Expert Systems With Applications 217 (2023) 119509

Contents lists available at ScienceDirect

Expert Systems With Applications

journal homepage: www.elsevier.com/locate/eswa

Review

News-based intelligent prediction of financial markets using text mining and machine learning: A systematic literature review

Matin N. Ashtiani *, Bijan Raahemi

Knowledge Discovery and Data Mining Lab, Telfer School of Management, University of Ottawa, 55 Laurier Avenue East, Ottawa ON K1N6N5, Canada

ARTICLE INFO

Keywords:
Stock market prediction
Systematic literature review
Natural language processing
Text mining
Machine learning

ABSTRACT

Researchers and practitioners have attempted to predict the financial market by analyzing textual (e.g., news articles and social media) and numeric data (e.g., hourly stock prices, and moving averages). Among textual data, while many papers have been published that analyze social media, news content has gained limited attention in predicting the stock market. Acknowledging that news is critical in predicting the stock market, the focus of this systematic review is on papers investigating machine learning and text mining techniques to predict the stock market using news. Using Kitchenham's methodology, we present a systematic review of the literature on intelligent financial market prediction, examining data mining and machine learning approaches and the employed datasets. From five digital libraries, we identified 61 studies from 2015–2022 for synthesis and interpretation. We present notable gaps and barriers to predicting financial markets, then recommend future research scopes. Various input data, including numerical (stock prices and technical indicators) and textual data (news text and sentiment), have been employed for news-based stock market prediction. News data collection can be costly and time-consuming: most studies have used custom crawlers to gather news articles; however, there are financial news databases available that could significantly facilitate news collection. Furthermore, although most datasets have covered fewer than 100K records, deep learning and more sophisticated artificial neural networks can process enormous datasets faster, improving future model performance. There is a growing trend toward using artificial neural networks, particularly recurrent neural networks and deep learning models, from 2018 to 2021. Furthermore, regression and gradient-boosting models have been developed for stock market prediction during the last four years. Although word embedding approaches for feature representation have been employed recently with good accuracy, emerging language models may be a focus for future research. Advanced natural language processing methods like transformers have undeniably contributed to intelligent stock market prediction. However, stock market prediction has not yet taken full advantage of them.

1. Introduction

A stock is an investment that represents fractional ownership in a company. The stock markets are public markets for issuing, selling, and buying stocks. They serve two essential purposes: providing funds to companies that they can use to support and grow their businesses and allowing investors to partake in companies' profits. The economy of a country heavily relies on the stock market sector. An investor may profit from intelligent investments in financial markets or lose all their assets through inappropriate trading. Stock market forecasting is the process of attempting to determine future stock price movements. The successful prediction of a stock's future price can result in significant profit.

Two main theoretical hypotheses define market behavior: the efficient market hypothesis (EMH) (Malkiel & Fama, 1970) and the adaptive market hypothesis (AMH) (Lo, 2005). The notion that markets are random and not predictable is firmly established in the random walk theory Bollen, Mao, and Zeng (2011). The EMH asserts that "financial markets are informationally efficient". It claims that all known information about stocks is already factored into their prices. For a long time, the EMH has been the prevailing theory. Later, behavioral finance emerged to challenge this theory, remarking that investors are not always rational, and a stock does not always trade at its asset value. Thus, behavioral economics seeks to explain stock market anomalies through psychology-based principles (Picasso, Merello, Ma, Oneto, & Cambria, 2019). The AMH is based on three basic tenets: people are motivated by self-interest, they naturally make mistakes, and they adapt and learn from these mistakes (Lo, 2005).

* Correspondence to: Knowledge Discovery and Data Mining Lab, Telfer School of Management, University of Ottawa, Ottawa, Canada.
E-mail addresses: mnaja036@uottawa.ca (M.N. Ashtiani), braahemi@uottawa.ca (B. Raahemi).

https://doi.org/10.1016/j.eswa.2023.119509
Received 28 July 2022; Received in revised form 25 December 2022; Accepted 3 January 2023
Available online 9 January 2023
0957-4174/

Over the recent decades, the predictability of the financial markets has been acknowledged by researchers around the world. However, they have argued that the present prediction techniques are unsatisfactory, making stock market prediction still one of the most sophisticated subjects in this area. Consequently, it has continued to be a very attractive research direction.

Stock market prices can be modeled using two main approaches: technical and fundamental (Gerencser, Torma, & Orlovits, 2009). A combination of both methods has also been applied in some studies.

The technical approach is characterized by statistical analysis of stock prices, whose future values are predicted by investigating their movement trend (Ahmadi et al., 2018; Anbalagan & Maheswari, 2015). Technical analysts attempt to forecast the financial market by focusing on charts that represent technical indicators and historical market prices (Wei, Chen, & Ho, 2011). They preprocess historical stock prices, calculate appropriate indicators, and eventually feed that data into a predictive model. The major technical indicators employed in the literature for technical analysis include the exponential moving average (EMA), moving average convergence/divergence rules (MACD), the simple moving average (SMA), on-balance volume (OBV), and the relative strength index (RSI) (Anbalagan & Maheswari, 2015; Bisoi & Dash, 2014; Dash, Dash, & Bisoi, 2014).

Fundamental analysis is a procedure for assessing a stock's value by interpreting external and internal determining components of its price (Selvin, Vinayakumar, Gopalakrishnan, Menon, & Soman, 2017). The fundamental aspects are economic data, financial performance, political and social behaviors, the business environment, and the firm's financial ratios (Beyaz, Tekiner, Zeng, & Keane, 2018). We point out market capitalization (MC), earnings per share (EPS), the price/sales ratio (P/S), and the debt/equity ratio (D/E) as some of the notable financial ratios. Another important aspect of fundamental analysis is data extraction techniques, the operating field of natural language processing (NLP), and text mining. Since textual data is multisource and unstructured, NLP approaches are extensively employed to extract as much information as possible, which is the most challenging research aspect. Examples of unstructured textual data are social media, news, blogs, forums, and financial reports. This data has been determined to be an appropriate predictor for financial market forecasting (Li, Wang, et al., 2014; Li, Xie, Chen, Wang, & Deng, 2014; Liu, Wu, Li, & Li, 2015; Zhang, Fuehres, & Gloor, 2011).

News texts have proved to be a rich data source for stock market prediction. Various studies have proved the effectiveness of this information in forecasting stock market prices (Shah, Isah, & Zulkernine, 2018; Zhang, Cui, Xu, Li, & Li, 2018). To be able to leverage textual data to predict the market using machine learning models, text transformation and representation techniques are employed. The purpose of text transformation and representation methods is to convert text into numerical vectors and prepare it for feeding into a machine learning model. Several text representation models have been introduced and employed in different areas of research. Most traditional methods attempt to use the statistical characteristics of the text to convert it to a vector. For instance, counting the frequency of occurrences of a word in a corpus is one of the straightforward techniques, which is called the term frequency (TF) approach (Hagenau, Liebmann, & Neumann, 2013; Wong, 2002). However, the statistical methods suffer from discarding the word's context in a sentence or even a paragraph. Consequently, with the advances in the NLP and neural network models, learning-based methods are proposed. The learning-based models or word embedding approaches utilize a neural network model to train a language model and assign a unique vector to each word in a lexicon. Word2vec, GloVe, and fastText are examples of non-contextual word embedding models (Goldberg & Levy, 2014; Pennington, Socher, & Manning, 2014). Recently, more sophisticated language models have also been introduced and contributed to the field. They employ advanced deep learning models like a transformer, which enables them to understand the context of a word more accurately and effectively. For instance, bidirectional encoder representations from transformers (BERT) is a language model that consider the context of a word in both backward and forward directions while learning to improve the performance of the word embedding models (Devlin, Chang, Lee, & Toutanova, 2018).

Sentiment analysis, or opinion mining, refers to a set of NLP techniques for analyzing public sentiments, appraisals, attitudes, emotions, evaluations, and opinions about various subjects (Medhat, Hassan, & Korashy, 2014). Sentiment analysis has contributed significantly to many applications, especially in finance for stock market prediction. Generally, it attempts to categorize the emotional attitude, or valence, as positive or negative by identifying and processing positive and negative words in their context. For example, Maks and Vossen (2012) proposed a lexicon-based sentiment analysis model for opinion mining and deep sentiment analysis.

Sentiments are significant elements of stock markets. Analyzing sentiments from different data sources can explain how stock markets respond to various news categories in the immediate, medium, and long term. Consequently, sentiment analysis has emerged as a new method to examine the impact of news sentiment about the markets. Besides the manual sentiment classification by domain-specific experts in each area, there are two categories of sentiment analysis approaches: lexical-based and learning-based. The former assesses the frequencies of negative and positive words and phrases in a textual corpus to measure the sentiment score of a document (Hu & Liu, 2004). On the other hand, the latter is language-independent and develops machine learning models to analyze the sentiment of the text. The naïve Bayes (NB) model, support vector machine (SVM), max entropy classifier, and artificial neural networks (ANNs) are examples of models employed for learning-based sentiment analysis (Boiy & Moens, 2009).

Researchers have developed several sentiment analysis tools and systems to facilitate sentiment extraction from a textual corpus. OpinionFinder is a tool for the automatic processing of textual data and identifying different components of subjectivity, like sentiment expressions and the source of opinions. Bollen et al. (2011) utilized this system to binary classify the sentiment of daily Twitter feeds as positive or negative. Additionally, they employed the Google Profile of Mood States (GPOMS), a tool for measuring the content of tweets in terms of human moods and categorizing it into six states: Alert, Vital, Sure, Calm, Happy, and Kind. Li, Xie, et al. (2014) combined the Loughran–McDonald financial sentiment dictionary and the Harvard IV-4 sentiment dictionary with building sentiment dictionaries. They then used these dictionaries to extract sentiment from a Hong Kong financial news website called FINET. A novel so-called joint sentiment/topic method (JST) was also presented by Nguyen, Shirai, and Velcin (2015) to extract the corresponding mood of the stocks from the message board of Yahoo Finance.

At the same time, knowledge-based techniques and machine learning methods are used as the other building blocks of sentiment analysis. For sentiment analysis with machine learning models, a classification model like SVM, naïve Bayes, or an ANN is trained on a labeled dataset (Pang & Lee, 2008). This dataset can be constructed with manual annotation (Boldrini, Balahur, Martínez-Barco, & Montoyo, 2012; Wiebe, Wilson, & Cardie, 2005) or even sourced from available resources, such as the star ratings of user reviews (Dave, Lawrence, & Pennock, 2003). A wide range of research works has been carried out for prediction purposes using variants of ANNs and deep learning techniques, especially long short-term memory (LSTM), convolutional neural networks (CNNs), and restricted Boltzmann machines (RBMs) (Ding, Zhang, Liu, & Duan, 2015; Liu, Hoi, Zhao, & Sun, 2016). There are hierarchical application levels for either lexicon-based or learning-based sentiment analysis, namely, the document, paragraph, phrase, sentence, and word levels (Hatzivassiloglou & McKeown, 1997; O'Hare et al., 2009; Pang & Lee, 2008; Wilson, Wiebe, & Hoffmann, 2005; Xu & Cohen, 2018).

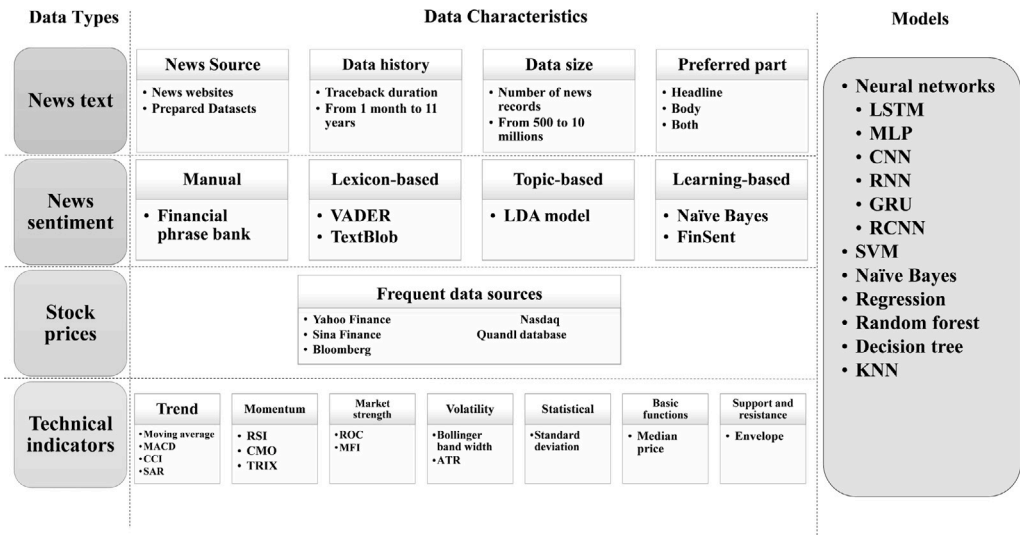

Fig. 1. This figure summarizes the data- and model-related information extracted from the reviewed literature. The left column presents the employed data types; the middle column describes the extracted characteristics corresponding to each data type; the last column includes the frequent machine learning models in the reviewed literature. All the summarized information is described in detail in the manuscript.

Fig. 1 summarizes the data- and model-related information extracted from the reviewed literature. This systematic literature review (SLR) aims to investigate, summarize, and synthesize the emerging text mining techniques used to forecast the stock market, the source and type of data used for prediction, and the applied machine learning models. The findings will be valuable for determining how well the most recent text mining methods deal with the available sources of unstructured news data. The main goal is to provide researchers with the current state of applying emerging text mining techniques, machine learning models, and data sources in stock market prediction. By identifying current gaps in the literature, we can detect the most challenging parts of the stock market prediction problem, which will help researchers focus on them in their future work.

In light of more sophisticated machine learning and data mining techniques arising in the last few decades, the stock market prediction topic has attracted more researchers. Consequently, there are several literature review articles in this area. To the best of our knowledge, there are several very recent review articles on stock market prediction focusing on deep learning approaches (Alzazah & Cheng, 2020; Hu, Zhao, & Khushi, 2021; Jiang, 2021; Nabipour, Nayyeri, Jabani, Mosavi, & Salwana, 2020; Thakkar & Chaudhari, 2021). Furthermore, Khadjeh Nassirtoussi, Aghabozorgi, Wah, and Ngo (2014) delivered an SLR to review the main frameworks for stock market prediction based on online text mining and identified a couple of significant gaps, and Xing, Cambria, and Welsch (2018) carried out a narrative review to illustrate the order and structure of the techniques and applications in the scope of natural-language-based financial forecasting. However, the current study can be distinguished from the other recently published review articles in the scope of stock market prediction in various aspects. First, this systematic review goes beyond deep neural networks and investigates all the most recently (from 2015 to September 2022) employed machine learning approaches for news-based stock market prediction. Second, we focused on news-based fundamental stock market prediction models instead of considering works examining other types of textual data, like data from social media (Twitter tweets) or the works which limited their study to technical approaches that merely employed historical stock prices and technical indicators. The reason behind discarding the social media textual data is that (i) the data from social networks like Twitter tweets or discussion board posts are pretty vulnerable to noise and bias and (ii) while many papers have been published that analyze social media, news content has gained limited attention in predicting the stock market. Third, we investigated and categorized the employed text representation models in the reviewed literature. Finally, this SLR keeps an eye on the utilized data sources and model performance for news-based stock market prediction, which distinguishes it from the other reviews that just highlight the specific machine learning models and ignore the data sources.

Furthermore, compared to a narrative literature review, this systematic review involved a comprehensive set of methodological research steps. It is based on the Kitchenham method, which is a rigorous and auditable study protocol for extracting, incorporating, and reporting results (Kitchenham, 2004). We adopted the guidelines presented by Kitchenham (2004), which delineate the step-wise schemes for taking an SLR forward.

To frame the study, we identified four research questions (Section 2.1). In the planning phase, we probed five widely known databases subject to the specified research questions and search strings. We retrieved the results of 293 journal articles and conference papers. We made use of DistillerSR (https://www.evidencepartners.com) to conduct the review. The primary retrieved articles were chosen based on the study selection processes during the planning phase. The final procedure identified 61 articles as the nominated papers. We derived the desired information to address the research questions and investigate and summarize the outcomes.

With the rapid growth of big data, the development of machine learning and text mining approaches, and the popularity of stock market prediction among researchers, new models, have recently been introduced to predict the stock market more accurately. The current study is not only an up-to-date SLR in the area of intelligent stock market prediction but also differentiates itself from the previous works in a few directions. The characteristics of the datasets, including data source, data size, data type, and the preferred part of the textual data, gained less attention in the other studies. Additionally, we mainly focused on the emerging text processing and machine learning techniques applied to news data (excluding social media data) to find novel algorithms and models and identify a benchmark dataset for the stock market prediction task (if it is available). Finally, this study aims to reveal the new findings, analyze the gaps, and identify directions for future work in intelligent stock market prediction.

The remainder of this paper is organized as follows: Section 2 describes the research method. Section 3 provides the synthesized results and answers the research questions. Finally, the study's outcomes are outlined in the Conclusion.

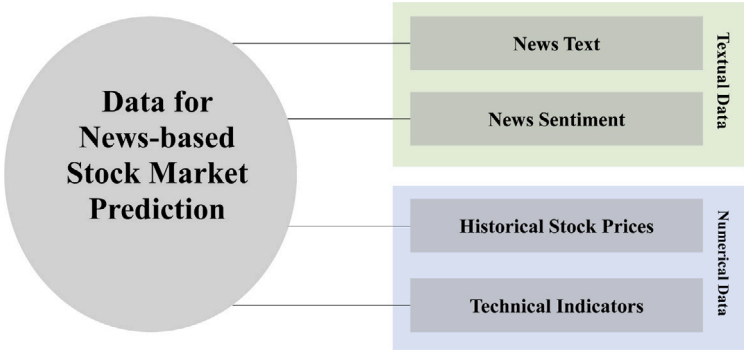

Fig. 2. Four employed data types for news-based stock market prediction.

2. Research method

A systematic literature review (SLR) was conducted to investigate the existing literature on financial market prediction using news data and address the research questions. This section represents the steps undertaken to develop our SLR protocol as per Kitchenham's methodology (Kitchenham, 2004).

2.1. Research questions

This study attempts to address the following research questions:

- (RQ1) What relevant data sources are used for intelligent prediction of the stock market?
- (RQ2) What are the most recent relevant text representation and machine learning techniques for intelligent stock market prediction?
- (RQ3) What kinds of evaluation criteria (performance measures) are used, and how well do the models predict the stock market?
- (RQ4) What are the trends, gaps, and future research directions?

The first question enables us to uncover the most popular sources of textual news data used for intelligent stock market forecasting. The second allows the identification of emerging text mining and machine learning techniques for financial market prediction. The third gives valuable knowledge about the models' performance and how it is calculated. Last but not least, the fourth question aims to identify the gaps and limitations of the current state-of-the-art works and lay down directions for future research in this area.

2.2. Developing the search strategy

Appropriate search concepts and keywords have an essential role in finding the most relevant studies. The following search string was developed to be queried in five different digital libraries:

"artificial intelligence" OR "machine learning" OR "data mining"
AND
"sentiment" OR "semantic"
AND
"stock market" OR "financial market"
AND
"news"

We explored the developed search expression in ACM Digital Library, Scopus, IEEE Xplore, ScienceDirect, and Web of Science. We took into account only peer-reviewed conference and journal papers. An 8-year timespan from January 2015 to September 2022 was considered in order to discover the emerging techniques and popular datasets. As Table 1 illustrates, a total of 293 articles were collected from the digital libraries. After removing 64 duplicate articles, 229 articles remained for screening. We then applied backward snowballing to discover additional relevant articles as a supplement to the automatically retrieved ones. A total of 6 papers were added with this step.

Table 1
Number of collected papers from the queried libraries.

No	Digital library	Number of papers
1	Scopus	173
2	Web of Science	64
3	IEEE Xplore	36
4	ACM Digital Library	8
5	ScienceDirect	12
Total number of collected papers		293

2.3. Selection of primary studies

We included and excluded studies as per the criteria presented in Table 2.

- **Duplicate removal:** The primary search involved 293 papers. After removing the 64 duplicate papers, 229 articles remained for subsequent screening stages.
- **Title and abstract screening:** We screened articles by refining their keywords, abstracts, and titles corresponding to the exclusion and inclusion criteria. This step excluded 95 articles.
- **Full-text screening:** Out of 132 articles remaining after the previous step, 79 studies were excluded based on applying the exclusion criteria to the full text. Hence, 55 articles remained after the full-text screening.
- **Snowballing inclusion:** Six articles were included from the backward snowballing technique.

2.4. Data extraction

This stage entailed compiling pertinent information from the retrieved articles based on the research questions. We extracted information about the input data categories and their specifications, text representation methods, and machine learning techniques (model types, evaluation criteria, model performance, and the best-performing model in each study).

3. Results

This section synthesizes and investigates the 61 discovered articles to address each research question.

3.1. RQ1: What relevant data sources are used for intelligent prediction of the stock market?

The data source is essential to forecasting and has a substantial impact on its performance. According to the reviewed literature, we identified four different sources of data employed for news-based stock market prediction: news text, news sentiment, historical stock prices, and technical indicators (see Fig. 2). Each article decided to either make

Table 2
Applied inclusion and exclusion criteria to decide the inclusion of each study.

Inclusion	I.1: Identified articles based on snowballing upon satisfying exclusion criteria
Exclusion	E1: Articles focused on using textual social media data (Tweets, blogs, etc.) to predict the stock market E2: Fully theoretical articles that do not implement any machine learning or data mining techniques in their studies E3: Articles that do not report their models' performance E4: Manuscripts in the form of short papers, posters, book chapters, and abstracts E5: Articles that do not report the source of data they employ E6: Articles in any language other than English E7: Survey and review articles E8: For articles observed to have been published two times with the same authors, the latest work was selected.

Table 3
Diversity of the data sources used by the researchers for news-based stock market prediction.

Historical stock prices	Technical indicators	Sentiment analysis	Text mining	#
No	No	No	Yes	4
		Yes	No	2
			Yes	8
	Yes	No	Yes	4
		Yes	No	2
Yes	No	No	Yes	8
		Yes	No	7
			Yes	13
	Yes	Yes	No	4
			Yes	10

use of only one of these data sources or employ a combination of them to tackle the news-based stock market prediction. As we discussed, in fundamental analysis, stock market prediction systems take at least two data sources as input: structured data from historical prices and technical indicators and unstructured textual data. However, the focus of the technical analysis is merely on the structured numerical data from historical stock prices and technical indicators.

Table 3 demonstrates various scenarios in which data sources are combined as well as the frequency of their use. As illustrated, a combination of historical stock prices along with news vectors resulting from text mining methods, as well as news sentiment, attracted the most attention (13 articles). Integrating all four input data types is the next most prevalent scenario (10 studies). In the following subsections, each data source and its characteristics are discussed in detail.

3.1.1. News text

Most traditional stock prediction methods characterize the stock market through past trading prices and try to discover signs that represent the price fluctuations based on this historical time series (Picasso et al., 2019). However, stock price prediction merely proceeding from technical data has proved inadequate. News articles are one of the determinant factors that affect the stock market. In particular, researchers expect better performance for predictors that take the influence of the news into account. Researchers have established that financial news can enhance stock price prediction performance. There is a tight correlation between a firm's stock price movements and the news articles associated with it (Shah et al., 2018; Zhang, Cui, et al., 2018). Therefore, news-based financial market prediction has attracted special interest in the past few years. Nevertheless, due to the ambiguity and complexity of natural languages, predicting the financial market accurately based on news articles is not straightforward. The news sources used in the studies are major news websites like Financial Times, The Wall Street Journal, Bloomberg, Reuters, Forbes, and Yahoo Finance (https://finance.yahoo.com/) (Antweiler & Frank, 2004; Chatrath, Miao, Ramchander, & Villupuram, 2014; Fung, Yu, & Lam, 2003; Rachlin, Last, Alberg, & Kandel, 2007; Schumaker & Chen, 2009; Schumaker, Zhang, Huang, & Chen, 2012; Wuthrich et al., 1998). The employed news data source in the literature could be either general news or domain-specific financial news. Most surveyed systems employed financial news under the assumption it is less noisy than general news. In particular, among the reviewed references, only (Lima, Portela, Santos, Abelha, & Machado, 2015) employed both economic and general news in their study. Although social media data (such as Twitter, message boards, and blogs) are generally used in research experiments to boost the performance of stock market prediction methods, as described in Section 1, they are beyond the scope of this SLR. Our review focuses exclusively on news data, any other kinds of textual information are excluded.

Data Source: Web scraping and financial news datasets. We discovered two data gathering approaches in the studies: (a) web scraping and (b) using available financial news datasets. Web scraping software enables the extraction of news data from the economic press. On the other hand, news datasets provide immediate access to the desired financial data. The advantage of using datasets is the that data can be used as a benchmark, and the results can be replicated and compared across various learning methodologies and prediction approaches. With crawled databases, access is limited: the data is available for a limited duration on the websites. However, the news datasets categorize and store the data for different time spans. Most of these datasets provide users with their application programming interfaces (APIs) to make the data retrieval procedures faster and easier. Furthermore, the available news datasets have more extensive data sources, making them more advantageous than limiting the study to only several press websites.

Most of the articles collected news data from well-known news websites. They chose these data sources based on the sites' popularity combined with the releases' frequency and coverage of economics and general subjects (Lima et al., 2015). Reuters is the most prevalent target news channel. Yahoo Finance, Bloomberg, Google Finance, and The Wall Street Journal are the next most common sources for gathering financial news data. It should be mentioned that 26 news websites that appeared only once in the studies are avoided to bring attention to the most prevalent news sources in the tree map in Fig. 3.

Generally, studies used a mixture of various news from several websites. For instance, Meyer, Bikdash, and Dai (2017) used the news

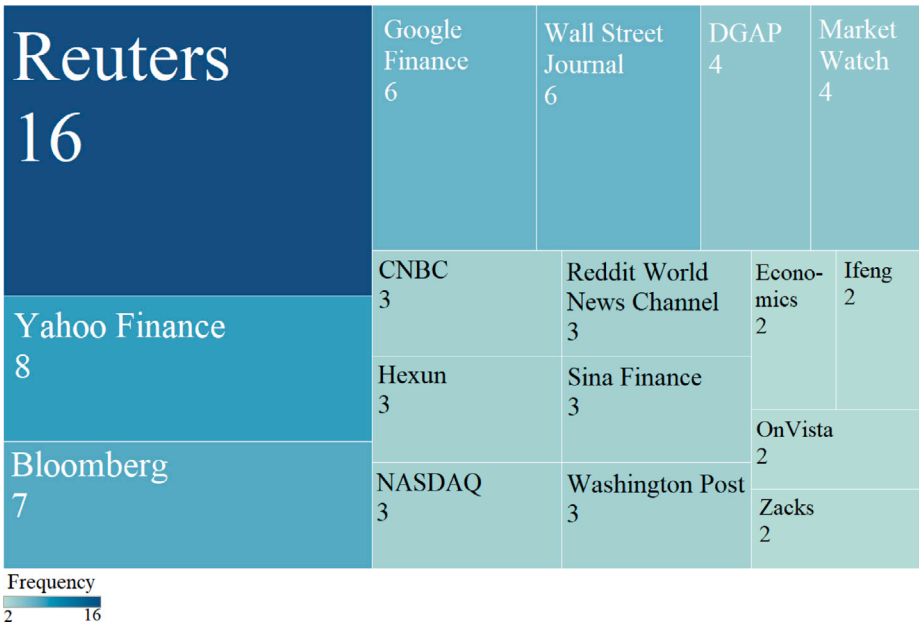

Fig. 3. The sources of news data for news-based stock market prediction. The size and the color of the rectangles are meaningful and represent the frequency of the articles using each of the sources of news.

feeds, including the principal financial news channels and websites such as MarketWatch, The New York Times, The Wall Street Journal, Reuters, Forbes, CNBC, and Yahoo Finance. Besides, additional feeds for every company in the S&P 500 were generated to collect firm-relevant articles. Ifeng news, which is a well-known professional financial channel in China, was the source of the financial news on stocks in another two studies (Chen, Luo, Xu, & Wang, 2016; Long, Song, & Tian, 2019). Minh, Sadeghi-Niaraki, Huy, Min, and Moon (2018) gathered daily financial news from Reuters and Bloomberg. Ma and Liang (2015) collected the data from www.eastmoney.com and www.cnstock.com via their specially devised web crawler program Similarly, Du and Zhang (2018) designed a crawler to retrieve periodic financial news. The gathered data came from real-time news, Federal Reserve speeches, and institution announcements.

In addition to news websites, we discovered several available news datasets aggregating financial news from even thousands of sources. Table 4 presents a list of these datasets, a brief description, and relevant references. These datasets benefit the researchers by facilitating access to thousands or millions of sources of English or even multilingual global financial news through their APIs. APIs make the data gathering procedure more straightforward and less time-consuming. In particular, they equip the researchers with readily available scripts for data gathering from news websites and save time in terms of developing scripts and functions.

Data history. There is no standard for determining the optimal traceback time frame of the data before starting to theorize and implement the machine learning models. We recognized contradicting literature regarding the length of time covered by the experiments' data. In particular, there is a wide variation from a month up to multiple years. Xing et al. (2018) suggested investigating a more extended period with less frequent data, such as ad hoc announcements. Furthermore, as an argument for studying a long period of 7 years, Feuerriegel and Prendinger (2016) indicated that the rationale behind this was to avoid the possibility of examining news predominantly covering a particular market event like the financial crisis. Table 5 presents how many months each study traced back for considering news and market data. As illustrated, most works considered a data history between 2 and 6 years. Among the studies that reported both the duration and the dataset size, the shortest time frame spans news within 1 month (Ingle & Deshmukh, 2021), followed by 8 months (Azizi, Abdolvand, Ghalibaf Asl, & Rajaee Harandi, 2021), 2 years (Zhang, Qu, Huang, Fang, & Yu, 2018), 3 years (Huang, 2020), and finally about 20 years (Feuerriegel & Gordon, 2018). However, in general, most of the articles included news covering multiple years (Table 5).

Dataset size. Dataset size refers to the number of textual data records in each study. Although several studies did not report the number of data records, those included a fairly broad spectrum varying from less than 1000 records (Alzazah et al., 2022; Devaraj et al., 2020; Kazemian et al., 2016) to even 10 million records (Ding et al., 2015). We discovered that most articles that reported dataset size used fewer than 100,000 news records (Table 5). Furthermore, when the textual input data was shorter, for example, news headlines, researchers adopted a larger number of data records than the studies in which the entire article was employed as the input data (Ding et al., 2015; Zhang, Qu, et al., 2018). However, this was not observed in all studies.

Preferred part of the textual data. After collecting the data from the specified sources, it is time to decide which part of the data is preferred to be carried forward for further processing. Although 25 articles did not mention the preferred part of the news employed, most studies considered news headlines in their works. In particular, 28 studies exclusively limited their textual input to news headlines, six articles took advantage of both headlines and the entire news body, and three research articles focused exclusively on the news body (Chiong, Fan, Hu, & Dhakal, 2022; Eck, Germani, Sharma, Seitz, & Ramdasi, 2021; Mohan et al., 2019).

3.1.2. News sentiment

The sentiment is the emotional stance that can be extracted from textual data through a computational procedure called sentiment analysis or opinion mining. Sentiment analysis is mainly based on identifying positive and negative words and processing text to classify its emotional stance into predefined categories (neutral, positive, and negative) and bring together insightful information regarding the context. We discovered various approaches for sentiment analysis for news-based stock market prediction. Forty-six articles used sentiment features to predict the market accordingly. Some studies designed a pure sentiment analysis approach to employ sentiment analysis from the news texts for stock market prediction only based on news sentiment (Chiong et al., 2018; Feuerriegel & Prendinger, 2016; Lima et al., 2015; Raman et al., 2022; Rao, Ramaraju, Smith, & Bansal, 2022). In contrast, the others

Table 4
Discovered financial news datasets and their description. The last column shows the references that employed these datasets as the news data sources in their studies.

News dataset	Dataset description	Ref.
Financial Phrase Bank	Consists of about 5000 annotated financial news English sentences from the LexisNexis database.	Fazlija and Harder (2022) Omarkhan, Kissymova, and Akhmetov (2021)
LexisNexis	Provides a news archive by major financial newspapers back up to 40 years.	Shynkevich, McGinnity, Coleman, and Belatreche (2016) Mishev, Gjorgjevikj, Vodenska, Chitkushev, and Trajanov (2020)
Event Registry	A news intelligence platform providing access to global news and blog content from over 150,000 sources.	Sarkar, Sahoo, Sah, and Pradhan (2020)
Intrinio	Includes real-time financial news data.	Picasso et al. (2019)
ProQuest	Provides a collection of digitized news content, spanning the latest headlines and developments in various fields, including finance.	Case and Clements (2021)
Pulse	Aggregated more than 200,000 Indian finance news headlines from various news websites like Business Standard, The Hindu Business, Reuter, and many other news websites.	Gite et al. (2021)
Quandl	A platform providing users with economic news data, which was acquired by Nasdaq in 2018.	Weng, Lu, Wang, Megahed, and Martinez (2018)

Table 5
Dataset time frame duration and the number of records for each study. Out of 61 reviewed studies, only the articles that reported these dataset characteristics are included in this table.

Ref.	Data span	Duration (Months)	# Records	Preferred part
Kazemian, Zhao, and Penn (2016)	March-1997 to March-2013	192	991	–
Vargas, de Lima, and Evsukoff (2017)	Oct-2006 to Nov-2013	96	106494	Headline
Liu, Zeng, Yang, and Carrio (2018)	Oct-2011 to July-2017	84	6423	Headline
Minh et al. (2018)	Oct-2006 to Nov-2013	96	553666	–
Zhang, Qu, et al. (2018)	Jan-2015 to Dec-2016	24	78195	Headline
Li, Jin, Xi, Liu, and Luo (2018)	May-2010 to Jan-2018	108	3000000	Headline and body
Wang, Ho, and Lin (2018)	Jan-2007 to Dec-2016	120	400000	–
Mohan, Mullapudi, Sammeta, Vijayvergia, and Anastasiu (2019)	Feb-2013 to March-2017	60	265463	Body
Khadjeh Nassirtoussi, Aghabozorgi, Wah, and Ngo (2015)	2008 to 2011	48	6906	Headline
Linardos, Kermanidis, and Maragoudakis (2015)	Nov-2007 to Feb-2010	48	3966	–
Feuerriegel and Gordon (2018)	July-1996 to April-2016	252	75927	–
Shynkevich et al. (2016)	Sep-2009 to Sep-2014	72	51435	–
Vicari and Gaspari (2021)	2008 to 2016	108	81000	Headline
Ding et al. (2015)	Oct-2006 to Nov-2013	96	10000000	Headline
Yoshihara, Seki, and Uehara (2016)	1999 to 2008	120	933549	–
Sarkar et al. (2020)	Jan-2014 to Dec-2018	60	18000	Headline
Devaraj et al. (2020)	Jan-2013 to Dec-2017	60	540	Headline
Ingle and Deshmukh (2021)	July-2016 to Aug-2016	1	1800	–
Fazlija and Harder (2022)	Jan-2007 to Nov-2016	120	9000000	Headline and body
Raman, Aljafari, Venkatesh, and Richardson (2022)	2013 to 2016	48	2500	–
Ulloa, Espezua, Villavicencio, Miranda, and Villanueva (2022)	2014 to 2021	96	13547	Headline
Alzazah, Cheng, and Gao (2022)	Jan-2015 to Dec-2020	72	483	–
Hinton, Osindero, and Teh (2006)	Jan-2017 to Dec-2020	48	2910	–
Azizi et al. (2021)	Sep-2017 to April-2018	8	67000	Headline
Bi, Liu, Wang, and Li (2021)	Aug-2008 to June-2016	98	50000	Headline
Case and Clements (2021)	Jan-2015 to Jan-2021	60	235040	–
Huang (2020)	2016 to 2018	36	15968	–
Sridhar and Sanagavarapu (2021)	Aug-2008 to July-2016	98	50301	Headline
Hu, Wang, Ho, and Tan (2021)	Oct-2006 to Nov-2013	86	109000	Headline and body

aggregated the sentiment features with the stock prices, technical indicators, and other textual semantic features to enrich the input for stock market prediction.

We categorized each sentiment analysis approach into one of four categories: manual, lexicon-based, topic-based, and learning-based. Table 6 presents the sentiment analysis approaches in the reviewed literature and briefly describes them. Manual sentiment categorization is a traditional and time-consuming way to decide the sentiment of a piece of text. The sentiment of the text should be categorized based on human opinion. Lexicon-based models were developed to benefit from a pre-defined annotated dictionary to facilitate deciding the sentimental orientation of each word in a sentence into positive and negative. In this approach, any text representation like bag-of-words (BOW) can be used to represent a sentence. Then, the sentiment score can be calculated by assigning an individual score to each word in that sentence. Finally, a pooling operation will be applied over all word sentiments to calculate the final score. Lexicon-based methods decide the sentiment label of a word based solely on a dictionary and discard the topic and the context of the word. Consequently, topic-based approaches were proposed to address this issue (Zhao, Jiang, Yan, & Li, 2010). Due to the limitation of lexicon-based models, including the limited procedure of developing and validating a comprehensive lexicon, learning-based sentiment classification models were devised. The aim of learning-based approaches is to train on a pair of text and sentiment labels and

Table 6
Sentiment analysis approach categories and the approaches employed in the reviewed literature. A brief description of each approach is provided.

Category	Employed approach	Description
Lexicon-based	Dictionary based polarity	Several dictionaries, including the LoughranMcDonald dictionary, Harvard IV-4, and the How-net emotional word set.
	VADER	A specific lexicon-based and rule-based sentiment analysis tool relying on a dictionary that maps lexical features to sentiment scores.
	Textblob	A library built upon NLTK which provides polarity and subjectivity sentiment outputs for each input sentence.
	SentiWordNet	A tool for measuring the sentiment of sentences built upon WordNet lexical database of English words.
	Polyglot	A library that performs sentiment analysis using polarity lexicons for 136 languages.
Learning-based	Naïve Bayes	Uses conditional probabilities of each lexical feature occurring in either positive or negative text in the training data to calculate the sentiment.
	Finsent News Sentiment	Employs FinBERT model for sentiment classification.
	Transformers	Employs BERT for sentiment classification.
	Google Cloud Natural Language API	A sentiment analysis API powered by Google, which categorizes the sentiment of the text into three categories using pre-trained machine learning models.
	BiLSTM	Employs BiLSTM for sentiment classification.
	Random forest	Employs Random forest for sentiment classification.
Manual	Manual annotation	
	Financial Phrase-Bank	English Financial news polarity for 4840 sentences labeled and agreed by 5–8 annotators.
Topic-based	LDA-S model	A proposed extension of the Latent Dirichlet Allocation (LDA) model to obtain the textual topic-specific sentiments.

learn to predict the corresponding sentiment label for new words more intelligently. However, since the supervised learning models rely on labeled datasets, their bias and subjectivity might be reflected in the predicted sentiments.

A comparison of the effectiveness of two sentiment analysis methods for stock market prediction was performed by Alzazah et al. (2022). Although the main purpose of their work was to investigate the reliability of using video news sentiment for stock market prediction, they compared the efficacy of TextBlob and Google Cloud Natural Language API. The results revealed that Google Cloud Natural Language API performed better than TextBlob in determining the sentiment score for each news video.

3.1.3. Historical stock prices

Market data or historical stock prices are numeric values of price points or indexes of past stock prices. Examples of numeric data are the opening price, previous closing price, current closing price, closing bid price, price change, closing offer price, and volume. Although the early attempts at stock market prediction were exclusively based on historical stock prices, researchers still take advantage of this critical data source as a supplement to the more complex unstructured data sources like textual data. Researchers have studied historical stock prices, including the daily opening and closing prices of each stock, to achieve a pattern that can be used for forecasting the future prices of stock indices.

The historical prices are gathered from various sources depending on the target stock indices for prediction. Based on the reviewed articles, Yahoo Finance (https://finance.yahoo.com/) is by far the most prevalent source. It provides financial data and news, including stock quotes, press releases, and financial reports. In particular, Yahoo Finance is used to gather the historical stock prices for the S&P 500, NASDAQ, Dow Jones Industrial Average index (DIJA), New York Stock Exchange (NYSE), and Athens Stock Exchange (ATHEX) (see Table 7). Sina Finance and wind.com were the sources for gathering the historical stock prices of the Shanghai Stock Exchange (SSE). Furthermore, Kaushal and Chaudhary (2017) used the Quandl database, a large public API incorporating millions of financial time-series data points, to gather market data. Vanstone, Gepp, and Harris (2019) and Feuerriegel and Gordon (2018) employed Bloomberg to provide the acquired financial data to predict Australia's 20 largest stocks and the German Stock index (DAX), respectively. The Central Reserve Bank of Peru API Tool for Developers (BCRPData API) was used to gather the historical prices of the Lima Stock Exchange by Ulloa et al. (2022). It is worth mentioning that some of the articles were not report the predicted financial market or the source of the historical stock prices.

3.1.4. Technical indicators

Technical indicators result from mathematical calculations based on historical market data like volume, price, and security interest. They are one of the most important information sources for technical analysts to analyze and predict future stock prices (Jabbarzadeh, Shavalpour, Khanjarpanah, & Dourvash, 2016). Studies revealed that technical indicators as a complementary input improved the performance of prediction models for news-based stock market prediction (Daradkeh, 2022; Vargas et al., 2017). Consequently, researchers combined these indicators to examine their importance for news-based stock market prediction.

The remarkable technical indicators for stock market prediction are listed in several studies (Göçken, Özçalıcı, Boru, & Dosdoğru, 2016; jae Kim & Han, 2000; Tsai & Hsiao, 2010). We aggregated and compiled a list of these indicators (see Table 8). Studies employed R programming packages or the StockStats library to calculate technical indicators (Picasso et al., 2019). We employed the same indicator categories as (Linardos et al., 2015) to classify technical indicators into seven categories: basic functions, market strength indicators, momentum indicators, statistical functions, support and resistance indicators, trend indicators, and volatility indicators. Table 8 lists the technical indicators, as well as their categories, and provides a brief description of them.

According to the supplementary information provided in Table 11 in Appendix, technical indicators are employed in combination with sentiment scores (16 studies), stock prices (14 studies), and text-driven features (10 studies). However, technical indicators are rarely combined with the word embedding approaches for news-based stock market prediction (5 studies). Furthermore, 9 studies used all these types of features as their input data.

3.2. RQ2: What are the most recent relevant text representation and machine learning techniques for intelligent stock market prediction?

In the following discussions, we look at the techniques used for preprocessing data and the methods applied for stock market prediction.

3.2.1. Text representation

Inaccurate representational inputs undoubtedly lead to pointless outputs. Therefore, determining the attributes whereby a piece of text can be interpreted is critical. Textual feature representation involves transforming unstructured textual data into numerical variables to make them mathematically computable. Every textual feature

Table 7
Financial instrument and the source of historical stock prices for reviewed articles. It should be noted that S&P 500, NASDAQ, DIJA, SSE, ATHEX, NYSE, BSE, DAX, and SZSE stand for Standard and Poor's 500, National Association of Securities Dealers Automated Quotations, Dow Jones Industrial Average, Shanghai Stock Exchange, Athens Stock Exchange, New York Stock Exchange, Bombay Stock Exchange, German Stock Index, and Shenzhen Stock Exchange, respectively.

Financial instrument	Financial data source	Ref.
S&P 500	Yahoo Finance	Devaraj et al. (2020), Ding et al. (2015), Minh et al. (2018) Alzazah et al. (2022), Case and Clements (2021), Hu, Wang, et al. (2021)
NASDAQ	Yahoo Finance	Li (2020), Nabil and Magdi (2022)
NASDAQ	Google Finance	Picasso et al. (2019), Sridhar and Sanagavarapu (2021)
DIJA	Yahoo Finance	Picasso et al. (2019), Sridhar and Sanagavarapu (2021)
DIJA	FactSet Platform	Bi et al. (2021)
SSE	Wind.com	Zhang, Qu, et al. (2018)
SSE	Sina Finance	Li et al. (2018)
Apple Index	Yahoo Finance	Liu et al. (2018)
Apple Index	Quandl	Kaushal and Chaudhary (2017)
ATHEX and NYSE	Yahoo Finance	Linardos et al. (2015)
NASDAQ and DJIA	Yahoo Finance	Weng et al. (2018)
Nifty100 and S&P 500	Yahoo Finance	Kamal et al. (2022)
Nifty50, and BSE	Yahoo Finance	Sharma, Tiwari, Gupta, and Garg (2021)
BSE and NIFTY50	Yahoo Finance	Gite et al. (2021)
NSE	Yahoo Finance	Srivastava, Tiwari, Bhardwaj, and Gupta (2022)
S&P ASX20	Bloomberg	Li et al. (2018)
DAX	Bloomberg	Feuerriegel and Gordon (2018)
Lima Stock Exchange	BCRPData API	Ulloa et al. (2022)
DFM	DFM Dataset	Daradkeh (2022)
SSE and SZSE	Sina Finance	Xu, Chai, Luo, and Li (2022)

Table 8
Technical indicators, their categories, and a brief description of them. The references that employed the indicators are listed in the last column.

Technical indicator	Indicator category	Description
Moving average	Trend	Smoothes the time series to form a trend following indicator.
MACD	Trend	Moving average convergence or divergence oscillator for trend following.
Stochastic oscillator	Momentum	Shows the location of the close relative to the high-low range.
Relative strength index (RSI)	Momentum	Measures the speed and change of price movements.
Williams R	Momentum	Inverse of the Fast stochastic oscillator.
Chande momentum oscillator (CMO)	Momentum	Captures the recent gains and losses to the price movement over the period.
Commodity channel index (CCI)	Trend	Identifies a new trend or warns of extreme conditions.
Rate of change (ROC)	Market strength	Measures the percent change from one period to the next.
Percentage price oscillator	Momentum	Measures the difference between two moving averages as a percentage.
Bollinger band width	Volatility	Measures the difference between the upper and lower Bollinger bands, divided by the value of the moving average.
Stop and reverse (SAR)	Trend	Determines the price direction of an asset.
Trend line	Momentum	Measures the rate of increase in the share price over time.
Median price	Basic functions	Calculates median price of stock prices.
Standard deviation	Statistical functions	Calculates standard deviation of stock prices.
Average true range (ATR)	Volatility	14-day simple moving average of a series of true range indicators.
Triple exponential average (TRIX)	Momentum	Presents the percentage change in a moving average.
Cutler's relative strength index	Momentum	Measures the speed and magnitude of price movements by analyzing average losses and gains.
Directional movement indicator (DMI)	Momentum	Measures the strength and price movement to reduce false signals.
Chaikin A/D oscillator	Market strength	Uses MACD to measure the momentum of the accumulation distribution line.
Market facilitation index (MFI)	Market strength	Measures the strength or weakness behind movements of an asset price.
Envelope	Support and resistance	Plots over a price chart with upper and lower bounds.

should be represented by a numeric value that machine learning algorithms can process. We divided the textual feature representation approaches into three main categories: sentiment-based, statistics-based, and learning-based. Traditionally, researchers leveraged sentiment-based or statistical approaches to represent the data. However, learning-based models like word embedding and transformers have recently been involved due to the unstructured or semi-structured nature of such data (Wang, Li, Song, Wei, & Li, 2011). Fig. 4 presents the categorizations of text representation approaches and examples of the discovered techniques in each category according to the reviewed literature. Sentiment-based text representation is discussed in Section 3.1.2 in detail. Therefore, the focus of the following is on the statistics-based and learning-based textual feature representation methods.

Statistic-based feature representation works with the number of occurrences of a word to measure the word's relevancy to a document. The common statistical techniques are bag-of-words and term frequency-inverse document frequency (TF-IDF). In bag-of-words, the frequency of a word in a document is considered a feature, while the order of the words and the grammatical structure of the document are disregarded. Du and Zhang (2018) and Chen et al. (2016) employed this method to transform the text into features in their work. Like bag-of-words, TF-IDF also considers the word's frequency. The TF-IDF weight for each word can be calculated by dividing the occurrence frequency of a word in a document by the frequency of the word in a collection of documents (Hagenau et al., 2013; Wong, 2002). The weight directly correlates with the appearance frequency of a word

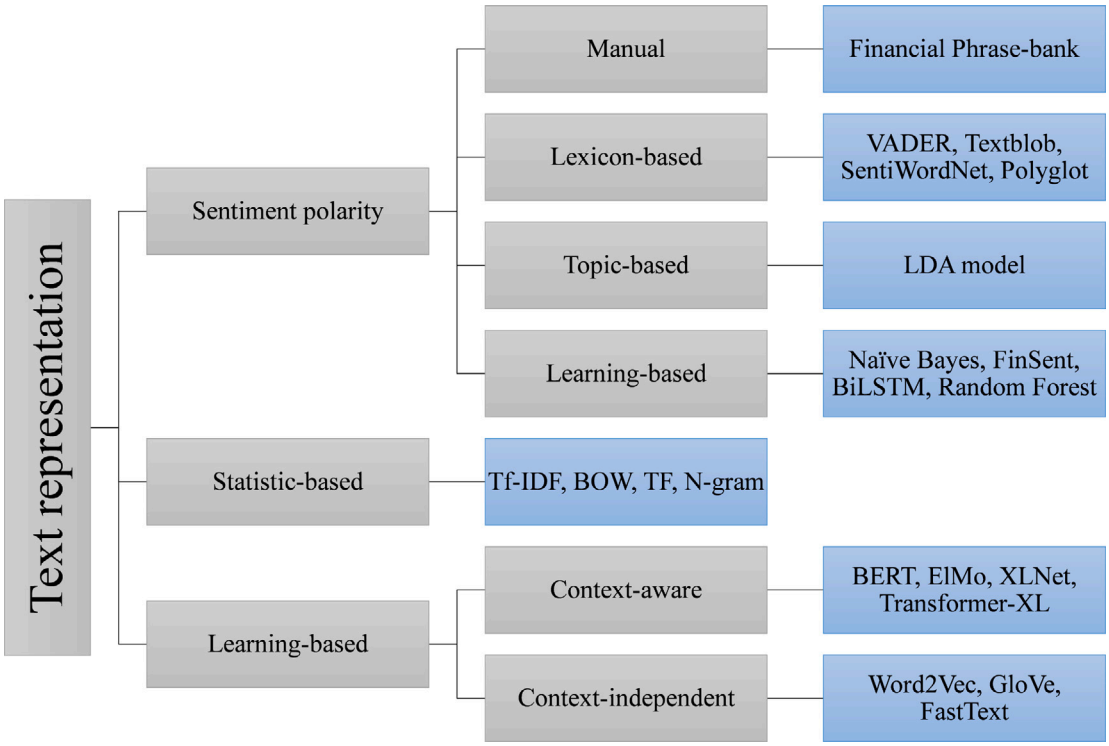

Fig. 4. Categorizations of text representation approaches. The blue rectangles include examples of the discovered techniques for each approach in the reviewed literature.

in the document. Nevertheless, considering the word's recurrence in the corpus neutralizes the overall prevalence of some words. Term frequency-category discrimination (TF-CDF) is a similar measure originating from category frequency. Although it has proved to be more convincing than TF-IDF (Wong, 2002), it was not employed in the reviewed articles.

Word embedding is the general name for the learning-based family of text representation methods. Unlike statistic-based representations, learning-based methods perform unsupervised learning procedures to create a model that assigns a vector to each word in a document. Contrary to the statistic-based methods, the vector size does not equal the number of distinct words in the corpus. Word embedding transforms tokens into numerical vectors. In recent years, these techniques have become notably important for natural language processing tasks, allowing numerous machine learning algorithms to consider vector representations as input data to experience more satisfying representations of textual input. Word embedding models make up a class of artificial neural network algorithms that can be effectively applied to create distributed and condensed representations of words (Bengio, Ducharme, Vincent, & Janvin, 2003). These models are artificial neural networks (ANNs) trained to generate the linguistic contexts of words. With the recent advancements in deep learning, these vector representations are better formed and enable researchers to compute semantic similarities (Chaturvedi, Ong, Tsang, Welsch, & Cambria, 2016).

Discarding the context of a word is the main drawback of statistics-based models. Word embedding approaches have been developed to address this issue. They can be grouped into context-independent (Word2Vec, GloVe, FastText, etc.) and context-aware (BERT, ELMo, Transformer-XL, XLNet, etc.) categories. The aim of context-independent representations is to encode properties related to single tokens, discarding the syntactic relations between them. However, a dynamic representation considering the information of the nearby words is the advantage of the context-aware methods. In particular, contextualized word embedding models benefit from effectively modeling long-term dependencies across tokens in a temporal sequence. Furthermore, another advantage of them is eliminating the sequential dependency on previous tokens, which leads to training a more efficient model.

Word2vec was produced, published, and patented in 2013 by a group of researchers at Google headed by Tomas Mikolov (Goldberg & Levy, 2014). It produces a vector matrix from a large input text. Word2vec assigns a corresponding vector of a predetermined size to each individual word in a corpus. It has recently been widely applied in the area of financial market forecasting (Li et al., 2018; Liu et al., 2018; Minh et al., 2018; Mohan et al., 2019; Vargas et al., 2017). Contrary to word2vec, which uses the word itself to generate the representations, fastText is another word embedding model that uses a bag-of-characters n-gram beside the word itself. In particular, fastText and word2vec are different in terms of the granularity level while learning the representations. While word2vec and fastText are predictive feedforward neural networks, GloVe draws attention to the co-occurrence of the words over a corpus and then is a count-based model. Dharma, Gaol, Leslie, Warnars, and Soewito (2022) examined the performance of these word embedding models on a text classification task. They recorded an accuracy of 97.2%, 95.8%, and 92.5% for the fastText, GloVe, and word2vec, respectively, revealing the superiority of fastText word embedding. However, the difference in accuracy was not notably significant.

Devlin et al. (2018) proposed BERT, which is a state-of-the-art architecture based on transformers applicable to several NLP areas, including text representation and language translation. It leverages an unsupervised deep bidirectional neural network to generate textual representations from a large unlabeled corpus. The advantage of BERT over the previous language representation models is that it enriches contextual information using a masked language model (MLM). In particular, BERT leverages predicting surrounding randomly masked words to establish context. Several variations of the BERT model are proposed and used by researchers, including, FinBert, RoBERTa, BART, ALBERT, and DistilBERT. ELMo stands for Embeddings from Language Models, which is another contextual word embedding model that uses bidirectional LSTM instead of transformers. Furthermore, it enables bidirectionality by concatenating right-to-left and left-to-right LSTMs

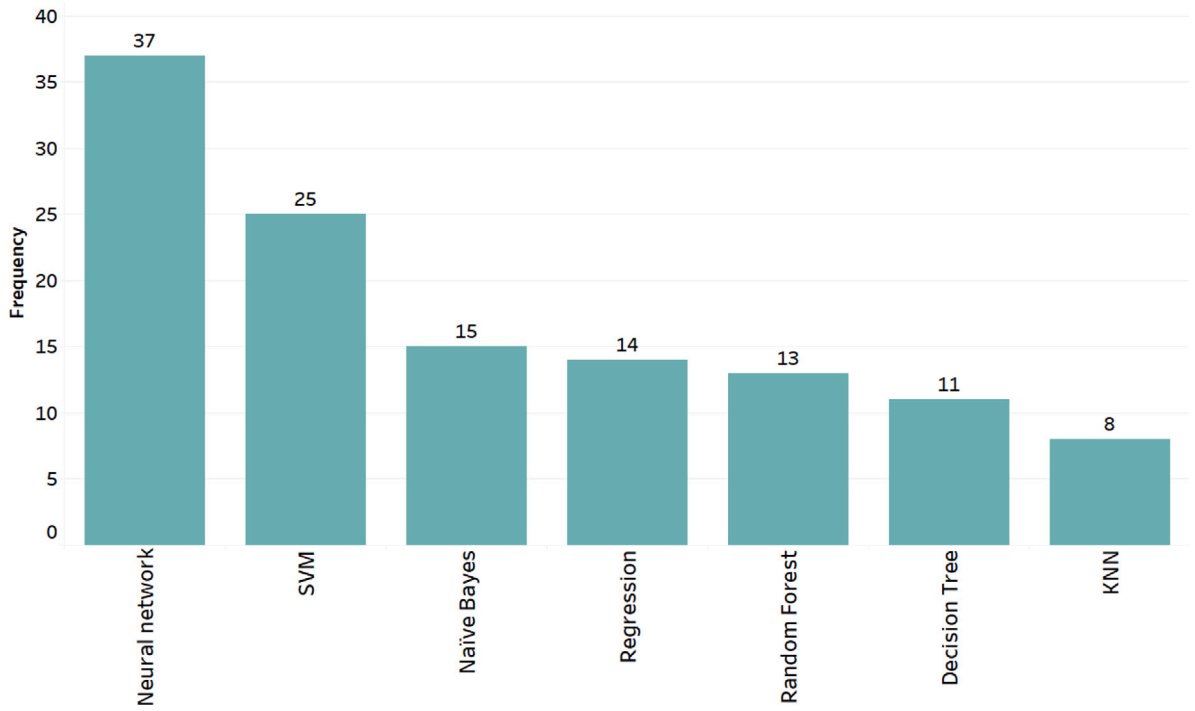

Fig. 5. Distribution of supervised learning models in the reviewed literature.

instead of masked language modeling, which makes it less effective than BERT. The generalized autoregressive language model or XLNet is another word embedding model which is categorized as a large contextual model. Integrating both an autoencoder and autoregressive architectures enables XLNet to outperform BERT in 20 different NLP tasks (Yang et al., 2019). In an interesting study, Mishev et al. (2020) comprehensively studied different test representations, particularly word embedding models, and compared their performance for a stock market prediction task using news articles. They reported that distilled variations of the BERT model outperformed BERT and XLNet for text classification purposes.

3.2.2. Machine learning models

Machine learning techniques should be involved in learning the relationship between textual features with a numeric representation and stock trends. What kinds of models are particularly suitable for the news-based financial market prediction problem? This is still an open question (Kumar & Ravi, 2016). In this section, we bring forward common machine learning algorithms favored in the examined literature and provide a brief description of the algorithms along with their specifications. Fig. 5 presents the distribution of various models in the reviewed articles. Each method has its advantages and disadvantages in response to the characteristics of stock market datasets, indicating that no particular model performs better in all cases. A method will function effectively if the problem is recognized and optimized properly. It is noteworthy that some studies employed more than one machine learning technique.

Comparing the algorithms used in the reviewed articles is not straightforward. The main objective of this section is to report what algorithms are used, which will help identify the gaps and areas for future research. Generally, all the exploited machine learning models are supervised algorithms. The systems use the input data to learn and classify the output in a supervised manner, assigning the input into up, down, and steady categories of stock market fluctuations. However, several works used regression methods such as logistic regression for making predictions, which is also a supervised approach. Regression algorithms are particularly suited to impact analysis problems.

ANNs are the most frequent models used in the literature (See Fig. 5). We identified the following ANNs in the reviewed literature: long short-term memory (LSTM), multi-layer feedforward perceptron (MLP), convolutional neural network (CNN), recurrent neural network (RNN), gated recurrent unit (GRU), recurrent-CNN, and deep belief network (DBN). LSTM is by far the most prevalent employed ANN. The LSTM architecture is a recurrent neural network (RNN) used for processing sequential data. This is why current research in stock market forecasting tends to use it.

MLP is the classic type of neural network comprising one to several layers of neurons. Generally, MLP is flexible and can be suited to many classification and regression problems using different data formats, including time series, images, and text. Attanasio, Cagliero, Garza, and Baralis (2019) reported MLP as the best-performing model compared to k-nearest neighbors (KNN), naïve Bayes, and random forest classifiers for predicting trend reversal in the stock market. They used historical stock prices, news, and technical indicators as the input data and examined the impact of news sentiment on the predictor's performance. The main findings of their study were twofold. First, they confirmed that taking the sentiment from news articles is worthwhile. In particular, feeding the models news and historical market prices significantly boosted the performance of the MLP model by 24% against using merely the historical prices data. Second, they confirmed the robustness of the classifiers for trend reversal prediction in the stock market.

Deep neural networks (DNNs) have attracted attention and been efficiently implemented in several research domains, including machine vision (He, Zhang, Ren, & Sun, 2016) and speech processing (Noda, Yamaguchi, Nakadai, Okuno, & Ogata, 2015). Furthermore, the effectiveness of sentiment extraction using deep-learning neural network models, including recurrent neural networks (RNNs) (Mishev et al., 2020; Rao et al., 2022; Yoshihara et al., 2016), convolutional neural networks (Ding et al., 2015; Li, 2020; Mishev et al., 2020) has been proven by many studies in the finance sector. Fig. 6 presents the frequency of the various deep neural network models in the literature.

CNN is one of the deep architectures of neural networks (Zhang & Wallace, 2015). Convolution, pooling, and fully connected layers are the building blocks of these networks. The convolution layer is a

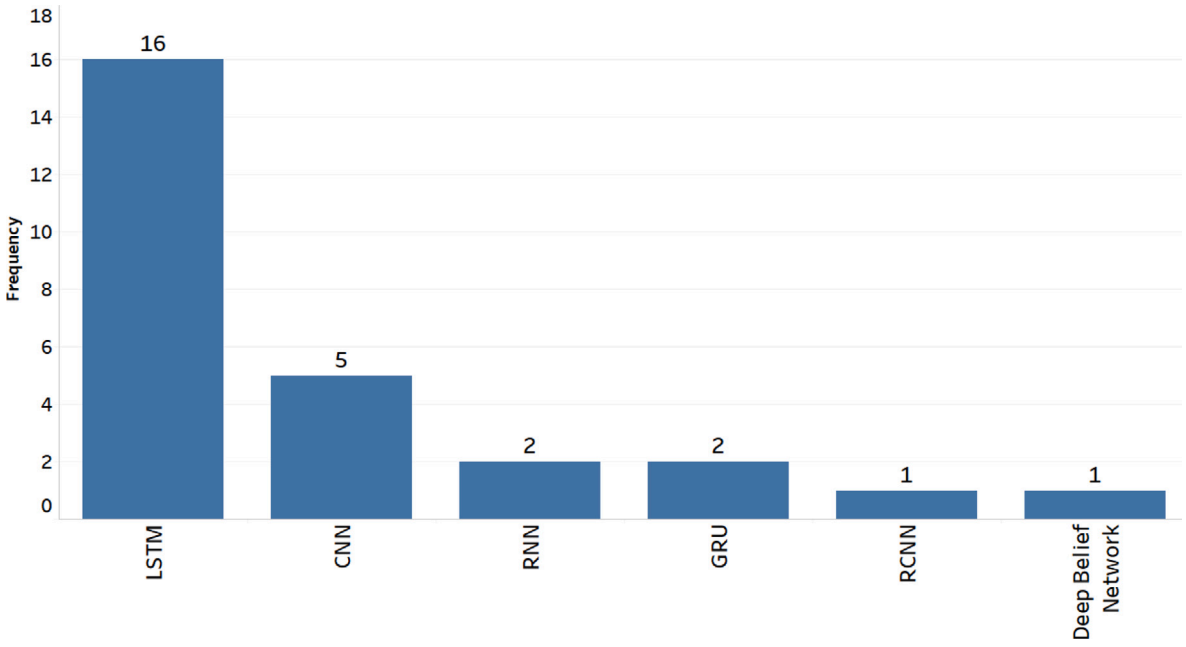

Fig. 6. Distribution of machine learning models in the reviewed literature.

mathematical operation that help with filtering the information and creating feature maps. Then, the extracted features will be generalized and made location-invariant using a pooling layer. Finally, a fully connected layer will be followed by several consecutive convolutions and pooling layers to map the computed features to the output. Although CNN models are extensively used a DNN for learning from grid shape data like images, they have also been explored for time series data processing. Financial time-series processing can benefit from convolution mechanisms to extract the locality of the sequence data and incorporate its spatial structure. Ding et al. (2015) built a deep CNN model for stock market prediction using financial news headlines and event embedding. They considered the standard MLP model as the baseline to compare the performance of the deep CNN model. They confirmed the superiority of the CNN model over MLP for news-based stock market prediction. This advantage is interpreted according to the ability of CNN models to analyze the impact of events over longer terms than MLP. Furthermore, it is indicated that CNN is capable of extracting the most representative parts of the feature for the forecasting task. However, Li (2020) compared the performance of the CNN with LSTM, which is a type of RNN, and concluded that the CNN model is less accurate than the LSTM model.

Although CNN models are capable of acquiring temporal, spatial, and dimensionality reduction using pooling mechanisms, RNNs operate effectively in capturing long-term dependencies and processing sequential data. RNNs are another type of neural network equipped with internal memory units, which distinguish them from the other types of deep neural networks (Liu, Qiu, & Huang, 2016). They are designed to work with large-scale time-series or sequential data, making them a good contributor to developing solutions for future event prediction problems during recent years (Li, Huang, Deng, & Zhu, 2014). Consequently, stock price prediction using financial news and historical stock prices has been motivated by RNNs' success in various fields in predicting sequence events. RNNs perform linear mappings of the input and output of the data and can also do a combination of linear and nonlinear transformations within the layers of the network. As the data flows through the network, weights and biases are adjusted, which act like linear transformations. Then, activation functions transform the input forwarded to the next layer of the network.

Mishev et al. (2020) compared the performance of RNN against other ANNs while feeding textual features into them for stock market prediction. Furthermore, they employed a novel, recently introduced mechanism for enhancing the performance of RNN models called attention (Bahdanau, Cho, & Bengio, 2014). The attention mechanism allows the strengthening of an RNN model by enabling it to identify the important parts of the input sequence and focus on those specific parts. Therefore, it avoids encoding irrelevant contextual parts, which makes the learning procedure more efficient. Mishev et al. (2020) indicated that variations of RNN models outperformed the other DNN models. They confirmed that bringing attention to the mechanism and developing bidirectional versions of the RNN models improves their performance. Bidirectional RNNs are often used to collect features from both directions, enabling the model to consider the contextual information before and after a word in a sentence.

Vargas et al. (2017) combined RNN and CNN together to introduce a recurrent convolutional neural network model (RCNN) for intraday news-based stock market prediction using technical indicators and news headlines. This architecture is capable of learning the temporal features of the news headlines and proved to be more efficient than the CNN model. Furthermore, the results confirmed that both financial news headlines and technical indicators contributed to the prediction model.

LSTM is an extension of RNN models that basically distinguishes itself from RNN with gated units in its hidden layer. In particular, these gated units allow the LSTM model to hold the information longer than RNN, making it useful for preserving long-term dependencies. LSTM is the most prevalent deep learning model employed in the reviewed literature (Fig. 6(b)), and it has proved to be successful in stock market prediction (Alzazah et al., 2022; Daradkeh, 2022; Hu, Wang, et al., 2021). We observed that most studies employed LSTM with the news headline instead of the body, and in most cases, the LSTM model was reported as the best-performing model in the studies. This is evidence of the privilege of LSTM over other deep networks employed in the scope of news-based stock market prediction. We look into the model's performance in detail in Section 3.3. Furthermore, the dataset size in the studies that applied LSTM was larger due to the necessity of having a lot of data to train an LSTM model.

Sarkar et al. (2020) suggested that news headline sentiment can significantly improve the performance of an LSTM model for stock market prediction compared to just using historical stock prices and technical indicators. Kulikovskikh and Voronkov (2020) managed to use the LSTM model on news headlines to predict the upward and downward

trends of the market efficiently. Sharma et al. (2021) suggested the high dependency of the Indian stock market on the sentiment extracted from the news headlines employing an LSTM model. Furthermore, several authors have incorporated LSTM with other deep networks to improve the efficiency of the models. In particular, Gite et al. (2021) employed the LSTM-CNN model to process news headlines and compared the results with a simple LSTM model using historical stock prices. Surprisingly, the simple LSTM model with the stock prices outperformed the LSTM-CNN with news headlines. Furthermore, in another combined study, Mohan et al. (2019) applied RNN-LSTM alongside stock prices, sentiment polarity, and textual information. They achieved outstanding results with LSTM, suggesting a solid correlation between the historical stock market price fluctuations and news articles in the financial sector. Nevertheless, the limitation of their model was weaker prediction performance for low stock prices or highly volatile markets.

Bidirectional extensions of the LSTM have also proved to be efficient for news-based stock market prediction. Merello, Ratto, Ma, Oneto, and Cambria (2018) assessed two probable components for financial market fluctuations: a single news source and multiple news sources. Initially, they determined the correlation between a single piece of news and related stock prices. They studied a set of recently published news articles as input and developed a model to select the most decisive record of news in the set by tracking the flow of derived information. Subsequently, they examined an accumulation of several news articles as input. A single vector represents a combination of recently published news stories, from the latest one back to a given point. As in the first approach, the model selects the most important news collection input and tracks its evolution over time. Their model includes three phases: extracting the most informative element, modeling the transformation of the information over various time intervals, and eventually, performing the prediction task using the BiLSTM model. They examined the performance of the BiLSTM model with and without news headline sentiments and compared the average stock price rate in both experimental settings. They used hybrid LSTM-CNN (Oncharoen & Vateekul, 2018) and GRU with GloVe embeddings (Liu, Trajkovic, Yeh, & Zhang, 2020) as the baseline models to compare their results and suggested that the proposed BiLSTM model with news headline sentiments outperformed the baseline models.

The last family of the deep RNN models in the literature was gated recurrent units (GRUs). Like LSTM, these models benefit from memory cells to preserve the context information in text processing tasks. However, they are less complex than LSTM since they have only two gates, reset and update, while in LSTM, each node has three gates: input, output, and forget. Although GRU models are more suitable for smaller datasets compared to LSTM, they have proved to be computationally more efficient and about 30% faster than LSTM on benchmark datasets (Cho et al., 2014). Inspired by bidirectional RNNs (Schuster & Paliwal, 1997), Minh et al. (2018) proposed a two-stream GRU (TGRU) model to be able to capture the contextual information in a sentence in both forward and backward directions. Although the TGRU model outperformed GRU and LSTM models, it was not compared to a BiLSTM model in their study.

SVM is the second most extensively used approach in the literature. Although deep neural networks have proved to be more effective when there is a large corpus to deal with for news-based stock market prediction, when there is a limited amount of data, SVM has proved to be more efficient than ANN models. Furthermore, another advantage of SVM is that its computational complexity is independent of the input space dimension. SVM is widely implemented in addressing classification problems, especially in the context of financial market prediction (Fig. 5). Many researchers have aimed to determine the future price of stocks by employing financial news articles and relying on the SVM machine learning model. The raw data is transfigured into the feature space using a transformation function known as the kernel, enabling researchers to operate in a high-dimensional feature space beyond calculating the coordinates of the data in the primary space. Derivatives of kernels were developed to support various data structures. In particular, kernels enable SVM to operate on strings without transforming them into fixed-length, real-valued feature vectors. Furthermore, graph kernels empower SVM to work with graphs directly. Nowadays, many researchers fuse SVM with various financial news bodies and kernels. Nevertheless, almost all the research has focused on news content, and information regarding the correlation among various news articles is rarely considered.

Several researchers have recorded the outperformance of the SVM model over other classifiers, including naïve Bayes and MLP (Preis, Reith, & Stanley, 2010), random forest (Eck et al., 2021), KNN (Khadjeh Nassirtoussi et al., 2015), and logistic regression (Kaushal & Chaudhary, 2017). In particular, Preis et al. (2010) compared the effectiveness of the SVM model to that of naïve Bayes and MLP models, using both positive and negative news sentiments, and showed that SVM delivered the best results. Furthermore, Long et al. (2019) proposed a new SVM kernel to derive the relationship among inputs and implemented this approach for stock market movement prediction using economic news. This additional knowledge was leveraged to enhance the prediction performance for stock market price trends. The performance of the new proposed kernel was compared to that of linear, radial basis function (RBF), sigmoid, and polynomial kernels and proved to propose more accurate results for stock market prediction.

The naïve Bayes model is another classification model employed for news-based stock market prediction which is built on the Bayes theorem. As with SVM, not much input data is needed to train a naïve Bayes prediction model. Unlike SVM, naïve Bayes holds a hypothesis of the independence among features, while SVM takes the interactions between features into consideration. Generally, articles have employed naïve Bayes as a text classification model for sentiment analysis to predict the polarity of a piece of news using a learning-based model. It was noted that this model performed accurately for text data classification (Jishag, Athira, Shailaja, & Thara, 2020). Jishag et al. (2020) proposed a model fusion architecture in which the results of the naïve Bayes model for news sentiment classification were combined with historical stock prices and fed into a KNN model to perform stock market prediction. Furthermore, Li et al. (2018) adopted a multinomial extension of the naïve Bayes classifier along with random forest and SVM as a baseline for stock market prediction. They used a dataset including commercial text (corporate announcements, research reports, stock news, and stock bulletin board system posts), the number of postings, and the historical stock prices from the Shanghai composite index as input. An interval of 10 days before the appearance of the trend was considered as the experiment period. However, the experiments showed that multinomial naïve Bayes did not perform as accurately as the LSTM model.

Members of regression models family, including logistic regression, support vector regression (SVR), random forest regression (RFR), principal component regression, LASSO regression, generalized linear model (GLM) regression, elastic net, boosted regression tree (BRT), neural networks regression (NNR) and the Bayesian ridge model, were employed 14 times among the literature. Generally, regression models attempt to explain the relationship between one dependent and one or several independent variables. Azizi et al. (2021) designed an architecture for news-based stock market prediction in which historical stock prices were fed into a regression model, and the results were combined with sentiment analysis to decide future stock prices. They also investigated the effect of the train/test split ratio on the performance of the implemented regression model and concluded that a train/test ratio of 80:20 led to the most accurate results. However, they were not able to achieve good performance using the regression model. Furthermore, Saxena, Bhagat, and Tamang (2021) evaluated the performance of various machine learning models, including naïve Bayes, generalized linear regression, logistic regression, decision tree, random forest, gradient boosted tree, and SVM. They used news headlines sentiments to predict the stock market and concluded that logistic

regression, naïve Bayes, SVM, and random forest delivered performance above 90% on this task. Furthermore, in terms of training time, the least training time was reported for the regression models.

Kaushal and Chaudhary (2017) introduced the news- and events-aware regression-based stock market prediction method. They examined the correlation between the historical market of a company and the news and events of the same company. They built an accurate stock price prediction model considering the impact of events and news on the stock price. The authors designed and implemented a sentiment polarity model for the news articles and converted text into feature vectors. Finally, they created a regression model to understand the influence of news and events on stock market price movements.

Regression and ensemble learning approaches, including boosting, bagging, and random forest, are integrated for stock market prediction. Generally, ensemble models seek to group several classifiers together to improve the performance of a simple predictor. Boosting methods aim to build strong learners by grouping weak learners together and minimizing the training error. Bagging or bootstrap aggregation is another ensemble method in which several base classifiers are fitted to various bootstrap samples. Last but not least, a random forest combines the output of multiple decision trees together. Weng et al. (2018) compared various ensemble models of the regression methods in their work. SVR is the regression version of the SVM classifier, which attempt to predict real values instead of class labels. Unlike linear regression, which trains a model based on minimizing the error between the actual and predicted labels, SVR attempts to fit the best line within a threshold. Like SVM, SVR also suffers from sensitivity to the kernel function. A bagging ensemble of SVR was used by Weng et al. (2018) for stock market prediction. BRT, RFR, and NNR regression are other types of regression ensembles used in work by Weng et al. (2018). They concluded that all the approaches except the SVR ensemble benefit from news sentiment data for stock market prediction. The SVR ensemble suffered from overfitting when the news data were added. Furthermore, the BRT and RFR ensembles delivered better performance compared to the rest of the models. A positive effect of principal components analysis (PCA) was also reported from their experiments.

The last study we want to discuss is a work by Ranibaran, Moin, Alizadeh, and Koochari (2021) in which several regression models, including SVR with various kernels, regression tree, random forest, Bayesian ridge, and LASSO regression, were compared. The best reported performance was obtained by a random forest classifier. Furthermore, employing different kernels for the SVR model had no significant impact on the results of this regression model.

Random forest is one of the common ensemble models which have attracted the researchers's attention for stock market prediction. However, generally, it was used as the baseline classifier for performance comparison of the proposed models. The only study that reported the advantage of the random forest model was the work by Ranibaran et al. (2021), in which random forest outperformed several regression models with stock prices and news headlines.

The decision tree is a common supervised classifier with many extensions, including C4.5, J48, and gradient-boosted trees. The random forest is independent, while gradient boosting creates the trees in an additive manner. The model resulting from incorporating decision trees is called a gradient-boosted tree. Like the other boosting approaches, it creates a prediction model using an ensemble of weak models, mainly decision trees. In general, both random forest and gradient boosting are ensemble methods comprising multiple decision trees. However, the procedure of tree building in the random forest is independent, while gradient boosting creates the trees in an additive manner. Gradient-boosted tree models usually outperform random forests (Madeh Piryonesi & El-Diraby, 2021). As a piece of evidence from the reviewed studies, Naderi Semiromi, Lessmann, and Peters (2020) explored the predictive relationship between financial events and the financial market. They examined extreme gradient boosting, a derivation of an improved gradient boosting method, and observed it outperformed random forest and SVM. Furthermore, Feuerriegel and Gordon (2018) and Mishev et al. (2020) implemented gradient boosting with TF-IDF weight features and word embedding feature vectors, respectively, to predict stock market prices.

We can summarize our findings regarding the machine learning models in the following points:

- The reviewed literature attempted to predict the future value of the stock market or the stock trend using various supervised models.
- There has been a trend toward using deep neural networks for news-based stock market prediction in recent years. In particular, LSTM, which is a popular type of RNN model, has gained the most attention recently.
- There is a trend toward using ANN models for future stock price prediction using deep neural networks, in particular, RNN models like LSTM.
- MLP using historical stock prices, technical indicators, and financial news sentiments has proved to be a robust classifier for trend reversal prediction in the stock market (Attanasio et al., 2019).
- The attention mechanism improves the performance of RNN models in predicting future market prices by focusing on the most important parts of contextual information and ignoring the irrelevant parts (Mishev et al., 2020).
- Bidirectional variations of RNN models, particularly, BiLSTM and BiGRU, have proved to be more efficient for news-based stock market prediction. They preserve the input in both forward and backward directions, making them understand the context better (Mishev et al., 2020).
- Although (Ding et al., 2015) reported the superiority of CNN over standard MLP models, CNN models are unable to compete with RNN variations, including attention-based and bidirectional RNNs (Mishev et al., 2020).
- Borrowing the convolutional layer from CNN and combining it with RNN into a new architecture outperformed the CNN model using news headlines and technical indicators as input data (Vargas et al., 2017).
- LSTM has frequently outperformed the other deep learning models, particularly when using large numbers of news headlines.
- In highly volatile market situations and for low stock prices, the LSTM model performed less efficiently.
- The advantage of GRU models is their processing speed compared to LSTM models. However, they are more suitable for smaller datasets.
- The advantage of SVM for news-based stock market prediction is that (a) compared to deep learning models, it can work efficiently with a limited dataset size, and (b) its computational complexity is independent of the input space dimension.

3.3. RQ3: What kinds of evaluation criteria (performance measures) are used, and how well do the models predict the stock market?

3.3.1. Performance measures

Previous studies have reported their results in multiple forms. Common evaluation metrics include accuracy, F1 score, precision, recall, mean squared error (MSE), mean absolute percent error (MAPE), and root mean square error (RMSE). Xing et al. (2018) categorized the measurements into directional accuracy, closeness, and trading simulation. We also identified these types of measures in the reviewed articles, as shown in Table 9.

The first and most prevalent group of measures is directional accuracy, which contains accuracy, F1 score, precision, and recall. In theory, any accuracy rate substantially higher than 50% can confirm the functionality of the prediction model. However, accuracy improvements on a benchmark would be more persuasive. Precision, also called the true positive rate, is the number of true positives divided by the number

Table 9
Model evaluation measurements.

Measurement	Articles
Accuracy	Chen, Fan, Chen, and Hsieh (2019), Meyer et al. (2017), Picasso et al. (2019) Li et al. (2018), Long et al. (2019), Merello et al. (2018), Shah et al. (2018) Chiong et al. (2018), Du and Zhang (2018), Minh et al. (2018), Zhang, Qu, et al. (2018) Kaushal and Chaudhary (2017), Liu et al. (2018), Seif, Ramzy Hamed, and Abdel Ghfar Hegazy (2018), Vargas et al. (2017) Kazemian et al. (2016), Khedr, Salama, and Yaseen (2017), Pröllochs, Feuerriegel, and Neumann (2016) Khadjeh Nassirtoussi et al. (2015), Lima et al. (2015)
Precision	Du and Zhang (2018), Li et al. (2018), Ma and Liang (2015), Minh et al. (2018) Jin et al. (2017), Kaushal and Chaudhary (2017), Lima et al. (2015), Pröllochs et al. (2016) Khadjeh Nassirtoussi et al. (2015), Linardos et al. (2015)
Recall	Du and Zhang (2018), Li et al. (2018), Ma and Liang (2015), Minh et al. (2018), Picasso et al. (2019) Jin et al. (2017), Kaushal and Chaudhary (2017), Pröllochs et al. (2016)
F1-score	Jin et al. (2017), Li et al. (2018), Liu et al. (2018), Zhang, Qu, et al. (2018) (Lima et al., 2015; Linardos et al., 2015; Pröllochs et al., 2016)
MSE	Wang et al. (2018)
MAPE	Mohan et al. (2019), Weng et al. (2018)
RMSE	Feuerriegel and Gordon (2018), Weng et al. (2018)
MAE	Weng et al. (2018)
Trading simulation	Feuerriegel and Prendinger (2016), Kazemian et al. (2016), Linardos et al. (2015)

of positive predictions. Recall or sensitivity is the fraction of positives that are correctly classified. Some studies have considered precision and recall to analyze the false positive and false negative errors (Du & Zhang, 2018; Li et al., 2018; Ma & Liang, 2015; Minh et al., 2018; Picasso et al., 2019). The F1 score is another indicator of a prediction model's performance and reflects both recall and precision.

The next measurements calculate how close the predicted prices are to each other and the corresponding actual stock prices. Generally, studies that used function approximation tasks evaluated their work with closeness metrics such as MSE, MAPE, and RMSE.

The last metric is a trading simulation, which provides data such as portfolio performance or the profit ratio to show the performance of a trading strategy (Xing et al., 2018). For example, Linardos et al. (2015) proposed a forecasting system that participated in the Athens Virtual Trader competition, achieving a considerable increase in capital and offering a significant practical trading experience. More specifically, for the majority of cases during the two-week participation period (April 12, 2010, to April 23, 2010), the system's correct prediction of the trend and the future movement of the selected stock price was notably high (around 70%). Also, Feuerriegel and Prendinger (2016) designed a trading strategy that utilized textual news to obtain profit based on new information entering the market.

3.3.2. High performance models

For each study, a combination of aspects of the proposed framework directly impacts the performance of the final model. In particular, in the area of stock market prediction, the dataset and its characteristics (number of records, length of each news text), the procedure of decoding the textual data into numerical vectors known as the feature representation, and the employed machine learning classifier are the most influential elements. We synthesized the data extracted from the reviewed literature and provide a table presenting the top-performing frameworks. The ideal scenario to compare the models is to examine them against a benchmark with exact performance measurements. However, to give future researchers in this area some insights in terms of performance, we present the description of the high-performance models. It is worth mentioning that, evidently, the results from different studies are not directly comparable due to the different settings and datasets for each experiment. However, we provide a comparison of the accuracy values to summarize the performance of the well-performing models according to their own experimental settings. Since the most prevalent measurement was accuracy, we compared the studies reporting the accuracy of their models. Table 10 presents the accuracy, machine learning model type, text representation, and preferred part of the news for extracting textual features. Furthermore, it was specified if the studies' models made use of historical stock prices and technical indicators.

According to Table 10, the future literature in the area of news-based stock market prediction can still benefit from taking advantage of learning-based methods for sentiment analysis and text representation. Furthermore, there is room for proposing new forecasting frameworks by integrating the learning-based models with historical stock prices and effective technical indicators in this research area.

Regarding the models' performance, we can summarize the following points based on the reviewed literature.

- SVM is a well-performing model for stock market prediction. It has been the focus of both earlier and more recent studies. In particular, all three articles published before 2017 employed the SVM model in their studies (Chen et al., 2016; Kazemian et al., 2016; Khadjeh Nassirtoussi et al., 2015). Likewise, among the most recent studies, Kaushal and Chaudhary (2017) employed SVM with the polarity score features of news from Reuters for Apple stock. It performed well compared to naïve Bayes and logistic regression models. Moreover, comparing several models, Eck et al. (2021) reported that SVM (70%) delivered the best accuracy, followed by random forest (67%), logistic regression (66%), decision tree (59%), and naïve Bayes (66%).
- The earlier studies used more traditional statistics-based feature representation techniques to transform textual data into numeric vectors, such as TF-IDF Khadjeh Nassirtoussi et al. (2015), bag-of-words (Chen et al., 2016), and sentiment polarity scores (Kazemian et al., 2016). Recently, however, despite emerging learning-based embedding methods, statistic-based approaches have continuously attracted researchers. Furthermore, Devaraj et al. (2020) exploited a combination of NLP data gathering and processing methods. They used named entity recognition (NER) to filter news articles based on the name of the desired firm. Besides, a combination of fine-grained part of speech (POS) methods with TF-IDF was used to select the most informative features.
- Most of the articles in the category of word embedding feature representation preferred news headlines over the body of the

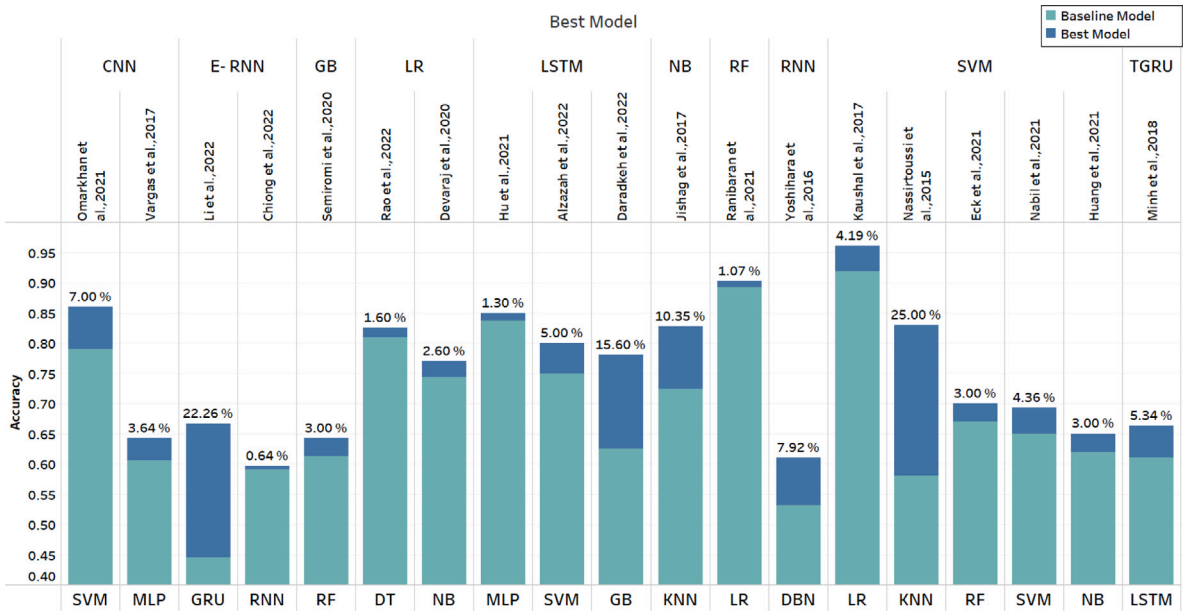

Fig. 7. A comparison of the compared models across the reviewed literature. E-RNN, LR, NB, RF, TGRU, GB, DBN, and DT stand for ensemble RNN, logistic regression, naïve Bayes, random forest, two-way GRU, gradient boosting, deep belief network, and decision tree, respectively.

Table 10
High-performance models. Regarding the dataset ratio, some studies divided the data into three parts–train/validation/test–while the rest only used only two splits: train/test.

Ref.	Model	Text representation	Preferred part	Acc.(%)	Dataset ratio	Sentiment analysis	Stock prices	Technical indicators
Gite et al. (2021)	LSTM	Word embedding	Headline	96.2	80:10:10	No	Yes	No
Kaushal and Chaudhary (2017)	SVM	TF-IDF	–	96.04	70:30	Lexicon-based	Yes	Yes
Case and Clements (2021)	LSTM	BERT	Headline & body	94.92	–	No	Yes	No
Azizi et al. (2021)	KNN	TF-IDF	Headline	91.7	90:10	Yes	Yes	No
Khedr et al. (2017)	KNN	TF-IDF	–	89.8	–	learning-based	No	No
Mishev et al. (2020)	BiLSTM	ElMo	Headline	88.8	–	No	No	Yes
Li et al. (2018)	LSTM	Word2Vec	Headline & body	88.3	–	Lexicon-based	Yes	No
Hu, Wang, et al. (2021)	LSTM	Word2Vec	Headline & body	85.02	70:30	Lexicon-based	Yes	No
Srivastava et al. (2022)	LSTM	Ngram/BOW	Headline	84.6	80:20	No	Yes	No
Kulikovskikh and Voronkov (2020)	LSTM	Word embedding	Headline	84.3	80:10:10	No	No	No
Khadjeh Nassirtoussi et al. (2015)	SVM	TF-IDF	Headline	83.33	80:20	Lexicon-based	Yes	No

articles (Kulikovskikh & Voronkov, 2020; Liu et al., 2018; Mishev et al., 2020). Mishev et al. (2020) presented a comprehensive study of the different word embedding approaches. This is one of the few studies in which the latest available text encoders known as transformers have been considered. They obtained an accuracy of 89.5% utilizing feature vectors from transformers and feeding them into a 12-layer deep neural network with 1024 hidden nodes and a total of 406M parameters.

3.3.3. Best-performing models compared to baseline models

Most studies compared the performance of the suggested state-of-the-art classifiers with that of some baseline models. To give a more accurate insight into the best-performing state-of-the-art models, a comparison of the best-performing models against the baseline models is provided in Fig. 7. In particular, for each study, the figure presents which model performed the best and how well it performed compared to the reported baseline accuracy. It is worth mentioning that Fig. 7 includes only the works in which a comparison based on accuracy measures is provided.

3.3.4. The effect of news part on model performance

Fig. 8 draws an intuitive relationship between the models' performance and the preferred part of the news. We considered the studies that reported their results using accuracy measures for consistency. Apparently, the results from different studies are not directly comparable due to the different settings for each experiment. Nonetheless, Fig. 8 brings attention to the impact of the preferred part of the news on the performance of the models and guides future researchers toward choosing the best part for developing a better predictor. The significance of news headlines for stock market prediction can be observed from the figure. Since news headlines are more straight to the point and less noisy than the whole body of the article, they have attracted more attention recently (Wong, 2002). Furthermore, computationally, they are less expensive than the news body due to their shorter length. The news body is more comprehensive than headlines and includes more detailed information about an event. Although news headlines presented impressive accuracy in several studies, the models benefit from using both news headlines and body text (Case & Clements, 2021; Hu, Wang, et al., 2021; Li et al., 2018). However, the news body itself appeared to be less effective in the works of Chiong et al. (2022), Eck et al. (2021).

3.4. RQ4: What are the trends, gaps, and future research directions?

This section discusses potential directions for future research according to the discovered gaps in the reviewed literature.

High correlation between news data and stock price movements. We observed a strong interrelationship between stock market prices and textual information from financial news articles. According to the studies, the embedded information in news data could be a reliable predictor of fluctuations in the stock market (Merello et al., 2018). Additionally, the impact of financial articles on identifying abnormal

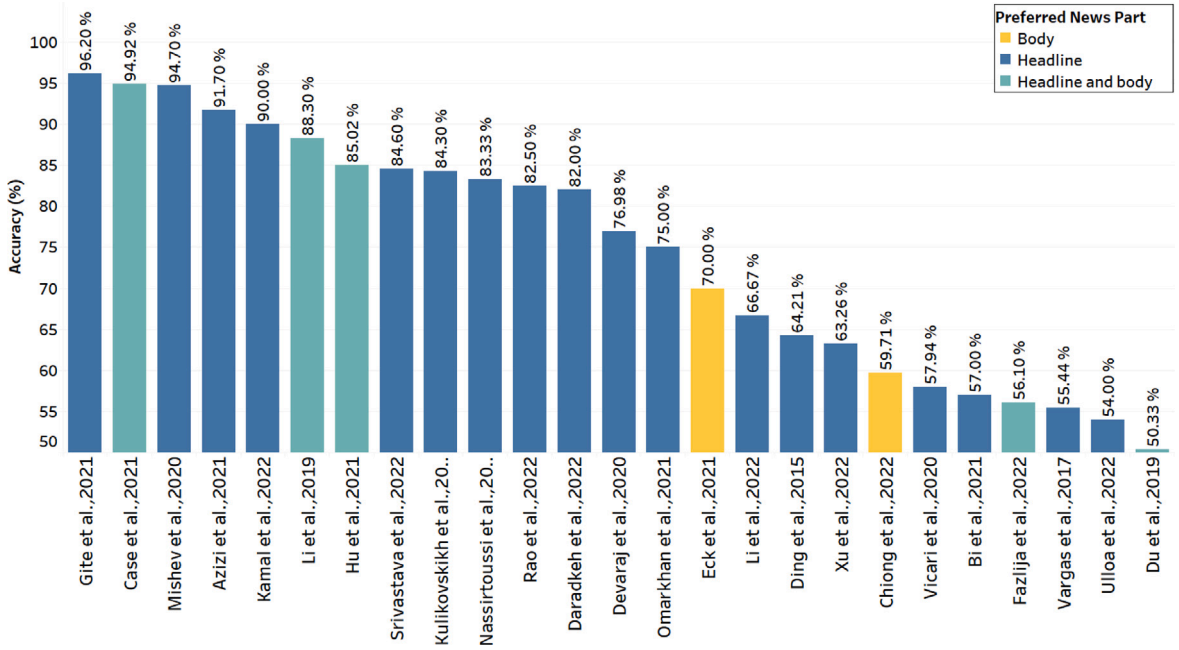

Fig. 8. The relationship between the models' accuracy and the preferred part of the news. The colors demonstrate the preferred news part. Out of 61 reviewed studies, the articles that reported their models' performance using accuracy measures and reported the preferred part of the news are included in this graph.

market trends was observed. Li et al. (2018) confirmed that their model successfully recognized unusual market trends by employing the abnormal sentiment information derived from financial news items, proving the effectiveness of financial data for stock market prediction.

Corporate disclosures as credible data types. One contribution of this study that distinguishes it from the other reviews conducted in this area is that we pay attention to the dataset as a determinant of the stock market forecasting problem. In terms of the publication frequency of the data, social media posts are continually generated, news articles are in the second position, and corporate disclosures are far less frequent. Although social media content such as tweets on Twitter are still in use for stock market prediction, many researchers excluded this data source from their studies due to the abundant noise in the data. However, even though the news articles used in the studies were from well-known newspapers and websites, they still have less authority and more noise than corporate disclosures. Corporate disclosures are more reliable data sources despite being generated less often. It is mandatory for companies to disclose materials related to the stock market immediately through standardized channels. Besides, the quality of the content should be verified and authorized by company executives. This makes corporate disclosures one of the most credible data sources. Furthermore, quarterly or annual financial statements, such as 10-K reports, can be used as another authorized source of textual data. Financial statements include a section known as "Management Discussion & Analysis (MD&A)", which provides comprehensive knowledge about a firm and its commercial status. Financial statements have been widely used for other fiscal applications, such as financial fraud detection using machine learning and data mining techniques (Ashtiani & Raahemi, 2021). However, we did not encounter any articles employing this information for the stock market prediction problem. This can be an interesting potential ground for future work.

Data gathering limitations. It is straightforward to collect historical market data for any given corporation from sources such as Yahoo Finance and Google Finance. While the APIs facilitate gathering such numerical data, collecting news articles is more challenging. Although we identified several APIs (Intrinio, Event Registry, FinSent, etc.) that can simplify the data gathering procedure and provide a more standard integrated data format to researchers, the importance and effectiveness of these news data sources are not fully appreciated. The researchers rarely relied on these financial news data sources. Most studies limited their news sources to one or several news websites and designed crawlers to scrape news from the online press (Picasso et al., 2019; Weng et al., 2018).

Inadequate textual data records. The number of instances of datasets is also a matter of discussion. Most researchers identified a limited number of news data records for specific firms and stock indexes as a shortcoming of their experiments. A small number of news records impairs the performance of the machine learning models significantly and diminishes their generalization. Besides, the advent of novel and more sophisticated deep neural networks made the impact of data size indisputable. In deep neural networks, there are many parameters to be tuned, which essentially increases the degrees of freedom of the model. Consequently, an inadequate number of data records undermines the accuracy of such classification models. Relying on news data APIs and larger prepared datasets could be a potential approach for growing the number of news data records.

The necessity of a benchmark dataset. To discover the state-of-the-art models for a stock market prediction task, we need to compare the results of the studies. As an illustration, we realized that although the target predicted index was generally among the most popular firms such as Apple, Google, IBM, and Yahoo, endeavors to compare the models' results with previous works have been rarely observed in the literature. Indeed, the diversity of datasets complicates the comparison between machine learning models. A benchmark news dataset would enable researchers to identify the best-performing models for stock market prediction. We believe creating a real-time comprehensive financial (and even political) news dataset for different companies and stock indexes might be valuable. An advantage of a comprehensive news dataset is that it can help researchers to investigate the correlation between different stock index prices. Graph representation learning methods help analyze these relations and consider them as a new feature for improving the performance of learning algorithms (Hamilton, 2020).

Transformers, the latest available word embedding techniques. To tackle the context-independence disadvantage of word embedding models like Word2Vec, transformers have been invented to produce

Table 11
The employed input data sources and the text representation approach for each reviewed article.

Ref.	Text representation	Stock prices	Sentiment	Technical indicators
Lima et al. (2015)	Sentiment polarity	No	Yes	No
Khadjeh Nassirtoussi et al. (2015)	TF-IDF	Yes	Yes	No
Linardos et al. (2015)	Sentiment polarity	Yes	Yes	Yes
Ding et al. (2015)	Word embedding	Yes	No	No
Kazemian et al. (2016)	TF	No	Yes	No
Feuerriegel and Prendinger (2016)	Sentiment polarity	No	Yes	Yes
Shynkevich et al. (2016)	TF-IDF	Yes	No	No
Chen et al. (2016)	BOW	No	Yes	No
Yoshihara et al. (2016)	BOW	No	No	Yes
Khedr et al. (2017)	TF-IDF	No	Yes	No
Vargas et al. (2017)	Word2Vec	No	No	Yes
Kaushal and Chaudhary (2017)	Sentiment polarity	Yes	Yes	Yes
Jishag et al. (2020)	TF-IDF	Yes	Yes	No
Liu et al. (2018)	Word Embedding	Yes	Yes	Yes
Chiong et al. (2018)	Sentiment polarity	No	Yes	No
Minh et al. (2018)	Word2vec GloVe	Yes	Yes	Yes
Zhang, Qu, et al. (2018)	Word2Vec	Yes	Yes	No
Weng et al. (2018)	Sentiment polarity	Yes	Yes	Yes
Wang et al. (2018)	Sentiment polarity	Yes	Yes	No
Du and Zhang (2018)	Sentiment polarity	Yes	Yes	No
Li et al. (2018)	Word2Vec	Yes	Yes	No
Merello et al. (2018)	TF	Yes	Yes	Yes
Long et al. (2019)	BOW	No	No	No
Mohan et al. (2019)	Word2Vec	Yes	Yes	Yes
Vanstone et al. (2019)	TF	Yes	Yes	No
Attanasio et al. (2019)	Sentiment polarity	Yes	Yes	Yes
Picasso et al. (2019)	Sentiment polarity	Yes	Yes	Yes
Naderi Semiromi et al. (2020)	TF-IDF	Yes	Yes	Yes
Vicari and Gaspari (2021)	Word Embedding	No	No	Yes
Mishev et al. (2020)	Word2Vec, Transformers, GloVe, FastText	No	Yes	No
Sarkar et al. (2020)	Sentiment polarity	Yes	Yes	Yes
Devaraj et al. (2020)	TF-IDF	Yes	Yes	Yes
Kulikovskikh and Voronkov (2020)	Word embedding	No	No	No
Li (2020)	Sentiment polarity	Yes	Yes	Yes
Ingle and Deshmukh (2021)	TF-IDF	No	No	No
Eck et al. (2021)	TF-IDF	Yes	No	No
Azizi et al. (2021)	TF-IDF	Yes	Yes	No
Bi et al. (2021)	BOW	Yes	Yes	No
Ranibaran et al. (2021)	Sentiment polarity	Yes	Yes	No
Omarkhan et al. (2021)	Word2Vec	No	No	No
Case and Clements (2021)	BERT	Yes	No	No
Sharma et al. (2021)	Sentiment polarity	Yes	Yes	No
Gite et al. (2021)	Word embedding	Yes	No	No
Huang (2020)	TF-IDF		Yes	No
Sridhar and Sanagavarapu (2021)	TF-IDF	Yes	Yes	No
Hu, Wang, et al. (2021)	Word2Vec	Yes	Yes	No
Nabil and Magdi (2022)	Sentiment polarity	Yes	Yes	No
Fazlija and Harder (2022)	Sentiment polarity	No	Yes	No
Kamal et al. (2022)	TF-IDF	Yes	Yes	No
Raman et al. (2022)	Sentiment polarity	No	Yes	Yes
Li and Pan (2022)	Sentiment polarity	Yes	Yes	No
Xu et al. (2022)	Word embedding	Yes	No	No
Chiong et al. (2022)	Sentiment polarity	Yes	Yes	No
Srivastava et al. (2022)	BOW	Yes	No	No
Ulloa et al. (2022)	BERT	Yes	Yes	No
Alzazah et al. (2022)	GloVe TF-IDF	Yes	Yes	No
Daradkeh (2022)	Word2Vec	Yes	Yes	Yes
Rao et al. (2022)	Sentiment polarity	No	Yes	No

better context-dependent representations of words. In particular, transformers generate multiple representations for the same word based on the context, as opposed to the embedding models, which create a unique vector for each word. Furthermore, another disadvantage of the word embedding models is that they are not able to generate representations for words that are not in their vocabulary. Mishev et al. (2020) suggested that the latest developments in deep neural networks and NLP transformers had improved the performance of sentiment analysis. Therefore, the financial stock market prediction problem can significantly take advantage of transformers for textual data representation. Several extensions of transformer-based models are employed by Mishev et al. (2020). The results indicated that the integration of transformer representations and deep learning classifiers outperformed the lexicon-based and statistics-based models for word representation. Hence, implementing variations of transformer models for text representation is a great avenue for future research.

Syntax is also of the utmost importance. Proper examination and implementation of syntax in conjunction with semantics (or even instead of it) can enhance the accuracy of textual classification. In particular, determining the part of speech (POS) with higher information is beneficial in text analysis. However, syntax processing is more complicated than semantic analysis, and it was probably for this reason that the reviewed studies did not leverage that aspect. Only one article used POS as a syntax analysis technique to preprocess textual data: Devaraj et al. (2020) marked words with the corresponding POS

Table 12
The best-performing model and reported performance of the reviewed articles.

Ref.	Best model	Measure	Performance
Lima et al. (2015)	SVM	Accuracy	99.8
Khadjeh Nassirtoussi et al. (2015)	SVM	Accuracy	83.33
Linardos et al. (2015)	SVM	F1	45.1
Ding et al. (2015)	EB-CNN	Accuracy	64.21
Kazemian et al. (2016)	SVM	Accuracy	80.164
Feuerriegel and Prendinger (2016)	RF	Return Mean	0.011807
Shynkevich et al. (2016)	SVM	Accuracy	81.31
Chen et al. (2016)	SVM	AUC	80.36
Yoshihara et al. (2016)	RNN	Accuracy	61
Khedr et al. (2017)	KNN	Accuracy	89.8
Vargas et al. (2017)	CNN	Accuracy	64.21
Kaushal and Chaudhary (2017)	SVM	Accuracy	96.04
Jishag et al. (2020)	NB	Accuracy	86.21
Liu et al. (2018)	LSTM	F1	71.33
Chiong et al. (2018)	SVM	Accuracy	59.15
Minh et al. (2018)	TGRU	Accuracy	66.32
Zhang, Qu, et al. (2018)	SVM	F1	62.1
Weng et al. (2018)	BRT	MAPE	1.89
Wang et al. (2018)	ANN	MSE	0.0000372
Du and Zhang (2018)	NB	Accuracy	50.33
Li et al. (2018)	LSTM	Accuracy	88.3
Merello et al. (2018)	LSTM	Accuracy	67
Long et al. (2019)	SVM	Accuracy	69.3
Mohan et al. (2019)	RNN	MAPE	2.03
Vanstone et al. (2019)	ANN	RMSE	0.5489
Attanasio et al. (2019)	MLP	RMSE	1.09
Picasso et al. (2019)	MLP	Accuracy	68
Naderi Semiromi et al. (2020)	XGB	Accuracy	64.3
Vicari and Gaspari (2021)	LSTM	Accuracy	57.94
Mishev et al. (2020)	LSTM	Accuracy	94.7
Sarkar et al. (2020)	LSTM	Price Prediction	–
Devaraj et al. (2020)	LR	Accuracy	76.98
Kulikovskikh and Voronkov (2020)	LSTM	Accuracy	84.3
Li (2020)	LSTM	Price Prediction	–
Ingle and Deshmukh (2021)	LR	Accuracy	85
Eck et al. (2021)	SVM	Accuracy	70
Azizi et al. (2021)	KNN	Accuracy	91.7
Bi et al. (2021)	Ensemble	Accuracy	57
Ranibaran et al. (2021)	RF	MAE	0.16
Omarkhan et al. (2021)	CNN	Accuracy	75
Case and Clements (2021)	BERT	Accuracy	94.92
Sharma et al. (2021)	LSTM	RMSE	0.34
Gite et al. (2021)	LSTM	Accuracy	96.2
Huang (2020)	SVM	Accuracy	65
Sridhar and Sanagavarapu (2021)	BiLSTM	R-squared	27.16
Hu, Wang, et al. (2021)	LSTM	Accuracy	85.02
Nabil and Magdi (2022)	LSTM	RMSE	0.014
Fazlija and Harder (2022)	RF	Accuracy	56.1
Kamal et al. (2022)	BERT	Accuracy	90
Raman et al. (2022)	Decision Forest Regression	RMSE	0.341
Li and Pan (2022)	Ensemble RNN	Accuracy	66.67
Xu et al. (2022)	BiGRU	Accuracy	63.26
Chiong et al. (2022)	Ensemble RNN	Accuracy	59.71
Srivastava et al. (2022)	LSTM	Accuracy	84.6
Ulloa et al. (2022)	MLP	Accuracy	54
Alzazah et al. (2022)	LSTM	Accuracy	80
Daradkeh (2022)	LSTM		82
Rao et al. (2022)	LR	Accuracy	82.5

together with sentiment analysis. They analyzed several combinations of POS in the preprocessing stage and then transformed the text into vectors using TF-IDF weights. They concluded that a mixture of nouns and verbs carries more information about the news articles, providing more context for sentiment analysis.

Stock market prediction can significantly benefit from machine learning and data mining techniques. However, barriers remain in terms of the cleaning, integration, modeling, and analytics required to derive actionable data from diverse data sources. The final purpose of the financial market prediction models is to imitate human behavior to predict the market and maximize profit effectively. The future directions discussed above are in line with this assumed goal: developing an intelligent system to predict stock prices with the aid of informative data sources in real time.

4. Conclusion

With the advent of novel natural language processing, data mining, and machine learning models, intelligent methods for financial market prediction are continually evolving. We conducted a systematic literature review to investigate the emerging state-of-the-art NLP-based stock market prediction techniques using news texts and uncovered gaps and avenues for future research. We identified the characteristics of the numerical and textual datasets, in particular, news text processing and text representation methods, to address four predefined

research questions based on a research protocol as per Kitchenham's methodology. Of 293 initially retrieved articles, 61 studies remained to be synthesized to address the research questions. In terms of the data, integration of four different data types, including textual (news text and news sentiment) and numerical (historical stock prices and technical indicators) data, was identified in the literature for news-based stock market prediction. Additionally, the data source and other characteristics of each data type were fully investigated. We discovered that news headlines were employed extensively for stock market prediction relative to news article bodies for financial market forecasting. Although researchers argued that news headlines are more straight-to-the-point than the whole body of the article, few studies preferred to use news article bodies over headlines. However, performance-wise, integrating the extra information from the news article bodies did not lead to better performance than merely using the news headlines. In particular, the advantage of using both news headlines and bodies over just news headlines was not observed in the reviewed literature.

The most popular machine learning models in terms of frequency are neural networks, support vector machines, naïve Bayes, regression, random forest, decision trees, and k-nearest neighbors. As expected, we uncovered a growing trend toward using artificial neural networks, particularly recurrent neural networks and deep learning models, from 2018 to 2021. Another interesting finding is the potential advantage of the bidirectional variations of recurrent neural networks like BiLSTM and BiGRU for news-based stock market prediction.

The contribution of natural language processing techniques to intelligent stock market prediction is undeniable. However, stock market prediction has still not taken full advantage of NLP techniques. In particular, there is great room for performance enhancement of the predictors by utilizing state-of-the-art text representation methods, that is, transformers like BERT or other contextual language models such as ELMo and XLNet models. Likewise, we encourage future studies to examine authorized credible textual data sources like the management's discussion and analysis (MD&A) section of the companies' quarterly and annual reports.

CRediT authorship contribution statement

Matin N. Ashtiani: Conceptualization, Investigation, Methodology, Writing – original draft, Visualization. **Bijan Raahemi:** Conceptualization, Supervision, Validation, Writing – review & editing.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: Bijan Raahemi reports financial support was provided by Natural Sciences and Engineering Research Council of Canada.

Data availability

No data was used for the research described in the article.

Acknowledgments

This work was supported by the Natural Sciences and Engineering Research Council of Canada (NSERC) Discovery Grant (Nbr RGPIN/341811-2012).

Appendix

Table 11 presents the employed data types and the text representation methods in each reviewed dataset. Furthermore, Table 12 shows the best-performing machine learning model and the corresponding reported performance for each of the studies.

References

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. http://dx.doi.org/10.1016/j.eswa.2017.10.023.

Alzazah, F. S., & Cheng, X. (2020). Recent advances in stock market prediction using text mining: A survey. In *E-business-Higher education and intelligence applications*. IntechOpen.

Alzazah, F., Cheng, X., & Gao, X. (2022). Predict market movements based on the sentiment of financial video news sites. In *2022 IEEE 16th international conference on semantic computing* (pp. 103–110). IEEE.

Anbalagan, T., & Maheswari, S. U. (2015). Classification and prediction of stock market index based on fuzzy metagraph. *Procedia Computer Science*, *47*, 214–221. http://dx.doi.org/10.1016/j.procs.2015.03.200.

Antweiler, W., & Frank, M. Z. (2004). Is all that talk just noise? The information content of internet stock message boards. *The Journal of Finance*, *59*(3), 1259–1294. http://dx.doi.org/10.1111/j.1540-6261.2004.00662.x.

Ashtiani, M. N., & Raahemi, B. (2021). Intelligent fraud detection in financial statements using machine learning and data mining: A systematic literature review. *IEEE Access*, http://dx.doi.org/10.1109/ACCESS.2021.3096799.

Attanasio, G., Cagliero, L., Garza, P., & Baralis, E. (2019). Combining news sentiment and technical analysis to predict stock trend reversal. In *2019 international conference on data mining workshops* (pp. 514–521). IEEE, http://dx.doi.org/10.1109/ICDMW.2019.00079.

Azizi, Z., Abdolvand, N., Ghalibaf Asl, H., & Rajaee Harandi, S. (2021). The impact of Persian news on stock returns through text mining techniques. *Iranian Journal of Management Studies*, *14*(4), 799–816.

Bahdanau, D., Cho, K., & Bengio, Y. (2014). Neural machine translation by jointly learning to align and translate. arXiv preprint arXiv:1409.0473.

Bengio, Y., Ducharme, R., Vincent, P., & Janvin, C. (2003). A neural probabilistic language model. *Journal of Machine Learning Research*, *3*, 1137–1155.

Beyaz, E., Tekiner, F., Zeng, X.-j., & Keane, J. (2018). Comparing technical and fundamental indicators in stock price forecasting. In *2018 IEEE 20th international conference on high performance computing and communications; IEEE 16th international conference on smart city; IEEE 4th international conference on data science and systems* (pp. 1607–1613). IEEE, http://dx.doi.org/10.1109/HPCC/SmartCity/DSS.2018.00262.

Bi, Y., Liu, H., Wang, R., & Li, S. (2021). Predicting stock market movements through daily news headlines sentiment analysis: US stock market. In *2021 2nd international conference on Big Data & artificial intelligence & software engineering* (pp. 642–648). IEEE.

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. http://dx.doi.org/10.1016/j.asoc.2014.01.039.

Boiy, E., & Moens, M.-F. (2009). A machine learning approach to sentiment analysis in multilingual web texts. *Information Retrieval*, *12*(5), 526–558. http://dx.doi.org/10.1007/s10791-008-9070-z.

Boldrini, E., Balahur, A., Martínez-Barco, P., & Montoyo, A. (2012). Using EmotiBlog to annotate and analyse subjectivity in the new textual genres. *Data Mining and Knowledge Discovery*, *25*(3), 603–634. http://dx.doi.org/10.1007/s10618-012-0259-9.

Bollen, J., Mao, H., & Zeng, X. (2011). Twitter mood predicts the stock market. *Journal of Computer Science*, *2*(1), 1–8. http://dx.doi.org/10.1016/j.jocs.2010.12.007.

Case, J., & Clements, A. (2021). The impact of sentiment in the news media on daily and monthly stock market returns. In *Australasian conference on data mining* (pp. 180–195). Springer.

Chatrath, A., Miao, H., Ramchander, S., & Villupuram, S. (2014). Currency jumps, cojumps and the role of macro news. *Journal of International Money and Finance*, *40*, 42–62. http://dx.doi.org/10.1016/j.jimonfin.2013.08.018.

Chaturvedi, I., Ong, Y.-S., Tsang, I. W., Welsch, R. E., & Cambria, E. (2016). Learning word dependencies in text by means of a deep recurrent belief network. *Knowledge-Based Systems*, *108*, 144–154. http://dx.doi.org/10.1016/j.knosys.2016.07.019.

Chen, M.-Y., Fan, M.-H., Chen, T.-H., & Hsieh, R.-P. (2019). Modeling public mood and emotion: Blog and news sentiment and politico-economic phenomena. In A. Visvizi, & M. D. Lytras (Eds.), *Politics and technology in the post-truth era* (pp. 57–71). Emerald Publishing Limited, http://dx.doi.org/10.1108/978-1-78756-983-620191005.

Chen, K., Luo, P., Xu, D., & Wang, H. (2016). The dynamic predictive power of company comparative networks for stock sector performance. *Information & Management*, *53*(8), 1006–1019. http://dx.doi.org/10.1016/j.im.2016.07.005.

Chiong, R., Fan, Z., Hu, Z., Adam, M. T., Lutz, B., & Neumann, D. (2018). A sentiment analysis-based machine learning approach for financial market prediction via news disclosures. In *Proceedings of the genetic and evolutionary computation conference companion* (pp. 278–279). ACM, http://dx.doi.org/10.1145/3205651.3205682.

Chiong, R., Fan, Z., Hu, Z., & Dhakal, S. (2022). A novel ensemble learning approach for stock market prediction based on sentiment analysis and the sliding window method. *IEEE Transactions on Computational Social Systems*.

Cho, K., Van Merriënboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., et al. (2014). Learning phrase representations using RNN encoder-decoder for statistical machine translation. arXiv preprint arXiv:1406.1078.

Daradkeh, M. K. (2022). A hybrid data analytics framework with sentiment convergence and multi-feature fusion for stock trend prediction. *Electronics*, *11*(2), 250.

Dash, R., Dash, P. K., & Bisoi, R. (2014). A self adaptive differential harmony search based optimized extreme learning machine for financial time series prediction. *Swarm and Evolutionary Computation*, *19*, 25–42. http://dx.doi.org/10.1016/j.swevo.2014.07.003.

Dave, K., Lawrence, S., & Pennock, D. M. (2003). Mining the peanut gallery: Opinion extraction and semantic classification of product reviews. In *Proceedings of the 12th international conference on world wide web* (pp. 519–528). http://dx.doi.org/10.1145/775152.775226.

Devaraj, M., et al. (2020). Analyzing news sentiments and their impact on stock market trends using POS and TF-IDF based approach. In *2020 IEEE 2nd international conference on artificial intelligence in engineering and technology* (pp. 1–6). IEEE, http://dx.doi.org/10.1109/IICAIET49801.2020.9257816.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). BERT: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805.

Dharma, E. M., Gaol, F. L., Leslie, H., Warnars, H., & Soewito, B. (2022). The accuracy comparison among word2vec, glove, and fasttext towards convolution neural network (cnn) text classification. *Journal of Theoretical Applied Information Technology*, *100*(2), 31.

Ding, X., Zhang, Y., Liu, T., & Duan, J. (2015). Deep learning for event-driven stock prediction. In *Twenty-fourth international joint conference on artificial intelligence* (pp. 2327–2333).

Du, W., & Zhang, M. (2018). Analysis and prediction about the relationship of foreign exchange market sentiment and exchange rate trend. In *Future of information and communication conference* (pp. 744–749). Springer, http://dx.doi.org/10.1007/978-3-030-03405-4_54.

Eck, M., Germani, J., Sharma, N., Seitz, J., & Ramdasi, P. P. (2021). Prediction of stock market performance based on financial news articles and their classification. In *Data management, analytics and innovation* (pp. 35–44). Springer, http://dx.doi.org/10.1007/978-981-15-5619-7_3.

Fazlija, B., & Harder, P. (2022). Using financial news sentiment for stock price direction prediction. *Mathematics*, *10*(13), 2156.

Feuerriegel, S., & Gordon, J. (2018). Long-term stock index forecasting based on text mining of regulatory disclosures. *Decision Support Systems*, *112*, 88–97. http://dx.doi.org/10.1016/j.dss.2018.06.008.

Feuerriegel, S., & Prendinger, H. (2016). News-based trading strategies. *Decision Support Systems*, *90*, 65–74. http://dx.doi.org/10.1016/j.dss.2016.06.020.

Fung, G. P. C., Yu, J. X., & Lam, W. (2003). Stock prediction: Integrating text mining approach using real-time news. In *2003 IEEE international conference on computational intelligence for financial engineering, 2003. Proceedings* (pp. 395–402). IEEE, http://dx.doi.org/10.1109/CIFER.2003.1196287.

Gerencser, L., Torma, B., & Orlovits, Z. (2009). Fundamental modelling of financial markets. *ERCIM News*, *78*, 16–17.

Gite, S., Khatavkar, H., Kotecha, K., Srivastava, S., Maheshwari, P., & Pandey, N. (2021). Explainable stock prices prediction from financial news articles using sentiment analysis. *PeerJ Computer Science*, *7*, Article e340.

Göçken, M., Özçalıcı, M., Boru, A., & Dosdoğru, A. T. (2016). Integrating metaheuristics and Artificial Neural Networks for improved stock price prediction. *Expert Systems with Applications*, *44*, 320–331. http://dx.doi.org/10.1016/j.eswa.2015.09.029, URL: https://www.sciencedirect.com/science/article/pii/S0957417415006570.

Goldberg, Y., & Levy, O. (2014). word2vec explained: deriving Mikolov et al.'s negative-sampling word-embedding method. arXiv preprint arXiv:1402.3722.

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. http://dx.doi.org/10.1016/j.dss.2013.02.006.

Hamilton, W. L. (2020). Graph representation learning. *Synthesis Lectures on Artifical Intelligence and Machine Learning*, *14*(3), 1–159.

Hatzivassiloglou, V., & McKeown, K. (1997). Predicting the semantic orientation of adjectives. In *35th annual meeting of the Association for Computational Linguistics and eighth conference of the European chapter of the Association for Computational Linguistics* (pp. 174–181).

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 770–778).

Hinton, G. E., Osindero, S., & Teh, Y.-W. (2006). A fast learning algorithm for deep belief nets. *Neural Computation*, *18*(7), 1527–1554.

Hu, M., & Liu, B. (2004). Mining and summarizing customer reviews. In *Proceedings of the tenth ACM SIGKDD international conference on knowledge discovery and data mining* (pp. 168–177). http://dx.doi.org/10.1145/1014052.1014073.

Hu, Z., Wang, Z., Ho, S.-B., & Tan, A.-H. (2021). Stock market trend forecasting based on multiple textual features: A deep learning method. In *2021 IEEE 33rd international conference on tools with artificial intelligence* (pp. 1002–1007). IEEE.

Hu, Z., Zhao, Y., & Khushi, M. (2021). A survey of forex and stock price prediction using deep learning. *Applied System Innovation*, *4*(1), 9.

Huang, X. (2020). An approach to stock price prediction based on news sentiment analysis. In *International conference on intelligent and interactive systems and applications* (pp. 179–185). Springer.

Ingle, V., & Deshmukh, S. (2021). Ensemble deep learning framework for stock market data prediction (EDLF-DP). *Global Transitions Proceedings*, *2*(1), 47–66. http://dx.doi.org/10.1016/j.gltp.2021.01.008.

Jabbarzadeh, A., Shavvalpour, S., Khanjarpanah, H., & Dourvash, D. (2016). A multiple-criteria approach for forecasting stock price direction: nonlinear probability models with application in S&P 500 index. *International Journal of Applied Engineering Research*, *11*(6), 3870–3878.

Jiang, W. (2021). Applications of deep learning in stock market prediction: recent progress. *Expert Systems with Applications*, *184*, Article 115537.

Jin, F., Wang, W., Chakraborty, P., Self, N., Chen, F., & Ramakrishnan, N. (2017). Tracking multiple social media for stock market event prediction. In *Industrial conference on data mining* (pp. 16–30). Springer, http://dx.doi.org/10.1007/978-3-319-62701-4_2.

Jishag, A., Athira, A., Shailaja, M., & Thara, S. (2020). Predicting the stock market behavior using historic data analysis and news sentiment analysis in R. In *First international conference on sustainable technologies for computational intelligence* (pp. 717–728). Springer, http://dx.doi.org/10.1007/978-981-15-0029-9_56.

Kamal, S., Sharma, S., Kumar, V., Alshazly, H., Hussein, H. S., & Martinetz, T. (2022). Trading stocks based on financial news using attention mechanism. *Mathematics*, *10*(12), 2001.

Kaushal, A., & Chaudhary, P. (2017). News and events aware stock price forecasting technique. In *2017 International conference on Big Data, IoT and data science* (pp. 8–13). IEEE.

Kazemian, S., Zhao, S., & Penn, G. (2016). Evaluating sentiment analysis in the context of securities trading. In *Proceedings of the 54th annual meeting of the association for computational linguistics (volume 1: Long papers)* (pp. 2094–2103). http://dx.doi.org/10.18653/v1/P16-1197.

Khadjeh Nassirtoussi, A., 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. http://dx.doi.org/10.1016/j.eswa.2014.06.009.

Khadjeh Nassirtoussi, A., Aghabozorgi, S., Wah, T. Y., & Ngo, D. C. L. (2015). Text mining of news-headlines for FOREX market prediction: A multi-layer dimension reduction algorithm with semantics and sentiment. *Expert Systems with Applications*, *42*(1), 306–324. http://dx.doi.org/10.1016/j.eswa.2014.08.004.

Khedr, A. E., Salama, S. E., & Yaseen, N. (2017). Predicting stock market behavior using data mining technique and news sentiment analysis. *International Journal of Intelligent Systems and Applications*, *9*(7), 22–30. http://dx.doi.org/10.5815/ijisa.2017.07.03.

jae Kim, K., & Han, I. (2000). Genetic algorithms approach to feature discretization in artificial neural networks for the prediction of stock price index. *Expert Systems with Applications*, *19*(2), 125–132. http://dx.doi.org/10.1016/S0957-4174(00)00027-0, URL: https://www.sciencedirect.com/science/article/pii/S0957417400000270.

Kitchenham, B. (2004). *Procedures for performing systematic reviews: Keele University technical report TRSE-0401*, Keele, UK.

Kulikovskikh, G. A., & Voronkov, I. M. (2020). Quotes forecasting method based on news analysis as part of an internet cloud service. In *2020 international scientific and technical conference modern computer network technologies* (pp. 1–5). IEEE, http://dx.doi.org/10.1109/MoNeTeC49726.2020.9257993.

Kumar, B. S., & Ravi, V. (2016). A survey of the applications of text mining in financial domain. *Knowledge-Based Systems*, *114*, 128–147. http://dx.doi.org/10.1016/j.knosys.2016.10.003.

Li, H. (2020). Related research on news sentiment tendency and stock price fluctuation. In *2020 international conference on energy Big Data and low-carbon development management, vol. 214* (p. 03001). EDP Sciences, http://dx.doi.org/10.1051/e3sconf/202021403001.

Li, X., Huang, X., Deng, X., & Zhu, S. (2014). Enhancing quantitative intra-day stock return prediction by integrating both market news and stock prices information. *Neurocomputing*, *142*, 228–238.

Li, Y., Jin, T., Xi, M., Liu, S., & Luo, Z. (2018). Massive text mining for abnormal market trend detection. In *2018 IEEE international conference on Big Data* (pp. 4135–4141). IEEE, http://dx.doi.org/10.1109/BigData.2018.8622450.

Li, Y., & Pan, Y. (2022). A novel ensemble deep learning model for stock prediction based on stock prices and news. *International Journal of Data Science and Analytics*, *13*(2), 139–149.

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. http://dx.doi.org/10.1016/j.ins.2014.03.096.

Li, X., Xie, H., Chen, L., Wang, J., & Deng, X. (2014). News impact on stock price return via sentiment analysis. *Knowledge-Based Systems*, *69*, 14–23. http://dx.doi.org/10.1016/j.knosys.2014.04.022.

Lima, L., Portela, F., Santos, M. F., Abelha, A., & Machado, J. (2015). Big data for stock market by means of mining techniques. In A. Rocha, A. M. Correia, S. Costanzo, & L. P. Reis (Eds.), *New contributions in information systems and technologies* (pp. 679–688). Springer, http://dx.doi.org/10.1007/978-3-319-16486-1_67.

Linardos, E., Kermanidis, K. L., & Maragoudakis, M. (2015). Using financial news articles with minimal linguistic resources to forecast stock behaviour. *International Journal of Data Mining, Modelling and Management*, *7*(3), 185–212.

Liu, C., Hoi, S. C., Zhao, P., & Sun, J. (2016). Online ARIMA algorithms for time series prediction. In *Proceedings of the AAAI conference on artificial intelligence, vol. 30* (pp. 1867–1873).

Liu, P., Qiu, X., & Huang, X. (2016). Recurrent neural network for text classification with multi-task learning. arXiv preprint arXiv:1605.05101.

Liu, Y., Trajkovic, J., Yeh, H.-G. H., & Zhang, W. (2020). Machine learning for predicting stock market movement using news headlines. In *2020 IEEE green energy and smart systems conference* (pp. 1–6). IEEE.

Liu, L., Wu, J., Li, P., & Li, Q. (2015). A social-media-based approach to predicting stock comovement. *Expert Systems with Applications*, *42*(8), 3893–3901. http://dx.doi.org/10.1016/j.eswa.2014.12.049.

Liu, Y., Zeng, Q., Yang, H., & Carrio, A. (2018). Stock price movement prediction from financial news with deep learning and knowledge graph embedding. In *Pacific rim knowledge acquisition workshop* (pp. 102–113). Springer, http://dx.doi.org/10.1007/978-3-319-97289-3_8.

Lo, A. W. (2005). Reconciling efficient markets with behavioral finance: The adaptive markets hypothesis. *Journal of Investment Consulting*, *7*(2), 21–44.

Long, W., Song, L., & Tian, Y. (2019). A new graphic kernel method of stock price trend prediction based on financial news semantic and structural similarity. *Expert Systems with Applications*, *118*, 411–424. http://dx.doi.org/10.1016/j.eswa.2018.10.008.

Ma, C., & Liang, X. (2015). Online mining in unstructured financial information: An empirical study in bulletin news. In *2015 12th international conference on service systems and service management* (pp. 1–6). IEEE, http://dx.doi.org/10.1109/ICSSSM.2015.7170151.

Madeh Piryonesi, S., & El-Diraby, T. E. (2021). Using machine learning to examine impact of type of performance indicator on flexible pavement deterioration modeling. *Journal of Infrastructure Systems*, *27*(2), Article 04021005. http://dx.doi.org/10.1061/(ASCE)IS.1943-555X.0000602.

Maks, I., & Vossen, P. (2012). A lexicon model for deep sentiment analysis and opinion mining applications. *Decision Support Systems*, *53*(4), 680–688. http://dx.doi.org/10.1016/j.dss.2012.05.025.

Malkiel, B. G., & Fama, E. F. (1970). Efficient capital markets: A review of theory and empirical work. *The Journal of Finance*, *25*(2), 383–417. http://dx.doi.org/10.1111/j.1540-6261.1970.tb00518.x.

Medhat, W., Hassan, A., & Korashy, H. (2014). Sentiment analysis algorithms and applications: A survey. *Ain Shams Engineering Journal*, *5*(4), 1093–1113. http://dx.doi.org/10.1016/j.asej.2014.04.011.

Merello, S., Ratto, A. P., Ma, Y., Oneto, L., & Cambria, E. (2018). Investigating timing and impact of news on the stock market. In *2018 IEEE international conference on data mining workshops* (pp. 1348–1354). IEEE, http://dx.doi.org/10.1109/ICDMW.2018.00191.

Meyer, B., Bikdash, M., & Dai, X. (2017). Fine-grained financial news sentiment analysis. In *SoutheastCon 2017* (pp. 1–8). IEEE, http://dx.doi.org/10.1109/SECON.2017.7925378.

Minh, D. L., Sadeghi-Niaraki, A., Huy, H. D., Min, K., & Moon, H. (2018). Deep learning approach for short-term stock trends prediction based on two-stream gated recurrent unit network. *IEEE Access*, *6*, 55392–55404. http://dx.doi.org/10.1109/ACCESS.2018.2868970.

Mishev, K., Gjorgjevikj, A., Vodenska, I., Chitkushev, L. T., & Trajanov, D. (2020). Evaluation of sentiment analysis in finance: From lexicons to transformers. *IEEE Access*, *8*, 131662–131682. http://dx.doi.org/10.1109/ACCESS.2020.3009626.

Mohan, S., Mullapudi, S., Sammeta, S., Vijayvergia, P., & Anastasiu, D. C. (2019). Stock price prediction using news sentiment analysis. In *2019 IEEE fifth international conference on Big Data computing service and applications* (pp. 205–208). IEEE, http://dx.doi.org/10.1109/BigDataService.2019.00035.

Nabil, A., & Magdi, N. (2022). A new model for stock market predication using a three-layer long short-term memory. In *2022 2nd international mobile, intelligent, and ubiquitous computing conference* (pp. 421–424). IEEE.

Nabipour, M., Nayyeri, P., Jabani, H., Mosavi, A., & Salwana, E. (2020). Deep learning for stock market prediction. *Entropy*, *22*(8), 840.

Naderi Semiromi, H., Lessmann, S., & Peters, W. (2020). News will tell: Forecasting foreign exchange rates based on news story events in the economy calendar. *The North American Journal of Economics and Finance*, *52*, Article 101181. http://dx.doi.org/10.1016/j.najef.2020.101181.

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. http://dx.doi.org/10.1016/j.eswa.2015.07.052.

Noda, K., Yamaguchi, Y., Nakadai, K., Okuno, H. G., & Ogata, T. (2015). Audio-visual speech recognition using deep learning. *Applied Intelligence*, *42*(4), 722–737.

O'Hare, N., Davy, M., Bermingham, A., Ferguson, P., Sheridan, P., Gurrin, C., et al. (2009). Topic-dependent sentiment analysis of financial blogs. In *Proceedings of the 1st international CIKM workshop on topic-sentiment analysis for mass opinion* (pp. 9–16). http://dx.doi.org/10.1145/1651461.1651464.

Omarkhan, M., Kissymova, G., & Akhmetov, I. (2021). Handling data imbalance using CNN and LSTM in financial news sentiment analysis. In *2021 16th international conference on electronics computer and computation* (pp. 1–8). IEEE.

Oncharoen, P., & Vateekul, P. (2018). Deep learning for stock market prediction using event embedding and technical indicators. In *2018 5th international conference on advanced informatics: concept theory and applications* (pp. 19–24). IEEE.

Pang, B., & Lee, L. (2008). Opinion mining and sentiment analysis. *Foundations and Trends in Information Retrieval*, *2*(1–2), 1–135. http://dx.doi.org/10.1561/1500000011.

Pennington, J., Socher, R., & Manning, C. D. (2014). Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing* (pp. 1532–1543).

Picasso, A., Merello, S., Ma, Y., Oneto, L., & Cambria, E. (2019). Technical analysis and sentiment embeddings for market trend prediction. *Expert Systems with Applications*, http://dx.doi.org/10.1016/j.eswa.2019.06.014.

Preis, T., Reith, D., & Stanley, H. E. (2010). Complex dynamics of our economic life on different scales: insights from search engine query data. *Philosophical Transactions of the Royal Society of London A (Mathematical and Physical Sciences)*, *368*(1933), 5707–5719.

Pröllochs, N., Feuerriegel, S., & Neumann, D. (2016). Negation scope detection in sentiment analysis: Decision support for news-driven trading. *Decision Support Systems*, *88*, 67–75. http://dx.doi.org/10.1016/j.dss.2016.05.009.

Rachlin, G., Last, M., Alberg, D., & Kandel, A. (2007). ADMIRAL: A data mining based financial trading system. In *2007 IEEE symposium on computational intelligence and data mining* (pp. 720–725). IEEE, http://dx.doi.org/10.1109/CIDM.2007.368947.

Raman, R., Aljafari, R., Venkatesh, V., & Richardson, V. (2022). Mixed-methods research in the age of analytics, an exemplar leveraging sentiments from news articles to predict firm performance. *International Journal of Information Management*, *64*, Article 102451.

Ranibaran, G., Moin, M.-S., Alizadeh, S. H., & Koochari, A. (2021). Analyzing effect of news polarity on stock market prediction: a machine learning approach. In *2021 12th international conference on information and knowledge technology* (pp. 102–106). IEEE.

Rao, J., Ramaraju, V., Smith, J., & Bansal, A. (2022). Agora: Introducing the internet's opinion to traditional stock analysis and prediction. In *2022 IEEE 16th international conference on semantic computing* (pp. 147–150). IEEE.

Sarkar, A., Sahoo, A. K., Sah, S., & Pradhan, C. (2020). LSTMSA: A novel approach for stock market prediction using LSTM and sentiment analysis. In *2020 international conference on computer science, engineering and applications* (pp. 1–6). IEEE, http://dx.doi.org/10.1109/ICCSEA49143.2020.9132928.

Saxena, A., Bhagat, V. V., & Tamang, A. (2021). Stock market trend analysis on Indian financial news headlines with natural language processing. In *2021 Asian conference on innovation in technology* (pp. 1–5). IEEE.

Schumaker, R. P., & Chen, H. (2009). Textual analysis of stock market prediction using breaking financial news: The AZFin text system. *ACM Transactions on Information Systems (TOIS)*, *27*(2), 1–19. http://dx.doi.org/10.1145/1462198.1462204.

Schumaker, R. P., Zhang, Y., Huang, C.-N., & Chen, H. (2012). Evaluating sentiment in financial news articles. *Decision Support Systems*, *53*(3), 458–464. http://dx.doi.org/10.1016/j.dss.2012.03.001.

Schuster, M., & Paliwal, K. K. (1997). Bidirectional recurrent neural networks. *IEEE Transactions on Signal Processing*, *45*(11), 2673–2681.

Seif, M. M., Ramzy Hamed, E. M., & Abdel Ghfar Hegazy, A. E. F. (2018). Stock market real time recommender model using apache spark framework. In *International conference on advanced machine learning technologies and applications* (pp. 671–683). Springer, http://dx.doi.org/10.1007/978-3-319-74690-6_66.

Selvin, S., Vinayakumar, R., Gopalakrishnan, E., Menon, V. K., & Soman, K. (2017). Stock price prediction using LSTM, RNN and CNN-sliding window model. In *2017 international conference on advances in computing, communications and informatics* (pp. 1643–1647). IEEE, http://dx.doi.org/10.1109/ICACCI.2017.8126078.

Shah, D., Isah, H., & Zulkernine, F. (2018). Predicting the effects of news sentiments on the stock market. In *2018 IEEE international conference on Big Data* (pp. 4705–4708). IEEE, http://dx.doi.org/10.1109/BigData.2018.8621884.

Sharma, A., Tiwari, P., Gupta, A., & Garg, P. (2021). Use of LSTM and ARIMAX algorithms to analyze impact of sentiment analysis in stock market prediction. In *Intelligent data communication technologies and internet of things* (pp. 377–394). Springer.

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. http://dx.doi.org/10.1016/j.dss.2016.03.001.

Sridhar, S., & Sanagavarapu, S. (2021). Effect of rate of change of stock prices with news sentiment analysis. In *2021 18th international conference on electrical engineering, computing science and automatic control* (pp. 1–6). IEEE.

Srivastava, S., Tiwari, R., Bhardwaj, R., & Gupta, D. (2022). Stock price prediction using LSTM and news sentiment analysis. In *2022 6th international conference on trends in electronics and informatics* (pp. 1660–1663). IEEE.

Thakkar, A., & Chaudhari, K. (2021). A comprehensive survey on deep neural networks for stock market: The need, challenges, and future directions. *Expert Systems with Applications*, *177*, Article 114800.

Tsai, C.-F., & Hsiao, Y.-C. (2010). Combining multiple feature selection methods for stock prediction: Union, intersection, and multi-intersection approaches. *Decision Support Systems*, *50*(1), 258–269. http://dx.doi.org/10.1016/j.dss.2010.08.028, URL: https://www.sciencedirect.com/science/article/pii/S0167923610001521.

Ulloa, A., Espezua, S., Villavicencio, J., Miranda, O., & Villanueva, E. (2022). Predicting daily trends in the lima stock exchange general index using economic indicators and financial news sentiments. In *Annual international conference on information management and Big Data* (pp. 34–49). Springer.

Vanstone, B. J., Gepp, A., & Harris, G. (2019). Do news and sentiment play a role in stock price prediction? *Applied Intelligence*, *49*(11), 3815–3820. http://dx.doi.org/10.1007/s10489-019-01458-9.

Vargas, M. R., de Lima, B. S., & Evsukoff, A. G. (2017). 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* (pp. 60–65). IEEE, http://dx.doi.org/10.1109/CIVEMSA.2017.7995302.

Vicari, M., & Gaspari, M. (2021). Analysis of news sentiments using natural language processing and deep learning. *AI & Society*, *36*(3), 931–937. http://dx.doi.org/10.1007/s00146-020-01111-x.

Wang, Z., Ho, S.-B., & Lin, Z. (2018). Stock market prediction analysis by incorporating social and news opinion and sentiment. In *2018 IEEE international conference on data mining workshops* (pp. 1375–1380). IEEE, http://dx.doi.org/10.1109/ICDMW.2018.00195.

Wang, S., Li, D., Song, X., Wei, Y., & Li, H. (2011). A feature selection method based on improved Fisher's discriminant ratio for text sentiment classification. *Expert Systems with Applications*, *38*(7), 8696–8702. http://dx.doi.org/10.1016/j.eswa.2011.01.077.

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. http://dx.doi.org/10.1016/j.eswa.2011.04.127.

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. http://dx.doi.org/10.1016/j.eswa.2018.06.016.

Wiebe, J., Wilson, T., & Cardie, C. (2005). Annotating expressions of opinions and emotions in language. *Language Resources and Evaluation*, *39*(2–3), 165–210. http://dx.doi.org/10.1007/s10579-005-7880-9.

Wilson, T., Wiebe, J., & Hoffmann, P. (2005). Recognizing contextual polarity in phrase-level sentiment analysis. In *Proceedings of human language technology conference and conference on empirical methods in natural language processing* (pp. 347–354). http://dx.doi.org/10.3115/1220575.1220619.

Wong, D. P. R. K. (2002). Currency exchange rate forecasting from news headlines. *Australian Computer Science Communications*, *24*(2), 131–139.

Wuthrich, B., Cho, V., Leung, S., Permunetilleke, D., Sankaran, K., & Zhang, J. (1998). Daily stock market forecast from textual web data. In *SMC'98 conference proceedings. 1998 IEEE international conference on systems, man, and cybernetics, vol. 3* (pp. 2720–2725). IEEE, http://dx.doi.org/10.1109/ICSMC.1998.725072.

Xing, F. Z., Cambria, E., & Welsch, R. E. (2018). Natural language based financial forecasting: a survey. *Artificial Intelligence Review*, *50*(1), 49–73. http://dx.doi.org/10.1007/s10462-017-9588-9.

Xu, H., Chai, L., Luo, Z., & Li, S. (2022). Stock movement prediction via gated recurrent unit network based on reinforcement learning with incorporated attention mechanisms. *Neurocomputing*, *467*, 214–228.

Xu, Y., & Cohen, S. B. (2018). Stock movement prediction from tweets and historical prices. In *Proceedings of the 56th annual meeting of the Association for Computational Linguistics (volume 1: Long papers)* (pp. 1970–1979). http://dx.doi.org/10.18653/v1/P18-1183.

Yang, Z., Dai, Z., Yang, Y., Carbonell, J., Salakhutdinov, R. R., & Le, Q. V. (2019). Xlnet: Generalized autoregressive pretraining for language understanding. *Advances in Neural Information Processing Systems*, *32*.

Yoshihara, A., Seki, K., & Uehara, K. (2016). Leveraging temporal properties of news events for stock market prediction. *Journal of Artificial Intelligence Research*, *5*(1), 103–110.

Zhang, J., Cui, S., Xu, Y., Li, Q., & Li, T. (2018). A novel data-driven stock price trend prediction system. *Expert Systems with Applications*, *97*, 60–69. http://dx.doi.org/10.1016/j.eswa.2017.12.026.

Zhang, X., Fuehres, H., & Gloor, P. A. (2011). Predicting stock market indicators through Twitter "I hope it is not as bad as I fear". *Procedia-Social and Behavioral Sciences*, *26*, 55–62. http://dx.doi.org/10.1016/j.sbspro.2011.10.562.

Zhang, X., Qu, S., Huang, J., Fang, B., & Yu, P. (2018). Stock market prediction via multi-source multiple instance learning. *IEEE Access*, *6*, 50720–50728. http://dx.doi.org/10.1109/ACCESS.2018.2869735.

Zhang, Y., & Wallace, B. (2015). A sensitivity analysis of (and practitioners' guide to) convolutional neural networks for sentence classification. arXiv preprint arXiv:1510.03820.

Zhao, X., Jiang, J., Yan, H., & Li, X. (2010). Jointly modeling aspects and opinions with a MaxEnt-LDA hybrid. In *Proceedings of the conference on empirical methods in natural language processing*. ACL.