Iranian Journal of Management Studies (IJMS) 2021, 14(4): 799-816
RESEARCH PAPER

The Impact of Persian News on Stock Returns Through Text Mining Techniques

Zahra Azizi, Neda Abdolvand*, Hassan Ghalibaf Asl, Saeedeh Rajaee Harandi

Department of Management, Faculty of Social Sciences and Economics, Alzahra University, Tehran, Iran

(Received: January 7, 2020; Revised: March 20, 2021; Accepted: April 3, 2021)

Abstract
The news contains information about the fundamentals of the company and can change the behavior of the stock market. However, most research in stock market prediction has relied on technical analysis, i.e., time series analysis, based on past stock data, and the impact of fundamental data – especially Persian news – on the stock prices has been neglected. Consequently, this study aimed to fill this gap. To this aim, the stock index values were collected from the Tehran Stock Exchange along with the news published during this period. Then, the semantic load of news sentences was determined using text mining and sentiments analysis techniques, and the news was classified into positive and negative categories using machine-learning algorithms. Finally, the relationship between news and stock index was evaluated using logistic regression. According to the results, published news has a positive or negative semantic burden, and is effective on the index value.

Keywords: Stock market index, Stock market prediction, Persian news, Text mining, Sentiment analysis, Technical and fundamental data.

Introduction

Stock markets considerably influence industries and individuals that ultimately affect the global economic growth (Rioja & Valev, 2014). The nature of the stock market has changed significantly due to the emergence of financial corporations and institutions, financial products, and government regulations on these products (De Fortuny et al., 2014; Mate et al. 2020). The stock market prediction has always been an attractive topic due to its vitality in the economic and financial sectors (Jishag et al., 2020). However, stock market trends are related to investors' investment behaviors (Meesad & Li, 2014). Besides, stock market behavior is influenced by several factors such as political situation, economic conditions, future corporate goals, investor expectations, the global stock exchange, and news (Hong, 2020). The dissemination of information from various media can also have an impact on stock prices and market behavior (Gunduz & Cataltepe, 2015; Sharma et al., 2017). Therefore, decision making in the stock market is challenging, due to the complexity and dynamic characteristics of the data it deals with. It depends heavily on what information is available to investors (Jishag et al., 2020).

To predict prices in the stock market, people are looking for methods and tools that reduce risk while increasing profits; therefore, forecasting plays an important role in stock market

* Corresponding Author Email: n.abdolvand@alzahra.ac.ir

business (Sharma et al., 2017). The vast volume of data generated by the stock market is considered to be a treasure trove of knowledge for investors. However, stock market prediction is a challenging task, even for the brightest and sharpest minds in the business. Predicting the stock market is not an easy task because of the complexity and dynamic features of the data it deals with (Jishag et al., 2020). News articles are among the sources of information that can modify expectations about a company's cash flow or investor's discount rates and can affect stock returns (Hagenau et al., 2013; Jishag et al., 2020). For this reason, information derived from news has recently become part of financial prediction systems (Shynkevich et al., 2016). Information about company principles, the activities in which a company is involved, and the expectations of other market participants are included in news articles (Hagenau et al., 2013; Li, Xie, et al., 2014; Li, Wang, et al., 2014; Shynkevich et al., 2015). However, the importance of different news topics and information sources is not the same in the eyes of the audience, and even differences in behavior and decision-making policies will lead to different reactions to the same news. Therefore, financial analysts are unable to access the entirety of the news about a set of stocks at the moment of their release (Zhang et al., 2018). Since disclosing the behavior of stock market data is essential for investors to avoid future investment risks (Jishag et al., 2020), it is necessary to analyze a large amount of textual data relevant to a particular stock, extract relevant information, and use it in financial prediction (Jishag et al., 2020; Shynkevich et al., 2015). With accurate and successful news analysis, investors can act based on the predictions and gain more profits (Hatefi Ghahfarrokhi & Shamsfard, 2020). Thus, a system that can simultaneously employ text mining techniques for rapid content analysis of news and economic techniques to predict fluctuations in financial stocks can help analysts and investors to predict the stock market behavior (Hagenau et al., 2013; Nizer & Nievola, 2012). Automatic text news classification involves text mining techniques that convert unstructured information into structured and machine-readable formats. It primarily uses machine learning techniques to classify information (Hagenau et al., 2013; Kalra & Agrawal, 2019; Tobback et al., 2018). Besides, polar messages reflect the investors' feeling about stocks. Therefore, sentiment analysis would help identify these kinds of messages and find their polarity (Meesad & Li, 2014).

Therefore, prediction based on news articles has received a great attention, and various studies have examined different news datasets, forecast indexes/markets, natural language processing, and forecasting algorithms (e.g., Feuerriegel & Gordon, 2018; Eck et al., 2020; Li et al., 2014; Meesad & Li, 2014; Seker et al., 2014; Thanh & Meesad, 2014; Xie & Jiang, 2019). Some of them have focused fundamentally on technical analysis, such as time series analysis, which is based on past stock data to predict price differentials with different phases (e.g., Alanyali et al., 2013; Thanh & Meesad, 2014), while others have turned their attention to the methods to improve the accuracy of the prediction based on sentiment analysis of news related to stock trends (e.g., Mate et al., 2020; Xie & Jiang, 2019).

Besides, it has been proven that the emotional aspects of news can affect the market, which is reflected in trade volume and returns (Yekrangi & Abdolvand, 2020). In order to investigate such emotional aspects, different emotion analysis algorithms aim to predict future market movements (Hagenau et al., 2013; Schumaker et al., 2012) and market return (Tabari et al., 2018). However, no satisfactory theoretical and technical framework has been developed for predicting financial markets using the combined approaches of both fundamental and technical data (Kumar & Ravi, 2016; Nassirtoussi et al., 2014).

Previous studies have indicated a strong correlation between the fluctuation of stock prices and the publication of stock-related news (Fung et al., 2005; Shynkevich et al., 2015). However, these studies have been in languages other than Persian. Recently, the use of Persian language by Internet users and the production of electronic news in this language have

increased significantly (Usage of content languages for websites, 2019). Persian content ranking is now among top 10 languages on the Internet, and the number of pages including Persian content on the Internet has reached the 8th rank in 2019 (W3Techs, 2019). This growth demonstrates the popularity and expanding significance of this language on the World Wide Web. Moreover, as indicated by W3Techs website, Persian language has the third rank in fastest growing content languages on the Internet. In addition, despite sanctions, foreign investors are active in the Tehran stock market and recently, their desire to participate in the stock market and production sectors of Iran has increased. Therefore, the question that arises is that if text mining techniques can be useful in predicting the impact of the fundamental data of Persian news on stock prices changes.

Despite the impact of news on the stock returns as one of the fundamental factors effective on stock market, no study has examined the impact of Persian political and economic news on the returns of total stock index using text mining and sentiment analysis techniques. In fact, the study of Moazeni et al. (2014) is the only research on the impact of news on the stock exchange index in Iran. Accordingly, this paper attempts to bridge this gap by combining two different components: sentiment analysis on stock-related Persian news reports and historical data analysis.

The study is organized as follows. First, the literature is reviewed. Then, the research method and results of data analysis are explained. Finally, conclusions and suggestions for future research are explained.

Literature Review

Stock market prediction is a way to understand the future fluctuations of a company's stock price (Jishag et al., 2020). Generally, two approaches are used to predict financial markets: the technical approach and the fundamental approach (Picasso et al., 2019). These approaches differ in their input data. Historical market data (such as the stock prices for a day, a week, and a month ago) are commonly used for technical analysis. Any other type of data from the country's economic structure, information, and news about the country, society, and company (e.g., inflation rates, trading volume, unemployment rates, and demand for a company's products) are used for fundamental analysis (Jishag et al., 2020). Fundamental data may also be extracted from numerical and structured sources, such as macroeconomic data, or regular financial reports of banks and government agencies (Cavalcante et al., 2016; Gunduz & Cataltepe, 2015; Li, Xie, et al., 2014; Nassirtoussi et al., 2014; Picasso et al., 2019; Tsai et al., 2011).

News contains information about the firm's fundamental principles and can change the behavior of the stock market (Li, Xie, et al., 2014; Li, Wang, et al., 2014). In addition, one of the factors that influence marketers' decisions is the news and its sentiments. Based on the Efficient-Market Hypothesis (EHM), market efficiency relies on the timely delivery of market information to investors and their timely response to this information (Fama, 1965). Emotions can also influence an individual's investment behavior and decisions (Zaleskiewicz & Traczyk, 2020). Since the structure of a language influences the way people convey their words to others, it is important to address this issue by using text mining and sentiment analysis techniques (Groth & Muntermann, 2011). Sentiment analysis is a new part of textual data analysis that combines language processing techniques and computational linguistics to identify and extract subjective terms and opinions from documents (Meesad & Li, 2014).Text mining is the process of extracting quality information from text documents using data mining techniques, statistics, information retrieval, machine learning, and computational linguistics (Pejić Bach et al., 2019; Tobback et al., 2018). Text mining converts unstructured information into machine-readable formats and mainly uses machine learning techniques to classify

information (Hagenau et al., 2013; Kalra & Agrawal, 2019; Pejić Bach et al., 2019), with one of its uses in the financial field being the stock market prediction (Kumar & Ravi, 2016). The application of text mining in financial forecasting includes FOREX rate prediction, stock market prediction, and hybrid prediction (Kumar & Ravi, 2016). Therefore, some researchers have used text mining mechanisms to analyze and classify financial news according to their content to help predict the future behavior of financial assets (Cavalcante et al., 2016). For example, Jishag et al. (2020) combined two components of sentiment analysis on news related to the stock market and historical data analysis to predict the trends in the stock market. Their proposed model had better prediction accuracy compared to other models.

The study by Mate et al. (2020) used sentiment analysis on stock market news to predict the changes in stock indices. They also used the output of sentiment analysis in machine learning algorithms to analyze the stock prices.

In another study, Hong (2020) proposed a prediction system based on the text mining method and stock market news. To respond to real-time stock market changes, it used LSTM and YTextMiner AI to reflect stock news in real time, and based on past-time series analysis data, found the closest situation to the time when the stock price had risen by mathematical calculation.

Eck et al. (2020) used financial news to predict the stock market performance. Their results indicated that support vector machines can deliver better results than other algorithms in predicting the stock market performance.

Lutz et al. (2020) developed a new machine learning approach to predict the sentence-level polarity labels in financial news. Their method used distributed text representations and multi-instance learning to transfer information from the document-level to the sentence-level. The proposed expert system could assist investors in their decision making and might help them in communicating their messages as intended.

In a similar study, Xie and Jiang (2019) used text mining and sentiment analysis in Chinese online financial news to predict the price trend of Chinese stocks and the stock price based on the support vector machine (SVM) algorithm. Their results indicated that the quality of news and the number of audiences have a significant effect on the source impact factor. In addition, for Chinese investors, traditional media has more influence than digital media.

In another study, Feuerriegel and Gordon (2018) used text mining and sentiment analysis techniques to study the impact of financial news on long-term stock market trends. They indicated the better performance of text-based models in predicting the stock trends as well as reducing the prediction errors.

Narayan and Bannigidadmath (2017) used a time series method to evaluate the impact of financial news on Islamic stock returns compared to non-Islamic (conventional) stocks. Their results indicated that financial news is effective on the prediction of some stocks. They also confirmed the better impact of positive words on both types of stock returns. Besides, they indicated that investing in Islamic stocks is more profitable than investing in conventional stock.

Weng et al. (2017) also presented an increasingly efficient smart business system by combining diverse online resources with temporal data and stock technical indicators. They used machine learning, decision tree, neural networks, and support vector machines as the basis of the inference engine. They also used AAPL (NASDAQ Apple) shares to evaluate the performance of their proposed system and indicated that diversifying the knowledge base by combining data from different sources can help improve the performance of specialized financial systems. Besides, the combination of online data sources with traditional technical indicators was found to provide greater predictive power than any of those sources alone. In another study, Shynkevich et al. (2016) examined the impact of financial news on stock prices

to improve the financial prediction and to support investors and traders in the decision-making process. Their results indicated that the performance of prediction systems improves with the increasing number of stock-related newsgroups.

Gunduz and Cataltepe (2015) used the analysis of news articles and stock prices to predict future market movements. In their study, Turkish-language text mining techniques were used to convert news articles into feature vectors. The balanced mutual information (BMI) method was used to identify features more relevant to determining the market orientation. The Bayesian Navigation algorithm was also used to model feature vectors and stock prices as well as to predict future market movements.

Nassirtoussi et al. (2014) predicted the FOREX market using news headlines. They used a multilayer model to perform the semantic analysis of news phrases and the analysis of investors' feelings about the market, and to reduce the dimensions of the extracted attributes. Their results indicated that there is a relationship between news headlines, stock movements, and the FOREX market.

Thanh and Meesad (2014) used time series data analysis and text mining techniques to predict stock market trends. They also used the combination of Linear Support Vector Machine Weight and Support Vector Machine algorithms to increase the accuracy of prediction. Their results indicated that one-against-one method outperforms one-against-all method and its accuracy is higher.

In another study, Li, Wang, et al. (2014) examined the impact of media on the stock market by weighing news related to companies listed in the Chinese stock index. Their results indicated that investor exchange activities are influenced by fundamental information on the news.

The study of Seker et al. (2014) was based on one of the most widely published newspapers in Turkey that have special pages for economic news. In this study, SVM and the nearest neighbor to k algorithms were applied. The authors concluded that analyzing time series and examining its relationship with economic news would help understand the financial market power in Turkey.

Combining words and nouns instead of single words and using the SVM classifier, Hagenau et al. (2013) predicted stock prices based on contextual information in financial news. They indicated that using combinations of words and noun phrases improves models' accuracy.

Likewise, Nizer and Nievola (2012) used text mining techniques and the GARCH model to predict fluctuations in the stock market. They analyzed a model based on the content of Portuguese news about their companies' stocks and their impact on the Brazilian stock market.

Schumaker and Chen (2009) studied the impact of breaking financial news on stock market prediction using several different textual representations, including Bag of Words, Noun Phrases, and Named Entities. They indicated the better performance of Noun scheme in stock market prediction.

As our brief review here clarifies, most research in stock market prediction has relied on technical analysis, i.e., time series analysis based on the past stock data. In addition, despite the use of hybrid approaches derived from both fundamental and technical data, there is no good theoretical and technical framework for predicting financial markets based on the best available knowledge (Kumar & Ravi, 2016; Nassirtoussi et al., 2014). In addition, studies show that no research has been conducted in Iran so far about the impact of political and economic news on stock index returns through text mining techniques, especially on Persian news. Therefore, in this study, the effect of Persian news on total stock indexes was investigated using text mining and sentiment analysis techniques. For this purpose, the stock

index values were first collected from the Tehran Stock Exchange and the news published during the related period was collected from a news database. Then, the semantic load of news sentences was determined using text mining and sentiments analysis techniques, and the news was classified into positive and negative categories using machine learning algorithms.

Methodology

This study aimed to investigate the impact of news Tehran Stock Exchange trend using text mining approaches. Similar to the studies of Jishag et al. (2020), Xie and Jiang (2019), Feuerriegel and Gordon (2018), Ritesh et al., (2017), Shynkevich et al., (2015), Gunduz and Cataltepe (2015), Meesad and Li (2014), Li, Wang, et al. (2014), Seker et al. (2014), Hagenau et al. (2013), Schumaker et al. (2012), and Schumaker and Chen (2009), this study consisted of two parts, namely sentiment analysis and historical data analysis. To this end, first the semantic loads of news headlines were analyzed using text mining and sentiment analysis techniques. In this study the impact of only one news source was used. The advantage of choosing a particular web domain is that all articles have the same structure. This makes the process of data clearing much easier than extracting data from different sources (Ritesh et al., 2017). In addition, only the impact of news headlines on the overall index was examined, because using news headlines instead of news text and focusing on one type of news instead of different types of news reduces data confusion (Nassirtoussi et al., 2014). After news preparation, news were preprocessed through the tokenization, removing prepositions, removing stop words, and stemming to convert them into a structured form. These steps were widely used in previous studies such as Feuerriegel and Gordon (2018), Meesad and Li (2014), and Nizer and Nievola (2012). Next, the output file from the preprocessing phase was used to determine the semantic load of the words and finally to label the news headlines. This was similar to the studies of Jishag et al. (2020), Xie and Jiang (2019), Feuerriegel and Gordon (2018), Ritesh et al., (2017), Shynkevich et al., (2015), Gunduz and Cataltepe (2015), Meesad and Li (2014), Li, Wang, et al. (2014), Seker et al. (2014), Hagenau et al. (2013), Schumaker et al. (2012), and Schumaker and Chen (2009). Then, the method of the number of repetitions of words in the sentence that had been used in the study of Raschka and Mirjalili (2017) was used for text representation, and the data was prepared for modeling.

After preparing the data, the news was classified into positive and negative categories using machine learning algorithms including Support Vector Machine (SVM), Naive Bayesian (NB), and K-Nearest Neighbor (KNN), similar to the method previously adopted by Meesad and Li (2014), Hagenau et al. (2013), Jishag et al. (2020), and Groth and Muntermann (2011) to identify patterns in the textual data. Then, the logistic regression was used to evaluate the relationship between news and stock index and its accuracy (Huang & Liu, 2020). Logistic regression is the most widely used type of regression in the industry, and is used to find the probability of successful and unsuccessful events (Dutta et al., 2012).

In order to evaluate the results of each algorithm, the confusion matrix and its derived criteria including accuracy, precision, recall, and F-measure were used (Meesad & Li, 2014). The confusion matrix has been widely used to evaluate text mining and sentiment analysis approaches (Nassirtoussi et al., 2014).

The Python programming language and its libraries were used to implement the research phases. Python is an interpretive, high-level, object-oriented, and open-source language that can be used to solve and implement many problems, including data science issues (Terra, 2021). Figure 1 indicates the research framework of the study.

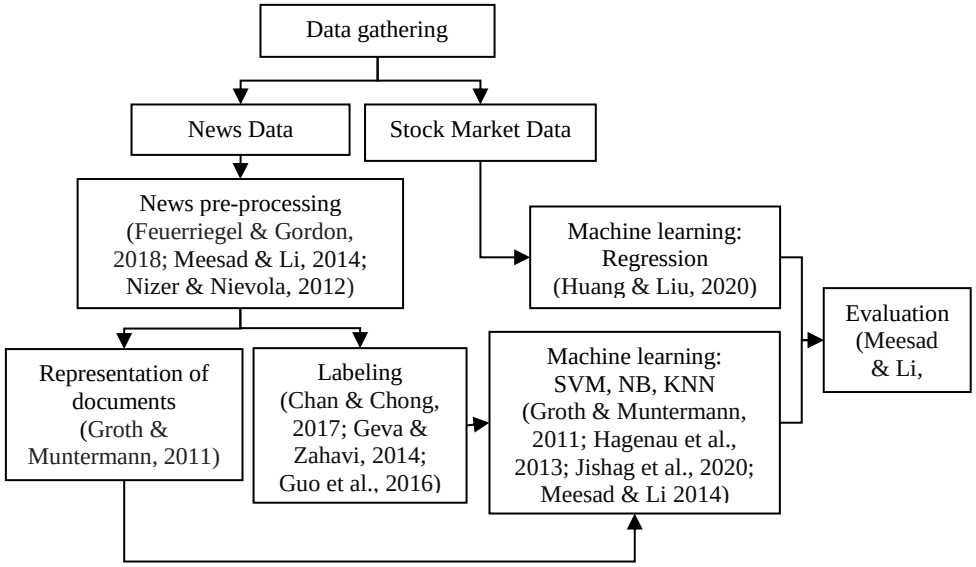

Fig. 1. Research Framework

Data Analysis

News Preparation and Preprocessing

In this part, the news was analyzed and classified as either positive or negative. To do so, several data preprocessing methods were performed on news, which are explained in the following lines.

This research focused on the prediction of Tehran's stock index over a six month period from September 2017 to April 2018. The moment stock index data were obtained from the Tehran Stock Exchange, including the index value and the time of its registration. In addition, news headlines from the Economics website (https://www.eghtesadonline.com), including the date and time of the news release, were collected using crawling methods in the Python programming environment. For this purpose, the Scrapy library and the parsing method were connected to the site server and according to the CSS selectors, the news headlines and their publication time were collected. The Scrapy module is an open-source module for extracting information from websites through web crawling. It can also extract data through web services programming interfaces (Scrapy, 2019). The extracted news data contained 67,000 news headlines along with the date and time of their publication.

As text is unstructured data, preprocessing is required to convert unstructured text data into a structured form (Meesad & Li, 2014). This was done through the following steps (adapted from Feuerriegel & Gordon, 2018).

Tokenization: Since the main data were text type, at the preprocessing phase, the news headlines were first broken into their constituent tokens through the Hazm library and the word tokenize method. The Hazm library was used for processing the Persian language in Python. This module has functions for text processing, classification, tokenization, rooting, tagging, parsing, and semantic reasoning (Hazm, 2019).

Removal of Prepositions: in this step, using regular expression and sub-method as well as string function and punctuation method, punctuation marks such as semicolons, question marks, exclamation marks, etc., were cleared from the document.

Removal of stop words: stop words are used frequently in a monolingual language and usually have no significant meaning (Meesad & Li, 2014; Nizer & Nievola, 2012). Therefore, these words were removed before further processing to be able to do more classification tasks.

To this end, the stop words file in the Hazm library was used. Finally, the roots of the verbs as well as bin words were identified using the Hazm library and the *st* function.

Stemming: some words may have the same meaning, but only the feature is different. In machine learning, it is better not to have similar features (Meesad & Li, 2014). Therefore, stemming the tokens to their root type was done.

Table 1 indicates the examples of the text preprocessing.

Table 1. Examples of Text Pre-Processing

News header before pre-processing	Pre-Processing			
	Tokenization	**Remove prepositions**	**Remove stop words**	**Stemming:**
ترامپ، با تـأخیر پیـروزی پـوتین را تبریـک گفت. Trump, said congratulation to Putin on his victory with delay.	ترامپ، با تـاخیر پیـروزی پوتین را تبریک گفت. Trump, said congratulation to Putin on His victory with delay.	ترامپ با تاخیر پیروزی پوتین را تبریک گفت Trump Said congratulation Putin on his victory with delay	ترامپ تاخیر پیروزی پـوتین تبریک گفتن Trump Say Congratulation Putin Victory Delay	ترامـپ تـاخیر پیـروز پوتین تبریک گفتن Trump Say Congratulation Putin Victorious Delay
فعالیت رییس سـازمان برنامـه و بودجـه، تـا لحظات پایانی سال+عکس Activities of the head of Planning and Budget Organization, until the end of year + photo	فعایـت ریـیس سـازمان برنامه و بودجه، تا لحظات پایانی سال + عکس Activities of The Head of Planning and Budget Organization, until The End of Year + photo	فعالیت رییس سـازمان برنامـه و بودجـه تـا لحظـات پایـانی سال عکس Activities head Planning and Budget Organization Year photo	فعالیت رییس سازمان برنامـه بودجه عکس Activities head Planning Budget Organization Year photo	فعال ریـیس سـازمان برنامه بودجه عکس Active head Planning Budget Organization Year photo
ایران در اعتراض به اظهارات اردوغـان، سـفیر ترکیه را احضار کرد. Iran summons Turkish ambassador over Erdogan's remarks.	ایـران در اعتـراض بـه اظهارات اردوغـان، سـفیر ترکیه را احضار کرد. Iran Summons Turkish Ambassador over Erdogan’s Remarks.	ایران در اعتراض بـه اظهـارات اردوغان سفیر ترکیه را احضـار کرد Iran Summons Turkish Ambassador over Erdogan Remarks	ایـران اعتـراض اظهـارات اردوغان سفیر ترکیـه احضـار کرد Iran summons Turkish ambassador Erdogan remarks	ایـران اعتـراض اظهـار اردوغان سـفیر ترکیـه احضار کرد Iran summon Turkish ambassador Erdogan remark

Sentiment Analysis and Data Labeling

In this step, the output file from the preprocessing phase was used to determine the semantic load of the words and finally to label the news headlines, similar to the studies of Jishag et al. (2020), Xie and Jiang (2019), Feuerriegel and Gordon (2018), Ritesh et al. (2017), Shynkevich et al. (2015), Gunduz and Cataltepe (2015), Meesad and Li (2014), Li, Wang, et al. (2014), Seker et al. (2014), Hagenau et al. (2013), Schumaker et al. (2012), and Schumaker and Chen (2009). For this purpose, each of the extracted tokens from the previous step was first assigned a semantic load in the numerical range from -1 to +1 by the polyglot library. Polyglot has a polarization lexicon for 136 languages. The polarity of words consists of three degrees: +1 for positive words, -1 for negative words, and zero for neutral words. The polarity package was used to check the polarity of a word (Polyglot, 2019). Then, the semantic load of each sentence was determined using the sum of the semantic load of the constituent tokens (Chan & Chong, 2017; Geva & Zahavi, 2014; Guo et al., 2016).

To this end, first the sentences were split to smaller units (words) and then the polarity of each word was determined. The labels 1, 0, and -1 were assigned to positive, neutral and

negative sentiments respectively. If in a news headline the number of tokens with negative semantic load of -1 be more than the number of tokens with semantic load of +1, the news headline's semantic load is equal to the total negative semantic load. This way, the semantic load of news headlines was determined, and a 0 or 1 label was assigned to each sentence. Label 1 indicates a positive semantic load and label 0 indicates a negative semantic load. If the semantic load of a sentence equals zero, the corresponding sentence will be deleted at this stage (Chan & Chong, 2017; Guo et al., 2016).

The semantic load of the financial words was also examined. Thus, if the news headline contained a word with a positive (or negative) financial load, this financial load would be added (or subtracted) to the calculated result, and the final label of sentences could be calculated by considering these words. Figure 2 indicates the polarity distribution between classes, which states the positive and negative values.

For example, the polarity of the sentence "ترامپ ، با تاخیر پیروزی پوتین را تبریک گفت" (Trump, said congratulation to Putin on his victory with delay.) is indicated in table in Table 2, and its sentiment analysis pseudo code is as follows:

```
from __future__ import unicode_literals
from hazm import POSTagger, Chunker
from hazm import word_tokenize
from polyglot.text import Text
tagger = POSTagger(model = '/home/mrrobot/Documents/resources/postagger.model')
tagged = tagger.tag(word_tokenize('ترامپ  تاخیر  پیروز  پوتین  تبریک  گفت'))
print(tagged)
[('ترامپ', 'N'), ('تاخیر', 'Ne'), ('پیروز', 'AJe'), ('پوتین', 'N'), ('تبریک', 'N'), ('گفت', 'V')]
text = Text('ترامپ  تاخیر  پیروز  پوتین  تبریک  گفت')
print("{:<16}{}".format("Word", "Polarity")+"\n"+"-"*30)
for w in text.words:
print("{:<16}{:>2}".format(w, w.polarity))
```

Table 2. The Polarity of the Example Sentence

Word	Polarity
ترامپ (Trump)	0
تاخیر (delay)	-1
پیروز (Victorious)	1
پوتین (Putin)	0
تبریک (congratulation)	1
گفتن (say)	0

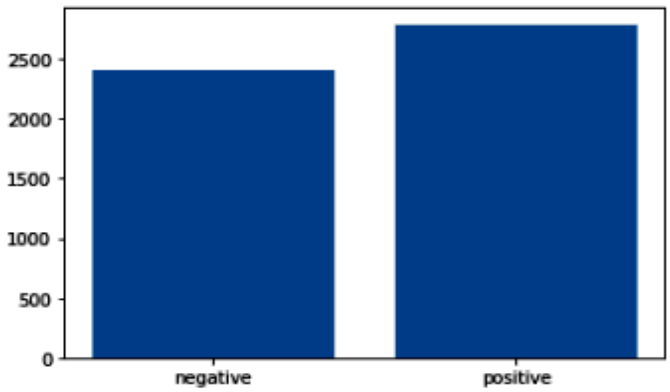

Fig. 2. Polarity Distribution between Classes

II. Historical Data Analysis

Textual Representation

In this step, news articles were represented in machine-friendly form. To this end, the stock trend was processed and a graph predicting the upcoming market trend was modeled. To do so, the method of the number of repetitions of words in the sentence that has been used in the study of Raschka and Mirjalili (2017) was used for text representation. In this method, an integer was assigned to the words with lower repetitions. Moreover, the zero value was added to the beginning or end of the words to make their length equal to the classification algorithm (Raschka & Mirjalili, 2017). In this technique, harmony of words was completely ignored (Shynkevich et al., 2015). Finally, the data were prepared for modeling.

Modeling and Model Evaluation

At this phase, common machine learning algorithms, such as SVM, KNN, and NB, were used to categorize news into positive and negative categories. To do this, the stock price reactions caused by news events were transformed to a binary measure, i.e., 0 for negative price effect and 1 for positive price effect (Hagenau et al., 2013). According to the previous studies, SVM is the best machine learning method for text classification tasks (Forman, 2003; Hagenau et al., 2013). When using SVM to financial forecast, the most important thing to consider is the choice of kernel function. The kernel function is a linear separator based on the internal multiplication of vectors. Since the dynamics of the financial time series are highly nonlinear, using nonlinear kernel functions can perform better than linear kernels. In this study, in the SVM classification algorithm, the Radial Basis Function (RBF) kernel was used. RBF is a real value function whose value depends only on the distance from the source (Sharma et al., 2017; Weng et al., 2017). In addition, the polynomial NB classifier was used because it performs better when the text is represented by the number of words occurring and classifying them accordingly (Kavuluru et al., 2013). Although the number of neighbors in the KNN classifier is 5, due to the better results of classification with 7 neighbors, in this study 7 neighbors were used.

Data segmentation for training and testing was done with three types of segmentation: 30–70, 20–80, and 10–90, to allow the comparison and selection of the best segmentation. The results of data segmentation are presented in Table 3. As the results indicate, the best data-splitting ratio is 10–90.

Table 3. The Accuracy of the Classifiers with Different Split Ratios

Data Split Ratio	Criterion /Classifier	SVM	KNN	NB
30–70	Accuracy	0.840	0.513	0.840
	Precision	0.810	0.573	0.522
	Recall	0.892	0.578	0.526
	F-Measure	0.822	0.573	0.514
20–80	Accuracy	0.894	0.824	0.894
	Precision	0.868	0.848	0.868
	Recall	0.925	0.911	0.925
	F-Measure	0.883	0.874	0.883
10–90	Accuracy	0.917	0.924	0.924
	Precision	0.901	0.911	0.911
	Recall	0.937	0.942	0.942
	F-Measure	0.912	0.920	0.920

Logistic Regression

At this point, the extracted labels and index numerical data were given as inputs to the logistic regression algorithm to investigate the impact of the news on the total index. In logistic regression, the dependent variable (X) is the numerical data of stock index and the independent variable (Y) is the label extracted from contextual data. The maximum impact of news on market data is in the 20-minute timeframe (Nuij et al., 2014). Therefore, the impact of each news item on the index was investigated for only 20 minutes. The news on holidays were ignored in this study. In addition, given that the stock exchange market in Iran is active from 9 am to 12:30 pm, the news released on hours when the stock market is inactive (which includes 12:30 am to 9 am of the next day) is effective in the index value at the beginning of each exchanging day. Therefore, the impact of the news on this study was only examined on the 9:00 a.m. index. As the index value of each working day was available for every 5-minute, there were 43 stock index values per working day. Therefore, the output of the labels extracted during the labeling phase was calculated for 5 min intervals from 9:00 to 12:30. For the period when the market is closed, i.e., from 12:30 to 9:00 the following day, the outcome of the published news was calculated along with its effect on the stock index at 9:00 the next day. This way, there were 43 values for the index and 43 values for the semantic load of the news releases each day. The impact of these two was examined by logistic regression. Table 4 indicates an example of stock index values in 5-minute intervals.

Table 4. An Example of Stock Index Values in 5 Minutes Intervals

Date	Index	5 minute interval	The final index value
2017-09-23	Total index	09:00	85.832
		09:05	85.873
		09:10	85.867
		09:15	85.829
		09:20	85.832
		09:25	85.822
		09:30	85.824
		09:35	85.819
		09:40	85.816
		09:45	85.812
		09:50	85.814
		09:55	85.813
		10:00	85.806

In the logistic regression, Python's default values were used, except for one variable called solver, which was set to liblinear by default, but only for small datasets. Therefore, the value of this variable was changed to lbfgs. In addition, the multi_class parameter was set to ovr, because it is ideal for binary classification problems. In logistic regression, all three types of split ratios, i.e., 30–70, 20–80, and 10–90, were performed. The results are shown in Table 5. As Table 5 indicates, the best accuracy for logistic regression was obtained in the 20–80 modes.

Table 5. Results of Logistic Regression with Different Split Ratios

Split Ratio	Criterion	Results
30–70	Accuracy	0.542
	Precision	0.500
	Recall	0.271
	F-Measure	0.351
20–80	Accuracy	0.549
	Precision	0.500
	Recall	0.275
	F-Measure	0.355
10–90	Accuracy	0.526
	Precision	0.500
	Recall	0.263
	F-Measure	0.345

Evaluation

In order to evaluate the results of each of the algorithms, the criteria extracted from the confusion matrix, including accuracy, precision, and F-measurement were used. The results of each algorithm evaluation are presented in Table 6. As Table 6 indicates, the classification accuracy of the logistic regression is 52%, meaning that 52% of samples were correctly classified. The precision is 50%, i.e., there was a 50% probability that the documents with a positive label were classified correctly. The recall is 26%, and there may be more positive samples. As the results indicate, the accuracy of the regression and other evaluation criteria for this algorithm is 52.2%, which is significant, because the prediction accuracy of stock price after a news release is rarely more than 58% (Hagenau et al., 2013).

Table 6. Comparison of the Evaluation Results of the Applied Algorithms

Model		P	N	Accuracy	Precision	Recall	F-Measure
SVM	P	1416	0	0.917	0.901	0.937	0.912
	N	204	827				
NB	P	1020	396	0.917	0.901	0.937	0.912
	N	695	336				
KNN	P	997	419	0.917	0.901	0.937	0.912
	N	547	484				
Logistic	P	0	498	0.522	0.500	0.261	0.343
Regression	N	0	543				

Discussion and Conclusions

Stock market is an essential part of rapidly emerging economies. Therefore, this study analyzed the impact of Persian news on stock prices by proposing a novel stock market prediction model. In this study, the semantic load of news sentences was determined using text mining techniques. In addition, the news was categorized into positive and negative categories using machine learning algorithms including SVM, NB, and KNN, which were given the best accuracy by dividing the data into 10-90 for training and test. Finally, the relationship between news and stock index was evaluated using logistic regression. A potential advantage of the proposed model is the possibility of detecting market movements that are too complex for humans to identify.

Because the structure of a language is effective on the way people interpret their words to others, using text mining and sentiments analysis techniques would help to address this subject (Groth & Muntermann, 2011). Such techniques not only can reflect the trend of public sentiments in the media, but also create clues to analyze the potential impact of these trends and reduce the risk of trading in volatile financial markets (Schumaker & Chen, 2009; Tsai et al., 2011).

According to the efficient market hypothesis (EMH) (Fama, 1965) and random walk theory (RW) (Bachelier & Cootner, 1964), stock prices cannot be predicted, because instead of existing/past prices, they are driven by news. Thus, stock prices in the market follow a random course and its forecast accuracy cannot be more than 50% (Bollen et al., 2011). However, the accuracy of the proposed model was 52.2%, a number that shows the good accuracy of the model. Therefore, based on the results, the hybrid feature selection method improves the accuracy of the stock trend prediction, and the model can provide a reference point for the stock investment and help with the prediction of the trend of stock price change in a period of time after news happens so as to guide investors to make correct investment decisions.

This study combined technical data with unstructured data, which is a main innovation of this study. Technical indicators are useful tools that indicate the real market situation. Using

the values of the technical indicators before and after the news release can be more useful than using net prices.

In addition, this study used text mining techniques as well as unstructured fundamental data to analyze the impact of Persian news on stock returns. This was done for the first time in the Persian language and is another innovation of this study. The proposed text mining approach supports decision making in financial markets and therefore has many consequences for management and individuals. The machine learning framework (i.e., SVM) used in this study contributes to automated trading in financial markets, and helps managers make profitable investment decisions as well. Given the popularity of the exchange-traded funds as a massive investment tool for private households, the method used in this study also can help this group of stakeholders in stock market trend prediction and their investment decision making.

The results of the study indicated the strong relation between the sentiment analysis report of the stock and the historical stock market trend. Besides, the study proved the relationship between economic news and stock prices, and indicated that stock trends can be predicted using news articles and previous price history. Government-related economic news is always important for the investors because they indicate the strength or weakness of the economy, the consumer, and the key sectors of the industry. Therefore, analyzing these kinds of news can help investors with their decisions about investment in the stock market. Therefore, the study adds to the literature on the prediction of stock market returns by automating the classification of political and economic news articles, using text mining and sentiment analysis techniques. In addition, it helps decision makers and investors discover the news that has the greatest impact on stock market trends. The results of the study are compatible with previous studies that have indicated the strong relationship between stock news and changes in stock prices and stock returns.

Considering the country's economic problem, which is caused by sanctions and the Covid-19 pandemic, stock market fluctuations are very high. This has caused individual investors to suffer more investment losses than foreign investors and investment institutions, despite the large number of transactions by foreign investors and investment institutions. Therefore, analyzing the economics news and their impact on stock market trends will help people in in deciding to invest in this market..

To benefit more from investing in the stock market, the investor tends to keep himself/herself updated through stock market news. However, there is a lot of news about price fluctuations floating in the financial markets every day, which makes it difficult for investors to make decisions. Therefore, the proposed model of this study will help investors with their investment decisions. Beside, professional stock traders spend most of their time on predicting the next news cycle, so that they can buy or sell stocks before the real numbers are released. They use several sources of information in this effort including economic reports. However, given the large amount of news that is published daily, it is difficult to check all the news by humans. Therefore, analyzing the impact of such news using the proposed model of this study would be helpful for stock traders in predicting the future stock prices and will help them in their buying or selling behaviors in the stock market.

In addition, stock traders are always interested in being informed of the next stock price trend and looking for ways to increase their capital gains by predicting future stock prices. Therefore, the model proposed in this study will help stock traders with using appropriate, correct, and scientific principles in determining the future stock price of investors

Given the fact that foreign investors are active in the Tehran Stock Market and their desire to participate in the stock market and production sectors of Iran has increased, the results of this study would be a guide for decisions made by both Iranian and foreign investors in their investments in the stock market.

According to the sentiment scores obtained from the analysis of the news articles, the published news has a "positive" or "negative" semantic load such that the negative news usually causes people to sell their stocks. Bad earnings reports, poor corporate governance, economic and political uncertainty, as well as unexpected and unpleasant events mean selling pressure and falling stock prices. On the other hand, positive news usually cause people to buy stocks. Good earnings reports, increased corporate governance, new products and acquisitions, as well as positive overall economic and political indicators turn into buying pressure and an increase in stock prices. Therefore, an investor who recognizes sentiments early would make a significant profit from the expected direction. Thus, the results of the study will help market participants and traders in understanding the approximate reaction of the index to the published news. Besides, prediction made in this study can be useful in designing and improving the exchange trading systems or decision support systems.

Having such intelligent systems that are developed using the insights gained through such purposeful text mining research efforts as this work would help investment banks and financial institutions as well as brokerage firms that are investing and trading in financial markets to make better financial decisions, which leads to significant financial returns on their investments and prevents severe losses.

In this study, the impact of news headlines from an economic news site on the total stock index was investigated; therefore, this research can be done using different news sources at different times. In addition, it is suggested that future research, in addition to news headlines, can also examine the news text. In addition, it is suggested to review various financial, sports, public, etc., news on the stock index.

The legal disclosures of companies or comments made on other media such as Twitter can also be used to investigate the impact of news on the stock index. Moreover, future research can do similar research on specific industry index, specific companies, top 30 industry indexes, top 50 companies, etc.

In addition, this study categorized news into two categories – i.e., "positive" and "negative" – which can be classified in a more detailed manner into "very positive", "positive", "ineffective", "very negative" and "negative." That is to say, it is recommended to the future studies to classify news into five categories and study their impact on stock market returns.

This study only predicted the upward or downward trend of the news, while the future research can predict discrete values of the news trend. In addition, it is suggested that future research might predict other elements such as risk in financial markets, including the stock market, using text mining techniques. Deep learning is potentially useful when dealing with high dimensional space and a large number of features in the text. Therefore, its application can be effective in the classification phase. In addition, fuzzy logic-based techniques such as fuzzy rule classifications, fuzzy clustering, etc., can be developed in this field. Therefore, it is suggested that similar research be performed using neural networks and deep learning.

One of the major limitations of working with the Persian language is the lack of relevant tools. For example, the only library available in Python for preprocessing Persian text is the Hazm library, which can only be used on the Linux operating system. Therefore, the development of relevant tools for further research in the Persian language seems necessary.

Given the inefficiency of the Iranian market and the neutralization of the impact of negative news on the stock index by large industries, the values obtained in this study are significant. However, with a better data set and the development of Persian language processing tools, the accuracy of predictions will increase.

Other possible areas for future research could be the use of expressions and phrases instead of single words. The advantage of this method is that the semantic relationship between the intended words can improve the accuracy of the model.

References

Alanyali, M., Moat, H. S., & Preis, T. (2013). Quantifying the relationship between financial news and the stock market. *Scientific Reports, 3*(1), 1-6. https://doi.org/10.1038/srep03578

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

Cavalcante, R. C., Brasileiro, R. C., Souza, V. L., Nobrega, J. P., & Oliveira, A. L. (2016). Computational intelligence and financial markets: A survey and future directions. *Expert Systems with Applications, 55*, 194-211. https://doi.org/10. 1016/j.eswa.2016.02.006

Chan, S. W., & Chong, M. W. (2017). Sentiment analysis in financial texts. *Decision Support Systems, 94*, 53-64. https://doi.org/10. 1016/j.dss. 2016.10.006

Bachelier, L. L., & Cootner, P. H. (1964). *The random character of stock market prices. Theorie de la speculation, Gauthiers*, MIT Press. Cambridge.

De Fortuny, E. J., De Smedt, T., Martens, D., & Daelemans, W. (2014). Evaluating and understanding text-based stock price prediction models. *Information Processing & Management, 50*(2), 426-441. https://doi.org/10. 1016/j.ipm.2013.12.002

Dutta, A., Bandopadhyay, G., & Sengupta, S. (2012). Prediction of stock performance in the Indian stock market using logistic regression. *International Journal of Business and Information, 7*(1), 105-136.

Eck, M., Germani, J., Sharma, N., Seitz, J., & Ramdasi, P. P. (2020). Prediction of stock market performance based on financial news articles and their classification. In Sharma N., Chakrabarti A., Balas V.E., Martinovic J. (Eds.), *Data Management, Analytics and Innovation Advances in Intelligent Systems and Computing, vol 1175.* (pp. 35-44). Springer. Singapore. https://doi.org/10.1007/978-981-15-5619-7_3

Fama, E. F. (1965). The behavior of stock-market prices. *The journal of Business, 38*(1), 34-105. https://www.jstor.org/stable/2350752

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

Forman, G. (2003). An extensive empirical study of feature selection metrics for text classification. *Journal of Machine Learning Research, 3*, 1289-1305.

Fung, G. P. C., Yu, J. X., & Lu, H. (2005). The Predicting power of textual information on financial markets. *IEEE Intell. Informatics Bull, 5*(1), 1-10.

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

Groth, S. S., & Muntermann, J. (2011). An intraday market risk management approach based on textual analysis. *Decision Support Systems, 50*(4), 680-691. https://doi.org/10. 1016/j.dss.2010.08.019

Gunduz, H., & Cataltepe, Z. (2015). Borsa Istanbul (BIST) daily prediction using financial news and balanced feature selection. *Expert Systems with Applications, 42*(22), 9001-9011. https://doi.org/10. 1016/j.eswa.2015.07.058

Guo, L., Shi, F., & Tu, J. (2016). Textual analysis and machine leaning: Crack unstructured data in finance and accounting. *The Journal of Finance and Data Science, 2*(3), 153-170. https://doi.org/10. 1016/j.jfds.2017.02.001

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. https://doi.org/10. 1016/j.dss.2013.02.006

Huang, J. Y., & Liu, J. H. (2020). Using social media mining technology to improve stock price forecast accuracy. *Journal of Forecasting, 39*(1), 104-116. https://doi.org/10.1002/for.2616

Hatefi Ghahfarrokhi, A., & Shamsfard, M. (2020). Tehran stock exchange prediction using sentiment analysis of online textual opinions. *Intelligent Systems in Accounting, Finance and Management, 27*(1), 22-37. https://doi.org/10.1002/isaf.1465

Hazm. (2019). https://pypi.org/project/hazm/. Retrieved January 29, 2019, from https://pypi.org/project/PyPrind/

Hong, S. (2020). A study on stock price prediction system based on text mining method using LSTM and stock market news. *Journal of Digital Convergence, 18*(7), 223-228. https://doi.org/10.14400/JDC.2020.18.7.223

Jishag, A. C., Athira, A. P., Shailaja, M., & Thara, S. (2020). Predicting the stock market behavior using historic data analysis and news sentiment analysis in R. In In: Luhach A., Kosa J., Poonia R., Gao XZ., Singh D. (Eds.), *First International Conference on Sustainable Technologies for Computational Intelligence*. Advances in Intelligent Systems and Computing, vol 1045. (pp. 717-728). Springer, Singapore. https://doi.org/10.1007/978-981-15-0029-9_56

Kalra, V., & Agrawal, R. (2019). *Challenges of text analytics in opinion mining*. Extracting Knowledge from Opinion Mining (pp. 268-282). IGI Global. https://doi.org/10.4018/978-1-5225-6117-0.ch012

Kavuluru, R., Hands, I., Durbin, E. B., & Witt, L. (2013). Automatic extraction of ICD-O-3 primary sites from cancer pathology reports. *AMIA Summits on Translational Science Proceedings*, *2013*, 112-116.

Kumar, B. S., & Ravi, V. (2016). A survey of the applications of text mining in financial domain. *Knowledge-Based Systems, 114*, 128-147. https://doi.org/10. 1016/j.knosys.2016.10.003

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. https://doi.org/10. 1016/j.knosys. 2014.04.022

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. https://doi.org/10. 1016/j.ins. 2014.03.096

Lutz, B., Pröllochs, N., & Neumann, D. (2020). Predicting sentence-level polarity labels of financial news using abnormal stock returns. *Expert Systems with Applications, 148*, 1-11. https://doi.org/10.1016/j.eswa.2020.113223

Mate, G. S., Kulkarni, R., Amidwar, S., & Muthya, (2020). Stock prediction through news sentiment analysis. *Journal of Architecture & Technology*, *11*(8). 36-40.

Meesad, P., & Li, J. (2014, December). Stock trend prediction relying on text mining and sentiment analysis with tweets. In Choo, Y. H(Eds.), *4th World Congress on Information and Communication Technologies (WICT 2014)* (pp. 257-262). IEEE. Melaka, Malaysia. https://doi.org/10.1109/WICT.2014.7077275

Moazeni, B., Nemati, M., & Sayyadi Moghaddam, M. (2014, November). *Investigating the impact of political and economic news on changes in the Tehran Stock Exchange Index*. International Management Conference [Paper presentation]. Mobin Cultural Ambassadors Institute, Tehran, Iran (In Persian).

Nassirtoussi, A. K., Aghabozorgi, S., Wah, T. Y., & Ngo, D. C. L. (2014). Text mining for market prediction: A systematic review. *Expert Systems with Applications, 41*(16), 7653-7670. https://doi.org/10. 1016/j.eswa.2014.06.009

Narayan, P. K., & Bannigidadmath, D. (2017). Does financial news predict stock returns? New evidence from Islamic and non-Islamic stocks. *Pacific-Basin Finance Journal, 42*, 24-45. https://doi.org/10.1016/j.pacfin.2015.12.009

Nizer, P. S. M., & Nievola, J. C. (2012). Predicting published news effect in the Brazilian stock market. *Expert Systems with Applications, 39*(12), 10674-10680. https://doi.org/10. 1016/j.eswa.2012.02.162.

Nuij, W., Milea, V., Hogenboom, F., Frasincar, F., & Kaymak, U. (2013). An automated framework for incorporating news into stock trading strategies. *IEEE Transactions on Knowledge and Data Engineering, 26*(4), 823-835. https://doi.org/10.1109/TKDE.2013.133.

Pejić Bach, M., Krstić, Ž., Seljan, S., & Turulja, L. (2019). Text mining for big data analysis in financial sector: A literature review. *Sustainability*, *11*(5), 1277-1304. https://doi.org/10.3390/su11051277

Picasso, A., Merello, S., Ma, Y., Oneto, L., & Cambria, E. (2019). Technical analysis and sentiment embeddings for market trend prediction. *Expert Systems with Applications, 135*(201), 60-70. https://doi.org/10.1016/j.eswa.2019.06.014

Polyglot. (2019). Sentiment @ polyglot.readthedocs.io. Retrieved from https://polyglot.readthedocs.io/en/latest/Sentiment.html

Raschka, S., & Mirjalili, V. (2017). Machine learning mit Python und Scikit-Learn und TensorFlow: Das Praxis-Handbuch für Data Science, Predictive Analytics und Deep Learning. MITP Verlags GmbH & Company KG.

Rioja, F., & Valev, N. (2014). Stock markets, banks and the sources of economic growth in low and high income countries. *Journal of Economics and Finance*, *38*(2), 302-320. https://doi.org/10.1007/s12197-011-9218-3

Ritesh, B. R., Chethan, R., & Jani, H. S. (2017). Stock movement prediction using machine learning on news articles. *International Journal on Computer Science and Engineering*, *4*(3), 153-155.

Sharma, A., Bhuriya, D., & Singh, U. (2017, April). Survey of stock market prediction using machine learning approach. In Smys. S. (Eds.) 2017 *International conference of Electronics, Communication and Aerospace Technology (ICECA), 2,* 506-509. Coimbatore, India. https://doi.org/10.1109/ICECA.2017.8212715

Seker, S. E., Mert, C., Al-Naami, K., Ozalp, N., & Ayan, U. (2014). Time series analysis on stock market for text mining correlation of economy news. *International Journal of Social Sciences and Humanity Studies*, *6*(1), 69-91.

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. https://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. https://doi.org/10.1016/j.dss.2012.03.001.

Scrapy. (2019). index @ scrapy.org. Retrieved from https://scrapy.org/

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*(2016), 74–83. https://doi.org/10.1016/j.dss.2016.03.001

Shynkevich, Y., McGinnity, T. M., Coleman, S., & Belatreche, A. (2015, July). Stock price prediction based on stock-specific and sub-industry-specific news articles. In Honorary, A. H (Eds.), *2015 International Joint Conference on Neural Networks (IJCNN)* (pp. 1-8). IEEE. Killarney, Ireland. https://doi.org/10.1109/IJCNN.2015.7280517

Tabari, N., Seyeditabari, A., Peddi, T., Hadzikadic, M., & Zadrozny, W. (2018, September). A comparison of neural network methods for accurate sentiment analysis of stock market tweets. In Alzate C. et al. (Eds.), ECML PKDD 2018 Workshops. MIDAS 2018, PAP 2018. Lecture Notes in Computer Science, vol 11054. (pp. 51-65). Springer, Cham. https://doi.org/10.1007/978-3-030-13463-1_4.

Terra, J. (2021). Why Python is essential for data analysis and data science. *Simplilearn. https://www.simplilearn.com/why-python-is-essential-for-data-analysis-article.*

Thanh, H. T., & Meesad, P. (2014). Stock market trend prediction based on text mining of corporate web and time series data. *Journal of Advanced Computational Intelligence and Intelligent Informatics*, *18*(1), 22-31. https://doi.org/10.20965/jaciii.2014.p0022

Tobback, E., Naudts, H., Daelemans, W., de Fortuny, E. J., & Martens, D. (2018). Belgian economic policy uncertainty index: Improvement through text mining. *International Journal of Forecasting*, *34*(2), 355-365. https://doi.org/10.1016/j.ijforecast.2016.08.006.

Tsai, C. F., Lin, Y. C., Yen, D. C., & Chen, Y. M. (2011). Predicting stock returns by classifier ensembles. *Applied Soft Computing*, *11*(2), 2452-2459. https://doi.org/10.1016/j.asoc.2010.10.001.

Weng, B., Ahmed, M. A., & Megahed, F. M. (2017). Stock market one-day ahead movement prediction using disparate data sources. *Expert Systems with Applications*, *79*, 153-163. https://doi.org/10.1016/j.eswa.2017.02.041.

W3Techs. (2019). https://w3techs.com/technologies/overview/content_language

Xie, Y., & Jiang, H. (2019). Stock market forecasting based on text mining technology: A support vector machine method. *arXiv preprint arXiv:1909.12789*.

Yekrangi, M., & Abdolvand, N. (2020). Financial markets sentiment analysis: Developing a specialized Lexicon. *Journal of Intelligent Information Systems*, 2020, 1-20. https://doi.org/10.1007/s10844-020-00630-9

Zaleskiewicz, T., & Traczyk, J. (2020). Emotions and financial decision making. In Zaleskiewicz T., Traczyk J. (Eds.), *Psychological Perspectives on Financial Decision Making* (pp. 107-133). Springer, Cham. https://doi.org/10.1007/978-3-030-45500-2_6

Zhang, Z., Zhang, Y., Shen, D., & Zhang, W. (2018). The dynamic cross-correlations between mass media news, new media news, and stock returns. *Complexity*, special issue, 1-10, https://doi.org/10.1155/2018/7619494