



## REVIEW OF SENTIMENT ANALYSIS ON TWITTER DATA USING PYTHON

Prajakta Deshpande

Master of Computer Application  
Marathwada Institute of Technology  
Aurangabad, Maharashtra, India.

Vikrant Shaga

Master of Computer Application  
Marathwada Institute of Technology  
Aurangabad Maharashtra, India.

Surabhi Thorat

Master of Computer Application  
Marathwada Institute of Technology  
Aurangabad, Maharashtra, India.

**Abstract:** As we are aware of the social media is usually preferred to share their view towards an object, event. Social networking sites are very beneficial for organization for their business purpose. As people share their opinion about their product and service. Also customer want to know about existing user before they purchase that product. Form that data we can get the data on that data we can do processing and such processed data is used for sentiment analysis. It is very useful for business analysis, production, quality, sales. Processing such twitter or social media data is very challenging task but gaining knowledge from data represents business and benefits regarding product. In this paper we have tried to analyze opinion mining, Advantages, scope of data mining and tools required for it.

**Keywords:** Text mining, Sentiment analysis, Data mining, NLP, twitter data

## I. INTRODUCTION

Text mining refer to as data mining which is the process of deriving some quality information in which it must be followed with some pattern and analysis in that it must give meaningful information regarding the object .Text mining is the collecting all the text from twitter so that we can get the actual mind set of the people towards particular object, product or the event. These can be used to analyze the data also we can take certain decision from that data, so it would help the organization to imagine the reviews about that product. Data analysis of twitter data is very helpful for any of organization to reduce the efforts of taking feedback from people to go and take survey from society. By using twitter we can get data from twitter profile on that they can tweet in 140 words. From twitter data we can take people common approach towards particular product like 'computer companies' for the data analysis we have to take only one word belong to the particular subject such as 'Computer' in which those tweets extracted from the twitter like that they can classify with top most used or preferred computer companies "Dell", "HP", "Apple" ,"Lenovo". While going with text mining firstly we need to clean the given data such as removing the stop words ,preposition, emoji's ,special symbols like '#', '@',collecting the key words like 'computer', 'good', 'like' etc.[1].

As we have seen data mining can be used for analysis in which many of those have done some research regarding particular subject in which using twitter with help of new coding technology like 'Python' for data and real time environment analysis like newly done tweets on particular subject. For analysis we can use many of the techniques such as NLP, statistics, machine learning etc. From this we can get positive, negative, satisfactory or dissatisfactory opinion about the domain that we will going to analyze. Now a day's

people can share their views on social networking site such as face book, twitter, instagram, LinkedIn, Google ,YouTube many of objects like trends ,politics ,fashion, movies, technology etc. it is important to analyze for the business ,particular decision making area . As twitter has become more widely used micro-blogging site to share views about the domain [1].

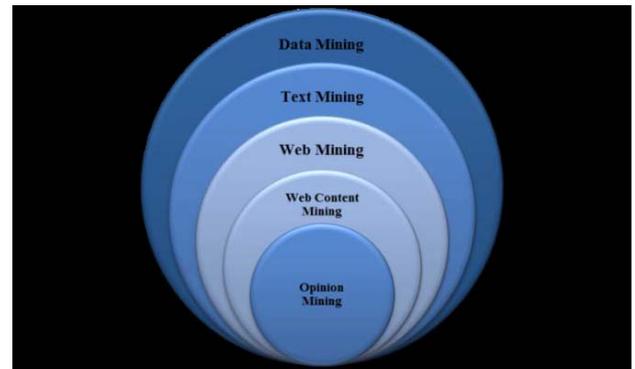


Figure 1: Evolution of Opinion Mining[ 2]

## II. WHY PYTHON IS RELEVANT TO DATA ANALYSIS?

Python is a excellent environment for building computationally intensive scientific application, highly concurrent, multithreaded applications. It is interpreted programming language used for testing, debugging also supporting for the twitter data analysis which must be followed by the natural language processing. All the flexibility regarding analysis could be provided by python [1].

### III. WHAT IS SENTIMENTAL DATA ANALYSIS?

Data analysis is used to sentimental analysis about particular subject from that we can find out the views or the opinion about the thing or the event. That review can be of positive, negative or neutral. For example , on twitter people can tweet about a apple product like I-phone series is that working properly?, Are they satisfied with the product, working of feature ,need some changes or not etc. From such data we can predict about the particular product that they are relevant for people or not. From data analysis it is very useful for any organization for their future prediction about product how long it will be in market, its business analysis, product quality and service of product. [5]

### IV. WHAT IS THE NEED OF SENTIMENT DATA ANALYSIS?

**1) Industrial Evolution-** By using sentimental data analysis they can complete find the useful data and add value to their gain using this customer to vendor everyone can get benefit.

**2) Decision making-** Social media mining will help to take decision for that also as it is increasing now a days this is very helpful for data analysis and decision making about any event, object.

**3)Understanding contextual** - Difficulty in understanding the human language such as change in culture ,variation time by time so it is the need to understand this facts as with the time.

**4) Internet marketing-** As now days increases in internet use most of the business are based on internet, blogs ,product .so tool of sentimental analysis are to be used .[1]

### V. METHODOLOGY FOR SENTIMENTAL ANALYSIS

For the data analysis we need some data for sentimental analysis data can be collected on the basis of real time .data can be extracted from twitter as it gives facility to extract such data from their database.

#### Data Collection:

Data is collected from twitter on the basis of tweets related to the computer companies such as HP, dell, Lenovo using this key words need to collect tweets related to it using twitter facility to collect data i.e. tweets from Twitter API . Twitter provides two types of API Rest API and Streaming API. Streaming API provides real time data or tweets done by people. For tweets collection need long and no limit data connection[1].

#### Twitter data collection:

Twitter provide a platform for data connection using the login id and API keys such as secret keys and Access tokens which gives u direct connection to twitter data using API by using this we firstly create Application which will be used for Streaming tweets provides the data for analysis. Now for twitter data analysis before creating script need to install **Tweepy** library on Python. [1].

#### Collecting data using Python:

Python is useful language now a days it provides access

to many libraries and services which enable python to communicate with twitter and API to collect tweets using Access tokens and secrete keys. For tweet collection we need to install library tweepy using command ‘pip install tweepy’ on command prompt and ready to get the data after script writing [1].

For getting data we need to set all the protocol i.e. ‘OAuth’ it is a standard protocol used for the authorization to user. We can get the data using the following script. Once we start getting data from twitter API we can collect data according to the computer companies on the basis of key words [1].

#### Data Pre-Processing-

Pre processing of data means the extracting key words from the given data such a way that it gives some meaningful or analytical information regarding the given event, object or thing. Preprocessing data means removing spaces, stop words, special symbols (!, @, \*),hash tags, duplicate words, emojis. Cleaning data is important as tweets contain many syntactical words which are not useful for analysis. Once the data is cleaned it is used for data analysis.[1] Also on social media we can share URL to identify such links and replace it with text or URL tags. As we are using streaming data process there are word like *writing*, *write*, *wrote*, *writer* all are reduce to one word *write*. [4]

#### Feature Determination –

Once the data is cleaned we need to extract feature related to the particular subject, event, object or thing such that which computer company is most reliable, preferred and good quality are extracted from tweets key words like good, satisfied ,processing is well, popular this are the positive key words towards the computer companies. There are some negative words related to the not working, bad, grumpy, stressed. Also there are emicons used in many tweets they can be classifies as smile, laugh , sad ,cry ,bad they are further classify into two types positive and negative.[1]

### VI. CLASSIFICATION-

Classifier consist of several machine learning classifier which are used to classify data according to the different python libraries like KNN, Naïve-bays classifier, rule based classifier, clustering analysis, regression analysis.

#### KNN algorithm

KNN is K-nearest neighbor algorithm used to find out the predictive measure problems and it is mostly used algorithm for computation of different analysis. It is very easy to calculate and apply it on given data. In this classifier algorithm input can be taken as object and its neighbors plays an important role in classes that will going to assign to the k nearest value of object. After the analysis of data using KNN results would be assumed as class membership values. It is used for distance measures values as follows : [5]

$$\sqrt{\sum_{i=1}^N (a_i - b_i)^2}$$

Euclidean distance function

as there are (a1, a2,...aN) and (b1, b2.... , bN) are the datasets these are the values on which we are going to apply KNN algorithm for distance calculation and by machine learning to train the dataset. As the class labels are to be used to define neighbors as K and positive or negative sign are to be used for assumption . Calculating this value is very tricky .choosing a correct values which give approximately correct value is agenda for KNN.[5]

Naïve bays classifier

Thomas Bayes he is the statistician discovered algorithm called the Bayesian theorem and naive bayes is based on this methods. Naive bayes is mostly used for classifying the textual data sets and this classification can be done on the basis of probabilities of each label column. It is very important to have large data set for calculation. In naive bayes theorem extracting such feature that are independent of each other.

Suppose we have D set of tuples and each tuples has attribute vector X(x1, x2, x3 , ... xn) of n dimensions. Let there are k number of classes C1, C2, C3... Ck. The classifier predicts X belongs to Ci if

Suppose there is set A contain tuples and they have attribute vector X and it belong to class Ci they are having number of classes 1,2,...,k.[5]

Bayes theorem given as follows,

$$P\left(\frac{C_i}{X}\right) = \frac{P\left(\frac{X}{C_i}\right) P(C_i)}{P(X)} \quad [4]$$

Statement can be evaluated as follows

1.  $P\left(\frac{C_i}{X}\right)$  is probability of occurrence Ci over the event X if it is true.
2. Ci and X are the events to be occurred .
3. P(X/Ci) is the probability event X over the Ci .

## VII. SCOPE OF SENTIMENTAL DATA ANALYSIS

Data analysis has large amount of applications in social media ,organization analysis is high on demand. Many of the companies preferring the sentimental analysis tools for business purpose , research as following:

- a) For the natural disasters such as earthquake can be monitored the deliver tweets and deliver notifications using constraints. They detect real time events using algorithms to monitor tweets using keywords they can be filtered according to the location.[4]
- b) Sentimental analysis can be done in event summarization using tweets gives continuous

representation for intermediate stage as per the relevant to the event.[4]

- c) For the joint topic modeling and their twitter feedback in such cases two different domain data can be collected and it can be associated with the same event such data is used for sentimental analysis accurately without misconception the events topic model is very helpful .[4]
- d) Unbalanced data sets are to use for the methods to solve the problem are under sampling & over sampling .it is very useful for reduction of number of documents of majority class by using sub methods remove by clustering. As we have seen classifier i.e. Naive Bayes and KNN are used to calculate accuracy of different datasets. Naive Bayes classifier seems insensitive to the unbalanced data and gives more accurate results than the support vector machine and K-NN which are sensitive to the unbalanced data. Multilingual sentiment classification is carried out successful.[4]
- e) Word of mouth is the process by which information is given from one person to another person . It would be useful for decision making, analysis, future scope ,related to the business , product review are to be shared by people. It can be analyze on the basis of review, blogs, social media and making it for decision making.[1]

## VIII. CONCLUSION

Sentimental analysis is used to identify people’s opinion, attitude and emotional state towards the particular event or object. Also for preprocessing and classifying data we have used different classifiers for data. Opinion mining is very useful for product reviews and analysis for the product buyer’s. Because of product reviews organization can find defects and modify according to the need of the customer. This would help to build customer relationship by satisfying their need. As we have seen that it will be extracting the feature from data and gaining knowledge from tweets of customer on product or event. Also from data we can find attracts and dislikes of customer. From data companies can save production cost as well as they can improve quality for product according to data analysis.

## IX. REFERENCES

- [1] Prateek Garg “Sentiment analysis of twitter data using NLTK in python”, <http://dspace.thapar.edu:8080/jspui/bitstream/10266/4273/4/4273.pdf>[online] (Nov 2017)
- [2] Surabhi Thorat “Opinion mining and sentimental Analysis-Its tools and Challenges” volumn 3 no.11, November 2014,ISSN 2319-4413.
- [3] Shruti wakade, Chandra shekhar, Kaythi J.Liszka, Chien-Chung Chan “Text mining for sentimental analysis of twitter data”,Worldcomp’12.July16-19
- [4] Bholane Savita datty, Prof.Deipali V. Gore “A survey on Sentimental Analysis on Twitter data using Different techniques.” Volume 6(6),2015,5358-5362.
- [5] Shubham Goyal “Sentimental Analysis of Twitter data using Text mining and Hybrid Classification Approach.” Volume 2 issue 5,ISSN 2454-123X.