2021 2nd International Conference on Big Data & Artificial Intelligence & Software Engineering (ICBASE) | 978-1-6654-2709-8/21/$31.00 ©2021 IEEE | DOI: 10.1109/ICBASE53849.2021.00127

Predicting Stock Market Movements Through Daily News Headlines Sentiment Analysis: US Stock Market

Yubo Bi[1]*
[1]Business school,
University of New South Wales Sydney,
Sydney, NSW, 2052, Australia
biyubo615@gmail.com

Hanting Liu[2]
[2]School of Information Science and Technology,
Xiamen University,
Xiamen, 361005, China

Ruiyang Wang[3]
[3]Nanjing Foreign Language School,
Nanjing,210018, China

Shiyou Li[4]
[4]Shanghai Pinghe School,
Shanghai, 201206, China

Abstract: This study leverages features extracted from daily world news headlines between August 2008 and June 2016 to predict next day's Dow Jones Industrial Average (DJIA) index movement. Other related index return, commodity price changes and trading volume data are used to improve the prediction accuracy. In this work, the predictive models are based on three machine learnings algorithms: Random Forest, Support Vector Machine and Naïve Bayes. Based on the prediction results of three predictive models, a majority vote is also conducted. The results show that: (1) general daily world news headlines are less correlated with next day's Dow Jones Industrial Average index movement; (2) sentiment implied in the news with indicators from technical analysis such as past return or trading volume can help to improve the prediction performance; (3) Naïve Bayes shows the superior ability in predicting next day's index change, while majority voting of different predictive models can further the prediction accuracy.

Keywords: Stock Market Prediction, Daily News Headlines, Machine Learning

I. Introduction

Stock market prediction has long been a challenge among investors, researchers and financial analysts. Nonetheless, this topic would be a continuous endeavor for stock market investors and financial institutions to maximize their investment returns. According to the efficient market hypothesis, the current stock prices of financial markets with weak form, semi-strong form and strong form efficiency have already reflected past price trends, all public information and all the information include public and private respectively [1,2]. That implies that a financial market with weak form efficiency, the analysis of past trend of stock returns is meaningless, meaning history is no indication of the future. Most recent stock price studies are conducted through two methods - fundamental analysis by predicting stock price through analyzing underlying businesses and forecasting future business performances, and technical analysis that predicting future stock prices based on past and present price trends [3,4].

However, the stock price movements are not only affected by its past trends or correlation with the financial markets, but also news, investors comments, current events or company announcements. Based on research, negative news has a significant impact over the Indian stock market [5]. Also, European Central Bank monetary policy announcements, sentiment implied in forward looking answers to the questions during press conferences significantly affect Euro stock market returns [6]. In addition, positive sentiment derived from political news will have a positive impact on the markets, while negative sentiment will have negative impacts [7]. Furthermore, publicly listed companies are required by the regulatory institutions to disclose any information that might have an impact on its stock price, which are also known as company announcement whose topics will have impacts on its stock price [8].

Nowadays, the improvement of internet influence not only the technical aspects of computer communications but also the whole society that include electronic commerce and community operations [9]. One of the most significant benefits offered by the internet evolution is information gathering. The internet speeds up the information communicated through the whole society which in turn also speeds up the reaction of stock prices. The internet also offers great quantity of information to conduct more comprehensive analysis of stock prices. As suggested by behavioral finance, the theoretical methodology to predict stock prices will include several sources of noise, considering not all investors are rationale. As a result, the information from Social Networking platforms will offer more effective indicators to predict stock price movements [10,11]. Because of such hard-to-quantify factors that influence stock prices, researchers start using machine learning techniques and sentiment analysis of the text information to predict stock prices [12,13]. Moreover, the evolution of Natural Language Processing offers the possibility to investigate the impact of news information on the stock market movements, while this strong technique is still less applied in finance field [14].

In this research, we derive the sentiment and topics implied

978-1-6654-2709-8/21/$31.00 ©2021 IEEE
DOI 10.1109/ICBASE53849.2021.00127

by each of daily news headlines for the purpose of predicting next day's DJIA index movement. We also incorporate past price trend and trading volume of S&P500, NASDAQ and Gold as variables to test whether they can improve the prediction results. Finally, we explore whether a specific standalone model can outperform an aggregate model in stock market movement prediction.

II. Literature Review

A. NLP based Sentiment analysis

In the realm of data science, NLP (Natural language processing) is one of theory-motivated machine learning technique that represents transformation of human-language and computer languages. Since the inception of NLP in the 1950s, previous research has been focusing on tasks such as machine translation, information retrieval, text summarization, information extraction, topic modeling, and more recently, opinion mining [5]. To understand the view or opinion implied in the text, NLP technique is essential for understanding the opinions or emotions that implicated or expressed in the text, which is a current research focus on the field of data science [15]. Under the scope of NLP, sentiment analysis is a widely used method to classify the subjective statements in the text, interchangeably throughout the document. It uses NLP to collect and examine opinion or sentiment contained in text while classifies textual data into categories, usually positive, negative, and neutral sentiments [16]. Kanakaraj and Guddeti proposed a system to gather textual data from the social network Twitter and to extract features from the tweets by leverages NLP techniques [15]. To increase the prediction accuracy, WordNe and WordNet Word Sense Disambiguation synsets are applied in the feature vector. All previous studies of sentiment analysis had a due to the complexity of dealing with raw data. Text preprocessing is a crucial step in sentiment analysis which is useful for preparing unstructured data for features extraction or information retrieval [17]. There are numerous previous approaches that based on sentiment analysis techniques to predict stock market trends. Some techniques use press releases or press conference transcripts to predict the next day's price change. While and others depend on sentiment extracted from social media sites such as Twitter, Reddit, or Facebook.

B. News Sentiment Analysis

The study proposed by Khedr leverages financial news related to stock markets, companies' news and financial reports [16]. The textual financial news, along with features extracted from historical stock prices are to predict the future behavior of the stock market using sentiment analysis. The prediction model uses Naïve Bayes as well as the k nearest neighbor (K-NN) technique which achieved an accuracy score of 89.80%. Cambria and White investigate that market participants will pick up on any revealed cues that allow them to draw inferences about the possible future path of monetary policy [18]. The study is mainly based on the word lists that contain both positive words and negative words which are deliberately selected for this particular study. Furthermore, based on grammatical and syntactical cues, the application of VADER allows adjustment of the sentiment scores of the text. Besides, three heuristics based on social media text are identified and are generalizable to intensifiers, contrastive conjunctions, and negations. Therefore, a relatively up-to-date and easy-to-adapt approach was proposed. Paramanik and Singhal use NRC word-emotion association lexicon constructed sentiment indices. In the meanwhile, an opposed augmented asymmetric GARCH model is proposed where the dominance of two traders' contradictory sentiments leaked in the text [5]. Costola indicated in their study that the variance of the sentiment and the volume of the news sources for Reuters and MarketWatch are negatively associated with market returns, suggesting that an increase in uncertainty results in an adverse impact on the stock market [19]. Shah developed a dictionary-based sentiment analysis model and a sentiment analysis dictionary for the financial [20].

C. Social Media Sentiment Analysis

Batra and Daudpota determined the positive relationship between people's opinions or attitude and market movement [21]. They perform sentiment analysis on tweets related to Apple products. The textual data are extracted from Stock Twits, and the sentiment scores are calculated through SVM algorithm, then are categorized as bullish or bearish, namely positive or negative. The presented model has an accuracy of 76.65% in stock prediction. Nguyen and Shirai introduced a new topic model TSLDA with a new feature that captures topics and their sentiments simultaneously to the proposed model [22]. The accuracy of the model in this study is better than LDA and JST-based methods by 6.43% and 6.07%, respectively. The results of this study indicate the positive influence of incorporation of sentiment data in social media on stock future behavior prediction. Sul analyzed the relationship between the sentiment in tweets about a specific firm from users with less than 171 followers (the median in the sample) and the stock's returns on the next trading day, as well as the next 10 days and 20 days [23]. The findings provide traders a profitable trading strategy idea that can produce around 15% annual return.

III. Data

News data and stock data were drawn from the period from August 12, 2008, to June 30, 2016. Figure 1 shows the number of different types of data collected.

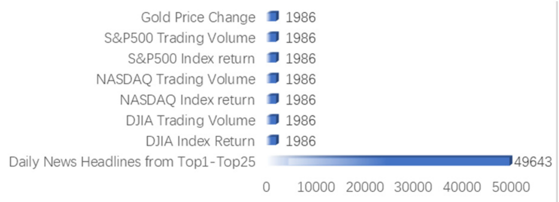

Figure 1 Dataset count (This chart represents the aggregate counts of our data sources. Index return, commodity price change and trading volume data corresponds to 1,986 trading days. News dataset contains total 49,643 news headlines.)

A. News data:

Roughly 50 thousand daily news headlines were provided by Reddit World News Channel, the largest forum on reddit that contains the latest news articles posted by redditors. Notably, US News and internal American politics are barred from being posted in this subreddit, between August 12, 2008, and June 30, 2016. They are ranked by reddit users' votes, and only the top 25 headlines are considered for a single date.

B. DJIA index data:

Stock data is represented by Dow Jones Industrial Average (DJIA) between August 12, 2008, and June 30, 2016, sourced from Yahoo Finance, a website that provides financial news, data and commentary including stock quotes, press releases, financial reports and original content.

C. Other related data:

Historical price data and trading volume for S&P 500, NASDAQ and Gold are derived from FactSet platform (https://www.factset.com/).

According to Houlihan and Creamer, message from Stock Twits was successfully used to predict future asset price movements [13]. In this research, we consider more general and original news topic that are not specifically related to any specific stock, rather general headlines. General news may affect trading actions of both rational and irrational investors which in turn may have more promising performance in predicting stock market movements [24]. By leveraging indicators from both news sentiment and technical analysis can actually improve the overall results regarding predicting stock price movements for three Japanese companies listed at US stock exchange [25]. However, in this research, we are considering the US financial market as a whole and using more extended period that corresponds to 1,986 trading days rather than two years in previous research.

IV. Methodology

The goal of the proposed model is to predict the stock market trends through predicting the DJIA index is either rising or falling. The proposed model combines the analysis of the most popular headlines from Reddit and the historical prices and compares different approaches to boost the accuracy of the prediction result. To achieve the required target, the following sections are included as depicted in Figure 2. 1) Initially, we gathered texts in daily news headlines from Reddit as well as DJIA index, other related indices and commodity price data from aforementioned sources. 2) the text message is preprocessed using Natural Language Toolkit (NLTK) to filter the text for feature extraction. The 25 headlines with the highest heat were integrated one for each day. 3) The next step is taking two different approaches to extract features from the text. 4) The fourth step is applying classification algorithms to construct predictive models. Selected algorithms are Random Forest (RF), Naive Bayes (NB), Support Vector Machine (SVM), and Voting Classifier which is used to conduct majority voting. 5) Lastly, after the predictive models are constructed and validated, we compare results obtained by two different feature extraction methods.

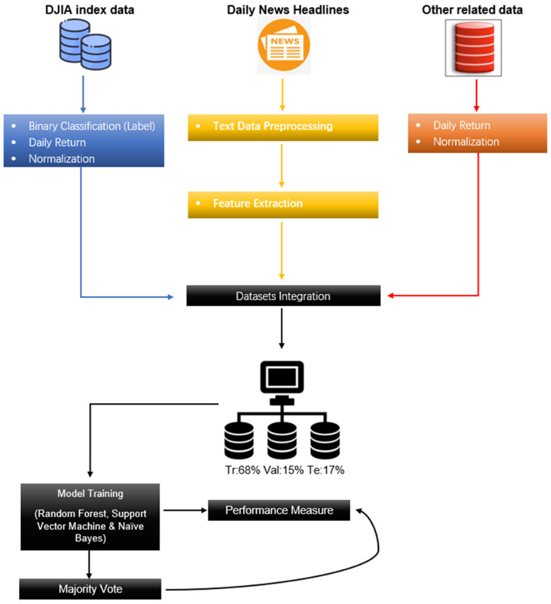

Figure 2 Process Flowchart (This chart shows the flow of combining the analysis of the most popular headlines from Reddit and the historical prices and comparing different approaches to boost the accuracy of the prediction result)

A. Dataset Pre-processing

1) DJIA Index

Since what we try to accomplish is to classify the movement direction of the overall stock market, we conducted a binary classification to label our DJIA index data according to the movement direction. Figure 3 shows the number of days index is decreasing and the number of days index is increasing or remaining stable, which are relatively evenly distributed.

- '1' when DJIA Adj Close value increased or stayed as the same.
- '0' when DJIA Adj Close value deceased.

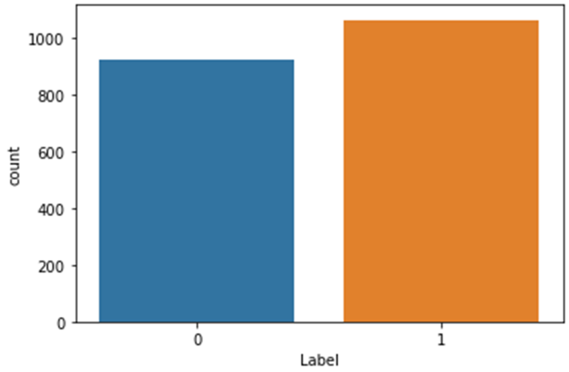

Figure 3 Label Count (This chart shows the number of days index is decreasing and the number of days index is increasing or remaining stable)

2) Other Related Index and Commodity price Data Pre-Processing

The classification task falls under the time series prediction problems since the movement of the overall stock market can be affected by the news headlines from the previous day and even the day before. Features such as past return and trading volume can cause a similar impact on the stock market as well. Therefore, the issue underlying the time series task would be that the data we gathered from the day merely have pertinence of stock market movement on the day. Our solution to prevent time lag is to shift the headlines in the dataset backward for a day as headlines should be used to predict the market trends of the next day. In the meanwhile, the return and trading volume of either DJIA and other related indices from the day before and two days ago are calculated and assigned to t, t-1, t-2. The steps included in pre-processing of related indices and generation of suitable features is 1) calculate the percentage of return of each trading day. 2) normalize the trading volume of each trading day.

For generating the rate of return, the formula is shown below.

$$\text{Rate of Return(r)} = \frac{Closing\ Value - Initial\ Value}{Initial\ Value} \times 100\% \quad (1)$$

The distribution of the volume data is not following a Gaussian distribution as Naive Bayes classification requests, hence the data should be rescaled to ranging between 0 and 1. We normalized the trading volume data by calculating the percentage change in trading volume relative to previous day. The results are further divided by 10 to mitigate the large volatility of trading volume changes and ensure consistency of scale.

$$\%\ change\ in\ Trading\ Volume = \frac{Trading\ Volume_t - Trading\ Volume_{t-1}}{Trading\ Volume_{t-1}} / 10 \quad (2)$$

3). Textual Data Pre-Processing

We can only move on to extract features from the previously combined text from headlines after it has been pre-processed. The textual data undergo the following processes:

1) Text Cleaning: all upper-case letters in the text are converted to lower case letter when cleaning the data.

2) Removing Null Value and meaningless letter: the letter ‘b’ in the head of each text is removed as well as the null value.

3) Removing Stop-words (where applicable): when applying to remove stop-words in Natural Language ToolKit (NLTK), each word in the list of words is compared to the dictionary, Words that do not have a significant meaning in the documents such as the, a, of are removed to reduce the number of features and improve the performance.

4) Lemmatizing: lemmatization from Wordnet would consider the content, identify the base form of words and convert them back to base form while stemming is the process of producing variations of the root of a word. Compare to stemming, lemmatization looks at the surrounding text to determine a given word’s part of speech. Generally speaking, both stemming and lemmatizing are applicable in the data pre-processing session. However, applying stemming may lead to an issue when the stemmed word will not be in the form of a word that can be calculated sentiment scores using VADER. Applying lemmatizing instead of stemming has been proven in this case is a better choice when pre-processing data.

5) Dictionary Check: We checked the news texts against dictionary to removing any meaningless words or wrong words.

B. Feature Extraction

Two different approaches are applied to extract features from textual data in the dataset, the first is adopting Count Vectorizer, the second is to calculate sentiment scores.

1) Count Vectorizer

Count Vectorizer is used to transform a given text into a count of vector-based frequency of each word that occurs throughout text, thus enables the pre-processing of text data before generating the vector representation. As result, a vectorized matrix is generated by Count Vectorizer, in which each unique word is represented by a column of the matrix.

2) 4Calculate Sentiment Scores

We used VADER (Valence Aware Dictionary and sentiment Reasoner) as a method to extract features by calculating sentiment scores implied in the headlines. VADER is a sentiment analysis toolkit that is specifically attuned to the text message sourced from social media. One benefit offered by VADER is that it will indicate both polarities, namely positive and negative, and intensity of emotion. VADER uses a combination of a sentiment lexicon which is a list of lexical features that are generally labeled according to their semantic orientation. The sentiment score of each headline is calculated by summing up the scores of each VADER-dictionary-listed word in the sentence, mapping the scores of categories of positive, negative, neutral, and deriving a compound sentiment score.

C. Proposed Model:

After the pre-processing of data and features extraction, two prediction models that share the same classification algorithms were constructed corresponding to two approaches to extract features. The original dataset was split into the train, test, and validation sets. The train set is at 68%, while the test set is at 17% and the validation set is 15%. In the first prediction model, we try doing nothing but using Count Vectorizer to transform text to a matrix. The second integrated prediction model combines sentiment score for each daily news headline along with other features to investigate their influence on stock market raising and falling.

1) Random Forest (RF)

Random forest is a combination of multiple decision trees. Decision trees often produce overfitting problems. Random forest has a total of m features and randomly selects k features to form n decision trees. For random forest is composed of numerous decision tress, Random variables are passed to decision tree to predict the outcome of each decision. It stores all the prediction results and uses the prediction target with a high number of votes as the final prediction.

2) Gaussian Naïve Bayes (GNB)

On accout of the simplicity and speed of Naïve Bayes classifier in textual classification, it can be adopted to predict the polarity of text. In scikit-learn libraries, there are 3 Naïve Bayes classification algorithm classes, which are Gaussian, Multinomial, and Bernoulli. Among them, we choose to apply Gaussian, which implements the classification by assuming the likelihood of the features to be Gaussian and continuous.

$$P(x_i|y) = \frac{1}{\sqrt{2\pi\sigma_y^2}} exp(\frac{(x_i-\mu_y)^2}{2\sigma_y^2}) \qquad (3)$$

3) Support Vector Machine (SVM)

SVM is a two-class linear classifier in the feature space. In SVM classification process, data is conversed to n-dimensional point in n-dimensional space, where n is equals to the count of features that extracted during preprocessing and the value equals to a particular coordinate. The classification of SVM is performed by finding hyper-plane in order to differentiate accurately. As an output, an optimal hyper-plane that classifies test data is generated by learning the training data.

4) Voting Classifier

We are hoping that combining decisions of different models can improve the overall performance of the proposed model. The Voting Classifier performs multiple computations for each prediction result, and then verify that a majority of the results agree. This classifier model can multiple times combine parallel outputs and returns the element which is the majority in the prediction results of each other models.

5) Grid Search CV for model tunning

Gird Search CV is applied to return the best parameters for the classification model, and then apply this parameter in our models to get better performances. Gird Search CV works by giving the optimized results and parameters when input the parameters. Within the specified parameter range, the parameters are adjusted step by step, and the parameter with the highest accuracy on the verification set is found from all the parameters. Gird Search CV can be divided into Grid Search and CV, namely grid search for best parameters and k-fold cross-validation.

D.Performance Evaluation Criteria

The classification models that we construct during the implementation stage were evaluated by calculating the precision, recall, F1 score, and accuracy. Results are presented in the form of a confusion matrix as Table 1 shows, true positives (TP) and true negatives (TN) are corresponded to correctly predicted positive and negative changes in next day's index movement. While false positives (FP) and false negatives (FN) represent incorrectly predicted tuples.

Table 1 Confusion matrix for TP, TN, FP, FN

Actual Predict	0	1
0	TN	FN
1	FP	TP

Below shows the explanations and equations for calculating the 4 performance evaluation criteria.

1) Precision

The accuracy is resolved in a sample that is identified as a positive category and is the proportion of the positive category.

$$Precision = \frac{TP}{TP + FP} \qquad (4)$$

2) Recall

Focus on the accuracy of correctly predicting the positive results out of total positive results.

$$Recall = \frac{TP}{TP + FN} \qquad (5)$$

3) F1score

F1 scores are the weighted average of accuracy and recall rates. β is used to balance the weights of Precision and Recall in F-score calculation. β is generally signed to be 1.

$$F1\ Score\ =\ (1\ + \beta^2) \times \frac{Presision \times Recall}{\beta^2 \times (\ Precision + Recall)} \qquad (6)$$

4) Accuracy

Accuracy is an indicator used to evaluate a classification model, i.e., the proportion of the total correctly predicted result that include both positive and negative changes.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \qquad (7)$$

V. Results

Table 2 Predictive model performance based on Count Vectorizer

Algorithms	Precision	Recall	F1Score	Accuracy
Random Forest	0.55	0.54	0.54	0.54
SVM	0.28	0.53	0.37	0.53
NaÏve Bayes	0.56	0.55	0.55	0.55
Voting Clasifier	0.59	0.57	0.55	0.57

Table 3 Predictive model performance based on sentiment score, past return and trading volume

Algorithms	Precision	Recall	F1Score	Accuracy
Random Forest	0.55	0.55	0.52	0.55
SVM	0.28	0.53	0.37	0.53
NaÏve Bayes	0.58	0.57	0.57	0.57
Voting Clasifier	0.61	0.58	0.53	0.58

This section illustrates the prediction results and discussions of this study. Table 2 and 3 shows the performance (Precision, Recall, F1 Score and Accuracy) of three different predictive models that are used in this research based on two different feature extraction method.

DJIA index movement prediction based on Count Vectorizer shows 54% accuracy for Random Forest, 53% for Support Vector Machine and 55% for Naïve Bayes. This result suggests that the specific word contained in the daily world news headlines does not have much prediction power to the next day's DJIA index movement. However, by considering the news sentiment with past return and trading volume of DJIA, S&P500, NASDAQ and gold, the prediction accuracy has improved to 55% for Random Forest and 57% for Naïve Bayes. This result suggests that leveraging both sentiment and variables from technical analysis can improve the outcome of predicting next day's DJIA index movement.

While the prediction accuracy of SVM have remained unchanged at 53% despite the different input features. In addition, the precision (28%), recall (53%), F1 Score (37%) and accuracy (53%) for SVM are consistently lower than other predictive models for both methods. The resulting low performance is caused by its inability to predict negative change for the DJIA index, which means SVM is least applicable to our input features. On the contrary, Naïve Bayes shows the most superior performance with highest precision, recall, F1 score and accuracy in both methods.

The performance of Voting Classifier which conducts a majority vote based on prediction results of three different models has further improve the precision, recall, F1 score and accuracy in both methods. Thus, despite the strongest performance of Naïve Bayes predictive model, the overall stock market prediction outcome can still be improved by combining decisions from different classifiers and reducing variance of estimation errors [26].

VI. Conclusion

In conclusion, this research utilizes daily world news headlines to predict next day's DJIA index movement with respect to whether it will increase or decrease. In addition, we incorporate other related stock index and commodity price data such as S&P500, NASDAQ and gold to improve the overall prediction outcome. Based on specific word contained in each daily news headline, the prediction result shows less correlation between the DJIA index and the daily world news headlines. Based on the sentiment implied in each daily news headline and some features from technical aspect, the prediction power has been improved. Naïve Bayes predictive model shows the most superior prediction power in both feature extraction methods. Furthermore, the prediction results can be improved by conducting a majority voting of three predictive models that are used in this research.

Nonetheless, the overall prediction accuracy remains relatively low despite the improvements. Therefore, we draw conclusion that the daily world news headlines are weakly correlated with next day's movement direction of DJIA index. The reasons can come from several aspects. Firstly, news information in our dataset contains not only news from US but also news from countries worldwide. In addition, the news information is related to wide dispersed topics which results in large noises. Secondly, the derived sentiments are largely composed with negative or neutral attitude which furthers the prediction difficulty.

Thus, future research will be based on information that are focused more specifically on the financial markets such as posts on Stock twits. On the other hand, this research only shows the low level of correlation between daily world news headlines and next day's index movement. However, the US stock market may absorb the information quickly once the news occurred or released.

References

[1] Fama, E. (1970). Efficient Capital Markets: A Review of Theory and Empirical Work. *The Journal Of Finance*, *25*(2), 383. doi: 10.2307/2325486

[2] Malkiel, B.G. (1989). Efficient Market Hypothesis. Eatwell J., Milgate M., Newman P. (eds) Finance. The New Palgrave. Palgrave Macmillan, London. https://doi.org/10.1007/978-1-349-20213-3_13

[3] Nti, I., Adekoya, A., & Weyori, B. (2019). A systematic review of fundamental and technical analysis of stock market predictions. *Artificial Intelligence Review*, *53*(4), 3007-3057. doi: 10.1007/s10462-019-09754-z

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

[5] Paramanik, R., & Singhal, V. (2020). Sentiment Analysis of Indian Stock Market Volatility. *Procedia Computer Science*, *176*, 330-338. doi: 10.1016/j.procs.2020.08.035

[6] Möller, R., & Reichmann, D. (2021). ECB language and stock returns – A textual analysis of ECB press conferences. *The Quarterly Review Of Economics And Finance*, *80*, 590-604. doi: 10.1016/j.qref.2021.04.003

[7] Suleman, M. T. (2010). Stock Market Reaction to Good and Bad Political News. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.1713804

[8] Ratku, A., Feuerriegel, S., & Neumann, D. (2014). Analysis of How Underlying Topics in Financial News Affect Stock Prices Using Latent Dirichlet Allocation. *SSRN Electronic Journal*. doi: 10.2139/ssrn.2529457

[9] Internet Society. (2021, June 1). Brief History of the Internet. https://www.internetsociety.org/internet/history-internet/brief-history-internet/.

[10] Chen, W., Cai, Y., Lai, K., & Xie, H. (2016). A topic-based sentiment analysis model to predict stock market price movement using Weibo mood. *Web Intelligence*, *14*(4), 287-300. doi: 10.3233/web-160345

[11] Li, B., Chan, K., Ou, C., & Ruifeng, S. (2017). Discovering public sentiment in social media for predicting stock movement of publicly listed companies. *Information Systems*, *69*, 81-92. doi: 10.1016/j.is.2016.10.001

[12] Khedr, A. E., S.E.Salama, & Yaseen, N. (2017). Predicting Stock Market Behavior using Data Mining Technique and News Sentiment Analysis. *International Journal of Intelligent Systems and Applications*, *9*(7), 22–30. https://doi.org/10.5815/ijisa.2017.07.03

[13] Houlihan, P., & Creamer, G. (2019). Leveraging Social Media to Predict Continuation and Reversal in Asset Prices. *Computational Economics*, *57*(2), 433-453. doi: 10.1007/s10614-019-09932-9

[14] Chahine, S., & Malhotra, N. (2018). Impact of social media strategies on stock price: the case of Twitter. *European Journal Of Marketing*, *52*(7/8), 1526-1549. doi: 10.1108/ejm-10-2017-0718

[15] Kanakaraj, M., & Guddeti, R. M. R. (2015, March). NLP based sentiment analysis on Twitter data using ensemble classifiers. In *2015 3Rd international conference on signal processing, communication and networking (ICSCN)* (pp. 1-5). IEEE.

[16] Khedr, A. E., S.E.Salama, & Yaseen, N. (2017). Predicting Stock Market Behavior using Data Mining Technique and News Sentiment Analysis. *International Journal of Intelligent Systems and Applications*, *9*(7), 22–30. https://doi.org/10.5815/ijisa.2017.07.03

[17] Gonçalves, T., & Quaresma, P. (2005). Evaluating preprocessing techniques in a text classification problem. *São Leopoldo, RS, Brasil: SBC-Sociedade Brasileira de Computação*.

[18] Cambria, E., & White, B. (2014). Jumping NLP curves: A review of natural language processing research. *IEEE Computational intelligence magazine*, *9*(2), 48-57.

[19] Costola, M., Nofer, M., Hinz, O., & Pelizzon, L. (2020). Machine Learning Sentiment Analysis, Covid-19 News and Stock Market Reactions. *SSRN Electronic Journal.* https://doi.org/10.2139/ssrn.3690922

[20] Shah, D., Isah, H., & Zulkernine, F. (2018). Predicting the Effects of News Sentiments on the Stock Market. *2018 IEEE International Conference on Big Data (Big Data)*, 4705-4708. doi: 10.1109/BigData.2018.8621884

[21] Batra, R., & Daudpota, S. M. (2018). Integrating StockTwits with sentiment analysis for better prediction of stock price movement. *2018 International Conference on Computing, Mathematics and Engineering*

Technologies (ICoMET). https://doi.org/10.1109/icomet.2018.8346382

[22] Nguyen, T. H., & Shirai, K. (2015, July). Topic modeling based sentiment analysis on social media for stock market prediction. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)* (pp. 1354-1364)

[23] Sul, H. K., Dennis, A. R., & Yuan, L. (2017). Trading on twitter: Using social media sentiment to predict stock returns. *Decision Sciences*, *48*(3), 454-488.

[24] Hirshleifer, D. (2001). Investor Psychology and Asset Pricing. *SSRN Electronic Journal*. doi: 10.2139/ssrn.265132

[25] Deng, S., Mitsubuchi, T., Shioda, K., Shimada, T., & Sakurai, A. (2011). Combining Technical Analysis with Sentiment Analysis for Stock Price Prediction. *2011 IEEE Ninth International Conference on Dependable, Autonomic and Secure Computing*, pp. 800-807, doi: 10.1109/DASC.2011.138.

[26] Kim, M., Min, S., & Han, I. (2006). An evolutionary approach to the combination of multiple classifiers to predict a stock price index. *Expert Systems With Applications*, *31*(2), 241-247. doi: 10.1016/j.eswa.2005.09.020