Sentimental Analysis of Twitter for Stock Market Investment

K L Shreyas Kumar[1] , Akshaya K M[2], Suhas S[3]

[12] *Students, Dept of Computer Science and Engineering, The National Institute of Engineering, Mysuru, India*
[2] *Assistant Professor,* Dept of Computer Science and Engineering, The National Institute of Engineering, Mysuru, India

---***---

Abstract - *Sentimental Analysis is one of the most popular technique which is widely been used in every industry. Extraction of sentiments from user's comments is used in detecting the user view for a particular company. Sentimental Analysis can help in predicting the mood of people which affects the stock prices and thus can help in prediction of actual prices. In this paper sentimental analysis is performed on the data extracted from Twitter and Stock Twits. The data is analyzed to compute the mood of user's comment. These comments are categorized into four categories which are happy, up, down and rejected. The polarity index along with market data is supplied to an artificial neural network to predict the results.*

***Key Words*: Sentimental Analysis (National language processing, text analysis), Naive Bayes , Support vector machine, Stocks Markets, Media, Moods, Artificial Neural Network.**

1. INTRODUCTION

Sentimental Analysis also known as Opinion Mining is an area that uses Natural Language Processing and Text Analysis that helps in building a system that identifies and extract information in source material. An initial task in sentimental analysis is to determine the polarity of a specified text at the document level, sentence level or aspect level. In core, it is a process that helps in determining the emotional level behind a sequence of words, used to gain an insight of speaker's attitude, opinions and emotions expressed in a sentence. Sentimental Analysis is very useful in social media monitoring as it provides an insight of public opinions for certain topics. The uses of sentimental analysis are very extensive and powerful. Sentimental Analysis provides the ability to extract insight from social data which is broadly used by various organizations across the world [1]. Prediction of the market and stock prices for a company has always been a wide area for the researchers to work upon. A company can be successful in long run only if its consumers are happy with its performance and are giving positive feedback for its products. Expedia Canada used this technique to quickly understand consumer attitude which increased in a negative feedback towards one of their television advertisement.

[2]. Sentimental Analysis was applied to huge scale twitter data in order to find the collective mood states and an exactness of 86.7% was found of Dow Jones Industrial Regular daily directions

[3]. Sentimental Analysis is a widely used field giving many benefits to every industry. Thus if sentiments are correctly categorized and their polarity are correctly determined they can be helpful in enhancing a company's performance and making its investors happy.

In our research work we have performed analysis on sentiments collected from twitter and trained the artificial neural network with the results and stock prices of eight top I.T. companies to predict the future growth of particular stock in the market.

2. RELATED RESERCH

A growing amount of literature is devoted to developing new tools and models for sentimental analysis. Previous studies had concentrated on a large group of population using social information in prediction of consumer's attitude towards a company. The most common use of sentimental analysis is analyzing of twitter tweets and demonstrating the top trends in marketplace. Sentimental analysis is also used in sales forecast of a product by examining tweets.

3. LITERTURE REVIEW

In order to reduce noise, selection of tweets containing tags of top 100 companies was considered. Each tweet was classified using a Naive Bayes method and a set of 2,500 tweets were trained. Results displayed that sentiment indicators are related with unusual returns and stock volume is linked with trading volume. Sentimental Analysis was applied to tweets extracted from Twitter and news headlines to generate new predictors for investment. From the collected data, they choose a random sample and defined each tweet as bullish or bearish if it contains those terms. They displayed that Twitter sentiment indicator and the occurrence of monetary terms on Twitter are statistically significant predictors of regular market returns. Sentimental analysis was also performed on a micro blogging service entirely devoted to stock market The sentiment of the posts was classified using a machine learning algorithm known as J48 classifier to generate a learning model. They proved that the mined sentiment have strong analytical value for coming market directions. Sentimental Analysis was used to forecast the closing index of Tata Services and an accuracy of 85.99% was found in the process. Sentimental analysis is often used to build a social behavior graph on human's online behavior to find the correlation between trading and volume prices of stocks.

4. PROPOSED METHODOLOGY

In this project a method for predicting stock prices is developed using Twitter tweets about various companies. Sentiment analysis of the collected tweets is used for prediction model for finding and analysing correlation between contents of stock prices and then making predictions for future prices will be developed by using machine learning.

4.1 IMPLEMENTATION

4.1 Data collection

Tweets on Microsoft, Facebook, Apple, Google, Tesla are extracted from twitter API. The tweets will have collected using Twitter API and filtered using keywords like $ MSFT, # Microsoft, #Windows etc. Not only the opinion of public about the company's stock but also the opinions about products and services offered by the company. The keywords used for filtering are devised with extensive care and tweets are extracted in such a way that they represent the exact emotions of public about Microsoft over a period of time. The news on twitter about Microsoft and tweets regarding the product releases can also be included. Stock opening and closing prices of Microsoft are obtained from Google Finance.

4.2 Data pre-processing

Stock prices data collected is not complete understandably because of weekends and public holidays when the stock market does not function. The missing data is approximated using a simple technique. Stock data usually follows a concave function. So, if the stock value on a day is x and the next value present is y with some missing in between. The first missing value is approximated to be (y+x)/2 and the same method is followed to fill all the gaps.

Tweets consist of many acronyms, emoticons and unnecessary data like pictures and URL's. So, tweets are pre-processed to represent correct emotions of public. For pre-processing of tweets, we employed three stages of filtering: Tokenization, stop words removal and regex matching for removing special characters.

1. Tokenization: Tweets are split into individual words based on the space and irrelevant symbols like emoticons are removed. We form a list of individual words removed. Form a list of individual words for each tweet
2. Stop word Removal: Words that do not express any emotion are called Stop words. After splitting a tweet, words like a, is, the, with etc. are removed from the list of words.
3. Regex Matching for special character Removal: Regex matching in Python is performed to match URLs and are replaced by the term URL

4.3 Sentimental analysis

Sentiment analysis task is very much fielded specific. Tweets are classified as positive, negative and neutral based on the sentiment present. Out of the total tweets are examined by humans and annotated as 1 for Positive, 0 for Neutral and 2 for Negative emotions. For classification of nonhuman annotated tweets, a machine learning model is trained whose features are extracted from the human annotated tweets.

4.3.1 Word List Generation

We develop our own word list based on the well known Profile of Mood States (POMS) questionnaire. POMS is an established psychometric questionnaire which asks a person to rate his/her current mood by answering 65 different questions on a scale of 1 to 5 (For example, rate on a scale of 1 to 5 how tensed you feel today?). These 65 words are then mapped on to 6 standard POMS moods- Tension, Depression, Anger, Vigour, Fatigue and Confusion. In order to do automate this analysis for tweets, the word list needs to be appropriately extended use the Google n-grams data for the same. We followed a much simpler approach of extending the list by considering all commonly occuring synonyms of the base 65 words.

4.3.2 Daily Score Computation

We used a simple word counting algorithm to find the score for every POMS word for a given day

#of times the word matches tweets
in
Score of a word = in a day /
#of total matches of all words

The denominator accounts for the fact that the number of tweets could vary from one day to another. This works well for our problem because of the nature of tweets which contain simple sentence structures and only a maximum of 140 characters (in most cases much less). We tried using the Stanford core NLP software for word tagging and then using a word's position in the sentence to find its importance. However, similar to our experience working with Opinion Finder, we observed that this process, besides being extremely slow was not too beneficial.

4.3.2 Score Mapping

We map the score of each word to the six standard POMS states using the mapping techniques specified in the POMS questionnaire. We then map the POMS states to our four mood states using static correlation rules (for example, happy is taken as sum of vigor and negation of depression)

4.4 Feature Extraction

Textual representations can be done using n-grams.
N-gram Representation:
N-gram representation is known for its specificity to match the corpus of text being studied. In these techniques a full corpus of related text is parsed which are tweets in the present work, and every appearing word sequence of length n is extracted from the tweets to form a dictionary of words and phrases. For example, the text
"Microsoft is launching a new product" has the following 3-gram word features: "Microsoft is launching", "is launching a", "launching a new" and "a new product". In our case, N-grams for all the tweets form the corpus. In this representation, tweet is split into N-grams and the features to the model are a string of 1s and 0s where 1 represents the presence of that N-gram of the tweet in the corpus and a 0 indicates the absence.

4.5 Model Training

The features extracted using the above methods for the tweets are fed to the classifier and trained using classification methods like Logistic Regression, Decision Tree, SVM and KNN to estimate the movement of the change in stock market price vs. the volume as well as sentiment of news articles and tweets. Apply Linear Regression to find relation between the change in stock market price vs. the volume as well as sentiment of news articles and tweets.

5. System architecture

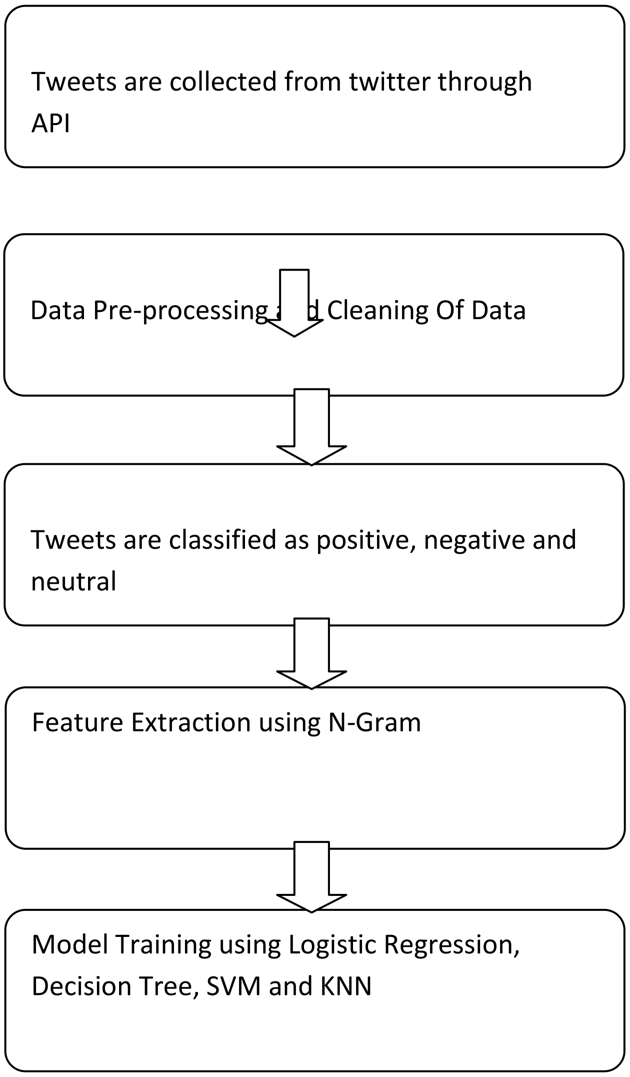
Tweets are collected from twitter through API
Data Pre-processing and Cleaning Of Data
Tweets are classified as positive, negative and neutral
Feature Extraction using N-Gram
Model Training using Logistic Regression, Decision Tree, SVM and KNN

Fig 5.1

6. SYSTEM REQUIREMENTS

6.1 Hardware Requirements

- Processor : Intel CORE i3 and above (Recommended because of high processing speed)
- Memory : 250 GB ROM, 2 GB RAM (Data is downloaded and stored in hard disk)

6.2. Software and Tools

- Operating System : Ubuntu 16.04,Windows
- Programming Languages : Java, python
- Tools and APIs: JDK, Java net beans, Libre Office Artificial Neural Network , Support vector machine , Machine learning , NLP(Natural Language Processing(Python)

7. Experimental Design

7.1 Datasets

1. Tweets from Twitter (twitter application interface)

2) Stock Information:
• Google Finance API Provides no delay, real time stock data in NYSE & NASDAQ provides historical day-by-day stock data

7.2 Evaluation Measures

1. Measure correlation between Volume of tweets vs change in stock price Sentiment of tweets vs change in stock price Volume of news articles vs change in stock price Sentiment of news article vs change in stock price

2. Mean Squared Error for Linear Regression Model Loss function and accuracy percentage for Classification model.

8. CONCLUSIONS

Sentimental Analysis provides the ability to analyze the opinions of people for a particular product or for a company. Prediction of stock market is really a hard nut to crack and requires lot of efforts. The market data if analyzed in a proper way can be very effectual in predicting a company's future. We have mined data and trained a neural network to predict the closing price. Though the closing prices are high of a company but due to sentence score, investment in that company will not be a good decision.

Instead of investing in a company whose closing prices are high, we will recommend you to invest in a company whose sentimental score is high and positive, there are high chances for its stock prices to go up in future. We have ensured that the error rate while performing all implementation is reducing to the least. This work can be extended for a better output if the data samples are taken for a much longer period.

9. REFERNCES

[1] Sunil Kumar Khatri, Himanshu Singhal and Prashant Johri. Sentimental analysis to Predict Bombay Stock Exchange Using Artificial Neural Network, Proc. Of ICRITO,2017.

[2] Sang-Hyun Cho and Hang-Bong Kang, "Text Sentiment Classification for SNS-based Marketing Using Domain Sentiment Dictionary", IEEE International Conference on Conference on consumer Electronics (ICCE), 2016

[3] Zimbra, David, M. Ghiassi, and Sean Lee. "Brand-Related Twitter Sentiment Analysis Using Feature Engineering and the Dynamic Architecture for Artificial Neural Networks." 49th Hawaii International Conference on System Sciences (HICSS). IEEE, 2018.

[4] Vincent Martin.Predicting the French Stock Market using Social Media Analysis, 8thInternational Workshop on semantic and social media adaption and Personalization, IEEE,2013.