MA902 Research Methods

Sentiment Analysis on Stock Market News

by Rathan Narayana Raju
Supervisor: Prof Peter Higgins
Department: Mathematics
University: University of Essex
April 29, 2021

***Abstract*—The stock market is a network of markets and exchanges where people buy, sell, and issue publicly traded company stock on a daily basis and it has become a source of wealth for many people from the past 100 years and will continue to be so in the future. Many financial analysts have been attempted to predict the growth and decline of the stock market based on the many news on the stock market from the Internet. Sentimental Analysis is a familiar approach that is commonly used in a variety of industries. The extraction of emotions from daily news is used to determine a company's stock price outlook. Sentimental Analysis may assist in determining the emotions, which influence stock markets, and hence in forecasting market prices. In this paper, sentimental analysis is performed on the data extracted from Kaggle which provides a comprehensive review of the market efficiency literature that investigates whether the stock prices inflated or deflated using sentimental analysis and supervised machine learning algorithms from the daily news headlines of the stock market.**

***Index Terms*—Sentiment analysis, Stocks Markets, Natural Language Processing.**

I. Introduction

There is a plethora of stock market news on the Internet these days, and investors need to recognize them right away if they want to participate in the stock market. Because of the enormous coverage, it is impossible for them to learn and comprehend any of it. As a result, in order to grasp all news quickly, we must build a support structure [1]. This would assist them in deciding the right time to purchase or sell securities in order to maximise their profits. The stock exchange can be traded both physically and online. When an individual purchases company stock, he or she becomes a shareholder of the company based on the holding ratio of the company's shares [2]. We also have the choice of choosing from a number of data sets for the analysis. This is attributed to the fact that a number of factors affect financial price volatility. The researchers are looking at a naive Bayesian text classifier, Support Vector Machines (SVM), Random Forest, and other machine learning models. Many tests have been carried out to compare the efficacy of these models [3]. The stock market's financial data is diverse, making it impossible to estimate or project the stock market activity. Data mining can be used to study large amounts of complex financial data, resulting in improved stock market prediction returns. Because of the financial market's significance in economics, using data analysis tools to assess it is a burgeoning area of study. Stronger prices contribute to a rise in a country's revenue. Stock market behavior is predicted using classification theory. We use a Nave Bayes classifier and a Random Forest classifier [2]. We can use the sentimental analysis approach that uses Natural Language Processing (NLP) instead of machine learning algorithms. This approach can be used to create templates for labelling ”news” into positive, negative, and neutral emotions. Text data, such as news stories, will have a direct effect on equity prices in the stock market [3].

In this paper, we would use sentimental analysis to see how the news data have an effect on the stock price. Because of the unstructured nature of reporting, stock market prediction focused on news mining is an appealing area of analysis with several challenges. The method of removing elusive, valuable, and theoretically unknown patterns from news data to acquire information is known as news mining. Text mining is a method for dealing with unstructured files. Text mining is also known as Knowledge Discovery of Text in data mining (KDT). Using Granger causality, explore the relationship between financial news and stock price fluctuations. According to the findings, there is a connection between news sentiment and market price shifts [2].

In data mining, natural language processing, and information retrieval, text preprocessing is crucial. It's used to organize unstructured data in order to derive information from it. Tokenization, stop-word elimination, and stemming are only a few examples of text preprocessing functions. Tokenization is the method of breaking down a document into tokens. Tokenization is a part of lexical analysis and is important in linguistics and computer science. The primary aim of tokenization is to identify relevant keywords. A stop word is a widely used word (such as ”the,” ”a,” ”an,” or ”in”) that a search engine has been designed to ignore, both when indexing and retrieving entries as the result of a search query. Text Normalization (or also called Word Normalization) methods used in Natural Language Processing to prepare text, sentences, and documents for further processing include stemming and lemmatization. Since the 1960s, computer scientists have researched stemming and lemmatization, and algorithms have been developed. The method of reducing a word to its word stem, which affixes to suffixes and prefixes or to the roots of words known as a lemma, is known as stemming.

In this paper, the suggested method predicts future stock market activity by combining sentiment analysis for financial news with features derived from historical stock markets from kaggle. Random forest and naïve Bayes algorithms are used in the prediction model. This is accomplished by taking into account various forms of business, industry, and financial news. For dealing with unstructured news data, various methods for numeric data preprocessing as well as text analysis are used. The strategic benefit of stock market trend predicting gained by data mining and sentiment analysis involves value maximisation, expense minimization, and improved consumer understanding of the stock market, both of which contribute to accurate investment decisions [2].

II. LITERATURE REVIEW

In the literature, some methods for forecasting stock market behaviour and price trends have been investigated. Any of these researches aim to improve the accuracy of predictions based on mood analyses of news or tweets, as well as asset markets [4]. Furthermore, several analysis methods have shown that there is a clear connection between financial news and stock price shifts. The difficulty of working with unstructured data has posed a problem in all previous research. To forecast stock market trend, both methods rely on text mining techniques; some rely on textual information compared to just closing prices, while others rely on textual information and stock prices charts screen tickers [5].

Through combining word association and lexical tools to interpret stock market news stories, Patric et al. [6] used many integrating text mining approaches for sentiment analysis of financial markets. SentiWS, a method for emotion analysis on various levels, is used in this thesis to interpret German language. The stock market screens are compared to the sentiment metrics model to obtain an investor's advice over a one-week period in order to assist them in avoiding investment risks. Previous studies were focused on textual data processing, and they attained stock price prediction accuracies that did not reach a range of 75 percent to 80 percent. The forecast accuracy range in news polarities does not exceed 76 percent. This paper's proposed thesis seeks to minimize damages by achieving high accuracy in estimation based on sentiment and historical numeric data collection.

Previous studies' forecast horizons differ; others forecast market fluctuations for 5 to 20 minutes, weekly, and monthly after news releases. Among the previous research aims is to receive investor recommendations, such as [5], while others are to forecast only news polarities based on historical evidence.

In this paper, we intend to employ NLP techniques to create a model for forecasting news opinion analysis. The proposed study proposes a new approach with improved prediction precision to reduce major investment losses and risks while simultaneously maximising stock market returns and minimising economic crises. The conceptual model aims to forecast stock market activity, whether it is rising or falling. The proposed architecture incorporates the study of stock market news and historical prices to improve the classification precision of stock market activity. The thesis suggests doing text analysis on stock market news to assess the polarity of the posts. In addition, stock exchange values are evaluated for the potential movements in opening up, strong, medium, and closing (OHLC) prices.

III. METHODOLOGY

In this study, we have examined the day-to-day news for predicting stock market behaviour, whether it decreases or increases.

A. Data Description

Initially, the data was pulled from the following two channels: News data: Crawled from Reddit WorldNews Channel's historical news headlines (/r/worldnews). Ranking by redesigned user votes, and for a single date, only the top 25 headlines are considered. (Availability: 2008-06-08 to 2016-07-01) Stock Data: ”proving an idea” is used by the Dow Jones Industrial Average (DJIA). (For the period 2008-08-08 to 2016-07-01).

The dataset consists of 4101 rows and 27 columns which has the top 25 headlines. This is a binary classification task. Therefore, there are only two labels ”1” when DJIA Adj Close value increased or remained the same. ”0” when DJIA Adj Close value decreased. The dataset consists of news headline from the year 2000 to 2016.

B. Data Pre-Processing

In the first stage of data pre-processing, the useful features is extracted from the text dataset, such as date, label, and the remaining top 25 headline columns. Next stage would be data cleaning, in which the dataset cleaned by removing punctuation, stop words, alphanumeric words, and and several other text pre-processing steps mentioned below and as shown in the Fig 1.

Removal of Punctuations - There is a set of punctuation which can be discarded based on our use case and the string module in Python eliminates all the punctuation in the top 25 news headlines.

Data standardization - Although it is often ignored, lowercasing all text data is one of the easiest and most powerful forms of text pre-processing.

Stop-word-removal - Words in the documentation that have little meaningful significance, such as the, a, of, etc., are omitted to reduce the amount of features and increase efficiency.

Stemming - Porter Stemmer is added to the data to return and term to its stem and delete suffixes such as (-Ed,-ing,-ion...etc.) to reduce document uncertainty and processing time, which improves model efficiency.

Tokenization - Tokenization is a process that divides long strings of text into smaller chunks, or tokens. Larger blocks of text can be tokenized into sentences, and sentences into words, and so on. Each news article is divided into tokens, which are meaningful terms called tokens.

Filter tokens - Words with two or less characters are filtered out.

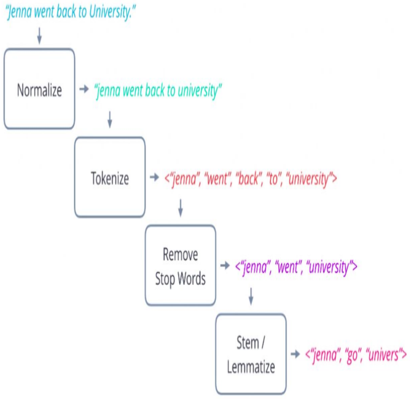
"Jenna went back to University."
Normalize
"jenna went back to university"
Tokenize
<"jenna", "went", "back", "to", "university">
Remove Stop Words
<"jenna", "went", "university">
Stem / Lemmatize
<"jenna", "go", "univers">

Fig. 1. Text Pre-Processing, adopted from Medium article [7]

C. Data Visualization

A wordcloud is used to visualise the data in order to get more familiar with it. Figure 1 shows the wordcloud with the top 25 headline columns together for the first row. The word cloud has been plotted for the word dictionary that we received from the news data; the word with the largest number of occurrences would have a larger font size.

From the word cloud shown in Figure 2, it is clear that the occurrence of words such as georgian, russia, us, help are very high as compared to the other words.

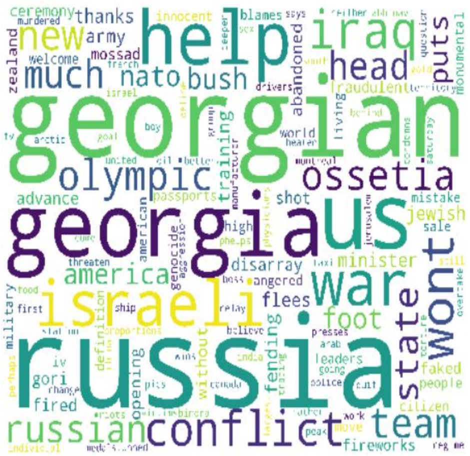

Fig. 2. Word Cloud of Top 25 Headlines.

positive words	negative words
Wide	small
high	weakness
steady	heavy
favourable	low
grow	down

Table 1. The Balance of contrary words.

We use 5 positive and negative seeds in this analysis, which are positive and negative terms in a dictionary. Table 1 depicts the seeds used in the experiment. The terms are manually selected and labelled correctly.

D. Model Building

After analyzing the data, we divided it into two sets training and testing. The Training Set is made up of data from 2008-08-08 to 2014-12-31, while the Testing Set is made up of data from 2015-01-02 to 2016-07-01. This represents an approximately 80/20 split.

1) TEXT DATA INTO VECTORS: For converting text data into vectors, the bag of words model and the tf-idf vectorizer is used. The Bag of Words (BoW) model is the most basic method of numerical text representation. Bag of words is a text modelling in Natural Language Processing technique. In technical terms, we may call it a tool for extracting features from text files. This method of removing features from records is clear and versatile. A bag of words is a text representation that represents the frequency of which words appear in a document. We just keep track of word counts and don't pay attention to grammatical nuances or word order. Since all detail about the order or arrangement of words in the text is discarded, it is referred to as a "bag" of words. The model is only concerned with whether known words appear in the document, not where they appear.

The "Term frequency–inverse document frequency" metric measures the value of a word in a collection or corpus of records. Term frequency –inverse text frequency is a feature weighting method that is used to determine the importance of words in a text or corpus array. The proposed model uses TF-Idf to measure the value of each word in a document depending on the ratio of Idf in a given document to the total number of documents in which the word appears. Words with a high significance have a strong connection to the text in which they occur.

Term Frequency (TF) is a metric that measures how often a term appears in the current text. Since any document is different in length, a word can occur even more often in long documents than in shorter ones. To normalise, the word frequency is often separated by the text volume as shown in the below Equation (1).

$$TF(t) = \frac{N}{T} \tag{1}$$

where:

N = Number of times term t appears in a document.

T = Total number of terms in the document.

The Inverse Document Frequency (IDF) is a measure for calculating the rarity of a word in a collection of documents. The IDF is a metric for calculating the rarity of an expression. As the phrase is used less often, the IDF score rises as shown in the Equation (2).

$$TF(t) = \log_e \frac{N}{n_t} \tag{2}$$

where: N = Total Number of documents.

n_t = Number of documents with term t in it.

Therefore,

$$TF - IDF_{score} = TF * IDF \tag{3}$$

2) RANDOM FOREST CLASSIFIER: A supervised learning algorithm, Random forests classifiers has the ability to be used for both classification and regression. It's also the most adaptable and user-friendly algorithm available. Trees are what make up a forest. A forest's strength is said to be proportional to the number of trees. Random forests generates decision trees from randomly chosen data sets, receives predictions from each tree, and votes on the best solution. It also gives a clear indication of the significance of the feature.

The secret is the low association between models. Uncorrelated models can generate ensemble forecasts that are more reliable than all of the individual predictions, similar to how low-correlation portfolios (like stocks and bonds) come together to create a portfolio that is larger than the sum of its components. The trees shield each other from their individual faults, which results in this wonderful outcome as shown in the Figure 3. The Random forest classifier is implemented by determining the forest's number of trees and the role for determining a split's quality. The conditions "gini" for Gini impurity and "entropy" for information gain are also supported. The training dataset is fitted using the fit function by the label variables in the training set. The test set is used for predicting the results.

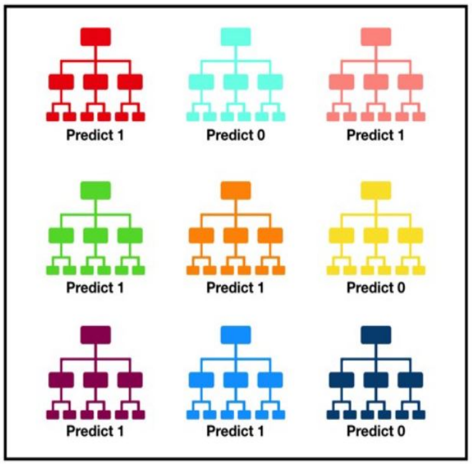

Fig. 3. Making a Prediction using a Random Forest Model Visualization, adopted from Towards Data Science article [8]

3) NAÏVE BAYESIAN CLASSIFIERS: The classification of the Naïve Bayes is to define inventory information as positive or negative based on TF-idf values. The Naïve Bayes Algorithm presumpt that it is independent of the values of other attributes for an attribute value on a certain type. Class conditional Independence is the term for this assumption. Because of its flexibility and pace in text classification, the Naïve Bayes classifier was used to predict the polarity of each message. It categorises each text as positive or negative [2].

$$PNB(c|d) = \frac{P(c) \times (\pi_{i=1}^{m} P(f_i/c)^{n_i(d)}}{P(d)} \tag{4}$$

where: f$_i =$ A feature that appears in the document.

n$_i(d) =$ The number of times a feature appears in the text.

$\pi =$ The circumference to diameter ratio of a circle

Equation (4) represents the calculation of the naïve classification Bayes where fi is the feature in the paper and ni (d) is the number of instances of the feature [2].

The sentiment analysis section examines the sentiment of recent stock news stories. Using pre-processing methods, the Naive Bayes algorithm is used in this phase to determine the sentiment for each news storey or company's financial performance.

The numeric data analysis component pre-processes numeric stock data. This method generates two feature vectors: one with marked news stories that are either positive or negative, and the other with numerical stock data as shown in the Figure 4. Just like we used Random forest classifier for prediction, we will use Naïve Bayes classifier for predicting the results. Here, The Naïve Bayes classifier is implemented using fit function by the train dataset and performing the same steps for the test dataset as well. The test set is used for predicting the results.

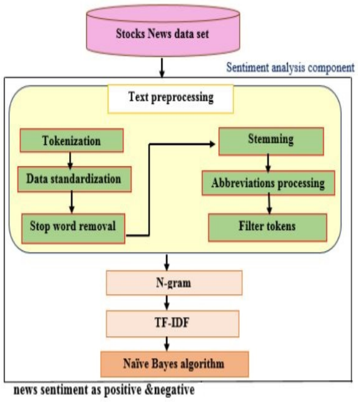

Fig. 4. Stock Trend Prediction using Naïve Bayes, adopted from [2]

IV. EXPERIMENT RESULTS

This section discusses the findings of the experiments conducted to forecast stock market activity using news sentiment analysis. The experiment explains the effects of the news sentiment analysis variable, which categorises news as good or bad, i.e. a rise or decline in the stock price of the particular company. During the sentiment analysis process, the news data is divided into training and test sets. The training set is used to train the algorithm, while the testing set is used to validate it. The training data for the model contains news articles with a positive or negative sentiment. The news content is contained in the testing data, and the model should be able to correctly categorise instances of testing data as positive or negative news. The findings of the experiments are detailed in the section that follow.

Random Forest Classifier - The findings reveal that the random forest algorithm achieves an accuracy of 84.39 percent. Our model's high precision as compared to previous studies is attributed to the use of NLP techniques and Tf-idf. The below figure 5 represents the confusion matrix for the Random forest classifier.

The precision metric is around 0.89 for label "0" and 0.81 for label "1". The re-call metric is around 0.78 for label "0" and 0.91 for label "1". The F1-score is about 0.83 for label "0" and 0.86 for label "1" for the random forest classifier model with tf-idf vectorizer.

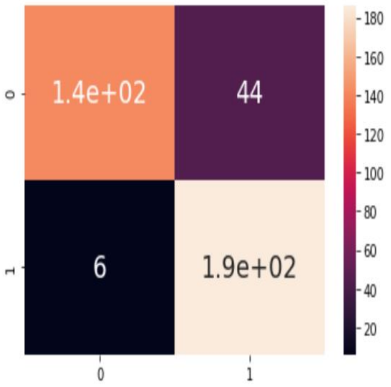

Fig. 5. Confusion Matrix for Random Forest Classifier.

Naïve Bayes classifiers - The results are consistent with previous research findings, which suggest a strong relationship between news and stock market fluctuations. In addition, text analysis is used, and closing prices are used to measure or analyse sequence outcomes. According to the findings, the naïve bayes algorithm has an accuracy of 85.15 percent. The below figure 6 represents the confusion matrix for the naïve bayes classifier.

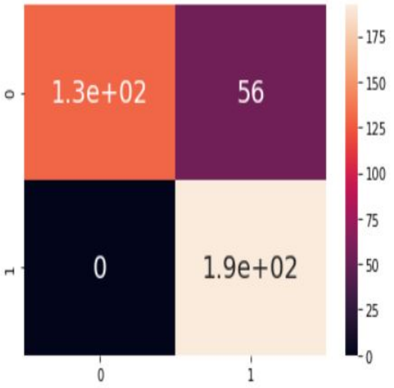

Fig. 6. Confusion Matrix for Naïve Bayes classifier.

The precision metric is around 1 for label ”0” and 0.77 for label ”1”. The re-call metric is around 0.70 for label ”0” and 1 for label ”1”. The F1-score is about 0.82 for label ”0” and 0.87 for label ”1” for the naïve bayes classifier model with tf-idf vectorizer.

V. CONCLUSION

In this paper, we forecast stock prices using emotional analysis based on news reports. An emotional analysis is used as a tool to construct a contemporary sentiment analysis based on news reports. The model improved its prediction accuracy for the future trend of the stock market by taking into account multiple sources of daily news with varying values of numeric attributes during the day. In this, we can see the pattern in stock price fluctuations. There are two supervised machine learning algorithm used Random Forest Classifier and Naïve Bayes Classifier. Random Forest Classifier gives an accuracy of 84.39 while Naïve Bayes Classifier gives an accuracy of 85.15. The optimal results are obtained by using all model approaches. The best accuracy is achieved using a Naïve Bayes Classifier with a tf-idf vectorizer. The model's results are consistent with previous research indicating a similar relationship between equity news and market changes. This model can be improved in the future by incorporating such technical testing metrics, as well as taking emotional sentences into account when determining news polarities and the effect of news that happens in news.

REFERENCES

[1] K. Mizumoto, H. Yanagimoto and M. Yoshioka *Sentiment Analysis of Stock Market News with Semi-supervised Learning*. 2012 IEEE/ACIS 11th International Conference on Computer and Information Science, 2012, pp. 325-328, doi: 10.1109/ICIS.2012.97.

[2] Ayman E. Khedr , S.E.Salama and Nagwa Yaseen *Predicting Stock Market Behavior using Data Mining Technique and News Sentiment Analysis*. I.J. Intelligent Systems and Applications, 2017, 7, 22-30 Published Online July 2017 in MECS, DOI: 10.5815/ijisa.2017.07.03.

[3] J. Kim, J. Seo, M. Lee and J. Seok *Stock Price Prediction Through the Sentimental Analysis of News Articles*. 2019 Eleventh International Conference on Ubiquitous and Future Networks (ICUFN), 2019, pp. 700-702, doi: 10.1109/ICUFN.2019.8806182.

[4] H. D. Hana Alostad *Directional Prediction of Stock Prices using Breaking News on Twitter*. IEEE/WIC/ACM Int. Conf. Web Intell. Intell. Agent Technol., pp. 0–7, 2015.

[5] W. Walter, K. Ho, W. R. Liu, and K. Tracy *The relation between news events and stock price jump: an analysis based on neural network*. 20th Int. Congr. Model. Simulation, Adelaide, Aust. 1–6 December 2013 www.mssanz.org.au/modsim2013, no. December, pp. 1–6.

[6] M. F. Patrick Uhr, Johannes Zenkert *Sentiment Analysis in Financial Markets*. IEEE Int. Conf. Syst. Man, Cybern., pp. 912–917, 2014.

[7] Medium, `https://medium.com/predict/how-does-nlp-pre-processing-actually-work-8d097c179af1`

[8] Towards Data Science, `https://towardsdatascience.com/understanding-random-forest-58381e0602d2`