**STOCK MARKET PREDICTION USING TEXT ANALYSIS**

*“Trading doesn’t just reveal your character, it also builds it if you stay in the game long enough.”*

November 29, 2018

***Group-2:***

1. Parth Doshi - (email) - (asuID)

2. Ammar Gandhi

3. Akhila Muthyala - 1213175494 - [amuthyal@asu.edu](mailto:amuthyal@asu.edu)

4. Saureen Parekh

5. Sucharitha Rumesh - 1213120166 - [srumesh@asu.edu](mailto:srumesh@asu.edu)

**INTRODUCTION**

Stock market has become one of the major components of economy not only in developed countries but also in third world developing countries. Making decision in stock market is not really easy because a lot of factors are involved with every choice we make. Therefore, a lot of analysis is required to make an optimal move on stock market which may involve price trend, market's nature, company's stability, different news and rumors about stocks etc. The objective of this study is to extract fundamental information from relevant news sources and use them to analyze or sometimes forecast the stock market from the common investor's viewpoint.

Making investment decisions in stock market is risky sometimes because it is the fastest and also easiest way of making money as well as losing money. Therefore, investing on stock market needs careful planning with deep analysis which now a days is possible using advanced technologies with large computational power, neural network, relational database etc. Stock market analysis can be separated into two categories. One is Technical and another one is fundamental. Technical analysis looks at the price trend of a security and uses this data to forecast its future price movements where fundamental analysis, on the other hand, looks at economic factors, known as fundamentals.

To understand the long term performance of a stock, to estimate the risk factors involved with an investment, to understand the entry and exit point in a volatile stock market, fundamental analysis is mandatory and news, records, rumors about stocks can provide necessary input for fundamental analysis.

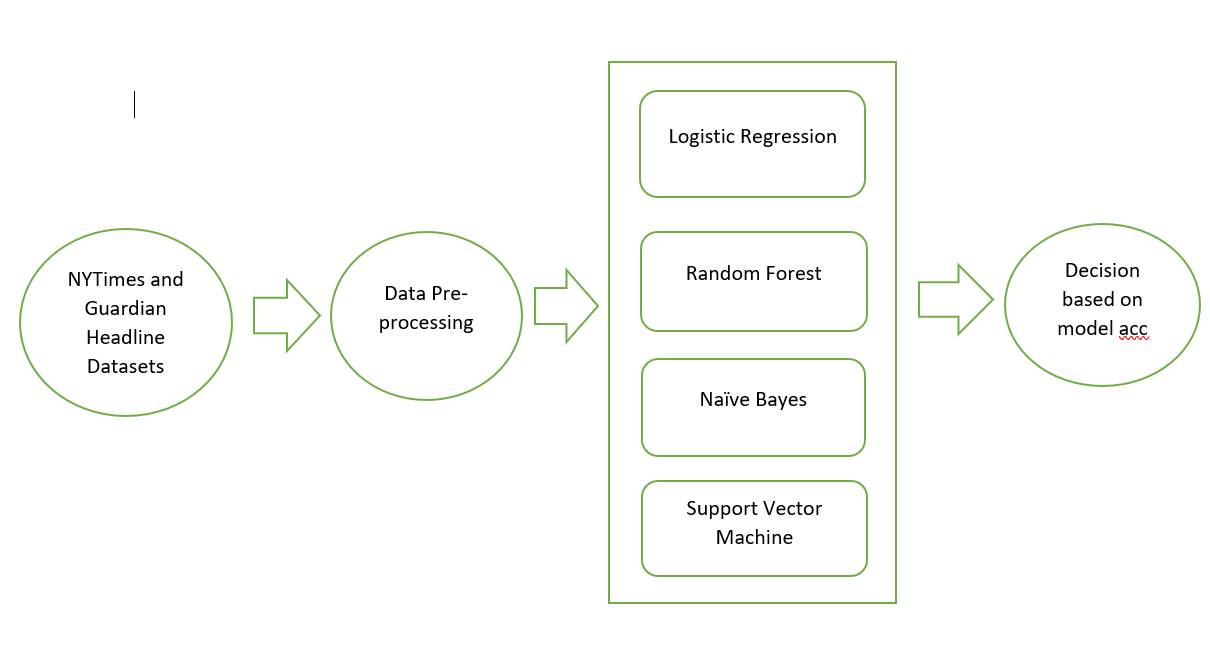


Fig: Framework of the problem

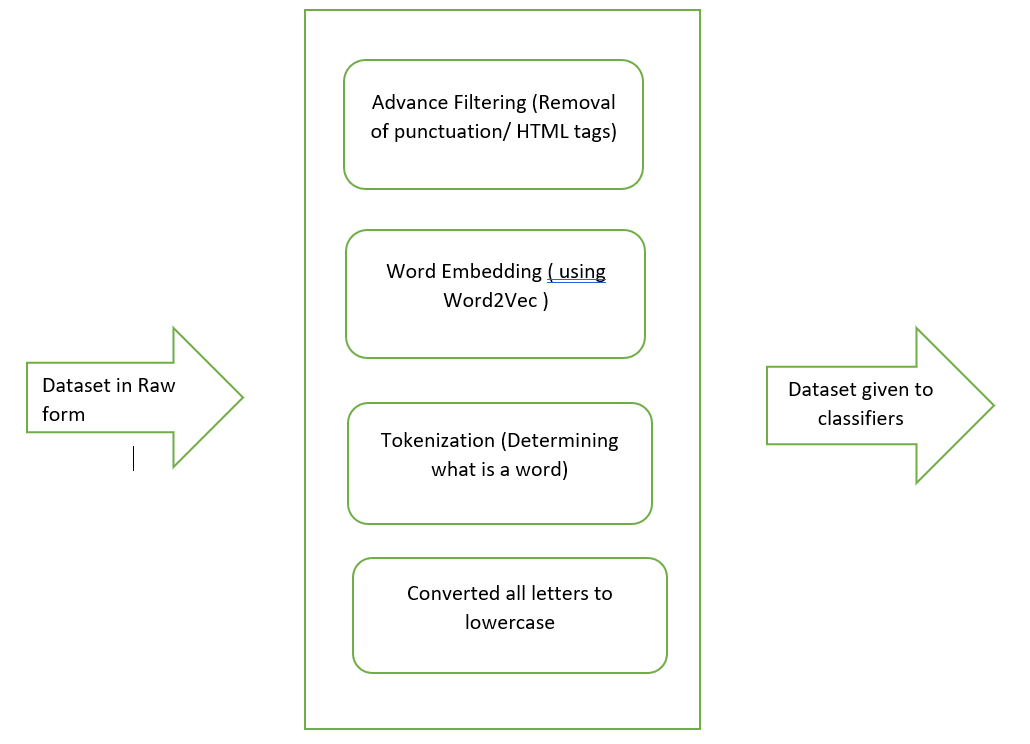


Fig: Data Preprocessing Steps

**PROBLEM DESCRIPTION**

Stock Market is a volatile economy which depends on a multitude of factors. However, the astonishing trend observed is the impact of sentiments of company stakeholders on stock prices. One such example is a tweet from Elon Musk, CEO- Tesla, Inc on August 7, 2018 that caused the stock to increase by 10 percent in a matter of minutes. Such sentiments can be analyzed to determine whether a stock should be held or sold without any human interference. Thus, in this project, we examine the utilization of Machine Learning in Natural Language Processing (NLP)/Text Analysis to have the capacity to direct an open sentiment analysis in Twitter Data to recognize the general state of mind of the public towards the organization. These sentiments could influence the stock costs of the company



To understand the objective of the project, we need to start with the Natural Language Processing and Sentiment Analysis. Natural language processing is concerned with interactions between computers and human languages, in particular how to program computers to process and analyze large amounts of natural language data.



Sentiment Analysis refers to the use of natural language processing, text analysis, computational linguistics, and bio-metrics to systematically identify, extract, quantify, and study effective states and subjective information. Sentiment analysis is widely applied to voice of the customer materials such as reviews and survey responses, online and social media, and health-care materials for applications that range from marketing to customer service to clinical medicine.

**METHODOLOGY**

Analysis of the stock market requires analysis of the raw, textual data. Certain prepossessing of the data set was required such as removal of stop word such as "the", "a", "and", etc. Another filtering required for this data set would be the names such as "America", "Obama", "IMF" etc. This is necessary in order to remove them as they have no effect in sentiment analysis and would just be considered as neutral words. Punctuation were also removed as they will not be considered by the system for analysis.

The raw data used for this system had to be further manipulated in order to use for our analysis. To obtain the numerical representation of the texts in the raw data we had to convert them into vectors. Natural Language has localized spatial correlations between the words. For Example, the word ’rainy’ can be more closely associated with the word ’weather’ in the same sentence than any other word. The technique used is called as "word embedding" where the Euclidean distance is used to represent a pair of related words. This distance is represented as their vector differences. These vectors were used for training and testing data.

The ’Bag of Words’ model was used to do so. This model basically gives us the number of occurrences of words in the headlines from the data set. This will give us a vector whose values are always positive. The number of times a certain word appears in a headline in the data set is divided by the number of times it appears in the entire data set. This will always give us a positive vector that needs to be put between values 0 and 1 which are Label 0 and Label 1 in our data set used for analysis.

After this process, we use the concept of N-gram in order to augment the information they contain. This is also used to store the order of the words in which they occur in the data set. The number of words stored in a order depends upon the value of n. For instance, say N=3 calls for a tri-gram model which will store 3 words in an order. After putting the data set through this model, we end up with a nXm vector where n is the size of the groups and m is the size of the vocabulary. Finally, the data set is split into a ratio of 4:1 where around 80% of the vector data was used for training and 20% for testing. The training data set was converted into numeric representation for machine learning purposes. The output depends on the algorithm selected. Logistic Regression, Random Forest, Naive Bayes and SVM were the algorithms chosen for the prediction.

In Logistic Regression, after the data preprocessing is done, ONE HOT encoding function is used to segregate them into labels as mentioned earlier. Posterior probabilities are then calculated. The data set which is store as a .csv file is then read and separated into training and testing data by splitting it based on the date. Logistic Regression is then performed on the resulting dataset and their accuracy is stored and projected. This model was chosen as it is used with an activation function that results in a binary target class as an output.

Random Forest is a supervised classification algorithm. It has the ability to build multiple decision trees and hence provides a more accurate result. This is because of the fact the decision trees have very low bias and high variance. Firstly, a set of n bootstrap samples are created. These bootstrap samples which are created are of the same size and they can be used again even if they have been already used. A decision tree is then built for every sample. Every new instance is later classified into n tress and then their prediction accuracy is calculated.

Support Vector Machine is chosen due to its popularity in both text classification and regression tasks. In this classifier, an optimal plane is identified in order to separate the data set into two different classes. Minimization of the margin between the two classes is done by solving the optimization problem to find the most optimal plane in SVM. After this, the data set is then used to identify weather the stock investments are “good” or “bad” and the prediction accuracy is recorded.

Finally, in the Naive Bayes classification technique, a Bayesian Network is generated for the data set based on the Bayes theorem. The given dataset , after going through the data pre processing, is converted into a frequency table where the probabilities of each event are calculated which is then used to create the likelihood table. Posterior probabilities are then calculated for each class based on the Naive Bayes equation and the highest outcome is projected.

**RESULTS**

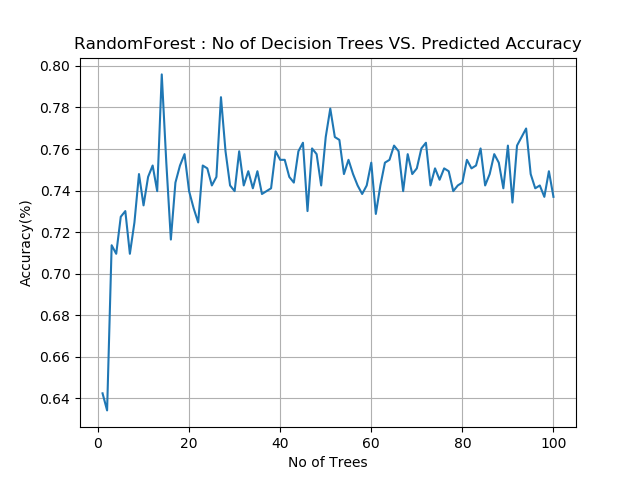
After testing the four Machine Learning classifiers on two different data set: NYTimes and Guardian’s data, the final accuracy of the prediction of the stock market trend has been calculated and projected. As you can see Random forest seems to have the highest prediction when compared to the other three classifier with 77.3% and 84.3% for NYTimes and Guardian data set respectively. The resulting graphs have also been generated( (Fig4, Fig5 & Fig6).

There are many reasons for this result. Random forest can be easily parallelised which indicates that they can be very effective on large dataset such as the ones chosen for this project. They also have the advantage of performing well on data with noises and outliers. Random forest

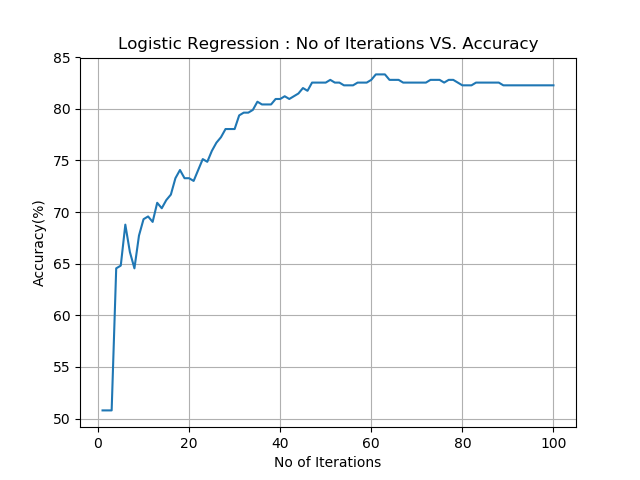
|  |  |  |
| --- | --- | --- |
| Machine Learning Classifiers | NYTime Data  Accuracy | Guardian’s Data  Accuracy |
| Logistic Regression | 74.1% | 83.33% |
| Random Forest | 77.3% | 84.3% |
| Support Vector Machine | 74.6% | 82% |
| Naive Bayes | 73% | 82.39% |

also generates multiple decision trees which have low bias and high variance and hence resulting in better accuracy.

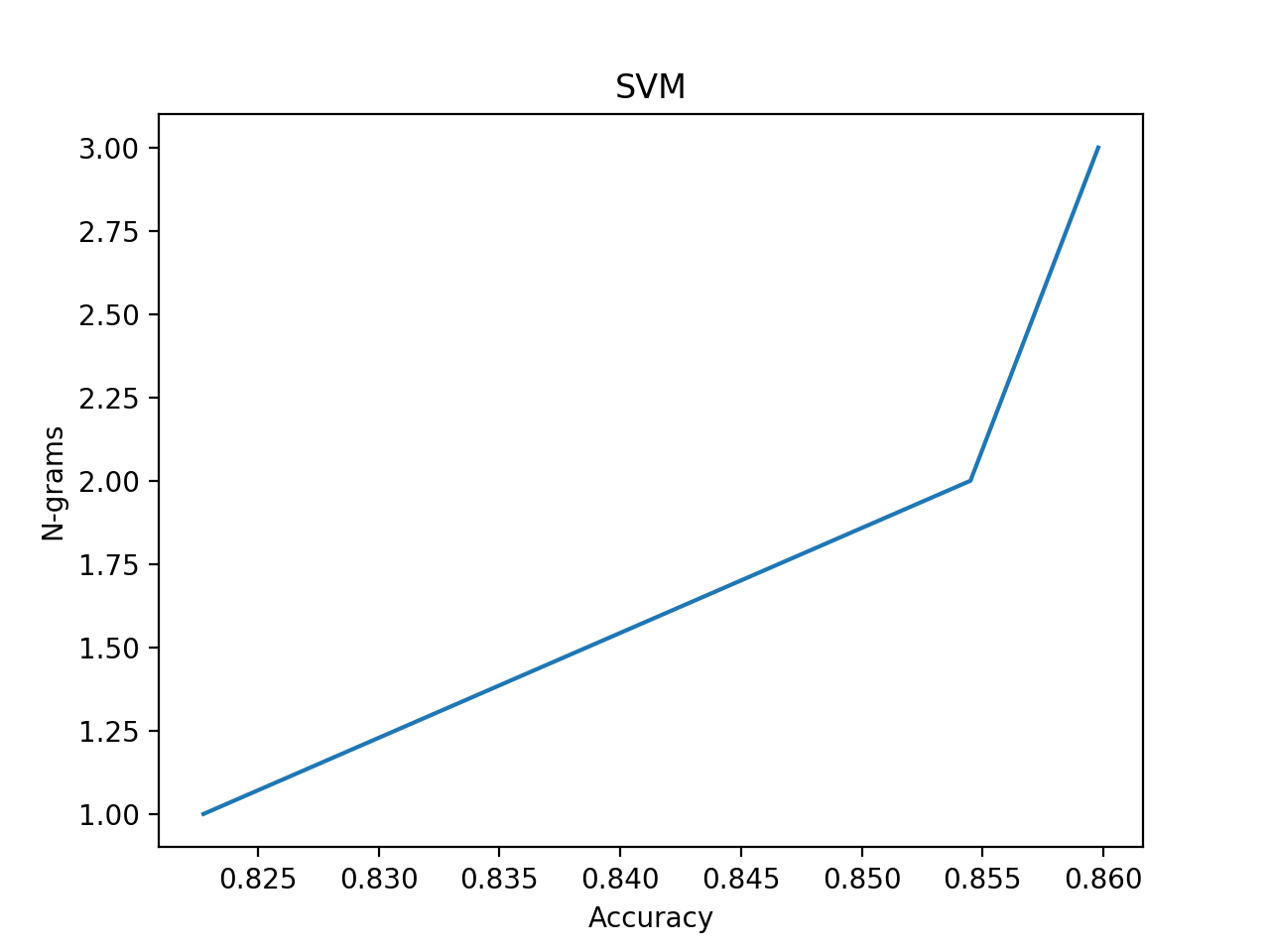
Support Vector Machine leads Logistic Regression in the NYTimes data however its the opposite in the Guardians data. The reason for such variation could be the fact that logistic regression is prone to overfitting. It is also prone to noises in the dataset as it depends on independent variables leading to wrong results. Although Logistic regression has the advantage by providing a quantified value for the variables involved, when compared in terms of data set size with random forest they usually fall behind.



On the other hand with SVM, overfitting is one of their major drawbacks. Defining a clear margin between two clusters especially when the data set is large is difficult. This will eventually reduce the generation ability of the classifier and thus resulting in a poor prediction rate. We speculate that the reason for this to happen would be that although SVM is a complex model and can handle large training data, it can also have the ability to increase the error due to small fluctuations in the training data thereby reducing its predictive ability.



For the Naive bayes classification, we split the training and testing data into 80% to 20% respectively. As we keep increasing the training dataset size, the model becomes more accurate because it starts to learn to represent all the training dataset examples. But if we keep increasing the training dataset, the accuracy of the model decreases because it will become more difficult for the learning algorithm or model to correctly represent all the training dataset examples. In order to learn properly, an algorithm requires enough data, just enough for us to get to the right accuracy. Once we reach that asymptote, we can not improve the accuracy by using more training data. Because of this, the accuracy of the model with the Guardian’s dataset is 82.39% but with the NYTimes’ dataset, the accuracy is only 73%.



**CONCLUSION & FUTURE WORK**

Social Media and Stock markets can be misleading while trying to deliver their messages. In this project we have collected the data from twitter feed as well as from news articles from different websites. In this project, we were able to understand the different techniques used in machine learning and the basics of Natural Language processing. We have also been able to understand the different encoding techniques for word embedding, conversion of text data into vectors and applying machine learning algorithms to help predict the stock market trends. From understanding the basic concepts of NLP in class as well as implementing it with real life examples help us gain a wider and deeper knowledge of Machine learning techniques.

We were also able to scrape our own data which helped us understand our project much more. From reading through various research papers and online resources, we were able to learn the usage of scikit to build the various models that were mentioned earlier. However, apart from the many advantages, we also have few disadvantages. Many data set that we considered had restrictions and cleaning of the entire data set was very tedious. It was also very challenging to find more data that were related to the topic.

One of the disadvantage is that it’s not clear weather the machine learning model trained on old news will be able to predict the future trend of the stock market or not. However, they cannot be resolved until unless the machine learning models can learn a broad context of unrelated events. Another one, for the SVM model, there were many error that kept occuring as it was difficult to provide a clear separation between the clusters since the datasets were typically very large. For example, few NaN value error kept occuring during the training period.

For future work, we would suggest to increase the number of features used in the dataset for better training of the classifier models. It is important to have as many features as possible for better learning. Another suggestion would be to expand this project to not one but multiple stock companies and try to make it a real time series analysis. This can be done through multi class classification of parameters of stocks.

**ACKOWLEDGMENT**

**REFERENCES**

[1] Bing, Li, Keith CC Chan, and Carol Ou. "Public Sentiment Analysis in Twitter Data for Prediction of a Company." *2014 Ieee 11th International Conference on E-Business Engineering (Icebe)*. IEEE, 2014.

[2] Jiang, Zhihua, Pei Chen, and Xixi Pan. "Announcement based stock prediction." *Computer, Consumer and Control (IS3C), 2016 International Symposium on*. IEEE, 2016.

[3] Xu, Feifei, and Vlado Keelj. "Collective Sentiment Mining of Microblogs in 24-Hour Stock Price Movement Prediction." *Business Informatics (CBI), 2014 IEEE 16th Conference on*. Vol. 2. IEEE, 2014.

[4]

|  |  |
| --- | --- |
|  |  |
|  |  |
|  |  |