Predictive Analytics for Stock Prices using Sentiment Analysis

Salma Elsayed
Faculty of Computer and Informatics,
Zagazig University, Zagazig, Egypt

ABSTRACT
Stock prediction is consider one of the most popular tasks in the last decade. Stock prediction can be achieved through the analysis of the numerical values (e.g., open price and close price) or the sentiment analysis of social media text (e.g., tweets). In this paper, we will discuss the several approach of stock prediction using sentiment analysis methods. These methods can be classified into four main categories, namely, machine learning, lexicon, graph, and hybrid based methods. Besides, we discussed the basic tools used to help in the task of sentiment analysis such as Term Frequency Inverse Document Frequency and word2cev algorithms.

Keywords
Stock price, stock market, sentiment analysis, stock movement, sentiment classification, machine learning, graph based, lexicon based, hybrid.

1. INTRODUCTION
Predicting stock prices is so difficult due to their extreme volatility, which is influenced by a variety of political and economic issues, as well as changes in leadership, investor sentiment, and a variety of other factors, So we can use Sentiment Analysis to forecastpeople's feelings, which have a significant effect on stock values,and therefore it aid in the forecasting of real stock movement. And because of people's reliance on social media increases, the interactions among stock market participants become more easy and regular.As a result, the sentiment exhibited by other investors and the opinions posted on social media may influence an investor's mood and decision-making, as well as influence the stock market to some extent [1-3].

Sentiment Classification approaches can be classified into: machine learning, lexicon-based, and hybrid. The Machine Learning technique(ML)employs well-known ML methods as well as language characteristics. The Lexicon-based is based on a sentiment lexicon which is acombination of commonly used and pre-compiled sentiment words. It is classified into two types of approaches: dictionary-based and corpus-based. Graph-based models for text processingmay be examined in a variety of ways to experiment with different features or associations between texts. One of the trending topics that add value to sentiment analysis is text ranking using a graph-based model. The hybrid includes both approachesand is widely used with sentiment lexicons performing an important role in the most of procedures[4-8].

The process of cleaning and preparing text for categorization is known as pre-processing. Online writings typically contain a lot of noise and unclear elements such as HTML tags, scripts, and ads. Furthermore, on the words level, many terms in the text have little effect on its overall direction. Reducing noise in the text should assist enhance the classifier's performance, hurry up the classification process and allowing for real-time sentiment analysis. The process includes several phases: online text cleaning, white space elimination, extending abbreviation, stop word removal, negation management, and finally extract features.There are various methods for determining the significance of each characteristic by assigning a weight to it in the text. The most common are feature frequency (FF) and Term Frequency Inverse Document Frequency (TF-IDF).This technique takes into account the frequency of the term in a single document as well as the distribution of the word in the documents. It can more accurately reflect the significance of a characteristic in categorization[9].The distributed word vector generated by Word2vec technology is the most widely used word representation. The semantic information of the word is contained in the word vector, which has a low dimension. However, emotion information about words is not included in distributed word vectors. Word embedding is a method that gets to know low-dimensional constantly vector representations of words and has attracted growing attention.The learnt embedding can hold syntactic and semantic connections and facilitating various NLP tasks including sentiment analysis, machine translation, word analogy and parsing[10, 11].

In [1], a unique Long short-term memory (LSTM) based model is proposed for stock market prediction. For one thing, the sentiment index is utilized to account for the investor's emotional tendency. Therefore, the proposed scheme shows great potential to assist the country by giving the government guidance on rationalizing and leading the stock market as well as offering profits to individuals by guiding investments.

In [2], They used time series models, neural networks, and a mixture of neural networks and financial news items to forecast stock values. According to the findings, there is a substantial correlation between stock prices and financial news items. They built prediction models using time series forecasting techniques such as ARIMA, RNN, and Facebook Prophet. They obtained superior results with RNN and discovered a relation between stock price direction and textual information. When stock prices are extremely volatile or low, the models do not perform well.

In [12], Their experimental system using stock data from three firms recorded on the Ghana stock exchange (from January 2010 to September 2019) demonstrates that the stock market is predicted using public attitudes. The proposed model improved accuracy in predicting future stock value for 1 day, 7 days, 30 days, 60 days, and 90 days relying on merged and separate datasets demonstrates that stock forecasting models' performance may be greatly enhanced through stock relevant data merging.

2. BACKGROUND
Sentiment analysis, also known as opinion mining or emotion AI, is the technique of evaluating web content to determine if it has a good, negative, or neutral emotional tone. Simply said,

sentiment analysis facilitates in determining the author's viewpoint about an issue. As a result, sentiment analysis saves time and energy because the sentiment extraction procedure is entirely automated. The human involvement becomes minimal because it is the procedure that processes sentiment analysis datasets [10].

It has recently been noted that the number of people actually participating in social media is quickly rising. People are sharing their viewpoints on numerous issues through reviews, postings, comments, and statuses. As a result, a massive amount of data is created on the Internet that may be examined for future studies. As a result, sentiment analysis has become a popular area with several applications[13].

The task of Sentiment Analysis is identified as a sentiment classification (SC)challenge. The initial stage in the SC issue is the extraction and selection of text characteristics. Some of these characteristics are as follows:

- Presence and frequency of terms: Individual words or word n-grams, as well as theirfrequency counts, are examples of these characteristics. It either offers binary weighting for words (zero if it appears, one if it does not) or utilizes term frequency weights to estimate the relative importance of features.
- Parts of speech (POS): identifying adjectives that are important markers of views.
- Viewpoint words and statements are expressions that are frequently usedto describe emotions such as good or negative, like or dislike. Some phrases, on the other hand, express viewpointsinstead of utilizing opinion words.
- Negations*:* the presence of negative phrases, such as not good is comparable to bad, may shift viewpoint orientation[10].

Preprocessing the data is done in two steps: Tokenization andstop word elimination.Tokenization is anoperation of converting context into useful data while preserving the substance of the text. Stop words are unwanted terms in text data that do not contribute meaning to the data. These words are removed from the text during feature extraction. Two techniques are used for feature extraction: TF-IDF and Doc2vec.

Term frequency-inverse document frequency is abbreviated as TF-IDF. The TF-IDF technique is widely used in information extraction and text mining. It is a weight measure that evaluates importance of words in the context.

The term frequency (TF) is the frequency of a particular term t in document d. When the term is repeated, its frequency rises. TF is computed by dividing the frequency of term t in document d by the number of terms in that document d[14-16].

$$\mathrm{TF(t,d)} = \frac{\text{Number of times term t appears in a document d}}{\text{Total number terms in a document d}}$$

IDF: TF only countsthe number of times of the term *t*. Several keywords, such as stop words, appear many times yet aren't always useful. As a result, the term's significance is measured using inverse document frequency (IDF). IDF allocate extra significance to the terms that appears only infrequently in the document d. IDF is computed as:

$$\mathrm{IDF}(t) = \log_e \frac{\text{Totalnumberofdocuments}}{\text{Totalnumberofdocumentswithtermtinit}}$$

For a term t in the document d, the final weight is computed as:

$$TF - IDF(t,d) = \mathrm{TF(t,d)} * \mathrm{l}IDF(t)$$

Doc2vec model is used to enhance embedding learning from word-to-word sequences.It's a more sophisticated version of word2vec. Doc2vec may be used to create word n-grams, sentences, paragraphs, and documents. Doc2vec refers to a collection of methods for expressing texts as low-dimensional vectors with fixed lengths. Doc2Vec is composed of a three layer neural network: an input layer, one hidden layer, and output layer. It is utilized to generate distributed representations of words.Deep learning is used to build two algorithms in Word2vec: a continuous bag of words (CBOW) and skip-gram (SG); in Doc2vec model, both methods equate to distributed bag of words (DBoW) and distributed memory (DM) [10].

Word embedding is a common approach natural language processing (NLP) which guides to learn words' low dimensional vector representations from the texts. It is able to hold semantic and syntactic relations. Word embedding methods can learn representations of distributional vector that encapsulate words' semantic meanings by utilizing statistical information like word repetition rates. The plurality of word embedding techniques are based solely on statistical data extracted from documents[11].

3. ML-BASED METHODS OF SA FOR STOCKPRICE PREDICTION

In [17],They predicted stock markets using news announcements by presenting sentiment analysis (SA) based machine learning technique. The findings demonstrated that SA is an effective method for extracting useful attributes from financial news, considerably decreasing feature dimensions. Using Particle Swarm Optimization for parameter optimization enhances the accuracy of the forecasting. They achieved an accuracy of 59.15% by using Support Vector Machine and Particle Swarm Optimization model.

In [18],They utilized two machine learning algorithms (Support vector Machine and Neural Network) to classify the days depending on the existence of events and apply the generated model for prediction. In their study, they tested the idea that sentiment analysis of twitter data set can give extra info and therefore enhance the forecasting accuracy of stock market. They applied learning algorithms to three different types of data. The initial piece of it that referred to as the fundamental set was the features of stock market in prior days. The second set was produces by combining an amount of twitter tweets with these keywords “Hope”, “Fear”, “Worry” to the initial group of data. The last group (Basic&8EMO) was generated by combining an amount of twitter data from these emotions: “loving”, “calm”, “energetic”, “happy”, “fearful”, “angry”, “sad” and “tired” . They predicted that the comparison of the expectations' performance depending on their sets would be varied.Based on their supposition about the appearance of extra data in twitter, they predict that the initial group would generate the minimum accuracy rate and depending on utilized set Basic&8EMO, the second set produces slightly greater rate of accuracy and obtains the maximum accuracy rate of prediction.Their first findings show that additional data from Twitter do not greatly improve efficiency. By using SVM algorithm to expect of the DJIA indicator, they obtained the highest average accuracy rate of 64.10%

In [19],They collected tweets through Twitter's Search API and handled them for additional analysis that included SA and Natural Language Processing (NLP). Following that, they used Support Vector Machine Algorithm and Naïve Bayes algorithm

to forecast sentiment of the tweets. They observed that Support Vector Machines provide greater accuracy via cross-validation by assessing each model for its appropriate sentiment classification. Regardless of this fact, they proceeded to keep in mind these approaches and compare the accuracy on a regular basis. They used the Yahoo Finance API to collect the historical stock data after expecting mood of all tweets.They developed a significant feature matrix for predicting stock price based on stock value change for each day and semantic meaning and finally suggested their own trading strategy.

In [20], They developed a new sentiment model to undertake context-sensitive sentiment analysis of online articles in the stock markets. Sina Finance, a model financial site, was chosen as the experimental system for collecting a corpus of financial reviews. According to empirical findings, there are substantial relationships between stock value variability patterns and stock community sentiment. According to the computational findings, the statistical machine learning technique outperformed the semantic technique in the classification efficiency. The findings implied that investor sentiment has an especially powerful effect on stock value according to stock growth. They applied sentiment analysis on public information concerning to the Moroccan stock market which was gathered using automated texts from a variety of internet sources. Two various sentiment analysis approaches were evaluated to determine which would offer the most accurate assessment of sentiments.According to the variety of the French language applied to evaluate stock value, they maintained the supervised machine learning of Naïve Bayes algorithm. This research investigates the relations between media content and stock value. This study investigates the interactions between media content and the stock market. They studied the casual correlation between public sentiment as assessed by a large-scale set of news and stock value of Casablanca Stock Exchange (CSE). Their findings indicate that the public sentiment may be collected from media websites using basic NLP approaches. Then, they observed that the negative sentiment has a substantial effect on the movement of stock market.

In [21], They applied sentiment analysis on public information concerning to the Moroccan stock market which was gathered using automated texts from a variety of internet sources. Two various sentiment analysis approaches were evaluated to determine which would offer the most accurate assessment of sentiments. According to the variety of the French language applied to evaluate stock value, they maintained the supervised machine learning of Naïve Bayes algorithm. This research investigates the relations between media content and stock value. This study investigates the interactions between media content and the stock market. They studied the casual correlation between public sentiment as assessed by a large-scale set of news and stock value of Casablanca Stock Exchange (CSE). Their findings indicate that the public sentiment may be collected from media websites using basic NLP approaches. Then, they observed that the negative sentiment has a significant effect on the movement of stock market.

4. LEXICON-BASED METHODS OF SA FOR STOCK PRICE PREDICTION

In [22], They predicted stock values with a 75% accuracy but unregrettably, only 10% of tweets in StockTwits(a financial social network was founded in 2009) are labeled. They needed a strong approach for sentiment analysis of senior authors to increase the efficiency of forecating stock value.They used a lexicon-based technique that did not require training data, and so it is valuable, especially in cases that include high-dimensional data. Sentiment analysis may be executed using a number of lexicon-based approaches. They used VADER(Valence Aware Dictionary for sEntiment Reasoning), SentiWordNet, and TextBlob on StockTwits data to determinewhether they could improveanefficiency of sentiment analysis. Naive Bayes ,Logistic regression, and Linear Support Vector Machine classification wereutilized as standard and compared to the outcomes of using lexicon-based approaches alongside machine learning methods. According to their findings, not only does VADER outperformed machine learning approaches in capturingemotions from financial social media, such asStockTwits, it is also quicker.

In [23], they created a lexicon-based technique to generate the sentiment form microblogs called FinLex. They build a financial lexicon to describe the emotional content of the microblogs because they are all associated with finance.They chose qualifying words from 2.6 million microblogs. the words of frequency, location, and part of speech tags are employed as attributes. In this procedure, they got 32,296 qualified words. And then they selected 40 words in both positive and negative sentiment as abase terms. Then they computed the so-PMI of each proposed word to evaluate whether the word is positive or negative. For every two emotions, they conducted the following analysis for word by word. they scanned the data and drew a boundary between the two words that appeared in thephrase.After analyzing the collection, they constructed a word relation graph, then they initialized all termsvaluewith 1.0, and run the SentiRank (equivalentto the PageRank, but the members of SentiRank are words) to determine the sentiment value of every words. Then they scaled thevalues and ensure that they all in a certain domain between (0, 1). finally, they dropped several words that have a bad value. They combined the stock historical information and outcome of sentiment analysis to forecast the stock movement of the next day.And They proposed a user-group model that pick advantage of users' data to enhance prediction efficiency even more.

In [24], They created the Thai Financial Probabilistic Lexicon (ThaiFinLex), a lexicon-based on terms discovered in Thai news stories and stock market closing values of stocks featured in the media. In contrast to existing lexicons, their suggested ThaiFinLex comprises of occurence words which have an influence on stock values and their related chances, which forecast stock valuechanges. They used split-validation to compare outcomes of their suggested PLSP model with the results of the other three traditional methods using 5-fold cross-validation: SVM, J48, BayesNet. The findings demonstrated that the suggested PLSP (Probabilistic Lexicon Based Stock Market Prediction) model outperformed the other mothods evaluated in this research. They attained up to 96.64 percent accuracy in particular.

In [25], They presented a technique for producing a domain-specific vocabulary by combining a probabilistic approach with financial-based knowledge, i.e., the stock price weight. This study differs from previous techniques in that the domain-specific lexicons are generated automatically, and the emotion scores are calculated by taking into consideration the previous stock price movement. They also did not have to modify their terminology from a general lexicon. They measure the usefulness of their domain-specific lexicons by comparing their results to those of commonly used generic lexicons such as SentiWordNet 3.0, Henry's financial lexicon, and Loughran's financial lexicon. According to the experimental findings, their domain-specific lexicon is 4.4 percent to 45.9 percent more accurate than Henry's lexicon, 2.5 percent to 8 percent more accurate than Loughran's lexicon, and 4.4 percent to 46.1

percent more accurate than SentiWordNet. Their domain-specific lexicons forecast stock price direction more accurately than the other three lexicons as the weighted price change grows (1% vs. 20% adjusted change in price following the call).

5. GRAPH-BASED METHODS OF SA FOR STOCK PRICE PREDICTION

In [26], Through opinion mining and graph-based semi-supervised learning, they create an algorithm to aid in the operation of making the decision in stock finance. From the standpoint of data analysis, their article definesfalsedata since the stock market is impacted by communication between participants and there is a lot of information devoid of integrity. This method not only filters out unneeded or irrelevant information but also filters out actual misleading data that can have an influence on stock investment decisions. Graph-based semi-supervised learning contributed to the more accurate categorization of the emotions of terms or texts in a circumstance, particularly the word2vec-based learning text in the stock market.In other words,particularterms and word relationships could have various meanings in different contexts. However, graph-based semi-supervised learning aids in the situational learning and classification of words or texts. Words with a high emotion ratefunctioned as annotated data, and their results were spread to nearby words without labels that were deemed to have unclear meanings.

In [27], They propose HyS3, a hybrid supervised semi-supervised model for predicting movement orientation. The graph-based semi-supervised component of HyS3 analyzes the markets' worldwide connections using a network built using ConKruG, a unique continuous Kruskal-based graph creation technique (Continuous Kruskal-based Graph). The ability of graph-based semi-supervised techniques to utilize world market data for forecasting is determined by the network that is built to simulate market interactions. A network that is constructed from the start by considering the sort of issue that it tries to solve is more suited for prediction. They developed a novel network design method that mimicked market falls and rises to provide the greatest potential forecast when compared to other networks in the study literature. According to their findings, using worldwide market information instead of using previous information of the market is forecasted would assist to have a good prediction than simply utilizing market's previous information However, oldinformation of the market can be valuable in projecting its future, but it must be utilized at the appropriate moment. They fed information collected from historical market data into the network as primary knowledge and used this knowledge to enhance prediction. The probabilities estimated in a supervised method allow for the injection of primary knowledge. This knowledge, which is intentionally put into many points of the ConKruG network through the HyS3 technique, could enhance forecating. They were able to create a prediction method that outperformed other current methods by utilizingpreviousinformation for everymarkets, as well as information from the other worldwide markets.

6. HYBRIDMETHODOF SA FOR STOCKPRICE PREDICTION

In [28],They suggested a hybrid technique for stock price prediction that combines elements of technical analysis with sentiment analysis (SA). The sentiment analysis characteristics are depend on a Pointwise Mutual Information (PMI) model, and they used neural network (NN) and -support vector regression (SVR) models to forecast the annual modification in the stock value. They proved that long-run stock price behavior can be effectively predicted using NNs and -SVRs. We use the PMI model in the sentimental analysis phase of the prediction to overcome the complexity of information representation. As a result, they used the PMI model to create a hybrid model that blends quantitative input variables (mainly fundamental analytical indicators) with a qualitative sentiment from company reports. Then, NNs and SVRs are used to predict stock returns one year in advance.

In [29], They introduced the SOM-GP procedure, a hybrid method based on a self-organizing map (SOM) neural network and genetic programming (GP) To anticipate stock values. The SOM-GP method is divided into three phases. The important historical stock exchange data, such as base price, maximum price, minimum price, closing price, transaction bulk, and so on, are first gathered in the first stage. Following that, the necessary technical indicators, such as the moving average (MA), The Williams overbought/oversold index (WMS percent R), psychological line (PSY), commodity channel index (CCI), and other independent input parameters utilized in the forecasting model of stock value are computed. In the second step, the next day's closing price is first normalized into a domain between -1 and 1 based on its highest and lowest prices. The normalized technicalindicators$\overline{x} = (\overline{x_1}, \overline{x_2}, \ldots\ldots\ldots, \overline{x_n})$ and the normalized closing value in the next day are divided into training,testing, and validation data, as depending on a pre-specificpercent.In the third step, the normalized technical indicators$\overline{x} = (\overline{x_1}, \overline{x_2}, \ldots\ldots\ldots, \overline{x_n})$ of each day in the validation data are sortedinto a cluster, designated as cluster c, by feeding$\overline{x}$ into the SOM neural model built in step 1. Therefore, the normalized technical indicators$\overline{x} = (\overline{x_1}, \overline{x_2}, \ldots\ldots\ldots, \overline{x_n})$ of every day in the validation data are inputted into the good trained genetic programming model relating to cluster c, as produces in Step 2, to obtainexpected normalized closing price in the next day. As a result, the forecast closing price of the next day, regarding the technical indicatorsof $(x_1, x_2, \ldots\ldots\ldots, x_n)$for a particular day then may be generatedby denormalizing. Finally, the efficiency of the suggested SOM-GP forecastingtechnique is verifiedby statistical measuressuch as the root mean squared error (RMSE), mean absolute error (MAE), and mean absolute percentage error (MAPE).

In [30], They suggested a hybrid system that predicts stock prices using a neural network LSTM and sentiment analysis to validate their forecasts. Sentiment analysis enabled them to examine different political and economic issues, which significantly impacted the stock market. Their findings demonstrated that there was a connection between public opinion and stock prices. On the other hand, LSTM has shown to be the best algorithm for predicting stock prices since it takes prior points into consideration but also utilized forget gates to eliminate extremely old data that was unlikely to influence the current outcome, making it incredibly efficient. Starting with this work, there were a number of further paths that might be followed. The first is to integrate additional social media sites in addition to Twitter to measure public sentiment. Second, our dataset only includes English persons; nevertheless, to map actual public mood, all languages must be included. Finally, LSTM may be further improved to predict more accurate results.

In [31], They presented a hybrid technique for forecasting future stock prices that combines the long-short term memory (LSTM) with the Empirical mode decomposition (EMD). They utilized comprehensive EMD to break down the complicated initial stock value time series into numerous subsequences that are smoother, more consistent, and more stable than the basic time series. Then, they used the Long Short Term

Memorytechnique to train and forecastof every sequence. Finally, they combined the prediction values of many subsequences to produce the prediction values of the initial stock price time sequences. They used five data points in their experiment to properly assess the method's performance. The results of the comparison with the other four prediction approaches reveal that the forecastingresults are more accurate. The hybrid prediction approach they suggested is useful and efficient in predicting future stock prices. As a result, the hybrid prediction approach has both operational and operational values.SA can be applied in different applications such as price prediction based on product description [32], improving the parallel applications by studying the log files [33, 34], bioinformatics [35], and data center log files analysis [36].

7. CONCLUSIONS

In this paper, we discussed the several methods of stock prediction based on the sentiment analysis method. We started by discussing the ML-based methods. The discussion show that the most used models are the NNs and SVM. A remarkable lexicon-based method utilized a domain-specific vocabulary by combining a probabilistic approach. For the graph-based methods, the most common techniques were opinion mining and graph-based semi-supervised learning. Finally, the hybrid methods main combine the lexicon methods with the ML methods.

8. REFERENCES

[1] Jin, Z., Y. Yang, and Y. Liu, *Stock closing price prediction based on sentiment analysis and LSTM.* Neural Computing and Applications, 2019: p. 1-17.

[2] Mohan, S., et al. *Stock price prediction using news sentiment analysis*. in *2019 IEEE Fifth International Conference on Big Data Computing Service and Applications (BigDataService)*. 2019. IEEE.

[3] Nemes, L. and A. Kiss, *Prediction of stock values changes using sentiment analysis of stock news headlines.* Journal of Information and Telecommunication, 2021: p. 1-20.

[4] Yue, L., et al., *A survey of sentiment analysis in social media.* Knowledge and Information Systems, 2019. **60**(2): p. 617-663.

[5] Feldman, R., *Techniques and applications for sentiment analysis.* Communications of the ACM, 2013. **56**(4): p. 82-89.

[6] Mukherjee, S., *Sentiment analysis*, in *ML. NET Revealed*. 2021, Springer. p. 113-127.

[7] Gopal, S. and M. Ramasamy, *Hybrid multiple structural break model for stock price trend prediction.* The Spanish Review of Financial Economics, 2017. **15**(2): p. 41-51.

[8] Patil, P., et al. *Stock market prediction using ensemble of graph theory, machine learning and deep learning models*. in *Proceedings of the 3rd International Conference on Software Engineering and Information Management*. 2020.

[9] Haddi, E., X. Liu, and Y. Shi, *The role of text pre-processing in sentiment analysis.* Procedia Computer Science, 2013. **17**: p. 26-32.

[10] Xu, G., et al., *Sentiment analysis of comment texts based on BiLSTM.* Ieee Access, 2019. **7**: p. 51522-51532.

[11] Tang, D., et al. *Learning sentiment-specific word embedding for twitter sentiment classification*. in *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 2014.

[12] Nti, I.K., A.F. Adekoya, and B.A. Weyori, *Predicting Stock Market Price Movement Using Sentiment Analysis: Evidence From Ghana.* Appl. Comput. Syst., 2020. **25**(1): p. 33-42.

[13] Yadav, A. and D.K. Vishwakarma, *Sentiment analysis using deep learning architectures: a review.* Artificial Intelligence Review, 2020. **53**(6): p. 4335-4385.

[14] Kumar, C.P. and L.D. Babu, *Novel text preprocessing framework for sentiment analysis*, in *Smart Intelligent Computing and Applications*. 2019, Springer. p. 309-317.

[15] Avinash, M. and E. Sivasankar, *A study of feature extraction techniques for sentiment analysis*, in *Emerging Technologies in Data Mining and Information Security*. 2019, Springer. p. 475-486.

[16] Duong, H.-T. and T.-A. Nguyen-Thi, *A review: preprocessing techniques and data augmentation for sentiment analysis.* Computational Social Networks, 2021. **8**(1): p. 1-16.

[17] Chiong, R., et al. *A sentiment analysis-based machine learning approach for financial market prediction via news disclosures*. in *Proceedings of the Genetic and Evolutionary Computation Conference Companion*. 2018.

[18] Porshnev, A., I. Redkin, and A. Shevchenko. *Machine learning in prediction of stock market indicators based on historical data and data from twitter sentiment analysis*. in *2013 IEEE 13th International Conference on Data Mining Workshops*. 2013. IEEE.

[19] Kordonis, J., S. Symeonidis, and A. Arampatzis. *Stock price forecasting via sentiment analysis on Twitter*. in *Proceedings of the 20th Pan-Hellenic Conference on Informatics*. 2016.

[20] Wu, D.D., L. Zheng, and D.L. Olson, *A decision support approach for online stock forum sentiment analysis.* IEEE transactions on systems, man, and cybernetics: systems, 2014. **44**(8): p. 1077-1087.

[21] Bourezk, H., et al. *Analyzing Moroccan Stock Market using Machine Learning and Sentiment Analysis*. in *2020 1st International Conference on Innovative Research in Applied Science, Engineering and Technology (IRASET)*. 2020. IEEE.

[22] Sohangir, S., N. Petty, and D. Wang. *Financial sentiment lexicon analysis*. in *2018 IEEE 12th international conference on semantic computing (ICSC)*. 2018. IEEE.

[23] Zhao, B., et al. *Stock market prediction exploiting microblog sentiment analysis*. in *2016 International Joint Conference on Neural Networks (IJCNN)*. 2016. IEEE.

[24] Sakphoowadon, S., N. Wisitpongphan, and C. Haruechaiyasak. *Probabilistic lexicon-based approach for stock market prediction: A case study of the Stock Exchange of Thailand (SET)*. in *2018 18th International symposium on communications and information technologies (ISCIT)*. 2018. IEEE.

[25] Turner, Z., K. Labille, and S. Gauch, *Lexicon-based sentiment analysis for stock movement prediction.* Journal of Construction Materials, 2021. **2**: p. 3-5.

[26] Yoon, B., Y. Jeong, and S. Kim, *Detecting a Risk Signal in Stock Investment Through Opinion Mining and Graph-Based Semi-Supervised Learning.* IEEE Access, 2020. **8**: p. 161943-161957.

[27] Kia, A.N., S. Haratizadeh, and S.B. Shouraki, *A hybrid supervised semi-supervised graph-based model to predict one-day ahead movement of global stock markets and commodity prices.* Expert Systems with Applications, 2018. **105**: p. 159-173.

[28] Wankhade, S.B., et al., *Hybrid model based on unification of technical analysis and sentiment analysis for stock price prediction.* INTERNATIONAL JOURNAL OF COMPUTERS & TECHNOLOGY, 2013. **11**(9): p. 3025-3033.

[29] Hsu, C.-M., *A hybrid procedure for stock price prediction by integrating self-organizing map and genetic programming.* Expert Systems with Applications, 2011. **38**(11): p. 14026-14036.

[30] Panday, H., et al., *Stock Prediction using Sentiment analysis and Long Short Term Memory.* European Journal of Molecular & Clinical Medicine, 2020. **7**(2): p. 5060-5069.

[31] Yujun, Y., Y. Yimei, and X. Jianhua, *A hybrid prediction method for stock price using LSTM and ensemble EMD.* Complexity, 2020. **2020**.

[32] Fathalla, A., et al., *Deep end-to-end learning for price prediction of second-hand items.* Knowledge and Information Systems, 2020. **62**(12): p. 4541-4568.

[33] Hosny, K.M., et al., *Fast computation of 2D and 3D Legendre moments using multi-core CPUs and GPU parallel architectures.* Journal of Real-Time Image Processing, 2019. **16**(6): p. 2027-2041.

[34] Salah, A., K. Li, and K. Li, *Lazy-Merge: A Novel Implementation for Indexed Parallel $ K $-Way In-Place Merging.* IEEE Transactions on Parallel and Distributed Systems, 2015. **27**(7): p. 2049-2061.

[35] Salah, A. and K. Li, *PAR-3D-BLAST: A parallel tool for searching and aligning protein structures.* Concurrency and Computation: Practice and Experience, 2014. **26**(10): p. 1705-1714.

[36] Al-Moalmi, A., et al., *A whale optimization system for energy-efficient container placement in data centers.* Expert Systems with Applications, 2021. 164: p. 113719.