Stock market movement prediction using Twitter Sentiment Analysis

Mayankkumar Tank
SEAS, Ahmedabad University
AU1841057
Ahmedabad, India
mayankbhai.t@ahduni.edu.in

Rahul Chocha
SEAS, Ahmedabad University
AU1841076
Jamnagar, India
rahul.c@ahduni.edu.in

Jeet Karia
SEAS, Ahmedabad University
AU1841109
Rajkot, India
jeet.k@ahduni.edu.in

Prince Dalsaniya
SEAS, Ahmedabad University
AU1841124
Rajkot, India
prince.d@ahduni.edu.in

***Abstract*—Nowadays, social media is heavily influencing the opinions of people and people's opinions play huge role in stock market movements. And we can gather people's opinions, insights, and views from various social media forums, blogs which includes twitter, stockwits, and many more. We are trying to find the correlation between sentiments on social media forums and stock market prices.**

***Index Terms*—Sentiment Analysis, Stock Market Prediction, Supervised learning, Decision tree, Naive Classifier, social network analysis**

I. Introduction

Twitter is almost like a micro-blogging service on which most investors and people put out their opinions. So, twitter is important platform which influences the stock market movements and decides the sentiments of the market. It will be useful to set the strategy for common people or to decide on which stock to invest and which to sell.

II. Problem Statement

As social media heavily influences the decisions of the people around the globe, it becomes potential bet to bet on and predict the stock market movements. So, we are finding the correlation that financial stock market data's features has and twitter's sentiments has. So, we are extracting features from the tweets and the financial data and correlating them and inferring fruitful conclusions about the future prices and making successful hedging strategies.

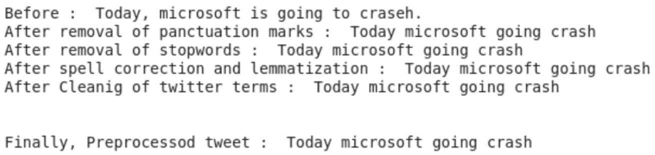
III. Literature Survey

T. Rao and S.Srivastava have used similar approach for correlation of financial data feature and tweets' features and then forecasting future prices or stock movements, they have carried out analysis using pearson correlation, Granger Casuality, EMMS, Forecast performance [1]. T.P. Souza and O.Kolchyna have done similar work where they are also considering the emotions of the tweets which includes calm, alert, sure, vital, kind and many more [3]. What T.Rao and S.Srivastava is doing differently is, they are taking particular time frames and analyzing results according to that time frame for e.g. they have taken features of tweets of 1 week and also of the financial data features which shows better accuracy and also gives better forecast results. A.Jain, P.Dandannavar have implemented different ML techniques like Decision Tree, Support Vector Machine, Classification Algorithms, Logistic Regression and compared their accuracy among different techniques [2].

IV. Implementation

We have implemented all of these in 'Google Colab' which provides us to use their GPUs for complex and resource extensive tasks. We have used 'python-3.8' for implementing all of these.

First, we started with the collection of dataset from the twitter API and stored it in the .csv file. Then, we started cleaning the dataset and pre-processing of the dataset which includes the removal of stopwords, removal of punctuation marks, spell correction, removal of twitter specific tags and mentions and lemmitization. Figure-1 shows the example of one of the tweets.

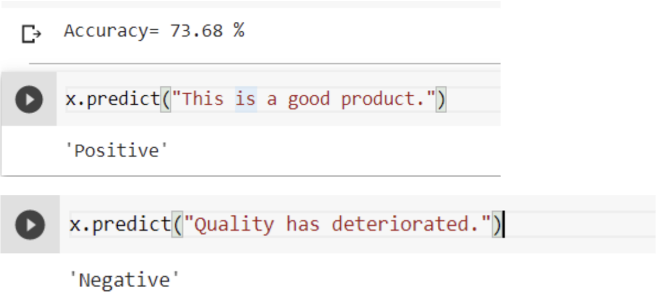


```
Before :  Today, microsoft is going to craseh.
After removal of panctuation marks :  Today microsoft going crash
After removal of stopwords :  Today microsoft going crash
After spell correction and lemmatization :  Today microsoft going crash
After Cleanig of twitter terms :  Today microsoft going crash


Finally, Preprocessod tweet :  Today microsoft going crash
```

Fig. 1. Output of Cleaning and Pre-Processing

Now comes the feature extraction from the tweets which includes implementation of different n-gram models. After the feature we have gone for different classification model like naive Bayesian classifier, decision tree, logistic regression and etc. We encountered the accuracy from different models and came to conclusion that naive Bayesian provides the best accuracy among all the techniques. So, we started to build the naive Bayesian from scratch. It provided good amount of accuracy. Figure-2 shows the accuracy and one of the example of it.

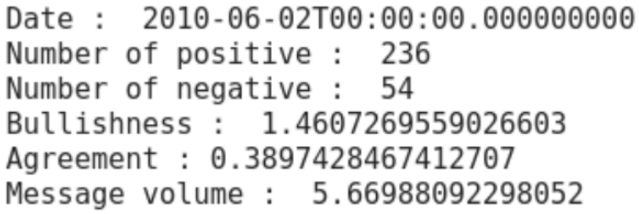
Then comes the feature extraction from the tweets related to the stock market like bullishness, Agreement, Message volume and etc. So, we started to calculate them from our dataset using the timestamp. We calculated different features like bullishness, bearishness, carried bullishness, etc from [1]. Finally concluded with features like Positive, Negative,



```
Accuracy= 73.68 %

x.predict("This is a good product.")

'Positive'

x.predict("Quality has deteriorated.")

'Negative'
```

Fig. 2. Output showing Accuracy

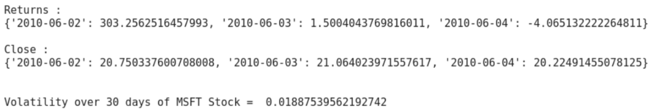
Bullishness, Agreement, Message volume, Carried Postive, Carried Negative.The output of that is shown in Figure-3.



```
Date :  2010-06-02T00:00:00.000000000
Number of positive :  236
Number of negative :  54
Bullishness :  1.4607269559026603
Agreement : 0.3897428467412707
Message volume :  5.66988092298052
```

Fig. 3. Stock market related features from tweets

Now comes the turn to play with the stock market data. We used the library called yfinance for the collection of stock market data and finding the features from the financial data which includes Returns, volatility, Closing price, etc. Then we have to find that how features of both are correlated with each other and forecasting future stock prices or movements. Figure-4 shows the output of this.



```
Returns :
{'2010-06-02': 303.2562516457993, '2010-06-03': 1.5004043769816011, '2010-06-04': -4.065132222264811}

Close :
{'2010-06-02': 20.750337600708008, '2010-06-03': 21.064023971557617, '2010-06-04': 20.22491455078125}


Volatility over 30 days of MSFT Stock =  0.018875395621927 42
```

Fig. 4. Financial data features

Conclusion

In this paper our first task was to find the sentiment of the different tweets taken from large data-set(1.6 million) and then this data is used for training the machine learning model to predict the different sentiment of tweets. This tweets from different time interval has been taken for the analyses purpose with relevant financial data. Then next step is to find relevant features from financial data and Twitter sentiment analysis. Then it has been used to correlate the market fluctuation. We have co-related features to find co-relation coefficient.

References

[1] Rao, Tushar and Srivastava, Saket, “Analyzing stock market movements using Twitter sentiment analysis.” In: Proceedings of the 2012 International Conference on Advances in Social Networks Analysis and Mining (ASONAM 2012).

[2] Jain, A. P., & Dandannavar, P. (2016), Application of Machine Learning Techniques to Sentiment Analysis, 2016 2nd International Conference on Applied and Theoretical Computing and Communication Technology

[3] T.Tuani, P.Souza, O.Kolchyna, P.Treleaven, T.Aste, “Twitter Sentiment Analysis Applied to Finance: A Case Study in the Retail Industry,” arXiv, 2015.

[4] “Sentiment Analysis for Stock Price Prediction” (towardsdatascience.com)

[5] “Time series Forecasting using Granger’s Causality and Vector Autoregressive Model” (towardsdatascience.com)