# Twitter Data Sentiment Analysis and Visualization

R. S. Gound
Professor
Dept. of IT
PCCOE, Nigdi,
Pune, India

Priyanka V. Tikone
Student: Dept. of IT
PCCOE, Nigdi
Pune, India

Shivani S. Suryawanshi
Student: Dept. of IT
PCCOE, Nigdi
Pune, India

Dipanshu Nagpal
Student: Dept. of IT
PCCOE, Nigdi
Pune, India

## ABSTRACT
Twitter is an online microblogging and social networking platform, which allows users to write short status, updates of maximum length 280 characters. These tweets reflect public sentiment about various topics and events happening. Analysing the public sentiment can help, firms trying to find out the response of their products in the market, predicting political elections and predicting socioeconomic phenomena like stock exchange. Sentiment analysis techniques are widely popular for this purpose. In this paper, we have tried to define and compare various sentiment classification approaches/methods for finding out the sentiments behind the tweet.

## Keywords
Lexicon, Machine Learning, Natural Language Processing, Sentiment Analysis, Twitter

## 1. INTRODUCTION
Twitter is a widely popular micro-blogging platform for users to express their opinions about governmental issues, product items, sports and so forth. A tweet is a text-based post and has only 280 characters. Twitter is a "what's happening-right-now" social network and hence tweets are valuable sources for businesses, government and individuals to determine public's opinion or sentiment about an entity (product, people, topic, event etc.) [1]. Tweets reflect those events as seen by the individuals tweeting, and can be aggregated to form the basis of event exploration and visualization [4]. However, the volume of tweets produced by Twitter every day is very vast. Hence, there is a need to automate the process of sentiment analysis to ease the tasks of determining public's opinions without having to read millions of tweets manually [1], [2]. This process of analysing and summarizing the opinions expressed in these huge opinionated user generated data is usually called Sentiment Analysis or Opinion Mining.
Sentiment analysis aims to determine the attitude of a speaker, writer, or other subject with respect to some topic or the overall contextual polarity or emotional reaction to a document, interaction, or event [7]. The attitude may be a judgment, affective state (the emotional state of the author or speaker), or the intended emotional communication (the emotional effect intended by the author or interlocutor). The sentiment can be classified as either positive, negative or neutral.

## 2. LEVELS OF SENTIMENT CLASSIFICATION

The sentiment classification can be done on three levels

### 2.1 Sentence Level Sentiment Classification
In sentence level sentiment classification, each sentence is first classified as subjective or objective. Only subjective sentences are useful for sentiment classification. Hence, the objective sentences are discarded and the polarity of subjective sentences is calculated. According to the polarity, the sentence is classified as positive, negative or neutral.

### 2.2 Document Level Sentiment Classification
In this approach, whole document is considered as one unit for classification into positive, negative or neutral category.

### 2.3 Aspect/Feature Level Sentiment Classification
Aspect or Feature level sentiment classification deals with identifying and extracting product features from the source data [1]. It is based on the idea that an opinion consists of a sentiment (positive or negative) and a target (of opinion) [3].

## 3. SENTIMENT CLASSIFICATION TECHNIQUES
There are mainly three techniques of sentiment classification. The Lexicon Based Approach uses dictionary, which contains positive and negative words to do the sentiment classification. The Machine Learning Based approach uses various supervised as well as unsupervised algorithms for classification purpose. The Hybrid Approach is the combination of both lexicon based approach and machine learning approach.

### 3.1 Lexicon Based Approach
Lexicon based approach for sentiment classification deals with classifying the sentiments using an opinion lexicon. An opinion lexicon is a collection of positive and negative words. To classify a sentence as either positive or negative, the number of positive and negative words in a sentence is calculated. If the sentence contains more number of positive words then the sentence is classified as positive. If the sentence contains more number of negative words, then the sentence is classified as negative. If the sentence contains equal number of positive and negative words, then the sentence is classified as neutral. The opinion lexicon can be constructed in several ways:

#### *3.1.1 Dictionary Based Approach*
In this approach, initially a small set of words with known orientations are collected manually [1]. This collection of words is called the seed list. Then this seed list is increased by searching the antonyms and synonyms of the seed words in known corpora like WordNet or thesaurus. This antonyms and synonyms are added into the seed list. This process continues until there are no words to be added. After that, the created

dictionary can be checked manually for any errors. The limitation of this approach is that the accuracy of classification depends on the size of dictionary. Also using this approach finding opinion words with context specific orientations is not possible.

#### *3.1.2 Corpus Based Approach*

Corpus based approach overcomes the limitation of dictionary based approach regarding the ability to find out opinion words of context specific orientations. They depend on large corpora for syntactic and semantic patterns of opinion words [1]. The limitation with this approach is that, it is difficult to prepare large corpora to cover all English words.

### 3.2 Machine Learning Based Approach

Machine learning approach for sentiment classification uses two datasets i.e. training dataset and testing dataset. The supervised and unsupervised machine learning algorithms are first applied on training dataset. The classifier trains itself with respect to differentiating attributes of text. The model obtained after training is applied on test data which is unseen. Machine learning technique for sentiment classification starts with collection of tweets. These tweets can be labelled or unlabeled. These tweets are noisy. They may contain words, punctuation marks or special characters, which do not express any sentiment. Hence, these tweets are first pre-processed to remove noise. Then the features relevant to sentiment analysis are extracted from the tweets. This forms the training data. A machine learning classifier is trained on this training dataset and it is then tested on the unseen test dataset.

### 3.3 Hybrid Approach

Hybrid approach of sentiment classification is the combination of both lexicon and machine learning techniques. It combines the best practices from both the approaches to come up with more efficient way of sentiment classification. The lexicon/learning combination has proven to improve accuracy. Lexicon based approach have high precision and low recall. Hence combining it with a machine learning classifier can improve the recall and accuracy of the algorithm [1], [2].

## 4. PROPOSED SYSTEM

In this system, we are using machine learning based approach for sentiment classification. For this, we are constructing dataset of tweets, which are obtained from Twitter using Tweepy API. After obtaining tweets, they are pre-processed to remove the noise. The tweets are labelled as either positive, negative or neutral. After pre-processing, useful *and* significant features are extracted from tweets. The machine learning classifiers are applied on the training dataset. The model obtained from training, is applied on unseen test dataset to check the accuracy of the model. A web application will be created which will display the results of the classification. The results are visualized and displayed on website for user convenience.

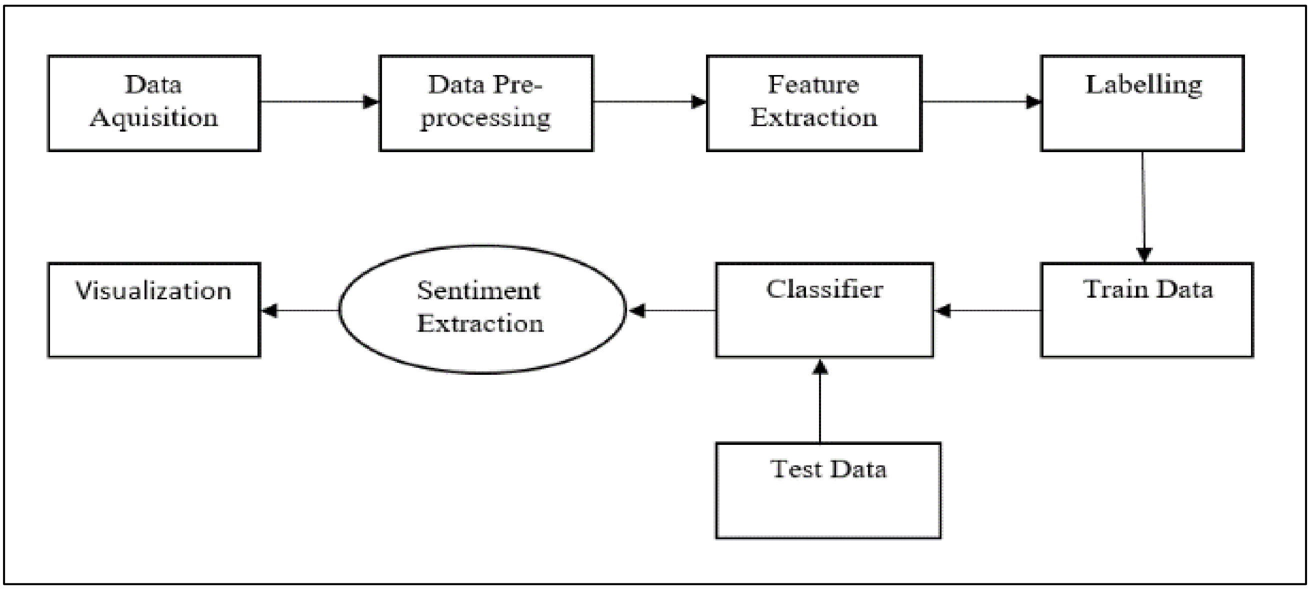


**Fig 1: System Architecture**

### 4.1 Data Collection

Twitter allows researchers to collect tweets by using a Twitter API. To collect tweets from twitter one must have a twitter account to obtain twitter credentials (i.e. API key, API secret, Access token and Access token secret) which can be obtained from twitter developer site. Then the user needs to install a library to connect to the twitter API using these credentials. Now tweets can be extracted from twitter.

### 4.2 Data Preprocessing

The tweets collected from twitter are noisy and boisterous. They often contain hyperlinks, stop words, punctuation marks or special characters, which do not portray any emotion and hence are of no use in sentiment classification. These surplus entities may affect the performance of the classifier adversely. Hence, we need to remove this noise first. Using Natural Language Processing techniques, this noise can be handled.

### 4.3 Feature Extraction

Once the tweets are pre-processed, we need to extract features relevant and significant to sentiment analysis. Extracting proper features from tweets is most important, as the performance of classification depends upon the features extracted.

### 4.4 Application of Machine Learning Classifier:

Once the features are extracted and training dataset is formed, next comes the step in which a machine learning classifier is applied on the training dataset. There are various supervised as well as unsupervised machine-learning classifiers for sentiment analysis. The model obtained from the training dataset is applied on the unseen test dataset, to check the accuracy and performance of the model. Supervised classifiers such as Naïve Bayes classifier, Support Vector Machine, decision tree algorithms can be used for classification of sentiments.

#### *4.4.1 Naïve Bayes Classifier:*

Naive Bayes classifier works very well for text classification as it computes the posterior probability of a class, based on the distribution of the words (features) in the document [1], [2]. The model uses the Bag of words feature extraction. Naïve Bayes classifier assumes that each feature is independent of each other. It uses Bayes Theorem to predict the probability that a given feature set belongs to a particular label.

$$P(\text{label} \mid \text{features}) = \frac{P(\text{features} \mid \text{label}) * P(\text{label})}{P(\text{features})}$$

P(label) is the prior probability of a label or the likelihood that a random feature set the label. P(features | label) is the prior probability that a given feature set is being classified as a label. P(features) is the prior probability that a given feature set is occurred.

#### *4.4.2 Support Vector Machine*

The main principle of SVMs is to determine linear separators in the search space, which can best separate the different classes [1], [2]. There can be several hyperplanes, but SVM classifier chooses the one, which gives the maximum distance for any point. The hyperplane chosen should depict maximum margin of separation. Text classification are perfectly suited for SVMs because of the sparse nature of text, in which few features are irrelevant, but they tend to be correlated with one another and generally organized into linearly separable categories[1], [2].

#### *4.4.3 Decision tree*

Decision tree classifier provides a hierarchical decomposition of the training data space in which a condition on the attribute value is used to divide the data. The condition or predicate is the presence or absence of one or more words. The division of the data space is done recursively until the leaf nodes contain certain minimum numbers of records, which are used for the purpose of classification [1].

### 4.5 Visualization

The results obtained from the experiment will be visualized in the form of bar graph, pie chart, time series graph etc. The visualized results will be made available on the website for end user.

## 5. CONCLUSION AND FUTURE SCOPE

This study proposes a sentiment analysis system using machine-learning approach. The study is based on sentence level sentiment classification. It is useful for finding out sentiments of people regarding any topic or event. It is also useful in predicting socio-economic phenomena like stock market prediction. As a future work, this study can be expanded to include feature/aspect level classification, which is useful in product review and recommendation system. The number of sentiment classes can be increased to get more refined sentiment prediction. The study is a prototype and is meant to present the potential use of social networking platforms such as Twitter for large scale information gathering and processing for future social media related applications. The study can be extended for applications such as emergency management, social unrest etc. The study can be extended.

## 6. REFERENCES

[1] Anuja P Jain, Padma Dandannavar "Application of Machine Learning Techniques to Sentiment Analysis", 2nd International Conference on Applied and Theoretical Computing and Communication Technology (iCATccT), July 2016.

[2] Walaa Medhat, Ahmed Hassan, "Sentiment analysis algorithms and applications:A survey" Shams Engineering Journal (2014) 5, 1093 – 1113.

[3] Bing Liu, "Sentiment Analysis and Opinion Mining", Morgan and Claypool Publishers, May 2012.p.18-19,27-28,44-45,47,90-101.

[4] Andrei Sechelea, Tien Do Huu, Evangelos Zimos, and Nikos Deligiannis "Twitter Data Clustering and Visualization", 23rd International Conference on Telecommunications (ICT).

[5] Martin Sarnovsky, Peter Butka, Andrea Huzvarova "Twitter data analysis and visualizations using the R language on top of the Hadoop platform", IEEE 15th International Symposium on Applied Machine Intelligence and Informatics January 26-28, 2017.

[6] Rohit Joshi, Rajkumar Tekchandani, "Comparative Analysis of Twitter Data Using Supervised Classifiers", proceedings of the 2016 International Conference on Inventive Computation Technologies (ICICT).

[7] Afroze Ibrahim Baqapuri, "Twitter Sentiment Analysis",Department of Electrical Engineering, School of Electrical Engineering & Computer Science, National University of Sciences & Technology, Islamabad, Pakistan 2012.