Analysis of Social Networking Platforms to Predict Stock Market Changes

Rohit Vincent, Rohini V
Department of Computer Science,
Christ University, Bangalore, India

ABSTRACT

This paper aims at creating a self-learning classifier model that classifies a text or phrase as positive or negative and provide a decision on the current stock market changes based on various information. The classifier plots a graph along with the user sentiments and the future prediction allowing the user to decide whether to invest in the stock or not.

Keywords: Naïve Bayes, Self-learning System, Feature selection, Stock Market Prediction, Linear Regression

I. INTRODUCTION

Sentiment analysis which is also known as opinion mining [1], mainly deals with the use of natural language processing and text analysis to identify the subjective sentiment of a phrase. There are two types of sentimental analysis namely subjective/objective identification and feature based sentimental analysis. Subjective/objective identification generally refers to the classification of a sentence into subjective or objective. Feature based sentimental analysis refers to the different features of an entity, meaning the attributes or components of the entity. In other words, sentimental analysis measures the emotional tone of text.

Once the sentimental analysis we is done, we can plot a graph which shows the patterns on user sentiments for a company's products and services and their direct impact on the stock market.

Also, a second line would be plotted on the same graph to show the stock quotes historically and the predicted value over the next week using linear regression.

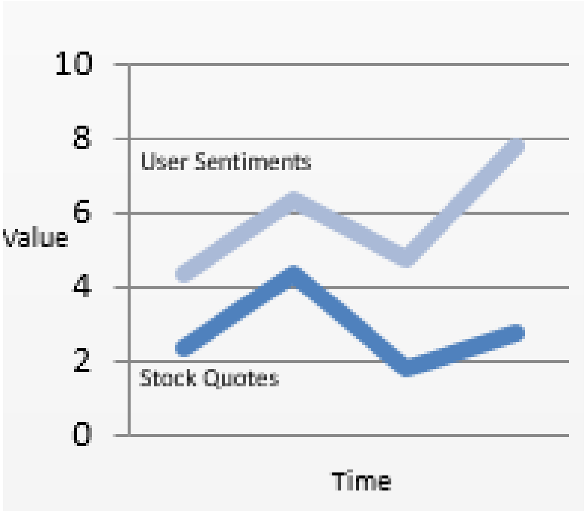

Fig 1: Proposed graph

II. PREDICTING STOCK MARKET CHANGES

The whole project would consist of three parts;

1. Data collection
2. Data classification
3. Data Prediction and regression.

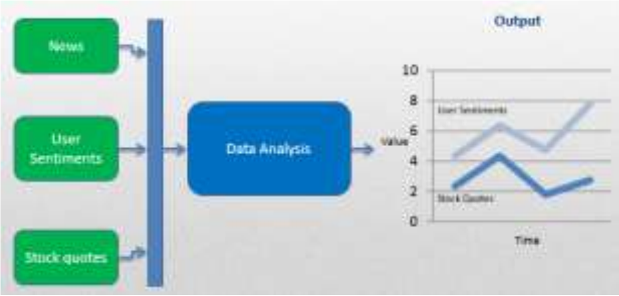

Fig 2: Data flow into graph

Data collection would be done from the following sources such as Yahoo finance for news and stock quotes, and user sentiments from social networking sites such as twitter. For twitter, there are two API's which provide streaming and searching from a specific date for a particular keyword. For example: We could livestream tweets on Apple Inc. using the steaming API and search all tweets related to apple from a previous date using the Search API where the keyword would be #APPL.

Fetching news and stock quotes are tricky since the number of free API's are rare to get such details from sites. YQI (Yahoo query Language) allows us to parse the DOM structure and fetch the required data in JSON format which can be easily read.

Data classification is done using a Bayesian classifier which will be built specifically by training data from tweets and news. These classifier should perform the below natural language techniques to achieve the maximum productivity and sentiment out of a news line or tweet.

- Natural Language Processor
- Remove Special Characters:Characters which are not needed are removed from the sentence before processing.
- Tokenize: Convert into separate words
- Language Correction: Correct language based on common mistakes.
- Stemming: Fetch the root word. Words such as running, playing are changed into run, play etc.
- Stop word Removal: Remove unwanted words. This involves removal of words such as the, this which does not convey any meaning or topic.
- Irrelevant data Removal
- Synonym Replacement: Words with similar meanings are replaced into a common word to increase the potential count of a sentiment.
- Antonym Replacement
- Smiley Correction: ☺ is good for the stock whereas ☹☺ is bad.
- #Tag Correction: Track good and bad hashtags and whether it is trending and also sayings like Bullish and Bearish which signify the market stand.

After this the classifier would classify the data based on the data already collected and manually classified. An example screenshot is shown in the next page.

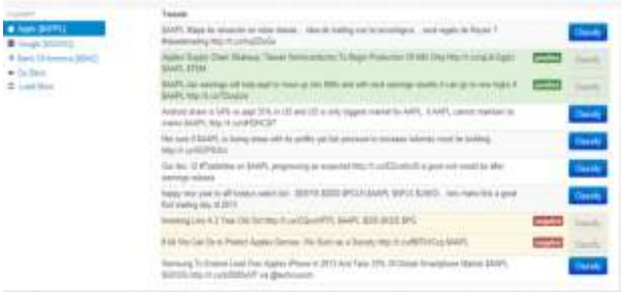

Fig 3: Proposed site to classify tweets

Another major source of information is to build a basic single level web parser which would retrieve important information from the links embedded in the tweets and show these as information related to the stand of the market.

To maximize the performance to classification of information, multithreading of sentences can be used to efficiently classify data. A producer consumer method can be used for efficiently classifying the data. A pictorial of the method is shown below:

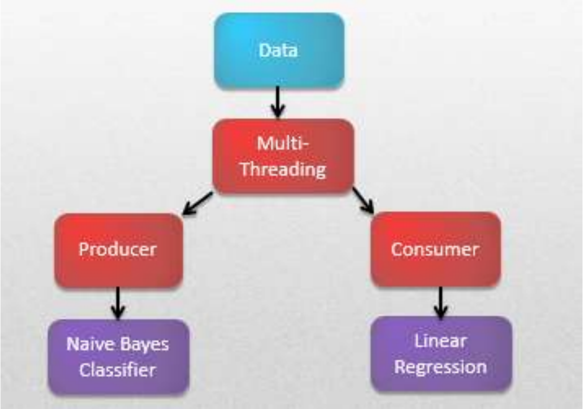

Figure 4: Multithreaded model

The Naïve Bayes Classifier is built on the Bayesian theorem. The classifier uses the same text pre-processing as shown above before checking the probability of the word occurring in the positive and negative dictionary. For each word the count of the positive/negative word is taken. This is then compared with the total length in the dictionary to classify the sentence.

Result = log (sum of positive word count/length of dictionary) +log (0.5)-log (sum of negative word count)

The formulae was modified to be used with log to simplify calculation and a value of log 0.5 was added to the positive log probability to adjust the classifier. The resulting value return by the formula ranged from 0 to 1 and a threshold was set to classify the sentence as positive or negative.

Data regression is done using linear regression. The Following formulae is used as the minimization function:

$$J(\theta) = \frac{1}{2}\sum_{i=1}^{m}(h_\theta x^{(i)} - y^{(i)})^2$$

And plotting the predicted point using the below formula:

$$\theta_j := \theta_j - \alpha\frac{1}{m}\sum_{i=1}^{m}(h_\theta(x^{(i)}) - y^{(i)})x_j^{(i)} \quad \text{(simultaneously update } \theta_j \text{ fo}$$

A proposed system architecture in shown in the figure below:

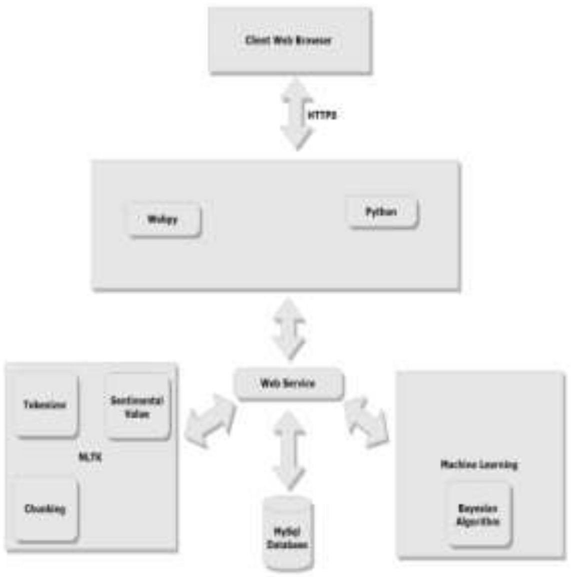

Figure 5: Proposed Architecture

III. CONCLUSION

Bayesian classifier with the right amount of training data will be able to classify tweets appropriately to their user sentiments. Since the stock market is what they think and the value of the stock has a direct correlation on the sentiment of the user's this application can help investors easily study the market and make appropriate stock mark investments.

REFERENCES

[1] A. Go, L. Huang, R. Bhayani, "Twitter sentiment classification using distant supervision", In: CS224N ProjectReport, Stanford, 2009.

[2] Safa Ben Hamouda, and JalelAkaich, "*Social Networks' Text Mining for Sentiment Classification: The case of Facebook' statuses updates in the "Arabic Spring" Era",*International Journal of Application or Innovation in Engineering & Management (IJAIEM), Vol. 2, 2013.

[3] B. Pang, and L. Lee, "Opinion Mining and sentiment analysis". Foundations and Trends in Information Retrieval, pp. 1–135, 2008.

[4] P.D. Turney, "Thumbs up or thumbs down? Semantic orientation applied tounsupervised classification of reviews", In Proceedings of the 40th annual meeting on association for computational linguistics, 2002.

[5] B. Pang, L. Lee, and S. Vaithyanathan, "Thumbs up? Sentiment classification using machine learning techniques". In P. Isabelle (Ed), In Proceeding of conference on empirical methods in natural language, pp.79-86, 2002,.

[6] X. Zhang, and F. Zhu, "The influence of Online consumer reviews on the demand for experience goods: The case of video games", In 27th international conference on information systems (ICIS), AISPress, 2006.

[7] Wikipedia, the free encyclopedia" *Sentimental analysis*".

[8] A. Esuli, and F. Sebastiani. "Determining the semantic orientation of terms through gloss analysis", In: Proceedings of the 14th ACM International Conference on Information and Knowledge Management (CIKM'05), 2005,

[9] K. T. Durant, and M. D. Smith, "Mining Sentiment Classification from Political Web Logs", WEBKDD'06, ACM 1-59593-444-8, 2006.

[10] M. Hurst and K. Nigam. "Retrieving topical sentiments from online document collections", in Document Recognition and Retrieval XI, pp. 27–34, 2004.