# Twitter Sentiment Analysis using bag of words

*Aayushi Somani[1], Vishaal Sharma[2], Kamalnath Irrukulla[3], Poornima Deshpande[4], Amal Krishna A[5], Herat Makwana[6], Ishank Jain[7], Ruturaj Chothe[8], Shubham Pawar[9], Rushikesh Khot[10], Anjali Shah[11], Aishwarya Salimath[12]*

*[1,2,3,4]Sinhgad College of Engineering, Pune*
*[5]National Institute of Technology, Rourkela*
*[6,7]National Institute of Technology, Surat*
*[8,9,10,11]RIT, Islampur*
*[12]WIT, Solapur*

**Abstract—** Social media analysis is the process of collecting data from popular social networking services and predicting the public view on any given domain based on the analysis of the collected data. This is achieved using machine learning and natural language processing techniques, along with various python libraries such as matplotlib, tweepy and textblob. The sentiment analysis tool has a simple user interface, which asks for a keyword based on which analyses of the tweets containing the keyword are segregated and statistically represented in terms of the sentiments being expressed as positive, negative or neutral.
The first step towards training the machine learning model begins with collecting the datasets that are made available using the Twitter API, which is achieved using the Tweepy library. It offers labelled datasets that can be used to efficiently train the machine learning model. The tweet processing and classification is done using the textblob library, which offers a simple API for natural language processing tasks such as sentiment analysis and classification. With this text classifier, we can label each Tweet as positive, negative or neutral of the sentimental value in a few minutes. However, human language is complex. Teaching a machine to analyse the various grammatical nuances along with diverse cultural variations, slang and misspellings that occur in social media make the process complex and teaching a machine to understand how the context affects the tone proves to be quite challenging. Sentiment analysis has its own limitations like any other ML predictive and is not used as a 100% accurate marker but with a little supervision it can be a great asset.

***Index Terms*— Sentiment analysis, Tweepy, TextBlob, Twitter**

## I. INTRODUCTION

Twitter is a popular microblogging service where users create status messages (called "tweets"). These tweets sometimes express opinions about different topics.

Sentiment analysis is the prediction of emotions in a word, sentences or corpus of documents. It is intended to serve as an application to understand the attitudes, opinions and emotions expressed within an online mention. Precisely, it is a paradigm of categorising conversations into positive, negative or neutral labels.

The purpose of this project is to build an algorithm that can accurately classify Twitter messages as positive or negative, with respect to a query term. Our hypothesis is that we can obtain high accuracy on classifying sentiment in Twitter messages using machine learning techniques.

Generally, this type of sentiment analysis is useful for consumers who are trying to research a product or service, or marketers researching public opinion of their company.

However, doing the analysis of tweets that express human emotions isn't an easy job. A lot of challenges are involved in terms of tonality, polarity, lexicon and grammar of the tweets. They tend to be highly unstructured and non-grammatical and therefore it get's difficult to interpret their meaning.

## II. PROCEDURES AND TOOLS

### *A.Introduction to the problem*

Every day massive amount of data is generated by social media users which can be used to analyse their opinion about any event, movie, product or politics. Sentiment analysis, also refers as opinion mining, is a sub machine learning task where we want to determine which is the general sentiment of a given document. Using machine learning techniques and natural language processing we can extract the subjective information of a document and try to classify it according to its polarity such as positive, neutral or negative.

In this project we are trying to classify tweets from Twitter into 'positive', 'negative' or 'neutral' sentiment by building a model based on probabilities. Twitter is a microblogging website where people can share their feelings quickly and spontaneously by sending tweets that are limited by 140 characters. You can directly address a tweet to someone by adding the target sign '@' or participate in a topic by adding a hashtag '#' to your tweet. Because of the current usage of Twitter in every field, it is the perfect source of data to determine the overall opinion about anything.

### *B.Libraries and Technologies*

- T*weepy*

It is the python client for the official twitter API. When we invoke an API method most of the time returned back to us will be a Tweepy model class instance. This will contain the data returned from Twitter which we can then use inside our application.

- *TextBlob*

It is a high level library built on top of the *NLTK library*. First, the clean_tweet method is called to remove links, special characters, etc from tweet using some simple *Regex*. Then, we pass tweet to create a *TextBlob* object.

- *Csv*

The Comma Separated Values format is a common import and export format for spreadsheet and databases. There exists no standards for csv operations, it is defined by many applications which read and write it. It implements classes to read and write tabular data in csv format.

- *re*

Python supports regular expressions by the library called *'re'*(though it's not fully Perl-compatible). *'Regular Expressions (RegEx)'* is one of the 'rules' based pattern search method. Instead of regular strings, search patterns are specified using raw strings "r", so that backslashes and meta characters are not interpreted by python but sent to RegEx directly.

- *sys*

System-specific parameters and functions. This module provides access to some variables used or maintained by the interpreter and to functions that interact strongly with the interpreter. It is always available. The list of command line arguments passed to a Python script.

- *Pandas*

Pandas is an open-source Python Library providing high-performance data manipulation and analysis tool using its powerful data structures.Python with Pandas is used in a wide range of fields including academic and commercial domains including finance, economics, Statistics, analytics, etc.

- *matplotlib*

It is used to 2D plot the arrays. It is a multi-platform data visualisation library built on *NumPy* arrays and designed to work with the broader *SciPy* stack. It allows us to visually access huge amount of data in some easy digestible visuals.

### *C.Classifier*

The list of word features need to be extracted from the tweets. It is a list with every distinct words ordered by frequency of appearance. We use the following function to get the list plus the two helper functions.To create a classifier, we need to decide what features are relevant. To do that, we first need a feature extractor. The one we are going to use returns a dictionary indicating what words are contained in the input passed. Here, the input is the tweet. We use the word features list defined above along with the input to create the dictionary. With our feature extractor, we can apply the features to our classifier using the method apply_features. We pass the feature extractor along with the tweets list.

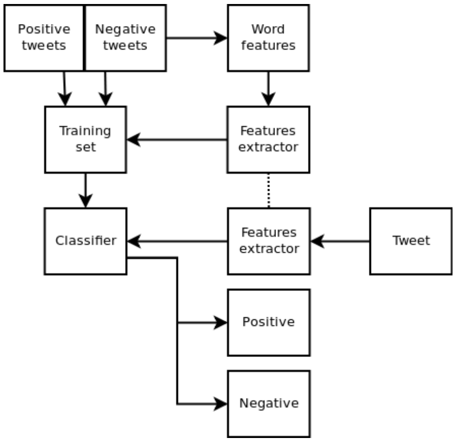


Fig. 1. The above illustration depicts the flow of the analysis process.

## III. LITERATURE REVIEW

A Pappu Rajan and S.P Victor [2014] presented a paper for web sentiment analysis for scoring positive or negative words using twitter data. Here, they are using the concept of Opinion Mining.

Elena Rudkowsky, Martin Haselmayer, Matthias Wastian, Marcelo Jenny, Stefan Emrich and Micheal Sedlmair [2017] presented a paper in which they have moved beyond the dominant approach of bag-of-words for sentiment analysis and introduced an alternative procedure based on distributed word embeddings.

Alec Go, Lei Huang, Richa Bhayani [2009] presented a paper for sentiment analysis where they have used the litmus test which is if the tweet could appear as a newspaper headline or as a sentence in Wikipedia, then it belongs in the neutral class.

Amandeep Kaur, Deepesh Khaneja, Khushboo Vyas, Ranjit Singh Saini [2016] presented a paper on Sentiment analysis on twitter using Apache Spark. Here, they are using Apache Spark to analyse real time tweets and their objective is to find the polarity of words in tweets as they are retrived.

Marc Lamberti [2015] presented a paper on Twitter Emotion analysis where he builds a model based on probabilities by analysing the tweets into 'positive' or 'negative'.

## IV. SOCIAL APPLICATIONS

The applications of sentiment analysis are broad and powerful. The ability to extract insights from social data is a practice that is being widely adopted by organisations across the world.

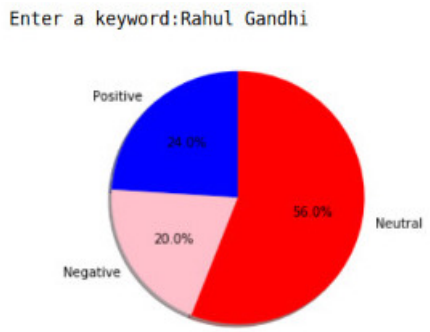
Sentiment analysis is extremely useful in social media monitoring as it allows us to gain an overview of the wider public opinion behind certain topics.

Shifts in sentiment on social media have been shown to correlate with shifts in the stock market.

The Obama administration used sentiment analysis to gauge public opinion to policy announcements and campaign messages ahead of 2012 presidential election. Being able to quickly see the sentiment behind everything from forum posts to news articles means being better able to strategise and plan for the future.

It can also be an essential part of your market research and customer service approach. Not only can you see what people think of your own products or services, you can see what they think about your competitors too. The overall customer experience of your users can be revealed quickly with sentiment analysis, but it can get far more granular too.

The ability to quickly understand consumer attitudes and react accordingly is something that Expedia Canada took advantage of when they noticed that there was a steady increase in negative feedback to the music used in one of their television adverts



```
Enter a keyword:Rahul Gandhi
```

Fig. 2. The above illustration depicts a pie chart of the twitter sentiment analysis of the keywords 'Rahul Gandhi'

## V. CONCLUSION

Twitter is a source of vast unstructured and noisy data sets that can be processed to locate interesting patterns and trends. Machine learning techniques perform reasonably well for classifying sentiment in tweets. This research introduce the theoretical basic of opinion mining. The proposed approach determines the sentiment of the text, whether it is positive or negative, which is extended to strength of polarity and also which was obtain the significant features and to Analysing the overall sentiment for each object by computing the weighted average for all the sentiments in the textual data.

## REFERENCES

1. Amandeep Kaur, Deepesh Khaneja, Khusbhoo Vyas, Ranjit Singh Saini, October' 2017 | Carlton University | Paper on Sentiment Analysis on twitter using Apache spark
2. Dr David Rossiter, Marc Lamberti, 21 July 2015 | Paper on Twitter Emotion Analysis
3. A Pappu Rajan, S.P.Victor, St.Xavier's College, 6 June 2014 | Web Sentiment Analysis for Scoring Positive or Negative Words using Twitter Data
4. Elena Rudkowsky, Martin Haselmayer, Matthias Wastian, Marcelo Jenny, Štefan Emrich, and Michael Sedlmair, University of Vienna, 2015 | More than Bags of Words: Sentiment Analysis with Word Embeddings