CSE 575 MASTER OF COMPUTER SCIENCE PORTFOLIO PROJECT REPORT

PARTH DOSHI
1215200012
pdoshi4@asu.edu

1. INTRODUCTION

Decision making in the domain of stock market has never been easy due to the multitude of factors involved and the impact of one false step could be devastating. Thus, an analysis focusing on stock price [1] trend, company's stability [2], the market's volatility and latest rumors regarding stock can be done to decide an optimal move. This study focuses on analyzing the relevant information from different news-streams and using them to make the stock market prediction.

The stock market can be analyzed in two distinct ways, Technical and Fundamental [3]. Technical analysis focuses on the price trend of a security and uses this data to forecast its future price movements whereas Fundamental analysis looks at the economic factors such as news, records, rumors [4] about the stock which are commonly known as fundamentals.

Figure 1: Tweet from Elon Musk

Consider this 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 [5]. Such sentiments can be examined to predict whether a stock should be held or sold. Therefore, in this project, we explore how Machine Learning can be utilized to perform sentimental analysis using top 25 headlines, thereby gauging the public opinion towards the company [8]. These sentiments [2] have the capacity to influence the stock costs of the company.

2. EXPLANATION OF THE SOLUTION

Let us first discuss Sentimental Analysis. Sentiment Analysis refers to the use of natural language processing [6], text analysis, computational linguistics, and bio-metrics to systematically identify, extract, quantify, and study effective states and subjective information. Sentiment analysis [7] is widely applied to the 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.

Figure 2: Stock Price of Tesla

First step involved collection of data. We found a pkl file online that contained stock opening prices and stock closing prices from 2008 to 2016. It also contained the relevant top 25 headlines scraped from ‘The New York Times’ website. We found another dataset containing top 25 headlines from ‘The Guardian’. The differences in stock

prices were used to label the headlines as positive or negative respectively. Next, data preprocessing [9] was carried out by lowercasing the headlines and aggregating them into single columns, stop words such as "a", "the", "and" etc. were also eliminated as they are considered neutral words. Punctuations were removed as they can be ignored.

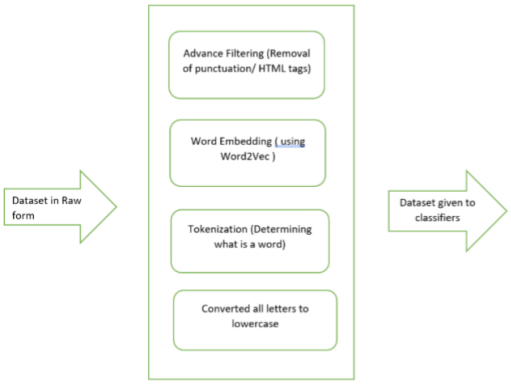


Figure 3: Data Pre-processing

Since raw words cannot be processed directly, we used 'Bag-of-Words' [10] model to convert them into a numeric representation. We use CountVectorizer to achieve this. Minimum document frequency was set to 3 to achieve accurate predictions. We get an N*M matrix where N is the number of records (in this case, number of days) and M is the size of the vocabulary. We split the data into 4:1 training set and test set respectively and run 4 algorithms, namely Naive Bayes, Logistic Regression, Support Vector Machine, and Random Forest.

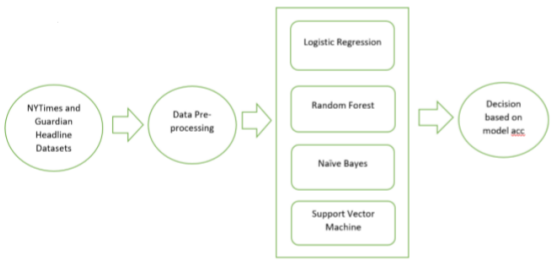
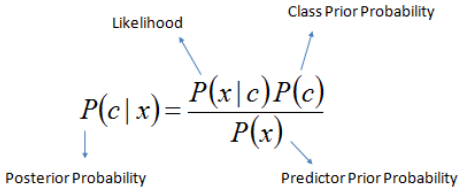


Figure 4: Framework of the problem

The first two algorithms, namely Naive Bayes and Logistic Regression, were implemented from scratch while the remaining two algorithms were implemented using sklearn library functions.

For Naive Bayes [14], we use the following equation to compute posterior probabilities –



$$P(c \mid X) = P(x_1 \mid c) \times P(x_2 \mid c) \times \cdots \times P(x_n \mid c) \times P(c)$$

For Logistic Regression [11], the following equation was used to compute posterior probabilities –

$$P(Y = 1|X =< X_1, ..., X_n >= \frac{1}{1+exp(w_0+\sum_{i=1}^{n} w_i X_i)}$$

And weights were updated using gradient ascent algorithms.

For Support Vector Machine [13], we use the following code to define a Linear Classifier –

```
from sklearn import svm
svm_model = svm.LinearSVC(C = 0.3,
                          class_weight='balanced')
```

For Random Forest [12], we use the following code with splitting criteria as 'entropy' –

```
from sklearn.ensemble import RandomForestClassifier
basicmodel = RandomForestClassifier(n_estimators=i+1,
criterion='entropy',max_features='auto')
```

3. RESULTS

We ran our algorithms on the two datasets and found that Random Forest performs well on both when compared to other classifiers with an accuracy of 77.3% and 84.3% respectively for NYTimes and Guardian dataset. The comparative accuracy graphs are also generated. (Figure 5 and ^)

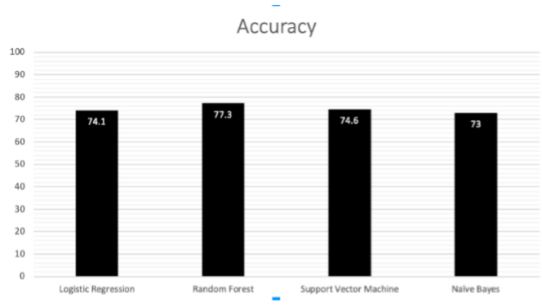


Figure 5: NYTimes Dataset Accuracy

We can see that we get consistent accuracies in the range of 73-77% for NYTimes and 82-84% for Guardian datasets.

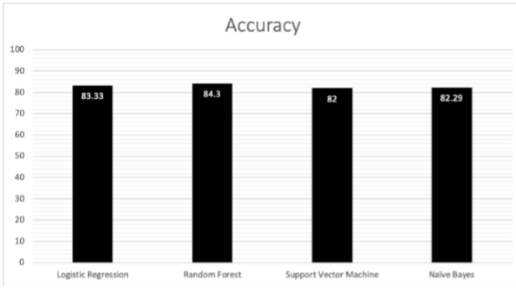


Figure 6: Guardian's Dataset Accuracy

Random forest can be very effective on large dataset such as the ones chosen for this project. They tend to perform well on data with noises and outliers. Random forest [15] also generates multiple decision trees which have low bias and high variance and hence resulting in better accuracy. Due to the large size of data, the three classifiers suffer due to them being prone to overfitting. After observing the result, it became apparent that we can leverage Machine Learning techniques to predict stock market moves successfully.

4. CONTRIBUTION TO THE PROJECT

The first few sessions were spent on collecting data and preprocessing it. I worked on processing the pkl file and labeling the headlines based on stock values. Further data cleaning tasks like punctuation removal and the lowercase transformation were done. I converted the resultant data into a csv file so that it would be easy to access and view. After that, the machine learning algorithms were decided after some brainstorming sessions.

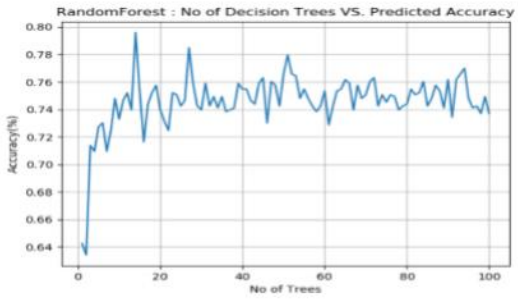


Figure 7: Random Forest

We finalized three algorithms namely Naïve Bayes, Logistic Regression and Support Vector Machine. I was assigned the Logistic Regression classifier and decided to build it from scratch. I computed the posterior probabilities for the training data and used the gradient ascent method to adjust the weight vector to attain optimal weights and then predicted labels of test set data using the optimal weights. Further accuracy metrics were implemented to have a better understanding of classifier performance. I suggested the idea of performing Random Forest classification due to its robustness and aptness at dealing with text data.

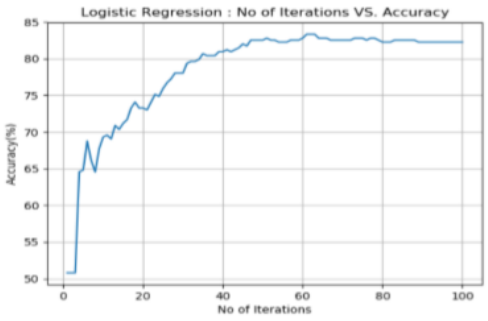


Figure 8: Logistic Regression

I assisted with the implementation of Random Forest and achieved good results with it. Lastly, I prepared the slides and helped documenting the report thoroughly.

5. NEW SKILLS, TECHNIQUES, AND KNOWLEDGE ACQUIRED

The project has helped me learn a lot more about Machine Learning techniques and its practical significance. I learned how to deal with different types of data sets and how to preprocess them effectively as the project involved dealing with extensive data set. Also, I learned to use python libraries such as sklearn, sci-kit and numpy efficiently. The implementation gave me a deeper understanding of Logistic Regression and Random Forest. I was able to deal with the new set of challenges effectively due to the continued guidance of Dr. Tong and his Teaching Assistants Mr. Jian Kang and Mr. Si Zhang. Working on this group project, I had to coordinate with people using tools like slack and git for consistency. It also helped me understand how to effectively utilize everyone's best attribute to yield the best results for the group. I would like to mention that this project proved to be an excellent learning curve and has equipped me with the skills required to tackle such problems effectively in my career.

6. TEAM MEMBERS

- Parth Doshi
- Saureen Parekh
- Ammar Gandhi
- Akhila Muthyala
- Sucharitha Rumesh

7. REFERENCES

[1] F. Xu and V. Keelj, "Collective sentiment mining of microblogs in 24-hour stock price movement prediction," in Business Informatics (CBI), 2014 IEEE 16th Conference on, vol. 2, pp. 60–67, IEEE, 2014.

[2] L. Bing, K. C. Chan, and C. Ou, "Public sentiment analysis in twitter data for prediction of a company," in 2014 Ieee 11th International Conference on E-Business Engineering (Icebe), pp. 232–239, IEEE, 2014.

[3] C. Janssen, C. Langager, and C. Murphy, "Technical analysis: Fundamental vs. technical analysis," Investopedia. com-Your Source For Investing Education. Web, vol. 27, 2011.

[4] X. Tang, C. Yang, and J. Zhou, "Stock price forecasting by combining news mining and time series analysis," in Web Intelligence and Intelligent Agent Technologies, 2009. WIIAT'09. IEEE/WIC/ACM International Joint Conferences on, vol. 1, pp. 279–282, IEEE, 2009.

[5] Z. Jiang, P. Chen, and X. Pan, "Announcement based stock prediction," in Computer, Consumer and Control (IS3C), 2016 International Symposium on, pp. 428–431, IEEE, 2016

[6] D. Rao, F. Deng, Z. Jiang, and G. Zhao, "Qualitative stock market predicting with common knowledge-based nature language processing: A unified view and procedure," in Intelligent Human-Machine Systems and Cybernetics (IHMSC), 2015 7th International Conference on, vol. 2, pp. 381–384, IEEE, 2015.

[7] M. Kaya, G. Fidan, and I. H. Toroslu, "Sentiment analysis of turkish political news," in Proceedings of the The 2012 IEEE/WIC/ACM International Joint Conferences on Web Intelligence and Intelligent Agent Technology-Volume 01, pp. 174–180, IEEE Computer Society, 2012

[8] M. Kanakaraj and R. M. R. Guddeti, "Performance analysis of ensemble methods on twitter sentiment analysis using nlp techniques," in Semantic Computing (ICSC), 2015 IEEE

International Conference on, pp. 169–170, IEEE, 2015.

[9] W. Bouachir, A. Torabi, G.-A. Bilodeau, and P. Blais, "A bag of words approach for semantic segmentation of monitored scenes," in Signal, Image, Video and Communications (ISIVC), International Symposium on, pp. 88–93, IEEE, 2016.

[10] R. Zhao and K. Mao, "Fuzzy bag-of-words model for document representation," IEEE Transactions on Fuzzy Systems, 2017.

[11] A. N. Refenes, A. Zapranis, and G. Francis, "Stock performance modeling using neural networks: a comparative study with regression models," Neural networks, vol. 7, no. 2, pp. 375–388, 1994.

[12] T. Manojlovi´c and I. Štajduhar, "Predicting stock market trends using random forests: A sample of the zagreb stock exchange," in Information and Communication Technology, Electronics and Microelectronics (MIPRO), 2015 38th International Convention on, pp. 1189–1193, IEEE, 2015.

[13] Z. Hu, J. Zhu, and K. Tse, "Stocks market prediction using support vector machine," in Information Management, Innovation Management and Industrial Engineering (ICIII), 2013 6th International Conference on, vol. 2, pp. 115–118, IEEE, 2013.

[14] D.Mahajan Shubhrata,V.Deshmukh Kaveri,R. Thite Pranit, Y. Samel Bhavana, P. Chate, et al., "Stock market prediction and analysis using naïve bayes," International Journal on Recent and Innovation Trends in Computing and Communication, vol.4, no.11, pp.121– 124, 2016.

[15] J. Jotheeswaran and S. Koteeswaran, "Feature selection using random forest method for sentiment analysis," Indian Journal of Science and Technology, vol. 9, no. 3, 2016.