis.* Paper presented at the International Conference on Industrial, Engineering and Other Applications of Applied Intelligent Systems.

Maqsood, H., Mehmood, I., Maqsood, M., Yasir, M., Afzal, S., Aadil, F., . . . Muhammad, K. (2020). A local and global event sentiment based efficient stock exchange forecasting using deep learning. *International Journal of Information Management, 50*, 432-451.

Mizuno, H., Kosaka, M., Yajima, H., & Komoda, N. (1998). Application of neural network to technical analysis of stock market prediction. *Studies in Informatic and control, 7*(3), 111-120.

Moat, H. S., Curme, C., Avakian, A., Kenett, D. Y., Stanley, H. E., & Preis, T. (2013). Quantifying Wikipedia usage patterns before stock market moves. *Scientific reports, 3*, 1801.

Mostafa, M. M. (2013). More than words: Social networks' text mining for consumer brand sentiments. *Expert Systems with Applications, 40*(10), 4241-4251.

Murphy, J. J. (1999). *Technical analysis of the financial markets: A comprehensive guide to trading methods and applications*: Penguin.

Nassirtoussi, A. K., Aghabozorgi, S., Wah, T. Y., & Ngo, D. C. L. (2014). Text mining for market prediction: A systematic review. *Expert Systems with Applications, 41*(16), 7653-7670.

Nazário, R. T. F., e Silva, J. L., Sobreiro, V. A., & Kimura, H. (2017). A literature review of technical analysis on stock markets. *The Quarterly Review of Economics and Finance, 66*, 115-126.

Nguyen, T. H., Shirai, K., & Velcin, J. (2015). Sentiment analysis on social media for stock movement prediction. *Expert Systems with Applications, 42*(24), 9603-9611.

Nisar, T. M., & Yeung, M. (2018). Twitter as a tool for forecasting stock market movements: A short-window event study. *The Journal of Finance and Data Science, 4*(2), 101-119.

Nofsinger, J. R. (2005). Social mood and financial economics. *The Journal of Behavioral Finance, 6*(3), 144-160.

Nofsinger, J. R. (2005). Social mood and financial economics. *The Journal of Behavioral Finance*, 6(3), 144-160.

Padmaja, S., & Fatima, S. S. (2013). Opinion mining and sentiment analysis-an assessment of peoples' belief: A survey. *International Journal of Ad hoc, Sensor & Ubiquitous Computing, 4*(1), 21.

Pagolu, V. S., Reddy, K. N., Panda, G., & Majhi, B. (2016). *Sentiment analysis of twitter data for predicting stock market movements.* Paper presented at the 2016 international conference on signal processing, communication, power and embedded system (SCOPES).

Peng, Y., & Jiang, H. (2015). Leverage financial news to predict stock price movements using word embeddings and deep neural networks. *arXiv preprint arXiv:1506.07220*.

Picasso, A., Merello, S., Ma, Y., Oneto, L., & Cambria, E. (2019). Technical analysis and sentiment embeddings for market trend prediction. *Expert Systems with Applications, 135*, 60-70.

Qian, B., & Rasheed, K. (2007). Stock market prediction with multiple classifiers. *Applied Intelligence, 26*(1), 25-33.

Regmi, U. R. (2012). Stock market development and economic growth: Empirical evidence from Nepal. *Administration and Management Review, 24*(1), 1-28.

Ruiz, E. J., Hristidis, V., Castillo, C., Gionis, A., & Jaimes, A. (2012). *Correlating financial time series with micro-blogging activity.* Paper presented at the Proceedings of the fifth ACM international conference on Web search and data mining.

Sang, C., & Di Pierro, M. (2019). Improving trading technical analysis with tensorflow long short-term memory (LSTM) neural network. *The Journal of Finance and Data Science, 5*(1), 1-11.

Schumaker, R. P., & Chen, H. (2009). A quantitative stock prediction system based on financial news. *Information Processing & Management, 45*(5), 571-583.

Sezer, O. B., Ozbayoglu, M., & Dogdu, E. (2017). A deep neural-network based stock trading system based on evolutionary optimized technical analysis parameters. *Procedia computer science, 114*, 473-480.

Sharma, D., Sabharwal, M., Goyal, V., & Vij, M. (2020). *Sentiment analysis techniques for social media data: A review.* Paper presented at the First International Conference on Sustainable Technologies for Computational Intelligence.

Shynkevich, Y., McGinnity, T. M., Coleman, S. A., & Belatreche, A. (2016). Forecasting movements of health-care stock prices based on different categories of news articles using multiple kernel learning. *Decision support systems, 85*, 74-83.

Smith, V. L. (2003). Constructivist and ecological rationality in economics. *American economic review, 93*(3), 465-508.

Taboada, M., Brooke, J., Tofiloski, M., Voll, K., & Stede, M. (2011). Lexicon-based methods for sentiment analysis. *Computational linguistics, 37*(2), 267-307.

Taylor, N. (2014). The rise and fall of technical trading rule success. *Journal of Banking & Finance, 40*, 286-302.

Ticknor, J. L. (2013). A Bayesian regularized artificial neural network for stock market forecasting. *Expert Systems with Applications, 40*(14), 5501-5506.

Vijh, M., Chandola, D., Tikkiwal, V. A., & Kumar, A. (2020). Stock closing price prediction using machine learning techniques. *Procedia computer science, 167*, 599-606.

Wang, J.-L., & Chan, S.-H. (2007). Stock market trading rule discovery using pattern recognition and technical analysis. *Expert Systems with Applications, 33*(2), 304-315.

Wei, L.-Y., Chen, T.-L., & Ho, T.-H. (2011). A hybrid model based on adaptive-network-based fuzzy inference system to forecast Taiwan stock market. *Expert Systems with Applications, 38*(11), 13625-13631.

Weng, B., Ahmed, M. A., & Megahed, F. M. (2017). Stock market one-day ahead movement prediction using disparate data sources. *Expert Systems with Applications, 79*, 153-163.

Weng, B., Lu, L., Wang, X., Megahed, F. M., & Martinez, W. (2018). Predicting short-term stock prices using ensemble methods and online data sources. *Expert Systems with Applications, 112*, 258-273.

Wong, W.-K. (2020). Review on behavioral economics and behavioral finance. *Studies in Economics and Finance*.

Yu, L. C., Wu, J. L., Chang, P. C., & Chu, H. S. (2013). Using a contextual entropy model to expand emotion words and their intensity for the sentiment classification of stock market news. *Knowledge-Based Systems*, 41, 89-97.

Zapranis, A., & Tsinaslanidis, P. E. (2012). A novel, rule-based technical pattern identification mechanism: Identifying and evaluating saucers and resistant levels in the US stock market. *Expert Systems with Applications, 39*(7), 6301-6308.

Zhang, L. (2013). *Sentiment analysis on Twitter with stock price and significant keyword correlation.*

Zhang, X., Shi, J., Wang, D., & Fang, B. (2018). Exploiting investors social network for stock prediction in China's market. *Journal of computational science, 28*, 294-303.

Zhang, X., Shi, J., Wang, D., & Fang, B. (2018). Exploiting investors social network for stock prediction in China's market. *Journal of computational science*, 28, 294-303.

Zhu, H., Jiang, Z.-Q., Li, S.-P., & Zhou, W.-X. (2015). Profitability of simple technical trading rules of Chinese stock exchange indexes. *Physica A: Statistical Mechanics and its Applications, 439*, 75-84.