Proceedings of the 5th International Conference on Computing and Data Science
DOI: 10.54254/2755-2721/29/20230903

Can natural language processing accurately predict stock market movements based on Reddit World News headlines?

Chen Cao[1, †], Xuang Yu[2,4, †] and Caiyi Qian[3, †]

[1]College of Art and Science, New York University, New York, 10003, United States, cc6963@nyu.edu
[2]College of Letters and Sciences, University of California Santa Barbara, 93106, United States, xuangyu@ucsb.edu
[3]College of Art and Science, University of San Francisco, San Francisco, 94117, United States, sophiqian2002@gmail.com

[4]xuangyu@ucsb.edu

[†]All authors contributed equally to this work and should be considered as co-first authors.

Abstract. This research examines the application of machine learning and natural language processing (NLP) methods to stock market movement forecasting. Many NLP approaches were used to gather and preprocess Dow Jones Industrial Average (DJIA) data and Reddit Global News headlines. The preprocessed data were then used to train three machine learning algorithms (Random Forest, Logistic Regression, and Naive Bayes) to forecast the daily trend of the DJIA. According to the study, the Naive Bayes algorithm, along with Textblob, fared better than the other two models, obtaining an accuracy of 68.59%, which is an improvement above previous research. These findings show how NLP and machine learning may be used to forecast stock market patterns and offer ideas for further study to boost the precision of these models.

Keywords: natural language processing, text opinion mining, stock market, DJIA, sentiment analysis, Reddit news, machine learning.

1. Introduction

Since the stock market plays a tremendous role in economic markets, predicting stock movements has become a trendy and challenging research topic. It is undeniable that traders face numerous risks when investing in stocks due to their uncertainties and complexity. There is a clear connection between stock market movements and stock investors' profitability. Developing models that can accurately predict the stock market is therefore required. The more accurate the model's prediction is, the less risky the investor will be and the higher the return will be.

The interplay between computer language and human language is the focus of the artificial intelligence (AI) field known as natural language processing (NLP). It makes it possible for computers to comprehend, translate, and create natural languages like English. In recent years, NLP has been increasingly utilized to analyze and predict trends and movements in the stock market. Using NLP to predict stock market behavior is a highly popular area of research, and it offers valuable insights into

DOI: 10.54254/2755-2721/29/20230903

sentiment, events, and financial data by processing massive volumes of textual-based data. The reason is that NLP frequently adopts machine learning algorithms to process and analyze large quantities of textual data. For example, in the task of sentiment analysis, machine learning algorithms typically perform training on data sets of large text documents with the labeled sentiment (e.g., positive, negative, neutral) to make predictions about the sentiment of previously unknown text data.

This paper will perform natural language processing on headlines from Reddit news, a go-to source for breaking news and updated events that contain influential platforms with a huge following of active users from all over the world, and use machine learning algorithms to predict fluctuations in stock market indices. Our method entails gathering and examining a sizable number of news headlines in order to determine the primary factors influencing stock price changes. We extract pertinent information from the text data using sophisticated NLP techniques, such as sentiment analysis and utilize it to train machine learning models. These models are then used to make predictions about stock market trends and make investment recommendations.

This research paper seeks to determine the degree of the association between news headlines and stock prices, as well as if news headlines have any predictive effect on the stock market. In this research, we will perform sentiment analysis of Reddit news headlines and use calculated sentiment scores to train different machine learning models in order to predict stock market movements. The results of this study will add to the body of knowledge on how news media and financial markets interact, and they may have implications for investors and financial analysts.

2. Literature review

Focused on using random-forest to predict future stock prices by using historical stock prices. A novel stock sequence array convolutional neural network model that focuses on feature extraction and price trend prediction using financial time series was also suggested by [1, 2]. Due to technological development leading to diversification and transparency of news media, it is gradually found that the method of analyzing historical stock prices to predict future stock prices cannot predict the market accurately. Thus, researchers consider more factors that affect stocks in the process of making models.

One of the most important sources of stock information is social media and news, which can provide the latest news about a company's activities, such as expansions, new products, and new policies. With the rise of social media, investor sentiment is becoming an increasingly significant element influencing the direction of the stock market: investors are increasingly willing to participate in stock reviews online. Sentiment analysis can be used to classify stock comments and thus predict stock market trends [3]. found that with a mean percentage error (MAPE) reduction of more than 6%, assessing the public mood from tweets had an accuracy of 86.7% in forecasting the daily change in the closing value of the Dow Jones Industrial Average. Additionally [4], mined user sentiment in Tweets and used Natural Language Processing to build a model to determine if sentiment is a proxy for stock price movements; the results show that the model is able to predict the stock market closing price with 76.12% accuracy. Moreover, according to [5], the prediction models were built using the Naive Bayes and Random Forest algorithms, as well as linear regression approaches, and had coefficients of determination of 0.9989 and 0.9983 for the best predictions. Furthermore, [6] proposed a data-driven pipeline that combines recent Twitter information about a firm into a time series stock price prediction model.

However, [7] emphasizes that as Twitter is frequently used by unscrupulous users to promote or disparage goods, services, ideas, and ideologies, which can negatively impact the economy and cause financial losses, quality control mechanisms must be in place before using Twitter data. This is why reliable and accurate news sources of information are important. [8] introduced a technique for mining textual opinions to assess Korean news and forecast changes in the KOSPI (Korea Composite Stock Price Index) and proved news sentiment can be utilized to forecast changes in stock prices. [9] analyzed 25 news headlines for each public market between 2008 and 2015 and predicted the end-of-day value of the DJIA index on the same day [7]. By performing a news sentiment analysis, a Polarity Dictionary was built, which predicted short-term stock price movements with an accuracy of 70.59%.

DOI: 10.54254/2755-2721/29/20230903

3. Data

3.1. Introduction

This paper combines NLP and machine learning methods in order to forecast stock market movements. We made use of information from the news and stock market data, respectively. The news headlines are achieved from Reddit World News, one of the biggest and most well-liked subreddits, as a part of the dataset. As of February 2023, Reddit news had more than 24 million subscribers. Using Reddit Global News headlines to predict stock market movements is common due to its accuracy in reflecting public emotions and tracking breaking news. The top 25 daily news headlines from the Reddit World News Channel (/r/worldnews) from August 8, 2008, to July 1, 2016, were used to compile statistics for these news headlines. The Dow Jones Industrial Average (DJIA), which was also compiled from August 8, 2008, to July 1, 2016 (totaling 1,990 trading days), provides the stock market data used in this study. The DJIA is a stock market index composed of the top 30 prominent companies that are widely used to gauge the performance of the US stock market.

3.2. Data collection

In this study, secondary data sources were predominantly used as a method for data collection. We collected both news headlines and DJIA data from Kaggle, a prominent online platform where data scientists and machine learning enthusiasts share ideology and datasets. The Reddit news headlines contain two categories: the first column is the date, the second through the 26th columns are the news headlines, and all news headlines are sorted by popularity.

The DJIA data contains the opening and adjusted closing prices for each trading day from 2008 until 2016. The information was further prepared to generate a new column that denoted whether the DJIA Adjusted Close was up (denoted by the symbol "1") or down (marked as "0"). We combined both datasets, and the example of our combined Data set is shown below in Table 1.

Table 1. Combined data set sample.

Date(1989 days)	DJIA indication	News Title 01	News Title	News Title 25
8/8/2008	0 (down)	New's Headline 01	…	New's Headline 25
…	…	…	…	…
7/1/2016	1(up or zero)	New's Headline 01	…	New's Headline 25

Note: This table is an example of how the data was pre-processed for this research. Date column represents the trading date, DJIA indication (simply the sign of PriceDiff :1 indicates stock was up and 0 indicates stock was down), and different News headlines for each respective trading day. This database resulted in 1,990 trading days and 49,750 (1990*25) news titles.

4. Methodology

4.1. Data preprocessing

Before processing the headline data, a number of preprocessing steps were performed to clean and prepare the text data. Stop words, punctuation, and numerals were taken out of the Reddit news headlines data before it was ready for analysis. The text was then tokenized, stemmed, and lemmatized.

The steps include:

1) Tokenization**:** breaking the text data into individual words or tokens.
2) Stopword removal**:** removing common words like "and," "the," and "is" that does not carry much meaning in the text.
3) Stemming**:** deleting suffixes to return words to their original form.

DOI: 10.54254/2755-2721/29/20230903

4) Lemmatization**:** reducing the words to their base form by using a dictionary to map words to their root form.

To minimize the dimensionality of the dataset and extract the pertinent data that could forecast stock price changes, tokenization, stemming, and lemmatization was employed, and the following Figure 1 shows the steps included in the data preprocessing.

Figure 1. Data preprocessing.

Note: This figure includes the steps in data preprocessing.

4.2. Sentiment score

One approach used frequently in NLP to determine the polarity of textual material is sentiment analysis. In this study, sentiment analysis using the Valence Aware Dictionary and Sentiment Reasoner (VADER), Sentiment, and TextBlob was done on the news headline data to determine a sentiment score between -1 and 1, signifying the sentiment polarity of the text. As can be seen in Table 2., Positive values suggest positive Sentiment, negative values suggest negative Sentiment, and 0 indicates neutral Sentiment based on the Sentiment conveyed in the text.

A rule-based program called VADER Sentiment can analyze text to ascertain the polarity of the Sentiment. Contrarily, TextBlob employs a machine learning system to categorize text into different sentiment groups.

To find any variations between the sentiment analysis performed by the two technologies, we shall compare their results. We seek to gain a more thorough understanding of the Sentiment expressed in the headlines, which we will utilize to train machine learning algorithms by employing two separate NLP tools.

Table 2. Combined data set sample.

Index	Date	News	Sentimentscore
284	10/29/2014	For 23rd time, U.N. na…	0.991
1893	9/20/2011	A US court has overtu…	0.9906
436	11/26/2010	the better for it Video…	0.9208
228	10/20/2009	b’ “Legalize it” lobby …	0.9027
878	3/18/2014	Putin announces that R…	0.8748
393	11/19/2008	b’ Bush ignored by the ...	0.8671

Note: This table shows the part of VADER Sentiment score.

4.3. Data transformation

In our research, we applied the term-frequency inverse document frequency (TF-IDF) transformation algorithm (formula 1), a method for vectorizing a set of text documents into a matrix of numerical data. The significance of each word in a document is considered using this method based on the frequency of the term both within the document and across the full corpus. In this study, a document-term matrix was produced from the Reddit news headlines data using the TfidfVectorizer function from the scikit-learn Python API. Each column in the matrix represents a distinct word or token that appears in the corpus, and each row represents a document (such as a news headline). The tf-idf scores for each word in each document are represented by the values in the matrix. Following that, machine learning algorithms for forecasting stock market trends were trained using the generated document-term matrix as input.

Formula 1.

DOI: 10.54254/2755-2721/29/20230903

$$w_{i,j} = tf_{i,j} \times log\left(\frac{N}{df_i}\right)$$

Note: TF-IDF = TF (term, document) * IDF (term) where TF (term, document) = (number of times the term appears in the document) / (total number of terms in the document) and IDF (term) = log_e (total number of documents/number of documents containing the term). In this formula, TF (term, document) represents the frequency of terms normalized in the document, and IDF (term) represents the inverse document frequency of the term in the whole corpus. The logarithmic function is used to reduce the impact of very frequent terms in the corpus.

4.4. Machine learning

Reddit news headlines data is vectorized, and thus the resulting feature matrix is divided into training and testing sets, with 80% of the data utilized for training and 20% for testing (holdout). We will be using various machine learning models, such as random forest, logistic regression, and naive bayes, to assess the effectiveness of our methods. We can compare the accuracy of each model by training and testing it on the same dataset, allowing us to choose the model that performs the best for our investment suggestions.

4.4.1. Random forest. An ensemble learning approach called the random forest model from [10] combines different decision trees to increase prediction accuracy. It is a popular option for a range of applications, including finance and marketing, because it is well-suited for managing high-dimensional data with noise and outliers.

4.4.2. Logistic regression. An effective linear model for binary classification tasks is logistic regression. It is straightforward and effective, and multiclass classification tasks can be simply added to it.

4.4.3. Naïve bayes. The probabilistic machine learning algorithm Naive Bayes is frequently employed for text categorization applications. With a lot of data, it can do text categorization tasks with a high degree of accuracy. In many text classification applications, such as sentiment analysis, spam detection, and document categorization, naive Bayes models are frequently utilized.

In view of the above, we can identify which machine learning algorithm offers the most accurate forecasts for stock market movements by comparing and testing these various models on our dataset of Reddit news headlines. Based on the underlying variables affecting changes in stock price, this study will enable us to offer more informed investment suggestions.

The testing set is used to assess the model's performance once it has been trained. For each news item in the testing set, the model produces predictions, and the anticipated labels are contrasted with the actual labels to calculate performance metrics, including accuracy, precision, recall, and F1 score.

Our models' entire objective is to forecast the DJIA's future direction (up or down). We assess the performance of the model using the test dataset once our model has been trained using the 80% training dataset, and we then calculate the model performance metrics. In conclusion, The following Figure 2 underlines the overall methodology process in this research.

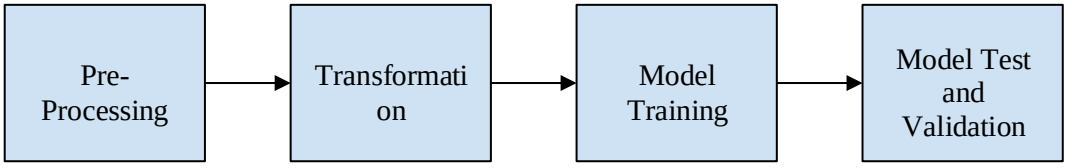

Figure 2. Overall methodology.

Note: This figure represents the main steps of the methodology used in this research.

DOI: 10.54254/2755-2721/29/20230903

5. Result

The sentiment scores of the Reddit Global News headlines were calculated using VADER Sentiment (Valence Aware Dictionary and sEntiment Reasoner) and TextBlob. Three machine learning algorithms—Random Forest, Logistic Regression, and Naive Bayes—were then trained using these sentiment scores. Accuracy, precision, recall, and F1-score were used to assess these models' performance. The results are shown in the following Table 3, and we evaluate the data using a confusion matrix explained in Table 4.

Table 3. Evaluation of different machine learning model.

Model	Sentiment Tool	Accuracy	Precision	Recall	F1-Score
Random Forest	VADER	0.54	0.56	0.68	0.68
Random Forest	TextBlob	0.52	0.56	0.69	0.68
Logistic Regression	VADER	0.53	0.52	0.89	0.69
Logistic Regression	TextBlob	0.52	0.52	0.89	0.69
Naive Bayes	VADER	0.54	0.53	0.91	0.69
Naive Bayes	TextBlob	0.69	0.68	0.96	0.81

Note: This table shows the Precision, Recall, Accuracy, and F1-score for different machine learning models, including Logistic Regression, Naive Bayes, and Random Forest that match with different NLP tools (VADER, Textblob).

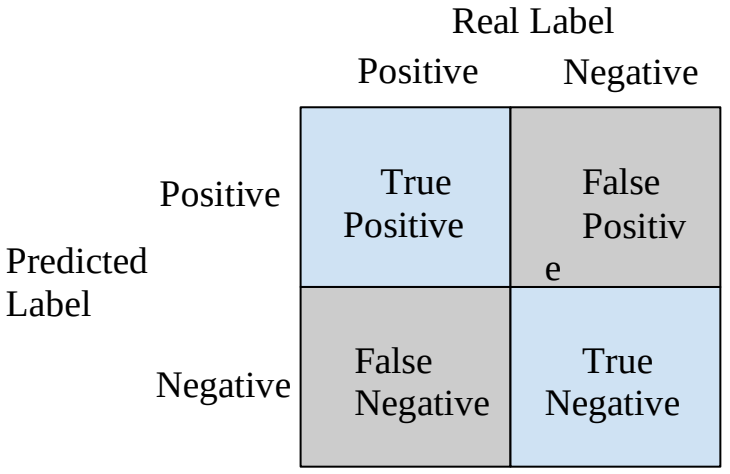

Figure 3. Confusion matrix.

Note: This confusion table is used to evaluate the performance of a classification model; the columns represent the actual values, while the rows represent the predicted values. Through this matrix, we could derive Precision, Recall, Accuracy, and F1-score.

Accuracy (Formula 2) is a measure of how well the sentiment ratings were used to forecast the daily trend of the Dow Jones Industrial Average (DJIA). It is determined by dividing the total number of predictions by the proportion of outcomes that were successfully anticipated. The most accurate models were Naive Bayes (69% for TextBlob and 54% for VADER), Logistic Regression (53% for VADER and 52% for TextBlob), and Random Forest (54% for VADER and 52% for TextBlob).

Formula 2

$$Accuracy = \frac{True\ Positives\ +\ True\ Negatives}{True\ Positives\ +\ False\ Positives\ +\ True\ Negatives\ +\ False\ Negatives}$$

DOI: 10.54254/2755-2721/29/20230903

Precision (Formula 3) is the percentage of the model's actual positive predictions among all of its positive forecasts. It assesses how successfully the model avoids making false positive predictions. The Naive Bayes model has the best accuracy (68% for TextBlob and 53% for VADER), followed by Logistic Regression (52% for TextBlob and VADER) and Random Forest (56% for VADER and TextBlob).

Formula 3

$$Precision = \frac{True\ Positives}{True\ Positives\ +\ False\ Positives}$$

Recall (Formula 4) is the percentage of accurate positive predictions among all positive results. It gauges how accurately the model predicts favorable outcomes. The Naive Bayes model has the greatest recall, 96% for TextBlob and 91% for VADER, followed by Logistical Regression with 89% for both TextBlob and VADER, and Random Forest with 69% for TextBlob and 68% for VADER.

Formula 4

$$Recall\ = \frac{True\ Positives}{True\ Positives\ +\ False\ Negatives}$$

F1-score (Formula 5), the weighted average of accuracy and recall, gives a general assessment of the model's performance. With scores of 68% for TextBlob and 69% for VADER, the Naive Bayes model had the greatest F1-score. Random Forest came in third with a score of 81%. (VADER and TextBlob).

Formula 5

$$F1\ Score = 2 \cdot \frac{Precision\ \cdot\ Recall}{Precision\ +\ Recall}$$

Overall, our findings imply that machine learning models trained using sentiment analysis and NLP tools can anticipate stock market movements with a fair amount of accuracy. Particularly, the Naive Bayes and TextBlob outperformed the other two models tested in terms of precision, accuracy, and F1-score.

6. Conclusion

The results of the study showed that the predictive models built using sentiment analysis and machine learning algorithms were able to predict stock market movements with a relatively high degree of accuracy. Our finding suggests that most scikit-learn machine learning classification algorithms cannot predict stock market trends accurately using news headlines alone, but the combination of Naive Bayes and TextBlob has shown promising results Since it is based on probability theory, which allows it to identify patterns in the data and make accurate predictions. Naive Bayes is known to perform well on text classification tasks, especially on small to medium-sized datasets, as Naive Bayes are relatively simple and can quickly learn from small amounts of data. In contrast, Logistic Regression and Random Forest models may require more data to learn effectively.

Our research has provided important light. However, it is important to acknowledge some of the limitations of our study. The low accuracy of most Sklearn machine learning models also suggests that news headlines may not be the best data source for predicting stock market trends using NLP. One possible reason for this is that news headlines are too short and lack the context and detail needed to make accurate predictions. Similar outcomes were obtained when two NLP techniques—TextBlob and VADER—were combined with several machine learning models in the research. This shows that selecting the best machine learning algorithm for the job at hand may be more important than selecting the appropriate NLP approach.

Despite these limitations, our study offers a valuable beginning point for additional study in this field. The experimental findings in this research demonstrate that stock market trends can indeed be predicted using Reddit WorldNews headlines, despite the fact that the majority of NLP techniques and machine

DOI: 10.54254/2755-2721/29/20230903

learning algorithms produce unsatisfactory results. Through continuous exploring and testing, researchers could possibly select the ideal fusion of NLP techniques and machine learning algorithms. Furthermore, in our future studies, we will use longer text segments to provide a more thorough perspective of the news narrative, which may be more effective and improve the prediction limitations imposed by the Reddit news headline.

Acknowledgement
All authors contributed equally to this work and should be considered as co-first authors.

References

[1] N. Sharma and A. Juneja, "Combining of random forest estimates using LSboost for stock market index prediction," 2017 2nd International Conference for Convergence in Technology (I2CT), Mumbai, India, 2017, pp. 1199-1202, doi: 10.1109/I2CT.2017.8226316.

[2] J. M. -T. Wu, Z. Li, G. Srivastava, J. Frnda, V. G. Diaz and J. C. -W. Lin, "A CNN-based Stock Price Trend Prediction with Futures and Historical Price," 2020 International Conference on Pervasive Artificial Intelligence (ICPAI), Taipei, Taiwan, 2020, pp. 134-139, doi: 10.1109/ICPAI51961.2020.00032.

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

[4] L. Bing, K. C. C. Chan and C. Ou, "Public Sentiment Analysis in Twitter Data for Prediction of a Company's Stock Price Movements," 2014 IEEE 11th International Conference on e-Business Engineering, Guangzhou, China, 2014, pp. 232-239, doi: 10.1109/ICEBE.2014.47.

[5] Cakra, Y. E., & Trisedya, B. D. (2015, October). Stock price prediction using linear regression based on sentiment analysis. In 2015 international conference on advanced computer science and information systems (ICACSIS) (pp. 147-154). IEEE.

[6] Dong, Y., Yan, D., Almudaifer, A. I., Yan, S., Jiang, Z., & Zhou, Y. (2020, December). Belt: A pipeline for stock price prediction using news. In 2020 IEEE International Conference on Big Data (Big Data) (pp. 1137-1146). IEEE.

[7] Shah, D., Isah, H., & Zulkernine, F. (2018, December). Predicting the effects of news sentiments on the stock market. In 2018 IEEE International Conference on Big Data (Big Data) (pp. 4705-4708). IEEE.

[8] Kim, Y., Jeong, S. R., & Ghani, I. (2014). Text opinion mining to analyze news for stock market prediction. Int. J. Advance. Soft Comput. Appl, 6(1), 2074-8523.

[9] Velay, M., & Daniel, F. (2018). Using NLP on news headlines to predict index trends. arXiv preprint arXiv:1806.09533.

[10] Ho, T. K. (1995, August). Random decision forests. In *Proceedings of 3rd international conference on document analysis and recognition* (Vol. 1, pp. 278-282). IEEE.