Sentiment Analysis Using Natural Language Processing

[1]Vaishali Mangrulkar, [2]Sameep Khandekar, [3]Mayur Parab, [4]Shreyans Shetty, [5]Siddesh Chaudhari

[1]Assistant Professor, [2,3,4,5]Students
EXTC Department
SIES, Graduate School of Technology
(University of Mumbai)
Navi Mumbai, India

***Abstract*: In this paper, stock prediction has been achieved using machine learning techniques and sentiment analysis. Use of web scraping techniques is done on tweets from twitter to collect large amounts of data required for sentiment analysis. Furthermore, implementation of modern machine learning techniques like Logistic regression for analysis and Random forests for making accurate decisions has been taken into account. This has given promising results with an accuracy of 91.96%.**

1. INTRODUCTION

For a long time, stock market prediction has been an area of research. The general assumption is that stock market trends take a random path. However, the research is getting closer to reliably predicting the stock market. The research is becoming more promising than ever, and it is getting very close to proving that the stock market responds to external stimuli.

The aim here is to see if public opinion influences market sentiment. The scraped data from Twitter is then analysed to perform sentiment analysis. The confusion matrix technique has been used to match expected values to test values in order to determine the project's accuracy.

2. LITERATURE SURVEY

The rise of e-commerce platforms has made online shopping mainstream. In this case, sentiment analysis can be used to serve customers in a more efficient and dynamic manner. SLCBAG is a sentiment lexicon-based model proposed by LiYang, Ying Li, Jin Wang, and R. Simon Sherratt that combines attention-based Bidirectional Gated Recurrent Unit (BiGRU) and Convolutional Neural Network (CNN). The SLCABG model employs both deep learning and sentiment analysis. To begin with, the sentiment lexicon improves the sentiment features in the reviews. The CNN and the Gated Recurrent Unit (GRU) network then extract the main sentiment features and context features in the reviews while using the attention mechanism to weight and finally classify the weighted sentiment features. Finally, the features of weight sentiment are classified. The data is derived from the real-world book evaluation of the well-known Chinese e-commerce website, dangdang.com. The model improves the performance of text sentiment analysis, according to the results.[1].

Koyel Chakraborty, Siddhartha Bhattacharyya, and Rajib Bag shed light on the evolution of data capture processes over time, as well as similarity detection based on similar choices of users in social networks. This article also examines the various methods for communalizing data. Data in various forms has also been analysed and presented in this article. Methods for evaluating sentiments have also been analysed, classified, and compared [2].

Nhan Cach Dang, Maria N. Moreno-Garcia, and Fernando De la Prieta shed light on the accuracy and efficiency of sentiment analysis being hampered by natural language processing challenges (NLP). Deep learning models are demonstrated to solve NLP problems. The most recent studies that have used deep learning techniques have been reviewed. [3].

S. A. El Rahman, F. A. AlOtaibi, and W. A. AlShehri proposed a method for performing sentiment analysis on real-world Twitter data. The model combined supervised and unsupervised machine learning algorithms. Using Twitter's API, data was extracted. Following that, data cleaning and discovery were carried out. Data was fed into several models for training. All tweets were classified as positive, negative, or neutral. Various machine learning algorithms were used, and the results of these models were tested using various metrics such as cross validation and f-score. The model demonstrates strong performance when mining texts are extracted directly from Twitter [6].

3. ALGORITHM

The first step is to feed values into the processor so that they can be processed. Simultaneously, the tweets are being fed into VADER for sentiment analysis [1].VADER is an acronym that stands for Valence Aware Dictionary for Sentiment Reasoning. VADER lexicon outputs different moods from the tweets, i.e. negative, positive, neutral and assigns a value to each of them. All these values are normalized into one compound value. Lastly, machine learning techniques have been used on the compound values and predicted values are visualized along with test values using matplotlib, a famous python library for data visualization.

4. DATASET

The stock price data for United Airlines from "yahoo.finance" was the only dataset used in this project. The data set contains values that are open, high, low, and close. Only the closing values [6] have been taken into account.

4.1. Data pre-processing

Only alphanumeric values were expected in the designed dataset. Hence, removing all the other characters except alphanumeric values was done to achieve the desired dataset. Some cells in the 'prices' column were empty. Hence, they were replaced with the mean value of the prices [3].

4.2. Creating data frames

The dataset that was created was further broken down into two data frames, one for testing and the other one for training. Dataset is ready for making predictions once fitting of independent values like sentiments and dependent values like prices are done.

5. SENTIMENT ANALYSIS

Sentiment analysis is often referred to as opinion mining. Opinion mining is a very popular area of study in the field of NLP (Natural Language Processing). Regardless of computational limitations, today, analysis of people's opinions, sentiments, attitudes, evaluations and emotions is done using various NLP techniques [4].

5.1. Sentiment Lexicons

Sentiment Lexicon consists of a list of lexical features such as words which are assigned a value ranging from negative to positive.

5.2. VADER Sentiment Analysis

VADER is a lexicon. It is a rule-based sentiment analysis tool that is especially focused on sentiments expressed on social media. It works well on texts from other domains. The VADER lexicon is empirically validated by multiple independent human judges. It consists of a "gold-standard" sentiment lexicon that is especially focused on microblog-like contexts. Here, VADER sentiment analysis precisely gives three moods of a particular tweet, i.e. 'negative', 'positive' and 'neutral' respectively. These three moods are later used to calculate normalized values known as 'compound'[5] as shown below.

Comp	Negative	Neutral	Positive
0.6234	0.037	0.86	0.103
0.9983	0.095	0.736	0.169
0.9995	0.075	0.75	0.175
0.9989	0.085	0.75	0.166

Fig 1 : VADER Sentiment Analysis Results

6. MACHINE LEARNING

To achieve maximum accuracy, various machine learning techniques were implemented.

6.1. Random Forest Regressor

A random forest is a meta-estimator that fits a number of classifying decision trees on various sub-samples of the dataset and uses averaging to improve the predictive accuracy and control over-fitting. With this approach, less than 30% accuracy can be achieved.

To remove the biases from the data, scrapping of nearly equal amounts of negative and positive tweets is done. In this way, the 'overfit' dataset has been removed which has made the predictions overly optimistic [6]. Additionally, the size of the dataset has been increased to make the model more sensitive to data.

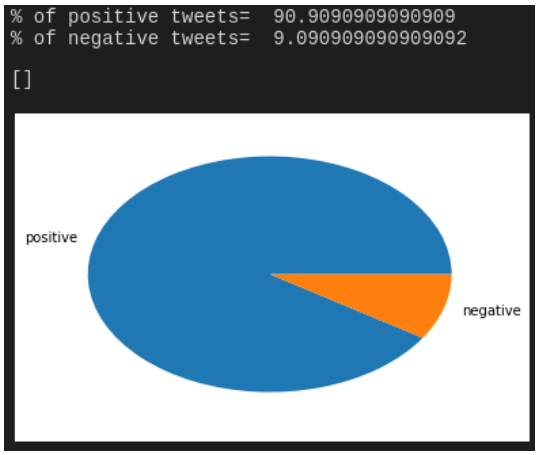

Fig 2: Before Removing Bias

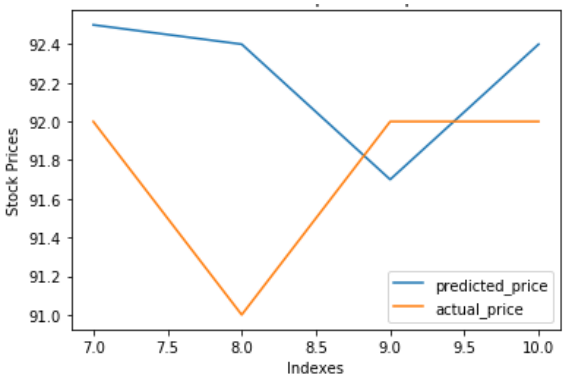

Fig 3: Random Forest Regressor with Bias

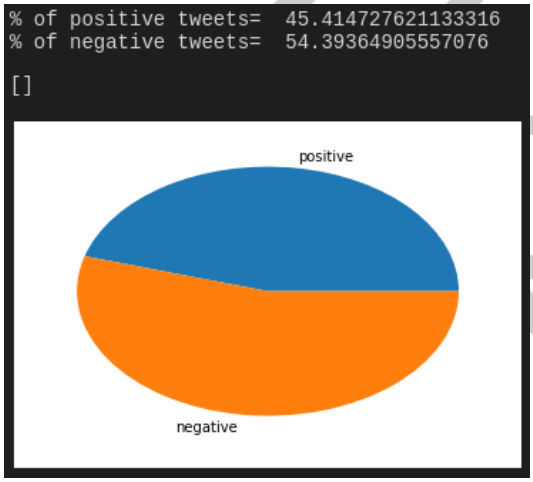

Fig 4: After Removing Bias

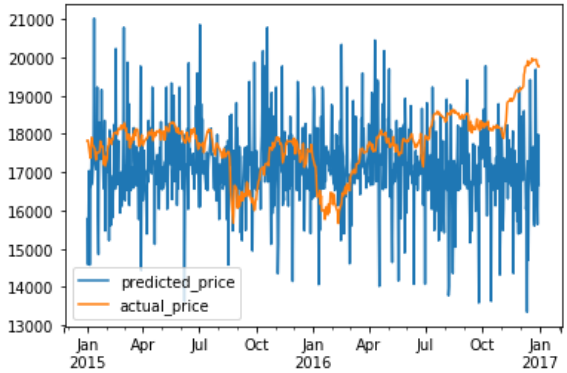

Fig 5: Results After Removing Bias

For the Random Forest Regressor, maximum accuracy of 28% has been achieved. To control the activation point of every lexicon, another machine learning technique, i.e. Logistic Regression is used.

6.2. Logistic Regression

Logistic regression at its core is based on a function called logistic function (also known as sigmoid function).

$$S(x) = \frac{1}{1 + e^{-x}}$$

The function rises quickly and also takes in any real valued number and maps it to a value between 0 and 1. The sigmoid function is never equal to 0 or 1.

Use of logistic regression in the model made it very sensitive to the slightest of changes, [7] hence resulting in a more accurate prediction. The final model achieved an accuracy of 91.96%.

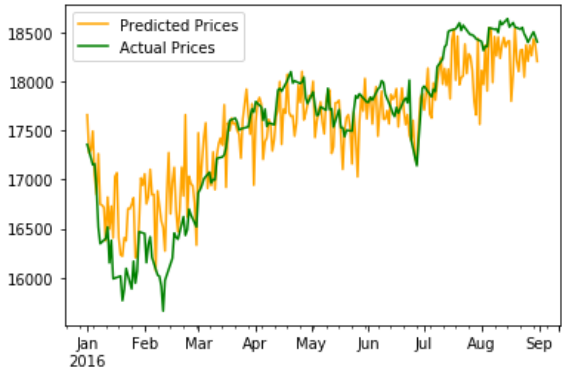

Fig 6: Logistic Regression on Random Forest

7. CONCLUSION

Investigation of the relation between the public sentiments and the market sentiments has been performed. The research establishes that it is possible to capture public sentiments through a complex corpus like twitter. The stock market prices have been predicted successfully with 91.96% accuracy establishing that market sentiment is indeed dependent on public sentiment.

8. APPLICATIONS AND SCOPE

Following are the applications and scope of sentiment analysis:

- Data mining used for stock analysis will be useful for investors to invest in the stock market based on the various factors considered by the software.
- It will help investors to make better predictions.
- Investing in the stock market would be easier for new investors.
- Improving NLP algorithms for better classification of texts.

9. ACKNOWLEDGMENTS

We are extremely grateful to the Electronics and Telecommunication department, SIES for their constant guidance and support.

REFERENCES

[1] LiYang, Ying Li, Jin Wang, R. Simon Sherratt,"https://ieeexplore.ieee.org/document/8970492", Digital Object Identifier 10.1109/ACCESS.2020.2969854

[2]Koyel Chakraborty, Siddhartha Bhattacharyya, Rajib Bag, "https://ieeexplore.ieee.org/document/8951256", IEEE Transactions on Computational Social Systems, vol. 7, no. 2, April 2020.

[3] Nhan Cach Dang, Maria N. Moreno-Garcia, Fernando De la Prieta, "https://www.researchgate.net/publication/341998176_Sentiment_Analysis_Based_on_Deep_Learning_A_Comparative_Study", Published: 14 March 2020.

[4] S. Poria, N. Majumder, D. Hazarika, E. Cambria, A. Gelbukh and A. Hussain, "Multimodal Sentiment Analysis: Addressing Key Issues and Setting Up the Baselines," in IEEE Intelligent Systems, vol. 33, no. 6, pp. 17-25, Nov.-Dec. 2018, doi: 10.1109/MIS.2018.2882362.".

[5] Aishwarya Mankar, Harshada Patil, Chetan Arage, Mahesh Gaikwad2, "International Journal of Scientific Research in Computer Science, Engineering and Information Technology (ijsrcseit.com)", IJSRCSEIT | Volume 3 | Issue 1 | ISSN : 2456-3307.

[6] S. A. El Rahman, F. A. AlOtaibi and W. A. AlShehri, "Sentiment Analysis of Twitter Data," 2019 International Conference on Computer and Information Sciences (ICCIS), 2019, pp. 1-4, doi: 10.1109/ICCISci.2019.8716464.

[7] S. Dhawan, K. Singh and P. Chauhan, "Sentiment Analysis of Twitter Data in Online Social Network," 2019 5th International Conference on Signal Processing, Computing and Control (ISPCC), 2019, pp. 255-259, doi: 10.1109/ISPCC48220.2019.8988450.