# Use of Machine Learning techniques for Predicting Stock Market Movement using News Headlines

Pranit Sawant, Sounak Sarkar, Simran Rishi

MSc Data Science

***Abstract*- There was news 'On the back of the 'Russia-Ukraine War' news, Indian markets witnessed one of the worst falls in recent times. NSE Nifty 50 Index fell by 4.78% or 815 points, and BSE Sensex 30 Index fell by 4.72% or 2702 points. There are many factors that affect the performance of the stock market, such as the global and local economy, political events, supply and demand, the COVID-19 pandemic, etc. Our project is about how the stock market is historically affected by such news. This project involves US-based data (DJIA index) to back our findings. We use top-25 news headlines data from Kaggle to create models to predict the movement of the Dow Jones Industrial Average. We explore the effect of word representation, using TF-IDF and vectorization (BOW) approaches. The project delves deep into the relationship between the stock movement and the news. The project analyses the data and tries to get insights from it to back up our findings. Our models show that the Decision tree classifier has the highest accuracy nearly 86 % by TF-IDF and 82% b count vectorization methods. We also got an accuracy of around 84% by the Logistic Regression model and other models that we can see in result part.**

***Keywords*: Stock market, TF-IDF, Vectorization, Logistic Regression, Random Forest, Decision Tree.**

## 1. Introduction

Stock market prediction is always challenging because it is highly volatile and dynamic. Many methods have been proposed to forecast the future directions of the stock market. Moreover, the news is one of the most significant factors impacting people's reactions to the stock market. Recently, the number of online news has rocketed (especially in the wake of Russia Ukraine war) which makes it hard for investors to cover all the latest information. Analyzing stock market performance and using the analysis for short and long-term prediction of stock market movement is an important and difficult problem. Stock market movement can be easily affected by political strategy, economic stability, trade war, unemployment rate, the expectations of investors, and some unprecedented events such as the recent pandemic, COVID-19, etc. Therefore, it is always essential and challenging for the stockholder to be capable of accurately predicting the stock values. The magnitude of change of each influencing factor, and the nature of change (positive or negative change) create an intricate dependency between the factor and the stock market movement. This research area is broad and includes risk assessment and portfolio management, but we will focus on the prediction of the Dow Jones Industrial Average (or Dow Index) that tracks and combines performance on stock markets for 30 large, publicly owned companies trading on the US stock market.

For years, the stock market prediction has depended on historical market data. Researchers applied a variety of algorithms and other techniques to analyze the stock market's behaviour. Although these researchers had a promising result, they could not predict the stock market accurately because researchers tried to predict the future prices from the historical prices with such random behaviour of the stock market and there is no justification for it. Some events will cause instant effects on the stock market. For example, if the price of gasoline dropped sharply, it would motivate investors to sell their shares in petroleum securities. Because of this, the stock prices in petroleum securities will decrease remarkably to reflect the bad event. We will adopt the Dow index as an indicator of stock market movement. Additionally, we propose to use the headlines to predict the movement of the Dow index. In this study, we will use the top-5 highest-rated news headlines from Reddit [2]. Figures 5 and 4 visualize frequent words from the headlines in the word cloud. In this paper, we will be focusing on comparing the shallow Machine Learning (ML) and Deep Learning (DL) approaches to predict the Dow Index movement based on the news headlines.

This article is arranged as follows: in section 2 we will discuss some previous works that we have found out through a literature survey section 3 describes how our system works. After that, section 4 explains how we collected data while section 5 shows the evaluation and analysis results. Finally, section 6 includes our conclusions with a brief discussion about what work can be done in this area.

## 2. Literature review

A great amount of literature is present in the field of sentiment analysis for the large-scale sentiment of news and blogs.

Extensive Research has been conducted in predicting the effect of the news on stock returns using neural networks like Thomson Reuters neural network analysis which proves that positive news and stories regarding stocks, companies, and the government have an immediate response whereas negative news tends to show a more prolonged and delayed response. Studies have addressed the issue of assessing the stock market and investor sentiment response to Earnings announcements.

Gang Li-Fei Liu (2012) in "Application of a clustering method on sentiment analysis" found that traditional clustering methods can also be effective in stock sentiment prediction. It used the TF- IDF weighting method, voting mechanism, and importing term scores to improve the accuracy of the clustering. It has proved to be an efficient and non-human intervening way to solve sentiment analysis problems.

Sheikh Shaugat Abdullah et al. (2013) in their paper "Analysis of stock market using text mining and natural language processing." used a framework that uses our text parser and analyzer algorithm with an open-source natural language processing tool to analyze, retrieve, and forecast investment decisions from any text data source on the stock market. They used the data from Dhaka Stock Exchange (DSE), the capital market of Bangladesh for their study.

Y. Kim et al. (2014) in their paper "Text opinion mining to analyze news for stock market prediction" have successfully implemented Natural Language Processing methodology in mining text opinions and unstructured big data to predict the increase and decrease (up and down) of KOSPI (Korean Composite Stock Price Index). This paper also proposes the use of Natural Language Processing methodology to infer the polarity of news articles.

Minh Dang and Duc Duong (2016) studied Improvement Methods for Stock Market Prediction using Financial News Articles. They proved the correlation between financial news and stock prices using the VN30 index. They also tried to predict the trend of individual stocks in the VN30 index. We chose 5 stocks that had the best technical indicators such as EPS, PE, ROA, and ROE among other stocks in the VN30 index: EIB, MSN, STB, VIC, and VNM using the SVM (Support Vector Machine) classification model.

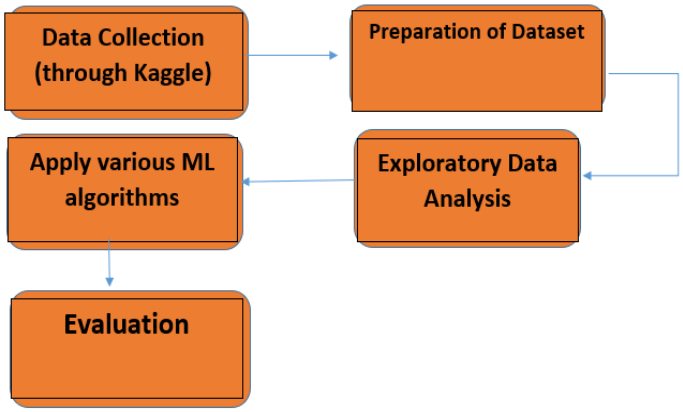
Max Sorto et al. (2017) in their paper "Feeling the Stock Market: A Study in the Prediction of Financial Markets Based on News Sentiment" used news articles from the Wall Street Journal and financial market data from the NASDAQ to find the polarity of news. They used a sentiment analysis system that used a summarization algorithm and the SenticNet 4.0 API. A modified version of the system was also tested using only news headlines, which omitted the summarization portion of the original system, and found that both the news headlines and the news article both give.

Keith Cortis et al. (2017) in "Fine-Grained Sentiment Analysis on Financial Microblogs and News "[8] studied how sentiments contained in financial microblogs and news could be used to predict the sentiment score of stock/companies. It used various techniques that could be used and compared all these techniques and algorithms. The scores were between -1 and 1 and mentioned that Deep Learning and traditional ML techniques provided the maximum contributions. A review of all the techniques used was done and one of the best methods was the hybrid (DL, Lexicon) technique.

Atul Singh, Aakash Patel, Arpit Sah, Mitanshu Khurana (2017) Computer Engineering Department of MPSTME, NMIMS Study the Relationship between Daily News and Stock Market Performance. They used VADER as a tool for quantifying the sentiment of daily news. By leveraging Twitter API, they were able to find the dynamic sentiment of any entity (proper nouns) in each news headline. Using the new compound sentiment score of the sentence, a correlation between the news sentiment and BSE SENSEX.

## 3. Methodology

Methodology to be followed:




Initially, we downloaded the new data in the CSV file format. Later we need to pre-process all the news in the collection in order to have an optimized dataset. Next, we label each piece of news into a specific class of positive, negative, and neutral by using the stock prices. Then the dataset will go through the Lemmatization phase including BOW (Bag of Words). The final step is the training and testing of different machine

learning algorithms to check their accuracy. Below is a detailed description of each step in our model.

Data Preparation:

Documents Labelling Generally, the news articles are believed to cause the movements based on delta closing price (closing price minus opening price for the same day) within the day the articles are established. Delta closing price for a specific day is formulated as a change in closing price from the previous day Pi-1 to closing price in the day Pi:

$\Delta P = Pi - Pi\text{-}1$ ----- (1)

The goal of processing the news data is to classify news into different classes. Based on the approaches so far, three classes have been defined for predicting the market directions:" Upward"," Downward" and "neutral". Here we will only consider the first two as the possibility of the third is very less. If the delta price of a day is greater than zero, all the news articles published in the day I am labelled as" Upward" (1). If the delta price is less than zero, all the news articles published in the day are labelled as" Downward" (0).

The label of the articles

=Upward $\Delta P > 0$

=Downward $\Delta P < 0$

After following the above techniques, we will get data as follows:

| Date | Open | Close | open | Label | Top1 |
|---|---|---|---|---|---|
| 29-12-2008 | 8515.87 | 8483.93 | -31.94 | 0 | Today Israel takes down an entire apt building of civilians to kill the family of one man. In Canada, we call that terrorism.' |
| 31-12-2008 | 8666.48 | 8776.39 | 109.91 | 1 | former Army Employee Pleads Guilty to Acting as Israeli Spy' |
| 02-01-2009 | 8772.25 | 9034.69 | 262.44 | 1 | Australia refuses Bush administration request to house Guantanamo detainees |
| 20-12-2010 | 11491.3 | 11478.13 | -13.17 | 0 | Visa, Mastercard, and PayPal all enable donations to be made to US-registered groups funding illegal Israeli settlements in the West Bank in defiance of international law. |
| 21-12-2010 | 11478.4 | 11533.16 | 54.8 | 1 | Reporters without borders to host mirror site of Wikileaks |

Data Pre-processing and Word Representation

As our input data is in the text format, we apply Natural Language Processing techniques that clean raw text in order to improve model performance. In this work, we use the following pre-processing techniques:

• converting every word to lowercase

• removing all non-ASCII characters

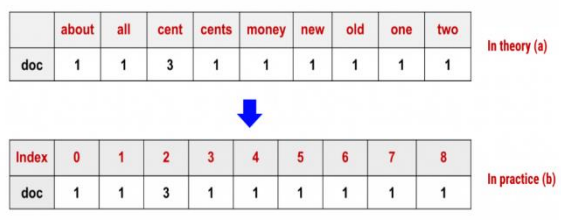
• removing punctuation

• implementing lemmatization that groups the different inflected forms of a word so that they can be analyzed as a single word (i.e., "rocks" will be grouped with "rock"; "better" will be grouped with "good"; "am", "our", will be grouped with "be").

The result of the pre-processing is in text format. We propose to explore the bag of words technique here. Text data is converted to a real-valued vector by various techniques. One such approach is a Bag of Words (BOW). While the former transforms the input text into a sparse vector, the latter uses a matrix format. Both techniques represent the text in numerical values and aim to extract unique words, global statistics, and relationships between the words in the text. The resulting dataset is the input for all Machine Learning algorithms that we implement and evaluate in this work.

CountVectorizer

Convert a collection of text documents to a matrix of token counts. This implementation produces a sparse representation of the counts using scipy.sparse.csr matrix. If you do not provide an a-priori dictionary and you do not use an analyzer that does feature selection, then the number of features will be equal to the vocabulary size found by analyzing the data.




TF-IDF

Term Frequency-Inverse Document Frequency (TF-IDF) is a numerical statistical tool for obtaining a matrix representation from converted text documents and is commonly used in

Natural Language Processing (NLP) applications such as information retrieval and text mining, etc. TF-IDF improves the basic Bag of Words (BOW). Approach for converting documents to vectors by accounting for the relevance of words to a particular document compared to other documents in the corpus. TF-IDF score of a word is the product of two statistics terms. The first one is term frequency, which accounts for the relevance of the word to a document. The second term is inverse document frequency, which accounts for how common a word is in the corpus. Thus, words that are common in every document will get low scores. The output of TF-IDF is a sparse vector with a high dimension for each document where the total number of non-zero elements is equal to the number of unique words in a document.

$$TFIDF\ score\ for\ term\ i\ in\ document\ j = TF(i,j) * IDF(i)$$

*where*

$IDF = Inverse\ Document\ Frequency$

$TF = Term\ Frequency$

$$TF(i,j) = \frac{\text{Term i frequency in document j}}{\text{Total words in document j}}$$

$$IDF(i) = \log_2\left(\frac{\text{Total documents}}{\text{documents with term i}}\right)$$

*and*

$t = Term$

$j = Document$

---

Machine Learning Algorithms

I) Logistic Regression

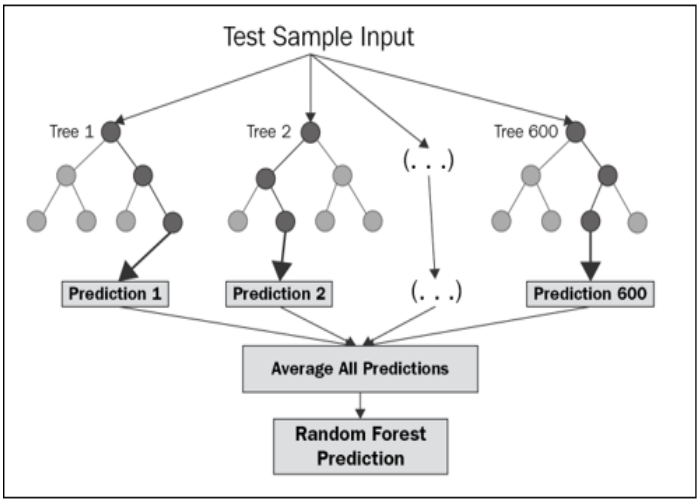
It is used in the biological sciences in the early twentieth century. It is used in many social science applications. It is used when the dependent variable (target) is categorical. For example,

- To predict whether an email is a spam (1) or (0)
- Whether the tumour is malignant (1) or not (0)

Consider a problem where we need to classify whether an email is spam or not. If we use linear regression for this problem, there is a need for setting up a critical value (threshold) based on which classification can be done. Suppose the actual class is malignant, the predicted continuous value is 0.4 and the threshold value is 0.5, the data point will be classified as not malignant which might lead to serious consequences in real time.

II) Random Forest Classifier:

It is also a "Tree"-based algorithm that uses the qualities and features of multiple Decision Trees for making decisions. It is referred to as a 'Forest' of trees and hence the name "Random Forest". The term 'Random' is given as this algorithm being a forest of 'Randomly created Decision Trees'. The Random Forest Algorithm is an enhancement to the existing Decision Tree Algorithm which suffers from a major problem of "overfitting". It is faster and more accurate in comparison with the Decision Tree Algorithm.




III) Decision Tree Classifier

Decision Trees usually represent human thinking ability while making a decision, so it is easy to understand. The logic behind the decision tree can be easily understood as it shows a tree-like structure. In a decision tree, for predicting the class of the given dataset, the algorithm begins from the root node of the tree. This algorithm compares the values of the root attribute with the record attribute and based on the comparison, follows the branch and jumps to the next node. For the next node, the algorithm again compares the attribute value with the other sub-nodes and follows the next node. The process continues the process until it reaches the leaf node of the tree.

Evaluation

Confusion matrices, Precision, recall, and F-measure accuracy were used to evaluate the proposed model. In the confusion matrices, TP, and TN indicate the right classification for the corresponding class, and FP, and FN indicate the false classification for the corresponding class.

Confusion Matrix

| | Actually Positive (1) | Actually Negative (0) |
|---|---|---|
| Predicted Positive (1) | True Positives (TPs) | False Positives (FPs) |
| Predicted Negative (0) | False Negatives (FNs) | True Negatives (TNs) |

Accuracy means the proportion of true positive (TP) and true negative (TN) in the test data. Precision is defined as the true positive (TP) against both true positive (TP) and false positive (FP). The recall is defined as the proportion of true positive (TP) against both true positive (TP) and false negative (FN). The formula is as follows:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

$$Precision = \frac{TP}{TP + FP}$$

$$Recall = \frac{TP}{TP + FN}$$

$$F1\text{-}score = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

## 4. Results:

| Method | Accuracy (Countvectorization) | Accuracy (TF-IDF) |
|---|---|---|
| Random Forest | 0.85 | 0.85 |
| Logistic Regression | 0.84 | 0.83 |
| Decision Tree | 0.82 | 0.86 |

## 5. Conclusion:

In this work, we analyze the stock market activity performance using daily world news headlines from Reddit. We implement our model on the movement of the Dow Jones Industrial Average from February 2000 to June 2020. We have proved the correlation between news and stock prices. To achieve that, the top news and the stock price data were gathered from the Kaggle site. We have achieved quite a high accuracy at 85% by vectorization method and 84 % by TF-IDF method using Random Forest Classifier. We also tried other machine learning algorithms. Moreover, we applied code without removing stopwords, as stopwords also have a significant effect on stock movement. In the future, we will improve the performance of the system by combining stock price prediction and technical analysis and by using more confident sources.

## 6. References:

[1] Dataset:
https://www.kaggle.com/datasets/aaron7sun/stocknews/code?datasetId=129

[2]. “Redditnews” https://www.reddit.com/r/news.

[3] Classification in Machine Learning https://medium.com/analytics-vidhya/classification-in-machine-learning-ed30753d9461

[4] Term Frequency and Inverse Document Frequency in Natural Language Processing

https://www.youtube.com/watch?v=D2V1okCEsiE&ab_channel=KrishNaik

[5] Feature Extraction techniques from the text - BOW and TF IDF|What is TF-IDF and bag of words in NLP https://www.geeksforgeeks.org/bag-of-words-bow-model-in-nlp/

[6] Natural Language Processing In 5 Minutes | What Is NLP And How Does It Work? | Simplilearn - YouTube

[7]https://www.researchgate.net/publication/346352752_Stock_Market_Prediction_using_Daily_News_Headlines

[8] Random Forest Algorithm https://www.javatpoint.com/machine-learning-random-forest-algorithm

[9] https://medium.com/swlh/predict-stock-market-trend-using-news-headlines-part0-introduction-60597ce477a6

[10] Introduction to NLTK: Tokenization, Stemming, Lemmatization, POS Tagging - GeeksforGeeks

[11] “Nltk library,” https://www.nltk.org/book/ch02.html.

[12] C. Huang, L. Huang, and T. Han, “Financial time series forecasting based on wavelet kernel support vector machine,” in 2012 8th International Conference on Natural Computation, 2012, pp. 79–83.

[13] ‘Study on the Relationship between Daily News and Stock Market Performance’ by Atul Singh, Aakash Patel, Arpit Sah, Mitanshu Khurana (Computer Engineering Department of MPSTME, NMIMS, V. L. Mehta Road, Vile Parle, 400056 Mumbai, India) https://www.irjet.net/.

[14] S. Lauren and S. D. Harlili. Stock trend prediction using simple moving averages supported by news classification. In the Advanced Informatics: Concept, Theory, and Application (ICAICTA), 2014 International Conference of, pages 135–139. IEEE, 2014.

[15] P. Mead and J. Li. Stock trend prediction relies on text mining and sentiment analysis with tweets. In the Int and Communication Technologies (WICT), 2014 Fourth World Congress on, pages 257–262. IEEE, 2014.

[16] Zhen Hu, Jibe Zhu, and Ken Tse “The Stocks Market Prediction Using Support Vector Machine”, The 6th International Conference on Information Management, Innovation Management and Industrial Engineering.

[17] Wei Huang, Yoshiteru Nakamori, Shou-Yang Wang, “Forecasting stock market movement direction with support vector machine”, Computers & Operations Research, Volume 32, Issue 10, October 2005, Pages 2513–2522.

[18] Debashish Das and Mohammad sharif Uddin data mining and neural network techniques in stock market prediction: a methodological review, international journal of artificial intelligence & applications, vol.4, no.1, January 2013

[19] N. Ancona, The Classification Properties of Support Vector Machines for Regression ML model, Technical Report, RIIESI/CNR Nr. 02/99.

[20] sklearn. ensemble.RandomForestClassifier — scikit-learn 1.2.0 documentation

[21] Machine learning basics: Random Forest Classification, Perform Random Forest on a dataset and visualize the results https://towardsdatascience.com/machine-learning-basics-random-forest-classification-499279bac51e

[22] Logistic Regression: Detailed overview https://medium.com/towards-data-science/logistic-regression-detailed-overview-46c4da4303bc