International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 07 Issue: 03 | Mar 2020 www.irjet.net p-ISSN: 2395-0072

# Implementation of Twitter Sentimental Analysis According to Hash Tag

**Ashmira Khan[1], Goshiya Sheikh[2], Ruchi Agrawal[3], Raima Rai[4], Nikita Soni[5], Mukesh Roy[6], Sanjay Kalamdhad[7]**

[1-7]*Dept. of Computer Science and Engineering, Shri Balaji Institute of Technology and Management, Betul, M.P*

---

**Abstract -** *Twitter is one of the most used applications by the people to express their opinion and show there sentiments towards different occasions. Sentiment analysis is an approach to retrieve the sentiment through the tweets of the public. Twitter sentiment analysis is application for sentiment analysis of data which are extracted from the twitter(tweets). With the help of this microblogging sites like twitter people get opinion about several things around the nation .Twitter is one such online social networking website where people post their views regarding to trending topics .It s huge platform having over 317 million users registered from all over the world. A good sentimental analysis of data of this huge platform can lead to achieve many new applications like – Movie reviews, Product reviews, Spam detection, Knowing consumer needs, etc. In this paper, we used two specific algorithm - Naïve Bayes Classifier Algorithm for polarity Classification & Hashtag classification for top modeling. This technique individually has some limitations for Sentiment analysis.*

***Key Words*: *Sentiment analysis, Naïve Bayes, Hashtag Classification, classification technique***

## 1. INTRODUCTION

Sentiment Analysis and Opinion Mining consists study of sentiments, attitudes, reactions, evaluation of the content of the text. Twitter is a microblogging media in real time to express the perception of a person or group about a particular topic to appear going on a timeline. The message which is displayed on Twitter is named as Tweet. The chronologically sorted collection of multiple tweets is the timeline. A person can express his view in front of the world in various forms like multimedia, text etc. Because of popularity of Twitter as an information source, it led to development of applications and research in many spheres. Twitter is used in predicting the happenings of earthquakes and identifying relevant users to follow to obtain disaster relevant information. Web search applications, Real world applications like world events, current trending topics in world, extracting latest information about incidents uses by the micro blog data for their analysis and conclusion making on the particular topics.

### 1.1 Problem Statement

A major benefit of social media is that we can see the good and bad things people say about the particular brand or personality. The bigger your company gets difficult it becomes to keep a handle on how everyone feels about your brand. For large companies with thousands of daily mentions on social media, news sites and blogs, it's extremely difficult to do this manually. To combat this problem, sentimental analysis software is necessary. This software's can be used to evaluate the people's sentiment about particular brand or personality.

## 2. Sentiment analysis

### 2.1 Definition

Sentiment analysis deals with identifying and classifying opinions or sentiments which are present in source text. Social media is generating a huge amount of sentiment rich data in the form of tweets. Sentiment analysis of this user generated data is very useful in knowing the opinion of the mass. 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.

## 3. TWITTER

### 3.1 Definition

The word 'micro' in microblogging specifies the limitation of content of the opinion expressed on it. A twitter user can compose at max 140 characters per each tweet. A tweet is not only a simple text message but it is a combination of text data and Meta data associated with the tweet. These attributes are the features of tweets. They expresses the content of the tweet or what is that tweet about. The Metadata can be utilized to find out the domain of the tweet. The Metadata of tweet are some entities and places. These entities include user mentions, hashtags, URLs, and media

Users, Twitter user ID. RT stands for retweet, '@' followed by a user identifier report the user, and '#' followed by a word characterizes a hashtag. Work on the Twitter in this paper is limited up to text data.

## 4. PROPOSED MODEL

Proposed architecture for sentiment classification. The system deals with the tweets extraction and sentiment classification. It consists of following modules.

1. Data collection
2. Data Preprocessing
3. Train the classifier
4. Data visualization
5. Sentiment Analysis

### 4.1 Data Collection

Accessing tweets from Twitter is the primary requirement for building a dataset to get processed and extract information. Twitter allows its users to retrieve real time streaming tweets by using twitter API. We propose to use the python library Tweepy which has the API to extract the tweets through authenticating connection with Twitter server. While collecting tweets we filter out retweets.

### 4.2 Data Preprocessing

The data extracted from twitter contains lot of special characters and unnecessary data which we not require. If data is not processed beforehand, it could affect the accuracy as well as performance of the network down the lane. So it is very important to process this data before training. We need to get rid of all the links, URLs and @ tags. Pre-processing also includes removal of stop words from the text to make analysis easier.

Extracting the sentiment from a tweet is not a huge matter as the data found on microblogging websites contains slang, abbreviations and Twitter specific symbols. The processed tweet requires to be cleared from URL, @ mentions and other Twitter specific symbols such as '#' whilst maintaining the text of the hashtag as it can contain an important reference to the sentiment of the tweet.

### 4.3 Train Classifier

To train the classifier model we will be using a labelled dataset in which every single tweet is labelled as positive or negative based on sentiment.

### 4.4 Data Visualization

The final step of this process is to take in the classified tweets and generate pie chart to visualize the results. The most frequent words in the dataset can be used to generate word cloud.

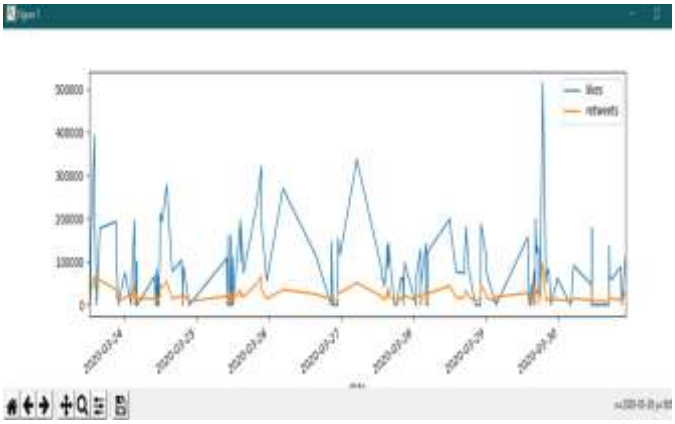

*Fig1. Graph Showing likes & retweets*

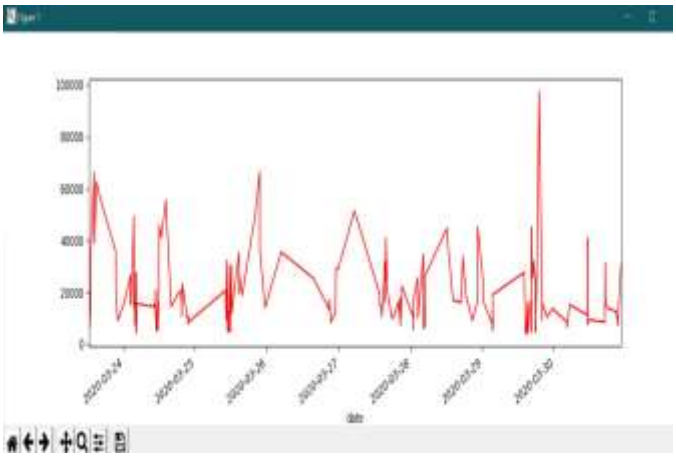

*Fig.2. Graph showing Real time tweets Data*

### 4.5 Sentiment Analysis of Twitter

Sentiment Analysis is classification of the polarity of a given text in the document, sentence or phrase. The goal is to determine whether the expressed opinion in the text is positive, negative or neutral.


*Fig 3. Sentimental Analysis: Step by step process*

### 4.5.1Tokenization:

Tokenization is the process by which big quantity of text is divided into smaller parts called tokens.

### 4.5.2 Cleaning Data:

By removing the numbers, punctuations, Lowercases, Part of speech tagging.

### 4.5.3 Remove Stop Words:

One of the major forms of pre-processing is to filter out useless data. In natural language processing, useless words (data), are referred to as stop words.

### 4.5.1 Classification:

- Rule-based systems that perform sentiment analysis based on a set of manually crafted rules.
- Automatic systems that rely on machine learning techniques to learn from data.
- Hybrid systems that combine both rule based and automatic approaches.

### 4.5.4 Naïve Bayes Theorem

Bayes' Theorem finds the probability of an event occurring given the probability of another event that has already occurred. Bayes' theorem is stated mathematically as the following equation:

$$P(y|X) = \frac{P(X|y)P(y)}{P(X)}$$

where X- Tuples, y-Hypothesis, P(y|X) represents Posterior probability of y conditioned on X i.e. Probability that Hypothesis holds true given the value of X, P(y) represents Prior probability of y i.e the Probability that H holds true irrespective of the tuple values, P(X|y) represents posterior probability of X conditioned on y i.e. the Probability that X will have certain values for a given Hypothesis, P(X) represents Prior probability of X.

The proposed system understands whether tweet is positive or negative with dictionary method. The formula is given below:

$$Accuracy = \frac{\sum True\ Positive + \sum True\ Negative}{\sum Total\ number\ of\ words}$$

Where True positive is number of tweets recognized as positive and true negative is number of tweets recognized as negative respectively.

### 4.5.5 Hashtag Classification

Hashtag classification is very important for topic modeling. While posting any message, the user uses a hash tag, for eg. #Covid19. So, from this we can know that the post is about the Corona Virus 19. This can help in classifying the preprocessed data in various topics. We do not change the hash tag words during preprocessing.

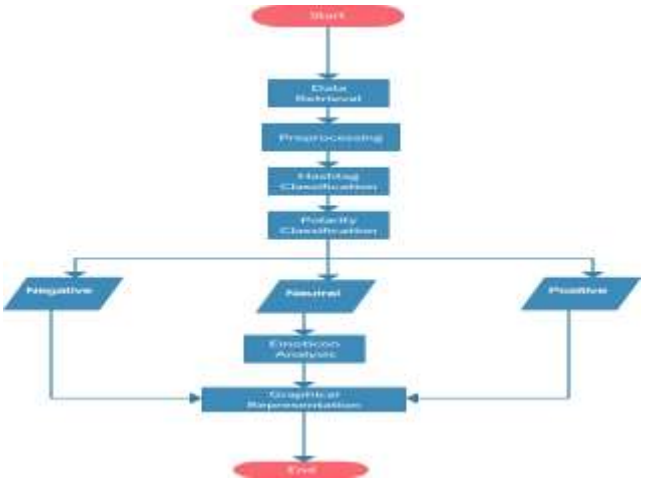


*Fig 4. Flowchart of Proposed System*

With the help of data parsing, the algorithm can identify the hash tagged words and with the help of that particular text message is classified into that group so that the data does not get mixed up and because of that accuracy increases. Our algorithm does not remove the less used hash tags instead it concentrates on the most used hash tags. Hash tag in the message can prove very much crucial for classifying the data.

In general, people write hash tags in a concatenated format. There are no white spaces or special character in between which parser can identify to split the text. For example Corona Virus 19, tweets related to all covid19 had a hash tag '#Covid19' or '#CoronavirusTruth' or'#CoronascaresDelhi'. First kind is used more than the second and the third kind. But, we cannot rely on people putting a capital letter at starting of every new word. So, we make a list of prepositions, conjunctions and 'wh' question words. Using that list, the parser searches for the word (irrespective of the case used) as given in the list, if it finds the word, it puts whitespace in front and rear of that particular word. If the word searched is having the first position i.e. immediately after the hash tag then the hash tag is removed and white space is inserted only in the rear. So, in our example the parser make the hash tag text as *'We WontGive It Back'* and then the tweets are classified according to it.

## 5. Twitter Sentiment Analyzer (Tweezer)

It will perform live analysis for any hashtag and it's related contexts and show you new tweets as they come in, along with a sentiment attached to it. We can search the tweets of particular person by entering screen name and the number of tweets we want to display.

*Fig5. Tweets extracting on real time of Donald Trump*

## 6. CONCLUSION

We will obtain a classification of polarities of sentiments into positive, negative or neutral. Naïve Bayes is simple, easy to train and has less execution time it shows the output in the form of pie charts. Thu s the basic knowledge required to do sentiment analysis of Twitter. Two methods are used in this project. The accuracy/ result of each method enable us to imagine the efficiency of applied technique in respective circumstances.

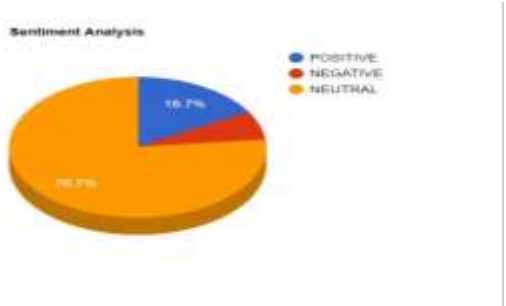


*Fig.6.Sentiment Analysis in the form of Pie Chart*

## REFERENCES

[1] Ramteke, Jyoti, Samarth Shah, Darshan Godhia, and Aadil Shaikh. "Election result prediction using Twitter sentiment analysis." In 2016 international conference on inventive computation technologies (ICICT), vol. 1, pp. 1-5. IEEE, 2016

[2] Sanjay Kalamdhad, Shivendra Dubey, Mukesh M. (2016) Feature based Sentiment Analysis of Product Reviews using modified PMI-IR Method, IJCTT ISSN 2231-2803 Volume 34, Number 2, April 2016.

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

[4]Bouazizi, Mondher, and Tomoaki Ohtsuki. "Sentiment analysis in twitter: From classification to quantification of sentiments within tweets." In 2016 IEEE Global Communications Conference (GLOBECOM), pp. 1-6. IEEE, 2016.

[5] Wang, Hao, Dogan Can, Abe Kazemzadeh, François Bar, and Shrikanth Narayanan. "A system for real-time twitter sentiment analysis of 2012 us presidential election cycle." In Proceedings of the ACL 2012 System Demonstrations, pp. 115-120. Association for Computational Linguistics, 2012.

[6]http://colah.github.io/posts/2015-08UnderstandingLSTM

[7]http://towardsdatascience.com/another-twitterposentiment-analysis

[8] 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

[9] 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.

[10] Suman, D.R, & Wenjun, Z., "Social Multimedia Signals: A Signal Processing Approach to Social Network Phenonmena", ISBN-13: 978-3319091167, Springer International Publishing Switzerland, 2015.

[11]Liu, B., Sentiment Analysis and Opinion Mining. Morgan & Claypool Publishers, 2012