**SENTIMENT ANALYSIS ON DEMONETIZATION TWEETS**

**Progress Report**

**BACHELOR OF TECHNOLOGY**

**In**

**Computer Science and Engineering**

**By**

**Abhishek Singh Adhikari (Roll no. 140180101001)**

**Ajay Rawat (Roll no. 140180101004)**

**Manoj Kumar (Roll no.140180101027)**

****

**Under the supervision of**

**Dr. Ajit Singh**

**Head Of the Department**

**Department of Computer Science and Engineering**

**CANDIDATE’S DECLARATION**

We hereby declare that the work, which is being presented in the report entitled, “**Sentiment Analysis On Demonetization Tweets”** in the fulfillment of the requirement for the award of the degree of Bachelor of Technology in Computer Science & Engineering, BTKIT Dwarahat, is an authentic record of our work carried out under the supervision of **Mr. Amit Paul**, Assistant Professor, Department of Computer Science & Engineering, BTKIT Dwarahat. The matter embodied in this report has not been submitted by us for the award of any other degree.

Date: (Abhishek Adhikari)

(Ajay Rawat)

(Manoj Kumar)

This is to certify that the above statement made by the candidate is correct to the best of my knowledge.

Date: Supervisor

(Mr. Amit Paul)

Assistant Professor

Department of Computer Science & Engineering

BTKIT Dwarahat

Forwarded by: Head

Department of Computer Science & Engineering

BTKIT Dwarahat

# 

# Acknowledgement



First of all, I would like to express my thanks to my guide **Dr Ajit Singh** HOD, Computer Science and Engineering Department, B.T.K.I.T, Dwarahat for being an excellent mentor for me during my whole course of project. His encouragement and valuable advice during the entire period has made it possible for me to complete my work.

I am also thankful to our coordinator **“Mr Vishal Kumar”** Assistant Professor Computer Science Engineering Department, B.T.K.I.T Dwarahat for setting high standards for his students and encouraging them time to time so that they can achieve them as well.

Lastly, I would like to thank my parents for their years of unyielding love and encourage. They have wanted the best for me and I admire their sacrifice and determination.

#### Abhishek Adhikari

#### Ajay Rawat

#### Manoj Kumar

# Abstract

In today’s world, Social Networking website like Twitter, Facebook, Tumbler, etc. plays a very significant role. Twitter is a micro-blogging platform which provides a tremendous amount of data which can be used for various applications of Sentiment Analysis like predictions, reviews, elections, marketing, etc. Sentiment Analysis is a process of extracting information from large amount of data, and classifies them into different classes called sentiments.

Python is simple yet powerful, high-level, interpreted and dynamic programming language, which is well known for its functionality of processing natural language data by using NLTK (Natural Language Toolkit). NLTK is a library of python, which provides a base for building programs and classification of data. NLTK also provide graphical demonstration for representing various results or trends and it also provide sample data to train and test various classifiers respectively.

The goal of this project is to classify twitter data into sentiments (positive or negative) by using different supervised machine learning classifiers on data collected for Demonetization occurred during ‘Nov 2016’ and to show how public reacted to this huge step of PMO India. We are also going to conclude which classifier gives more accuracy during classification.

**Table of Contents**



Front Page……………………………………………………………………………..1

Declaration…………………………………………………………………………….2

[Acknowledgement ……………………………………………………………………..3](#_TOC_250045)

[Abstract ………………………………………………………………………………..4](#_TOC_250044)

[Table of Contents 6](#_TOC_250043)

[List of figures 7](#_TOC_250042)

[List of Tables 8](#_TOC_250041)

[List of Abbreviations 9](#_TOC_250040)

[Chapter1:Introduction 1-11](#_TOC_250039)

* 1. [Introduction to Sentiment Analysis 1](#_TOC_250038)
  2. [Introduction to Python 2](#_TOC_250037)
  3. [Introduction to NLTK ……………………………………………………..2](#_TOC_250036)
  4. [Introduction to Supervised Machine learning Classifiers………………… 3](#_TOC_250035)
     1. Naïve-Bayes (NB)Classifier ……………………………………..3
  5. [Goal of Project ………………………………………………………….....8](#_TOC_250033)
  6. [Need of Sentimental Analysis ……………………………………………...9](#_TOC_250032)
     1. [Industry Evolution ………………………………………………..9](#_TOC_250031)
     2. [Research Demand ………………………………………………..9](#_TOC_250030)
     3. [Decision Making ……………………………………...…………10](#_TOC_250029)
     4. [Understanding Contextual ………………………………………10](#_TOC_250028)
     5. [Internet Marketing ………………………………………………10](#_TOC_250027)
  7. [Applications of Sentiment Analysis ………………………………………10](#_TOC_250026)
     1. [Word of Mouth(WOM) ………………………………………....10](#_TOC_250025)
     2. [Voice of Voters ………………………………………………….11](#_TOC_250024)
     3. [Online Commerce………………………………………………. 11](#_TOC_250023)
     4. [Voice of the Market(VOM) ……………………………………..11](#_TOC_250022)
     5. [Brand Reputation Management(BRM) …………………………11](#_TOC_250021)
     6. [Government ……………………………………………………..12](#_TOC_250020)

Chapter 2:ProblemStatement …...19-20

* 1. [Objectives 19](#_TOC_250019)
  2. [Methodology …………………………19](#_TOC_250018)

Chapter 3:Implementati on 21-31

* 1. [Proposed Architecture …………………………………………………….21](#_TOC_250017)
  2. [Data Collection 22](#_TOC_250016)
     1. [Twitter Data 22](#_TOC_250015)
     2. [Training Data 24](#_TOC_250014)
  3. [Data Storage 25](#_TOC_250013)
  4. [Data Pre-Processing 25](#_TOC_250012)
  5. [Classification 28](#_TOC_250011)
     1. [Feature Extraction 28](#_TOC_250010)

Chapter 4: Results and Analysis 32-36

* 1. [Tweets Collected 32](#_TOC_250009)
  2. [Extracted Features 32](#_TOC_250008)
  3. [Classifier Accuracy for Training Data ……………………………………33](#_TOC_250007)
  4. [Twitter Data Analysis 33](#_TOC_250006)

Chapter 5: Conclusion and Future Scope 37-38

* 1. [Conclusion ………………………………………………………………..37](#_TOC_250002)

5.2 [Future Scope 37](#_TOC_250001)

References 39-40

**List of Figures**



**Figure 3.1** Process to classify tweets using build classifier 21

**Figure 3.3** Database of collected tweets 25

**Figure 3.4** Code for extracting features from tweets 30

**Figure 3.5** Sample code for training and testing build classifier 31

**Figure 4.1** Sample Tweets Collected for demonetization 32

**Figure 4.2** Extracted features from training data 33

**Figure 4.3** Classifier accuracy for training data 33

**Figure 4.4** Sentiment analysis for each tweet, when Test Twitter data is analyzed 33

**Figure 4.5** Sentiment analysis overall results when Test Twitter data is analysis 34

# List of Tables

**Table 4.1** Sample movie reviews in NLTK Corpus ………………………………….24

**Table 4.2** Sample tweet and processed tweet ………………………………………...26

**Table 4.3** Removed and modified content ……………………………………………27

**Table 4.4** Sample cleaned data ……………………………………………………….27

## 

## List of Abbreviations



**NLTK**: Natural Language Toolkit

**NLP**: Natural Language Processing

**NB**: Naïve-Bayes

**OvR**: One-vs-Rest

**OvA:** One-vs-All

**WOM**: World of Mouth

**VOM**: Voice of the Market

**BRM**: Brand Reputation Management

**API**: Application programming Interface

**CSV**: Comma Separated Values

**URL**: Uniform Resource locator

**tf-idf**: term frequency-inverse document frequency

# Introduction

In this chapter we are going to give the introductions on Sentiment Analysis, Python and Natural Language Toolkit (NLTK). Then we are explaining the objective of our project. After this we will discuss why there is a need of sentiment analysis and some of the applications of Sentiment Analysis which are used in our daily life.

### Introduction to Sentiment Analysis

Sentiment Analysis is process of collecting and analyzing data based upon the person feelings, reviews and thoughts. Sentimental analysis often called as opinion mining as it mines the important feature from people opinions. Sentimental Analysis is done by using various machine learning techniques, statistical models and Natural Language Processing (NLP) for the extraction of feature from a large data.

Sentiment Analysis can be done at document, phrase and sentence level. In document level, summary of the entire document is taken first and then it is analyze whether the sentiment is positive, negative or neutral. In phrase level, analysis of phrases in a sentence is taken in account to check the polarity. In Sentence level, each sentence is classified in a particular class to provide the sentiment.

Sentimental Analysis has various applications. It is used to generate opinions for people of social media by analyzing their feelings or thoughts which they provide in form of text. Sentimental Analysis is used in many real life scenarios, to get reviews about any product or movies, to get the financial report of any company, for predictions or marketing.

Twitter is a micro blogging platform where anyone can read or write short form of message which is called tweets. The amount of data accumulated on twitter is very huge. This data is unstructured and written in natural language. Twitter Sentimental Analysis is the process of accessing tweets for a particular topic and predicts the sentiment of these tweets as positive, negative or neutral with the help of different machine learning algorithm.

### Introduction to Python

Python is a high level, dynamic programming language which is used for this project. Python3.4 version was used as it is a mature, versatile and robust programming language. It is an interpreted language which makes the testing and debugging extremely quickly as there is no compilation step. There are extensive open source libraries available for this version of python and a large community of users.

Python is simple yet powerful, interpreted and dynamic programming language, which is well known for its functionality of processing natural language data, i.e. spoken English using NLTK. Other high level programming languages such as ‘R’ and ‘Matlab’ were considered because they have many benefits such as ease of use but they do not offer the same flexibility and freedom that Python can deliver.

### Introduction to NLTK

Natural Language Toolkit (NLTK) is library in Python, which provides a base for building programs and classification of data. NLTK is a collection of resources for Python that can be used for text processing, classification, tagging and tokenization. This toolbox plays a key role in transforming the text data in the tweets into a format that can be used to extract sentiment from them.

NLTK provides various functions which are used in pre-processing of data so that data available from twitter become fit for mining and extracting features. NLTK support various machine learning algorithms which are used for training classifier and to calculate the accuracy of different classifier.

In our project we use Python as our base programming language which is used for writing code snippets. NLTK is a library of Python which plays a very important role in converting natural language text to a sentiment either positive or negative. NLTK also provides different sets of data which are used for training classifiers. These datasets are structured and stored in library of NLTK, which can be accessed easily with the help of Python.

### Introduction to Supervised Machine learning Classifiers

Supervised machine learning is a technique whose task is to deduce a function from tagged training samples. The training samples for supervised learning consist of large set of examples for a particular topic. In supervised learning, every example training data comes in a pair of input (vector quantity) and output value (desired result). These algorithms analyze data and generate an output function, which is used to mapped new data sets to respective classes. Different machine learning classifiers which we are going to use to build our classifier are:

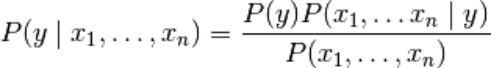
∑ Naïve-Bayes Classifier

∑ Support Vector Machine Classifier

#### Naïve-Bayes (NB) Classifier [1]

Naïve-Bayes classifiers are probabilistic classifiers which come under machine learning techniques. These classifiers are based on applying Bayes’ theorem with strong (naïve) assumption of independence between each pair of features. Let us

assume, there is a dependent vector from x1 to xn, and a class variable ‘y’. Therefore, according to Bayes’ :



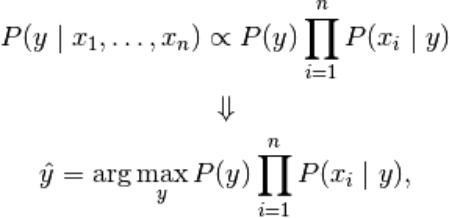
Now according to assumption of independence



For every ‘i’, this function becomes



In this  on given input is constant, hence we can apply classification ruleas:



And for estimating we can use MAP (Maximum A Posterior) estimation and; the of class ‘y’ in training sample is relative frequency.

### Goal of Project

With the emergence of social networking, many websites have evolved in the past decade like Twitter, Facebook, Tumbler, etc. Twitter is one the website which is widely used all over the world. According to Twitter it has been recorded that around 200 billion tweets posts every year. Twitter allows people to express their thoughts, feelings, emotions, opinions, reviews, etc. about any topic in natural language within 140 characters. Python is the standard high-level programming language which is best for NLP. Thus, for processing natural language data, Python uses one of its libraries called Natural Language Toolkit. NLTK provides large amount of corpora which helps in training classifiers and it helps in performing all NLP methodology like tokenizing, part-of-speech tagging, stemming, lemmatizing, parsing and performing sentiment analysis for given datasets.

It is a challenging task to deal with a large dataset, but with the use of NLTK we can easily classify our data and give more accurate results based on different classifiers. The goal of this project is to perform sentiment analysis on tweets related to demonetization. Public opinions on ‘Demonetization’ are mined from Twitter and then classified into sentiments, whether positive or negative or neutral by using supervised machine learning classifiers. These results will let us know about the reviews and opinions of people on demonetization.

To achieve this goal, a module is created which can perform live sentimental analysis. In live sentimental analysis user can obtain the trend of any live trending topic depicted by two sentiment category (positive and negative) in live graphs. Further accuracy and reliability of the module can be checked with the help of various machine learning classifiers.

### Need of Sentimental Analysis

#### Industry Evolution

Only the useful amount of data is required in the industry as compared to the set of complete unstructured form of the data. However the sentiment analysis done is useful for extracting the important feature from the data that will be needed solely for the purpose of industry. Sentimental Analysis will provide a great opportunity to the industries for providing value to their gain value and audience for themselves. Any of the industries with the business to consumer will get benefit from this whether it is restaurants, entertainment, hospitality, mobile customer, retail or being travel.

#### Research Demand

Another important reason that stands behind the growth of SA deals with the demand of research in evaluation, appraisals, opinion and their classification. Present solutions for the purpose of sentiment analysis and opinion mining are rapidly evolving, specifically by decreasing the amount of human effort that will be required to classify the comments. Also the research theme that will be based in the long established disciplines of computer science like as text mining, machine learning, natural language processing and artificial intelligence, voting advise applications, automated content analysis, etc.

#### Decision Making

Every person who stores information on the blogs, various web applications and the web social media, social websites for getting the relevant information you need a particular method that can be used to analyze data and consequently return some of the useful results. It is going to be very difficult for company to conduct the survey that will be on the regular basis so that there comes the need to analyze the data and locate the best of the products that will be based on user’s opinions, reviews and advices. The reviews and the opinions also help the people to take important decisions helping them in research and business areas.

#### Understanding Contextual

As human language is getting very complex day by day so it has become difficult for the machine to be able to understand human language that can be expressed in the slangs, misspelling, nuances, and the cultural variation. Thus, there will be a need of system that will make better understanding between the human and the machine language.

#### Internet Marketing

Another important reason behind the increase in the demand of sentimental analysis is the marketing done via internet by the business and companies organization. Now they regularly monitor the opinion of the user about their brand, product, or event on blog or the social post. Thus, we see that the sentimental Analysis could also work as a tool for marketing too.

### Applications of Sentiment Analysis

Sentiment analysis has large amount of applications in the NLP domain. Due to the increase in the sentiment analysis, social network data is on high demand. Many companies have already adopted the sentimental analysis for the process of betterment. Some of major applications are mentioned as following:

#### Word of Mouth (WOM)

Word of Mouth (WOM) is the process by which the information is given from one

person to another person. It would essentially help the people to take the decisions. Word of Mouth has given the information about the opinions, attitudes, reactions of consumers about the related business, services and the products or even the ones that can be shared with more than one person. Therefore, this is going to be where Sentiment Analysis comes into picture. As the online review blogs, sites, social networking sites have provided the large amount of opinions, it has helped in the process of decision-making so much easier for the user.

#### Voice of Voters

Each of the political parties usually spent a major chunk of the amount of money for the aim of campaigning for their party or for influencing the voters. Thus if the politicians know the people opinions, reviews, suggestions, these can be done with more effect. This is how process of Sentimental analysis does not only help political parties but on the other hand help the news analysts alongside. Also the British and the American administration had already used some of the similar techniques.

#### Online Commerce

There is vast number of websites related to ecommerce. Majority of them had the policy of getting the feedback from its users and customers. After getting information from various areas like service and quality details of the users of company users experience about features, product and any suggestions. These details and reviews have been collected by company and conversion of data into the geographical form with the updates of the recent online commerce websites who use these current techniques.

#### Voice of the Market (VOM)

Whenever a product is to be launched by a specific company, the customers would to know about the product ratings, reviews and detailed descriptions about it. Sentiment Analysis can help in analyzing marketing, advertising and for making new strategies for promoting the product. It provides the customer an opportunity to choose the best among the all.

#### Brand Reputation Management (BRM)

Sentiment analysis would help to determine how would be a company’s brand,

service and the service or product that would be perceived by the online community. Brand Reputation Management will be concerned about the management of the reputation of market. It has focuses on the company and product rather than customer. Thus the opportunities were created for the purpose of managing and strengthening the brand reputation of the organizations.

#### Government

Sentiment Analysis has helped the administration for the purpose of providing various services to the public. Fair results have to be generated for analyzing the negative and positive points of government. Thus sentiment analysis is helpful in many fields like decision making policies, recruitments, taxation and evaluating social strategies. Some of the similar techniques that provide the citizen oriented government model where the services and the priorities should be provided as per the citizens.

# Problem Statement

Sentiment Analysis is a process of extracting feature from user’s thoughts, views, feelings and opinions which they post on any social network websites. The result of sentiment analysis is classification of natural language text into classes such as positive, negative and neutral. The amount of data generated from social network sites is huge; this data is unstructured and cannot give any meaningful information until it is analyzed. Thus, to make this huge amount of data useful we perform sentiment analysis, i.e. extracting feature from this data and classify them. Sentiment analysis is very necessary in today’s world, as people always get affected by the thinking and opinions other people. Today, if any one wants to purchase a product or to give vote or to watch a movie, etc. then that person will first wants to know what are other people reviews, reactions and opinions about that product or candidate or movie on social media websites like Twitter, Facebook, Tumbler, etc. So there is a need of system that can automatically generate sentiment analysis from this huge amount of data.

### Objectives

The main objective of this project work is to perform the sentiment analysis on Demonetization tweets such that people opinions about this strategy can be obtained which are extracted from Twitter.

Thus to achieve this objective we build a classifier based on supervised learning and perform live sentiment analysis on data collected.

### Methodology

To achieve this objective discussed above in section 3.1, the following methodology is used:

∑ A thorough study of existing approaches and techniques in field of sentiment analysis.

∑ Collection of related data from Twitter with the help of Twitter API.

∑ Pre-processing of data collected from Twitter so that it can be fit for mining.

∑ To build a classifier based on different supervised machine learning techniques.

∑ Training and testing of build classifier using large datasets

∑ Computing the result of different classifier using dataset collected from Twitter.

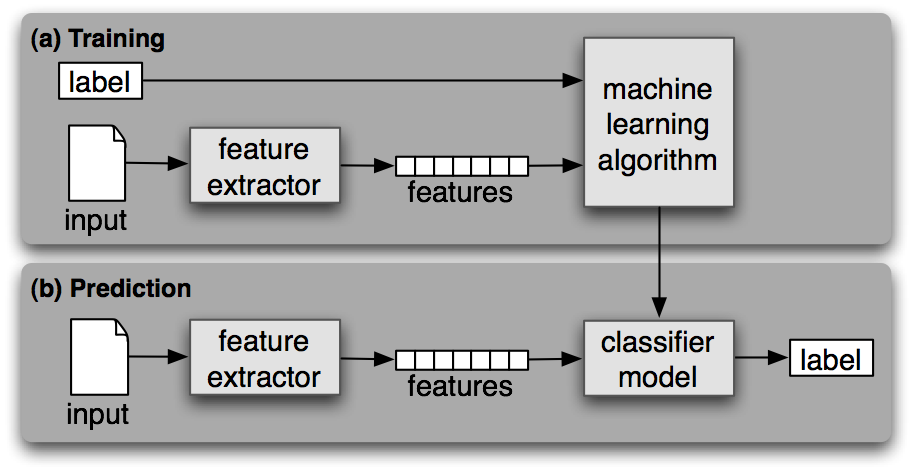
∑ Plotting a pie chart that show the trend of positive and negative sentiment.

# Implementation

Data collection is not a simple task, as it may seem. Various decisions have to be made for collecting data. For our project we maintain dataset for training, testing and for twitter sentiment analysis. In this chapter we are going to study how data is collected, stored, processed and classified. Before discussing these process and different dataset, let us discuss our proposed architecture.

### Proposed Architecture

As our goal is to achieve sentiment analysis for data provided from Twitter. We are going to build a classifier which consists of different machine learning classifiers. Once our classifier is ready and trained we are going to follow the steps shown in Figure 4.1



**Figure 3.1** Process to classify tweets using build classifier

#### 3.1.1 Data Collection

#### The data can be obtained using different-different APIs made for Twitter application but instead of doing this extra step, we managed to get already made dataset of demonetization tweets on [*www.kaggle.com/dataset*](http://www.kaggle.com/dataset)and then, this data is saved to our project directory for further processing.

#### Training Data

Other data which we collected for this project is training data. This data is used to train the classifier which we are going to build. To collect this data we use NLTK library of Python. NLTK consists of corpora, which is very large and consists of structured set of text files which are used to perform analysis. In these corpora there are various types of text files like quotes, reviews, chat, history, etc. From these corpora we will select files of movie reviews for our training purpose. Sample of these reviews is shown in Table4.1

**Table 4.1** Sample movie reviews in NLTK Corpus

|  |  |
| --- | --- |
| **Movie Reviews** | **CLASS** |
| foolish, idiotic and boring it's so lad dish and youngish , only  teenagers could find it funny | NEGATIVE |
| the rock is destined to be the 21st century's new conan and that he's  going to make a splash even greater than arnoldschwarzenegger | POSITIVE |
| Barry Sonnenfeld owes frank the pug big time the biggest problem  with roger avary's uproar against the map | NEGATIVE |
| the seaside splendor and shallow , beautiful people are nice to look  at while you wait for the story to get going | POSITIVE |

In movie reviews corpus there are around 5000 reviews each for positive and negative feedback. These reviews are short and arranged in text files which are easy to access. We train our classifier from around 80% of the data and then we test it with remaining 20% to check that trained classifier is working properly or not.

### 3.2 Data Pre-Processing

Data obtained from twitter is not fit for extracting features. Mostly tweets consists of message along with usernames, empty spaces, special characters, stop words, emoticons, abbreviations, hash tags, time stamps, URL’s ,etc. Thus to make this data fit for mining we pre-process this data by using various function of NLTK. In pre- processing we first extract our main message from the tweet, then we remove all empty spaces, stop words (like is, a, the, he, them, etc.), hash tags, repeating words, URL’s, etc. We then replace all emoticons and abbreviations with their corresponding meanings like :-), =D, =), LOL, Rolf, etc. are replaced with happy or laugh. Once we are done with it, we are ready with processed tweet which is provided to classifier for required results. A sample processed tweet is shown in Table 4.2

**Table 3.2** Sample Tweet and Processed Tweet

|  |  |
| --- | --- |
| **Tweet Type** | **Result** |
| Original tweet | RT @rssurjewala: Critical question: Was PayTM informed about #Demonetization edict by PM? It's clearly fishy and requires full disclosure &amp;… |
| Processed tweet | critical question paytm informed demonetization edict pm clearly fishy requires full disclosure |

Cleaning of Twitter data is necessary, since tweets contain several syntactic features that may not be useful for analysis. The pre-processing is done in such a way that data represented only in terms of words that can easily classify the class.

We create a code in Python in which we define a function which will be used to obtain processed tweet. This code is used to achieve the following functions:

∑ remove quotes - provides the user to remove quotes from the text

∑ remove @ - provides choice of removing the @ symbol, removing the @ along with the user name, or replace the @ and the user name with a word 'AT\_USER' and add it to stopwords

∑ remove URL (Uniform resource locator) - provides choices of removing URLs or replacing them with 'URL' word and add it to stop words

∑ remove RT (Re-Tweet) - removes the word RT from tweets

∑ remove Emoticons - remove emoticons from tweets and replace them with their specific meaning

∑ remove duplicates – remove all repeating words from text so that there will be no duplicates

∑ remove # - removes the hash tag class

∑ remove stop words – remove all stop words like a, he, the, and, etc which provides no meaning for classification

Table 4.3 shows the various types of contents that are included in tweets and also the actions performed on these contents. Some of the example of clean tweets is shown in Table 4.4

**Table 3.3** Removed and modified content

|  |  |
| --- | --- |
| **CONTENT** | **ACTION** |
| Punctuation (! ? , . ” : ; ) | Removed |
| #word | Removed #word |
| @any\_user | Remove @any\_user or replaced with  “AT\_USER” and then added in stop  words. |
| Uppercase characters | Lowercase all content |
| URLs and web links | Remove URLs or replaced with “URL”  and then added in stop words |
| Number | Removed |
| Word not starting with alphabets | Removed |
| All Word | Stemmed all word  (Converted into simple form) |
| Stop words | Removed |
| Emoticons | Replaced with respective meaning |
| White spaces | Removed |

**Table 3.4** Sample cleaned data

|  |  |
| --- | --- |
| **Raw data** | **Clean data** |
| RT @mrpatel1954: #Demonetization has long term benefits. Pain now is temporary. Let us help #nation to make #India #corruption free. | demonetization long term benefit pain temporary let help nation make india corruption free |
| RT @Atheist\_Krishna: BEFORE and AFTER Gandhi ji heard they are standing there against #Demonetization  . <https://t.co/9NheK63TPg> | gandhi heard standing against demonetization |

Once our data is cleaned and ready for processing our next step is to classify this cleaned data into different classes. For this we have to use supervise machine learning classifiers.

### 3.3 Classification

To classify tweets in different class (positive and negative) we build a classifier which consists of several machine learning classifiers. To build our classifier we used a library of Python called, Scikit-learn. Scikit-learn is a very powerful and most useful library in Python which provides many classification algorithms. Scikit-learn also include tools for classification, clustering, regression and visualization. To install Scikit-learn we simply use on line command in python which is ‘pip install scikit- learn’.

In order to build our classifier, we use one in-build classifiers which come in Scikit- learn library, which are:

∑ Naïve-Bayes Classifier

The reason we are using one classifiers, so that we can get the more reliable output. To use these classifiers, we write a script in Python, in which we first import the classifier and then we pass the training set to each classifier.

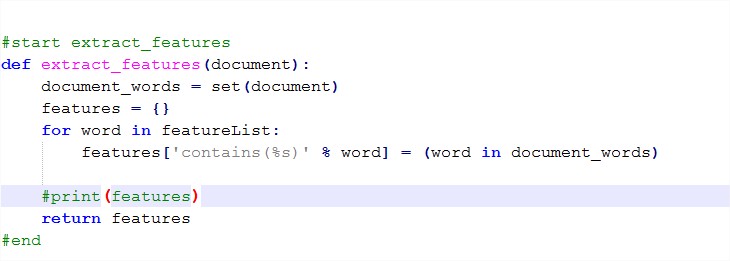
#### 3.4.1 Feature Extraction

As we already discussed, training and testing data is collected from NLTK corpus. We have round 5000 tweets on demonetization each for positive, negative and neutral class. We take first 4000 tweets as training set and remaining 1000 as testing sets.

Both the training and testing data must be represented in same order for learning. One of the ways that data can be represented is feature-based. By features, it is meant that some attributes that are thought to capture the pattern of the data are first selected and the entire dataset must be represented in terms of them before it is fed to a machine learning algorithm. Different features such as n-gram presence or n-gram frequency, POS (Part of Speech) tags, syntactic features, or semantic features can be used. For example, one can use the keyword lexicons as features. Then the dataset can be represented by these features using either their presence or frequency.

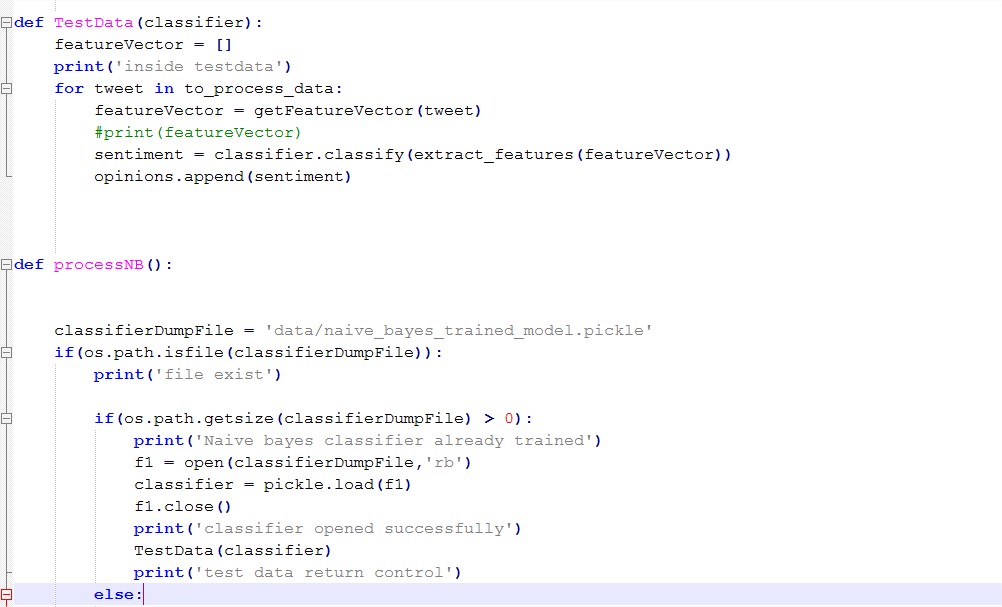
Attribute selection is the process of extracting features by which the data will be represented before any machine learning training takes place. Attribute selection is the first task when one intends to represent instances for machine learning. Once the attributes are selected, the data will be represented using the attributes. So attributes are the features. Although we used the entire data set in our selection of attributes, the representation of the data must be done on a per instance (Twitter post) basis.

Feature vector plays a very important role in classification and helps to determine the working of the build classifier. Feature vector also help in predicting the unknown data sample. There are many types of feature vectors, but in this process we used unigram approach. Each tweet words are added to generate the feature vectors. The presence/absence of sentimental word helps to indicate the polarity of the sentences. We create a python script to extract the features from the training data. Code snippet for extracting features is shown in Figure4.4



**Figure 3.4** Code for extracting features from tweets

Once we extract the features from training data, we are going to pass these in our build classifiers. A script in written in python which is used to pass training sets in classifier. Once, the classifier is trained we can also check the accuracy of each classifier by passing the testing set. Sample script of training and testing of classifier is shown in Figure4.5



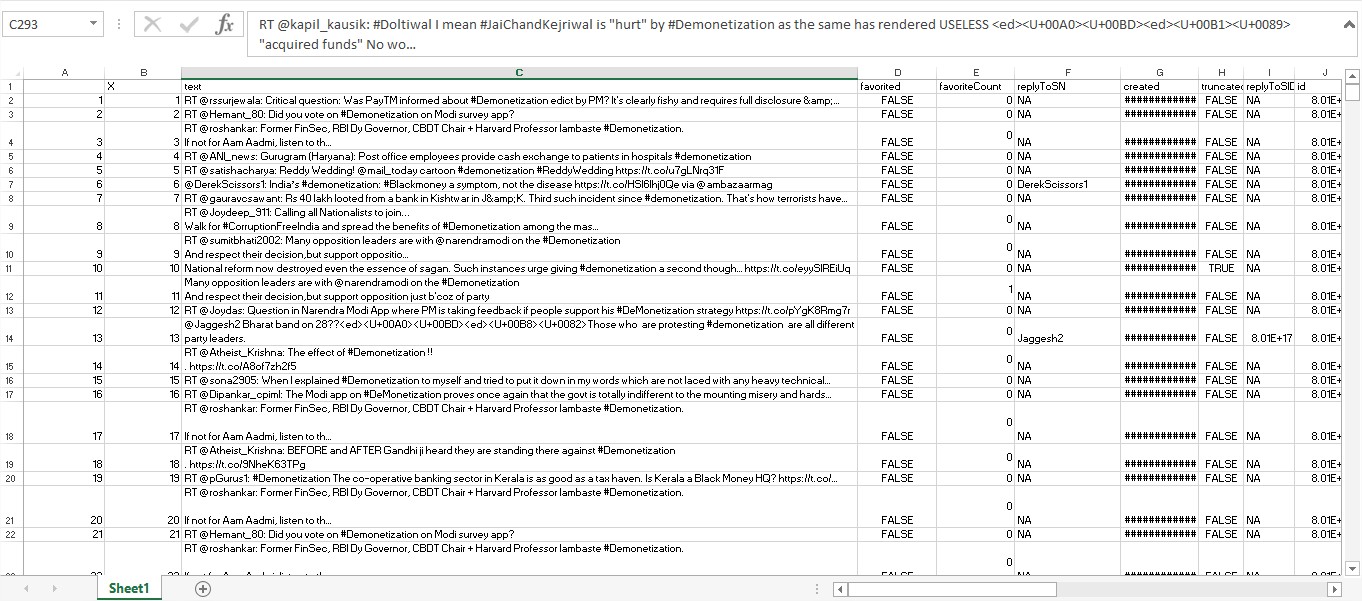
**Figure 3.5** Sample code for training and testing build classifier

# Chapter 4 Results and Analysis

In this chapter we are going to show various results that we have achieved in our implementation.

### Tweets Collected

Tweets are collected from a dataset available publicly on *www.kaggle.com*. A sample file for demonetization tweets is shown in Figure 5.1

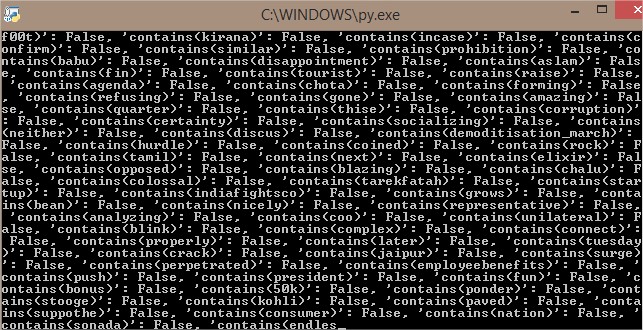


**Figure 4.1** Sample Tweets Collected for demonetization

Thus each such excel file is stored in a directory as shown in Figure 4.3

### Extracted Features

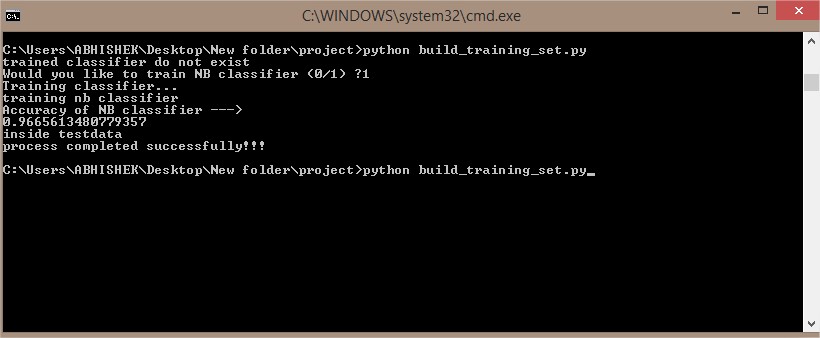
When we ran the script shown in Figure 4.4, it will extract the features from the training data and also apply a Boolean value to each attribute. The output for extracted features is shown in Figure 5.2



**Figure 4.2** Extracted features from training data

### Classifier Accuracy for Training Data

Once we ran the script shown in Figure 4.5, we get the accuracy of each classifier for movie reviews training data. The output is shown in Figure 5.3



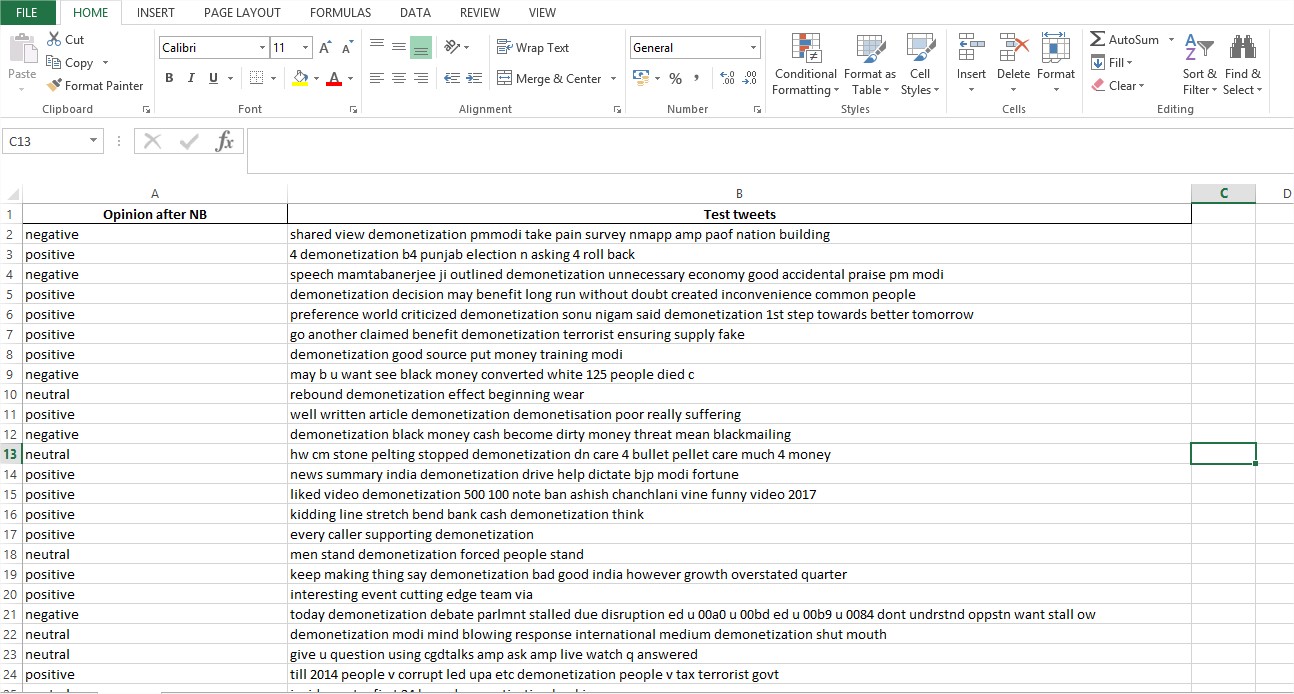
**Figure 4.3** Classifiers accuracy for training data

As, we can show from Figure 4.3, naïve bayes classifiers giving accuracy of average 96% and above. Thus our build classifier is fully trained and ready for sentiment analysis of twitter data.

### Twitter Data Analysis

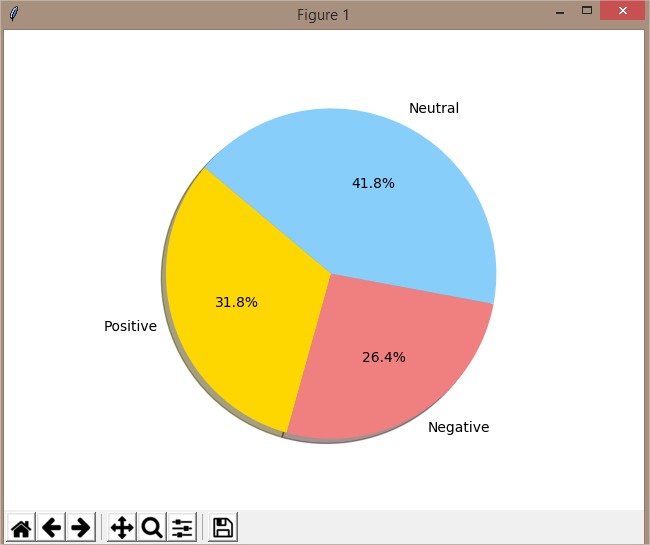
For our project, we collected tweets on demonetization from *www.kaggle.com*. We store all these tweets and pre-processed them so that they can be fit for mining. Once our data sets are ready we are going to passour dataset through our classifier and check the accuracy.

#### Analysis forTest Data

****

**Figure 4.4** Sentiment analysis for each tweet, when Test Twitter data is analyzed

#### Analysis forTest Data

****

**Figure 4.5** Sentiment analysis overall classification results when Test Twitter data is analysis

# Chapter 5 Conclusion and Future Scope

### Conclusion

Sentiment analysis is used to identifying people’s opinion, attitude and emotional states. The views of the people can be positive or negative. Commonly, parts of speech are used as feature to extract the sentiment of the text. An adjective plays a crucial role in identifying sentiment from parts of speech. Sometimes words having adjective and adverb are used together then it is difficult to identify sentiment and opinion.

To do the sentiment analysis of tweets, the proposed system first extracts the twitter posts from twitter by user. The system can also computes the frequency of each term in tweet. Using machine learning supervised approach help to obtain the results.

Twitter is large source of data, which make it more attractive for performing sentiment analysis. We perform analysis on around 5,000 tweets total for each party, so that we analyze the results, understand the patterns and give a review on people opinion. We also saw how any social event, speech or rally cause a fluctuation in sentiment of people. We also get to know which strategies are getting more support from people. It is not necessary that our classifier can only be used for twitter sentiment analysis. It is general classifier. It can be used for any purpose based on tweets we collect with the help of keyword. It can be used for finance, marketing, reviewing and many more.

### Future Scope

Some of future scopes that can be included in our research work are:

∑ Use of parser can be embedded into system to improve results.

∑ A web-based application can be made for our work infuture.

∑ We can improve our system that can deal with sentences of multiple meanings.

∑ We can also increase the classification categories so that we can get better results.

∑ We can start work on multi languages like Hindi, Spanish, and Arabic to provide sentiment analysis to morelocal.

# References



**Books:**

1. J. Han and M. Kamber, “Data Mining: Concepts and Techniques”, Harcourt India Morgan Kauffman, 2001.
2. Alex Berson and Stephen J. Smith, “Data Warehousing, Data mining and OLAP”, Tata McGraw-Hill, 2004.
3. Margaret H. Dunham, “Data Mining: Introductory and Advanced Topics”, Pearson Education, 2004.
4. Sam Anahory and Dennis Murry, “Data Warehousing in the Real World”, Pearson Education, 2003.

**Articles & Research Paper:**

1. H. Zang, “The Optimality Of Naïve Bayes”, Proc. FLAIRS, 2004.
2. P Pangand & L Lee, “Opinion mining and sentiment analysis. Foundation and trends in information retrieval” 2008.

**Blogs:**

1. [www.infotrie.com/sentiment-analysis-blog/](http://www.infotrie.com/sentiment-analysis-blog/)
2. sentiment-analysis.blogspot.com/
3. <https://blog.algorithmia.com/introduction-sentiment-analysis-algorithms/>
4. <https://www.ravikiranj.net/posts/2012/code/how-build-twitter-sentiment-analyzer/>

**Websites:**

1. <https://stackoverflow.com/>
2. <https://github.com/>
3. <https://docs.python.org/>