TWITTER SENTIMENT ANALYSIS USING R-STUDIO

[1]Devika Deshmukh, [2]Rahul Patil, [3]Shubham Chafale,[4]Shivani Basutiya
[1]Professor, [2]Student, [3]Student, [4]Student
[1]Department of Information Technology,
[1]Rajiv Gandhi College of Engineering and Research, Nagpur, India

Abstract: Twitter is a social media platform, a place where people from all parts of the world can make their opinions heard. Twitter produces around 500 million of tweets daily which amounts to about 8TB of data. The data generated in twitter can be very useful if analyzed as we can extract important information via opinion mining. Opinions about any news or launch of a product or a certain kind of trend can be observed well in twitter data. The main aim of sentiment analysis (or opinion mining) is to discover emotion, opinion, subjectivity and attitude from a natural text. In twitter sentiment analysis, we categorize tweets into positive and negative sentiment. The application of sentiment analysis is broad and powerful. The ability to extract Insight from social data is a practice that is being widely adopted by organization across the world. Shifts in sentiment on social media have been shown to call error rate with shifting in the stock market. The Political parties use sentiment analysis to catch public opinion to policy announcement and campaign message. The ability to quickly understand consumer attitude and react accordingly. Expedia Canada took advantages of when they notice that there was a study increase in negative feedback to the music needing one of their television adverts.

Index Terms - Sentiment analysis, R Studio, Twitter.

I. INTRODUCTION

With the recent growth of mobile information systems and the increased availability of smart phones, social media has become a large part of daily life in most societies. This development has entailed the creation of massive amounts of data which when analyzed can be used to extract valuable information about a variety of subjects [1].

Sentiment analysis is the computational task of automatically determining what feelings a writer is expressing in text. Sentiment is often framed as a binary distinction (positive vs. negative), but it can also be specific emotion an author is expressing (like fear, joy or anger), also known as emotion mining, the process of classifying the emotion conveyed by a text, for example as negative, positive or neutral. The data made available by social media has contributed to a burst of research activity within sentiment analysis in recent times and a shift in the focus of the field towards this type of data. Information gained from applying sentiment analysis to social media data has many potential usages, for instance, to help marketers evaluate the success of an ad campaign, to identify how different demographics have received a product release, to predict user behavior.

Some applications for sentiment analysis include:

- Analyzing the social media discussion around a certain topic.
- Evaluating survey responses.
- Determining whether product reviews are positive or negative.

As internet is growing bigger, its horizons are becoming wider. Social Media and Micro blogging platforms like Facebook, Twitter, and Tumblr dominate in spreading encapsulated news and trending topics across the globe at a rapid pace. A topic becomes trending if more and more users are contributing their opinion and judgments, thereby making it a valuable source of online perception. These topics generally intended to spread awareness or to promote public figures, political campaigns during elections, product endorsements and entertainment like movies, award shows. Large organizations and firms take advantage of people's feedback to improve their products and services which further help in enhancing marketing strategies. One such example can be leaking the pictures of upcoming iPhone to create a hype to extract people's emotions and market the product before its release. Thus, there is a huge potential of discovering and analyzing interesting patterns from the infinite social media data for business-driven applications.

Sentiment analysis is the prediction of emotions in a word, sentence or corpus of documents. It is intended to serve as an application to understand the attitudes, opinions and emotions expressed within an online mention. The intention is to gain an overview of the wider public opinion behind certain topics. Precisely, it is a paradigm of categorizing conversations into positive, negative or neutral labels. Many people use social media sites for networking with other people and to stay up-to-date with news and current events.

II RELATED WORK

Sentiment Analysis is an approach of studying people's emotions and sentiments based on a particular topic, service, event and its attributes. Sentiment analysis appears as a part of various business analysis systems to find opinions about their services or products. Both the availability of CPU resources and enormous amount of data generated by the users makes sentiment analysis an active research field in upcoming years. Most of the existing approaches focus on efficient feature extraction, at the same time some approaches focus on extracting semantic features, which makes mush contribution to sentiment analysis. This review paper gives a comprehensive overview on Sentiment analysis using statistical techniques. We start with these techniques which are generally needed for preprocessing the input data. Then, we illustrate recent trend in sentiment analysis and its related works.

Applying sentiment analysis on Twitter is the upcoming trend with researchers recognizing the scientific trials and its potential applications. The challenges unique to this problem area are largely attributed to the dominantly informal tone of the micro blogging. Pak and Paroubek [5] rationale the use microblogging and more particularly Twitter as a corpus for sentiment analysis. They cited:

- Microblogging platforms are used by different people to express their opinion about different topics,thus it is a valuable source of people's opinions.
- Twitter contains an enormous number of text posts and it grows every day. The collected corpus can be arbitrarily large.
- Twitter's audience varies from regular users to celebrities, company representatives, politicians, and even country presidents. Therefore, it is possible to collect text posts of users from different social and interests groups.
- Twitter's audience is represented by users from many countries.

Parikh and Movassate [6] implemented two Naive Bayes unigram models, a Naive Bayes bigram model and a Maximum Entropy model to classify tweets. They found that the Naive Bayes classifiers worked much better than the Maximum Entropy model could.

Go et al. [7] proposed a solution by using distant supervision, in which their training data consisted of tweets with emoticons. This approach was initially introduced by Read [8]. The emoticons served as noisy labels. They build models using Naive Bayes, MaxEnt and Support Vector Machines (SVM). Their feature space consisted of unigrams, bigrams and POS. The reported that SVM outperformed other models and that unigram were more effective as features. Pak and Paroubek [5] have done similar work but classify the tweets as objective, positive and negative. In order to collect a corpus of objective posts, they retrieved text messages from Twitter accounts of popular newspapers and magazine, such as "New York Times", "Washington Posts" etc. Their classifier is based on the multinomial Naïve Bayes classifier that uses N-gram and POS-tags as features.

Barbosa et al. [2] too classified tweets as objective or subjective and then the subjective tweets were classified as positive or negative. The feature space used included features of tweets like retweet, hashtags, link, punctuation and exclamation marks in conjunction with features like prior polarity of words and POS of words. Mining for entity opinions in Twitter, Batra and Rao [9] used a dataset of tweets spanning two months starting from June 2009. The dataset has roughly 60 million tweets. The entity was extracted using the Stanford NER, user tags and URLs were used to augment the entities found. A corpus of 200,000 product reviews that had been labeled as positive or negative was used to train the model. Using this corpus, the model computed the probability that a given unigram or bigram was being used in a positive context and the probability that it was being used in a negative context.

Bifet and Frank [3] used Twitter streaming data provided by Firehouse, which gave all messages from every user in real-time. They experimented with three fast incremental methods that were well-suited to deal with data streams: multinomial naive Bayes, stochastic gradient descent, and the Hoeffding tree. They concluded that SGD-based model, used with an appropriate learning rate was the best.

Agarwal et al. [10] approached the task of mining sentiment from twitter, as a 3-way task of classifying sentiment into positive, negative and neutral classes. They experimented with three types of models: unigram model, a feature-based model and a tree kernel-based model. For the tree kernel-based model they designed a new tree representation for tweets. The feature-based model that uses 100 features and the unigram model uses over 10,000 features. They concluded features that combine prior polarity of words with their parts-of-speech tags are most important for the classification task. The tree kernel-based model outperformed the other two. The Sentiment Analysis tasks can be done at several levels of granularity, namely, word level, phrase or sentence level, document level and feature level [11].

III RESEARCH METHODOLOGY

3.1Work done in the project is depicted as follows:

STEP 1: GETTING DATA FROM TWITTER:

Getting Twitter API keys

In order to access Twitter Streaming API, we need to get 4 pieces of information from Twitter: API key, API secret, Access token and Access token secret. Follow the steps below to get all 4 elements:

1. Create a twitter account if you do not already have one.
2. Go to https://apps.twitter.com/ and log in with your twitter credentials.
3. Click "Create New App"
4. Fill out the form, agree to the terms, and click "Create your Twitter application"
5. In the next page, click on "API keys" tab, and copy your "API key" and "API secret".
6. Scroll down and click "Create my access token" and copy your "Access token" and "Access token secret".

STEP 2: TWEETS PREPROCESS:

Sentiment analysis helps us gauge sentiment of tweets, however many of the tweets we get from the API might really not be 'classifiable' into some sentiment. In order to fit our model to our dataset we need to clean and process our data

Steps for data cleaning:

Removal of HTML tags and symbols: When we take data from web pages then some dynamic content is converted into html tags. The symbol @ is used for giving reference to any link or user.

```
cleanTweet=gsub("@\\w+", "", cleanTweet)
```

Removal of Punctuations :For example: “.”, “,”,””?” need to be removed to create text only file.

```
cleanTweet=gsub("[[:punct:]]", "", cleanTweet)
```

Removal of Retweets: We will remove retweets from the list to provide a "pure" set of tweets.

```
cleanTweet=gsub("rt|RT", "", SearchText)
```

Removal of URLs: URLs and hyperlinks in text data should be removed.

```
cleanTweet=gsub("http\\w+", "", cleanTweet)
```

STEP 3: SENTIMENT EXTRTACTIONS:

The package comes with four sentiment dictionaries and provides a method for accessing the robust, but computationally expensive, sentiment extraction so that you can quickly extract plot and sentiment data from your own text files.

```
library("syuzhet")
```

The NRC sentiment dictionary containing function to calculate the presence of eight different emotions and their corresponding valence in a text file. A data frame where each row represents a sentence from the original file. The columns include one for each emotion type as well as a positive or negative valence. The ten columns are as follows: "anger", "anticipation", "disgust", "fear", "joy", "sadness", "surprise", "trust", "negative”, “positive”.

```
d<- get_nrc_sentiment(cleanTweet)
```

This is a technical simplification, is compared with text word in text file then according to comparison based on word that these tokens combine with emotion-strengthening components to form sentiment then the system calculates sentiment value (positive/negative) and polarity score for tweets. This is how whole process of the extraction of text file in the sentiments is done.

Using R-Studio
According to comparison based on word that these tokens combine with emotion-strengthening components to form sentiment then the system calculates sentiment value (positive/negative) and polarity score for tweets. This is how whole process of the extraction of text file in the sentiments is done.

STEP 4: SENTIMENT ANALYSIS AND GRAPH

The count of sentiment evaluated from the tweets text is represented using graphical visualization. The R provides library “ggplot2” which contain below function to show the result of sentiment analysis.

```
library("ggplot2")
```

3.2 Proposed Architecture

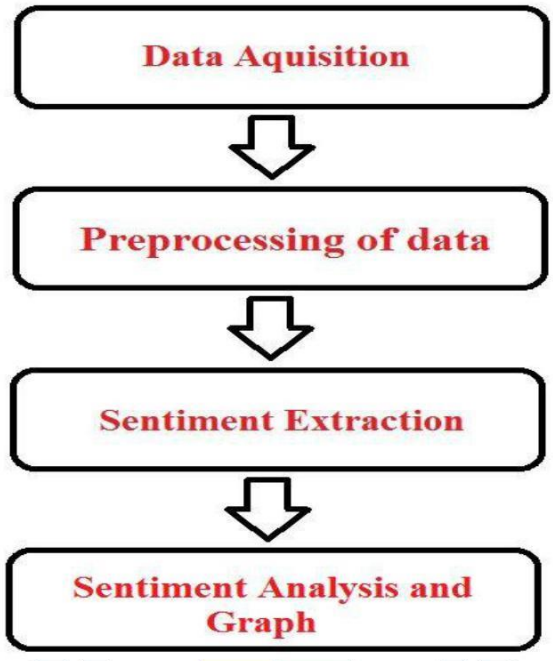

Fig 1: Sentiment Analysis Framework

3.3 Basic Step

Step -1. Data acquisition
Collecting the reviews and comments of social media sites. Data acquisition is the process of sampling signals that measure real world physical condition and converting the result samples into digital numeric values that can be manipulated by computer.

Step -2. Processing of data
This step involves filtration and removal of irrelevant content comments from the collected dataset. Data processing is the collection and manipulation of items of data to produce meaningful information. It is conversion of data into usable and desired form. This conversion is carried out using a predefined sequence of operations either manually or automatically.

Step -3. Sentiment Extraction
Now from the filtered dataset extraction of user sentiments and emotion using statistic tool. Sentiment is extracted on a particular and specific trending or intending topic. This extracted data can be used in different ways and in a wide range.

Step -4. Sentiment analysis and Visualization
Final from extracted result classifying the feedback as positive or negative and representing in graphical form. The accuracy of a sentiment analysis is in principle, how well it agrees with human judgments. According to our analysis we have polarity like positive and negative, but we also discuss about advance polarity like anger, happy, sad, anticipation etc.

DATA COLLECTION:
Sentiment analysis will be performed on user tweets of twitter. So, data required for analysis is tweets, it will be collected from twitter using API (Application Programming Interface). This API provides gateway for accessing data.

DATA PREPROCESSING:

Data collected from tweets cannot be used directly. It contains Emojis, special characters, retweets, user details and timestamps which is unnecessary. We need only text data for analysis. Text data will be separated is CSV (Comma Separated value) file.

R-STUDIO TOOL:

RStudio is a free and open source integrated development environment for R, a programming language for statistical computing and graphics. RStudio supports operating systems like windows, macOS, Ubuntu, Fedora, Red Hat Linux, openSUSE.

GRAPH VISUALIZATION:

Using the tweets positive and negative value, the graph will be plotted using graphic library. the graph will help to understand the summarized effect of post on general public about that event.

IV. RESULTS AND DISCUSSIONS

A live Twitter comment is collected under the keywords entered by the user. Approximately over 1500 tweets are then stored as a csv file for analysis. Sentiment analysis is computational task of automatically determining what feelings a writer expressing in text. The accuracy of a sentiment analysis is in principle, how well it agrees with human judgments. According to our analysis we have polarity like positive and negative but we also discuss about advance polarity like anger, happy, sad, anticipation etc. we have generate the keys and access token through which we access the tweets of twitter and after fetching that tweets we just done a text mining to analysis a text whether it is in positive side or negative side, the text mining is done my automation by giving an command to verify with the dictionary that in which field polarity is best suitable for it. After deciding the polarity, we plot the graph as shown below we get the total count of the polarity which help to easily analysis. Sentiment analysis is determining what feelings a writer expressing in text.

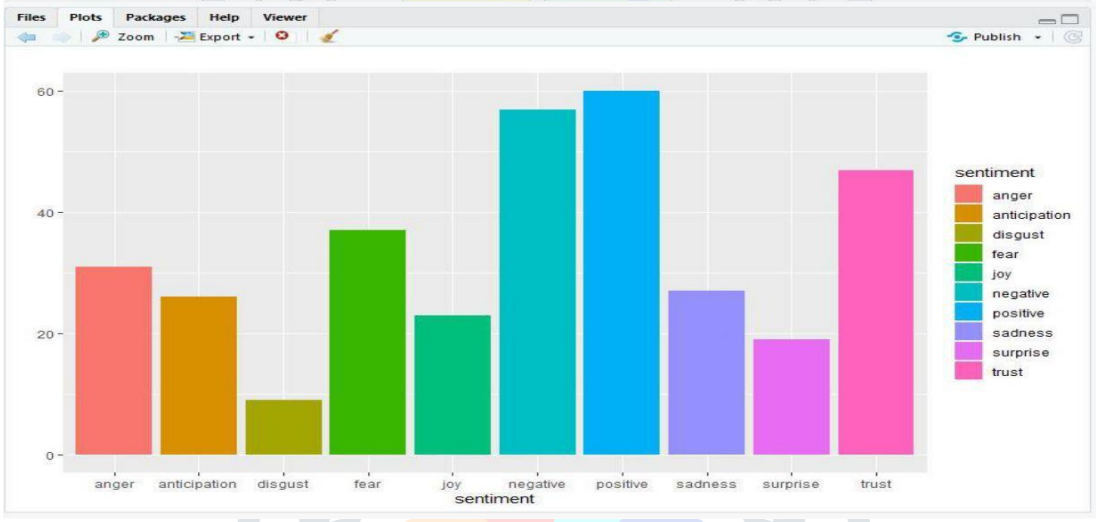

Fig 2: Graph of analyzed sentiments on “Pulwama Attack”.

The above graph shows the analysis of sentiment on the recent Pulwama Terror Attack. Towering amount of people have expressed their sentiment on twitter.

V. CONCLUSION AND FUTURE SCOPE

The task of sentiment analysis is still in the developing stage and far from completion. So, we propose a couple of ideas which we feel are worth exploring in the future and may result in further improved performance. In this project, we have done sentiment analysis on twitter trending topics and got a feedback about the sentiments of different users in graphical form. We successfully removed the redundancy of data that can affect the result of analysis.

In this research we are focusing on general sentiment analysis. There is potential of work in the field of sentiment analysis with partially known context. For example, we noticed that our analysis shows about the sentiments of peoples like joy, fear, trust, anticipation, etc. So, we can attempt to perform separate sentiment analysis on tweets that only belong to one of these classes (i.e. the training data would not be general but specific to one of these categories) and compare the results we get if we apply general sentiment analysis on it instead.

In future we can create an app in which front end will handled by C# and the back end will be handled by R. The app will extract tweets from the API and will give a graphical representation of sentimental analysis.

VI. ACNOWLEDGEMENT

I would like to thank our guide Ms. Devika Deshmukh who supported us through the different phases of project. Also we are grateful to Rajiv Gandhi College Of Engineering And Research for providing us the resources which led to successful implementation of the project.

VII. REFERENCES

[1]. Ms. Devika Deshmukh, Ms. Ashwini Yerkar, 2019. Investigating sentiment analysis using clustering and NPL tool.

[2]. Barbosa, L., and Feng, J. 2010. Robust sentiment detection on twitter from biased and noisy data. In Proc. of Coling.

[3]. Bifet, A., and Frank, E. 2010. Sentiment knowledge discovery in twitter streaming data. In Proc. of 13th International Conference on Discovery Science.

[4]. Kim, S.-M., and Hovy, E. 2004. Determining the sentiment of opinions. In Proceedings of Coling.

[5] A. Pak and P. Paroubek. "Twitter as a Corpus for Sentiment Analysis and Opinion Mining". In Proceedings of the Seventh
Conference on International Language Resources and Evaluation, 2010, pp.1320–1326.

[6] R. Parikh and M. Movassate, "Sentiment Analysis of UserGenerated Twitter Updates using Various Classification Techniques", CS224N Final Report, 2009

[7] A. Go, R. Bhayani, L.Huang. "Twitter Sentiment Classification Using Distant Supervision". Stanford University, Technical Paper ,2009

[8] J. Read. "Using emoticons to reduce dependency in machine learning techniques for sentiment classification". In Proceedings of ACL-05, 43nd Meeting of the Association for Computational Linguistics. Association for Computational Linguistics, 2005

[9] S. Batra and D. Rao, "Entity Based Sentiment Analysis on Twitter", Stanford University,2010

[10] A. Agarwal, B. Xie, I. Vovsha, O. Rambow, R. Passonneau, "Sentiment Analysis of Twitter Data", In Proceedings of the
ACL 2011 Workshop on Languages in Social Media,2011 , pp. 30–38

[11] A. Kumar. and T. M. Sebastian, "Sentiment Analysis: A Perspective on its Past, Present and Future", International Journal of Intelligent Systems and Applications (IJISA), MECS Publisher, 2012 (Accepted to be published) .