

Twitter Sentiment Analysis

Project

*Submitted in partial fulfillment of the
Requirements for the award of the degree of*

BACHELOR OF TECHNOLOGY (B. TECH)

Submitted by:

Vinay Vashistha

180060101119



Under the supervision of

Mrs. Neha Arora

(Assistant professor)

Mrs. Divya Mishra

(HOD IT)

Dr. Taresh Singh

(HOD CSE)

**DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING
COLLEGE OF ENGINEERING ROORKEE, ROORKEE
ROORKEE-247667 (UTTARAKHAND) INDIA
MAY, 2022**

ACKNOWLEDGEMENT

I would like to place on record my deep sense of gratitude to ***Mrs. Divya Mishra*** Department of Computer Science and Engineering for her generous guidance, help and useful suggestions. I express my sincere gratitude to **Dr. Taresh Singh, HoD** in Department of Computer Science and Engineering for his stimulating guidance, continuous encouragement and supervision throughout the course of present work.

I also wish to extend my thanks to Mrs. Neha Arora for their insightful comments and constructive suggestions to improve the quality of this research work.

I am extremely thankful to CSE lab incharge for providing me infrastructural facilities to work in, without which this work would not have been possible.

Vinay Vashistha

180060101119

.....

DECLARATION

I hereby certify that the work which is being presented in the project report entitled “**Twitter Sentiment Analysis**” by “VINAY VASHISTHA” in partial fulfillment of requirements for the award of degree of B.Tech. submitted in the Department of Computer Science and Engineering at “College of Engineering, Roorkee” under UTTARAKHAND TECHNICAL UNIVERSITY is an authentic record of my own work carried out during a period from 15th February 2022 to 15th May 2022 under the supervision of “MRS. DIVYA MISHRA”.

Vinay Vashistha

180060101119

.....

CERTIFICATE OF APPROVAL

This is to certify the report entitled the “**TWITTER SENTIMENT ANALYSIS**” is record of bonafide work, carried out by **Vinay Vashistha** under my supervision and guidance.

In my opinion, the report in its present form is in partial fulfilment of all the requirement for the award of Bachelor of Technology (B. Tech), in Computer Science and Engineering at College of Engineering Roorkee, is an authentic work carried by them under my supervision and guidance.

To the best of my knowledge, the matter embodied in the project has not been submitted for any other degree or diploma.

Mrs. Neha Arora

Mrs. Divya Mishra

Dr. Taresh Singh
(HOD CSE)

The B. Tech Viva –Voce Examination of Vinay Vashistha has been held on and is accepted.

External Examiner

ABSTRACT

Technology today has become a momentous driving vehicle for communication world-wide. Social media platforms like twitter, Facebook, Instagram are the most important arenas for expressing views on transformations happening in and around the world every day. Twitter is a rich origin of info for mining of user opinions. This paper reflects the idea of taking user opinions into consideration performing sentiment analysis and establishing conclusions on interested topics using Machine Learning algorithms. Random forest, Logical Regression, decision tree, XGB Classifier and Support Vectors Machines in Machine Learning are tuned-up using supervised learning to obtain outputs for sentiment analysis respectively. Sentiment analysis desires to obtain sentiment polarity (positive or negative) and from user data. Such analysis essentially serves a gateway for consumer needs and generates growth opportunities in businesses.

INDEX

Chapter 1	8
Introduction	8
Introduction to Sentiment Analysis	8
What is Sentiment Analysis?	9
What is Emotional Analysis?	10
Introduction to Python	11
Introduction to Supervised Machine learning Classifiers	11
Random Forest Algorithm	11
Logistic Regression	12
Decision Tree Classification Algorithm	12
Support Vector Machines	14
Chapter 2	15
Goal of project	15
The Project	15
Data	16
Chapter 3	19
Need of Sentimental Analysis	19
Industry Evolution	19
Research Demand	19
Decision Making	19
Understanding Contextual	20
Internet Marketing	20
Chapter 4	21
Applications of Sentiment Analysis	21
Word of Mouth (WOM)	21

Voice of Voters	21
Online Commerce	22
Voice of the Market (VOM)	22
Brand Reputation Management (BRM)	22
Government	23
Chapter 5	24
Problem Statement	24
Objectives	25
Methodology	25
Chapter 6	26
Implementation	26
Proposed Architecture	26
Data Collection	27
Twitter Data	27
Training Data	27
Chapter 7	41
Conclusion and Future Scope	41
Conclusion	41
Future Scope	41

Chapter 1

Introduction

In this chapter we are going to give the introductions on Sentiment Analysis and Python. Then we are explaining the objective of our thesis. 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. Sentiment Analysis is domain centered, i.e. results of one domain cannot be applied to other domain. 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.

Social media sentiment analysis has turn out to be a distinguished area of study and experimentation in current years..Twitter a micro-blogging site, has lion's share in social media info. Most research has been confined to classify tweets into positive,negative categories ignoring sarcasm.Human emotions are extremely diverse and cannot be restricted to certain metrics alone.Polarity analysis gives limited information on the actual intent of message delivered by

author and just positive or negative classes are not sufficient to understand nuances of underlying tone of a sentence. This brings the need to take one step above sentiment analysis leading to emotion analysis. In this paper we throw light on methods we have used to derive sentiment analysis considering sarcasm and how we have accomplished emotion analysis of user opinions.

A supervised learning technique provides labels to classifier to make it understand the insights among various features. Once the classifier gets familiarized with train data it can perform classification on unseen test data. We have chosen random forest classifier, Logistic Regression Classifier, decision Tree Classifier and Support Vector Machine classification algorithms to carry out sentiment analysis respectively.

Performing SA(sentiment analysis) will help organizations or companies to improve services, track products and obtain customer feedback in a normalized form. Gaining insights from large volumes of data is a mountain of a task for humans hence using an automated process will easily drill down into different customer feedback segments mentioned on social media or elsewhere. Effective business strategies can be built from results of sentiment analysis. Identifying clear emotions will establish a transparent meaning of text which potentially develops customer relationships, motivation and extends consumer expectations towards a brand or service.

Emotion detection involves a wide platter of emotions classified into states like joy, fear, anger, surprise and many more. We here examine sentiments and emotions of short texts coined as tweets from the famous social media, twitter. Generally people discuss a lot of things daily but it is difficult to get insights just by reading through each of their opinions so there should be a way that helps us to get insights of users opinions in an unbiased manner, So this model helps in drawing out Sentiment of users, classify them and finally present them to us. 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.

What is Sentiment Analysis?

Sentiment analysis (also known as opinion mining) refers to the use of natural language processing, text analysis, computational linguistics, and biometrics to systematically identify, extract, quantify, and study affective states and subjective information.

Sentiment analysis is contextual mining of text which identifies and extracts subjective information in source material, and helping a business to understand the social sentiment of their brand, product or service while monitoring online conversations. However, analysis of social media streams is usually restricted to just basic sentiment analysis and count based metrics. This is akin to just

scratching the surface and missing out on those high value insights that are waiting to be discovered.

What is Emotional Analysis?

It is the process of identifying human emotions, most typically from facial expressions as well as from verbal expressions. It relies on a deeper analysis of human emotions and sensitivities.

Emotions analytics (EA) software collects data on how a person communicates verbally and nonverbally to understand the person's mood or attitude. The technology, also referred to as emotional analytics, provides insights into how a customer perceives a product, the presentation of a product or their interactions with a customer service representative.

Just as with other data related to customer experience, emotions data is used to create strategies that will improve the business's customer relationship management (CRM). EA software programs can be used with companies' data collection, data classification, data analytics and data visualization initiatives.

"What other people think" has always been an important piece of information for most of us during the decision-making process. The Internet and the Web have now (among other things) made it possible to find out about the opinions and experiences of those in the vast pool of people that are neither our personal acquaintances nor well-known professional critics — that is, people we have never heard of. And conversely, more and more people are making their opinions available to strangers via the Internet. The interest that individual users show in online opinions about products and services, and the potential influence such opinions wield, is something that is driving force for this area of interest. And there are many challenges involved in this process which needs to be walked all over in order to attain proper outcomes out of them.

Introduction to Python

Python is a high level, dynamic programming language which is used for this thesis. 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 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:

- Random Forest Classifier
- Logistic Regression Classifier
- Decision Tree Classifier
- SVM (Support Vector Classifier)

Random Forest Algorithm

Random Forest is a popular machine learning algorithm that belongs to the supervised learning technique. It can be used for both Classification and Regression problems in ML. It is based on the concept of **ensemble learning**, which is a process of *combining multiple classifiers to solve a complex problem and to improve the performance of the model*.

As the name suggests, ***"Random Forest is a classifier that contains a number of decision trees on various subsets of the given dataset and takes the average to improve the predictive accuracy of that dataset."*** Instead of relying on one decision tree, the random forest takes the prediction from each tree and based on the majority votes of predictions, and it predicts the final output.

The greater number of trees in the forest leads to higher accuracy and prevents the problem of overfitting.

Logistic Regression

Logistic regression is a supervised learning classification algorithm used to predict the probability of a target variable. The nature of target or dependent variable is dichotomous, which means there would be only two possible classes.

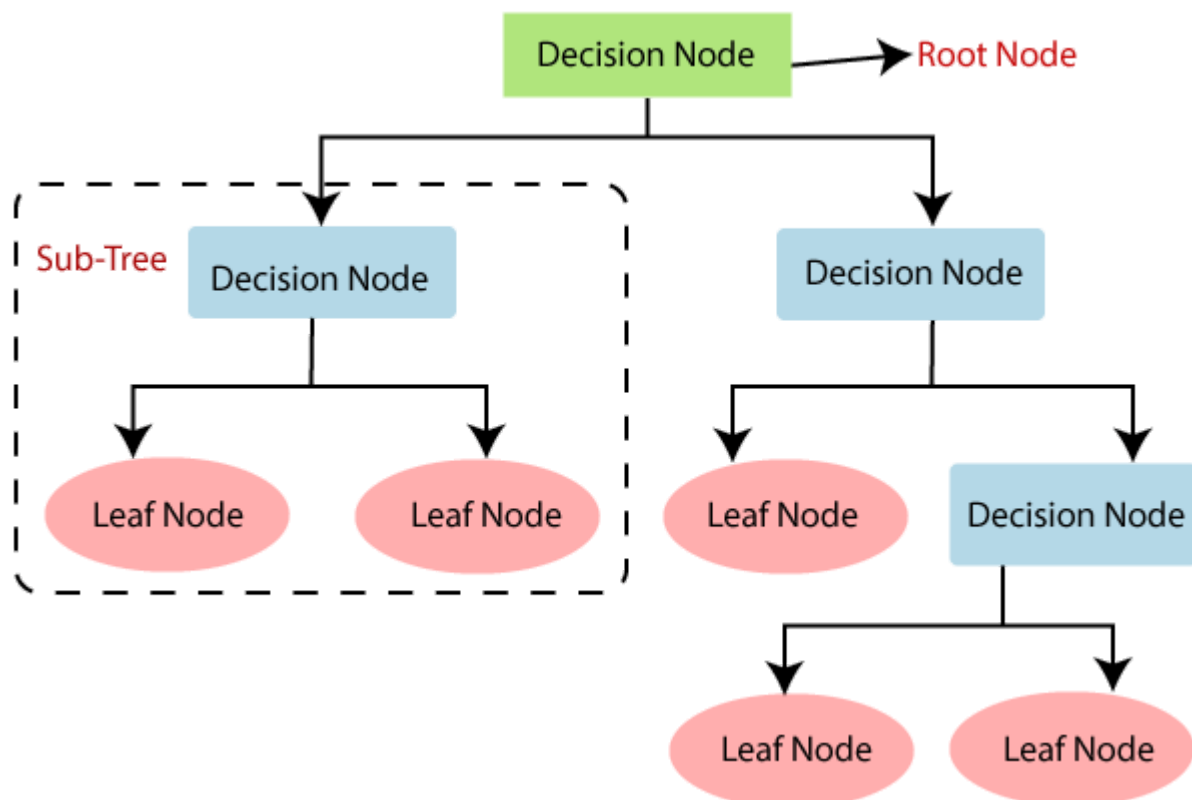
In simple words, the dependent variable is binary in nature having data coded as either 1 (stands for success/yes) or 0 (stands for failure/no).

Mathematically, a logistic regression model predicts $P(Y=1)$ as a function of X . It is one of the simplest ML algorithms that can be used for various classification problems such as spam detection, Diabetes prediction, cancer detection etc.

Decision Tree Classification Algorithm

- o Decision Tree is a **Supervised learning technique** that can be used for both classification and Regression problems, but mostly it is preferred for solving Classification problems. It is a tree-structured classifier, where **internal nodes represent the features of a dataset, branches represent the decision rules and each leaf node represents the outcome.**
- o In a Decision tree, there are two nodes, which are the **Decision Node** and **Leaf Node**. Decision nodes are used to make any decision and have multiple branches, whereas Leaf nodes are the output of those decisions and do not contain any further branches.
- o The decisions or the test are performed on the basis of features of the given dataset.
- o ***It is a graphical representation for getting all the possible solutions to a problem/decision based on given conditions.***
- o It is called a decision tree because, similar to a tree, it starts with the root node, which expands on further branches and constructs a tree-like structure.
- o In order to build a tree, we use the **CART algorithm**, which stands for **Classification and Regression Tree algorithm**. A decision tree simply asks a question, and based on the answer (Yes/No), it further split the tree into subtrees.

Below diagram explains the general structure of a decision tree:



Support Vector Machines

Support vector machines (SVMs) are a set of supervised learning methods used for [classification](#), [regression](#) and [outliers detection](#).

The advantages of support vector machines are:

- Effective in high dimensional spaces.
- Still effective in cases where number of dimensions is greater than the number of samples.
- Uses a subset of training points in the decision function (called support vectors), so it is also memory efficient.
- Versatile: different [Kernel functions](#) can be specified for the decision function. Common kernels are provided, but it is also possible to specify custom kernels.

The disadvantages of support vector machines include:

- If the number of features is much greater than the number of samples, avoid over-fitting in choosing [Kernel functions](#) and regularization term is crucial.
- SVMs do not directly provide probability estimates, these are calculated using an expensive five-fold cross-validation.

Support Vectors Classifier tries to find the best hyperplane to separate the different classes by maximizing the distance between sample points and the hyperplane.

Chapter 2

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.

The Project

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. It is a really useful analysis since we could possibly determine the overall opinion about a selling object, or predict stock markets for a given company like, if most people think positive about it, possibly its stock markets will increase, and so on. Sentiment analysis is actually far from to be solved since the language is very complex (objectivity/subjectivity, negation, vocabulary, grammar...) but it is also why it is very interesting to working on. In this project I choose to try to classify tweets from Twitter into “positive” or “negative” sentiment by building a model based on probabilities. Twitter is a microblogging website where people can share their feelings quickly and

spontaneously by sending a tweet limited by 140 characters. You can directly address a tweet to someone by adding the target sign “@” or participate to a topic by adding a hashtag “#” to your tweet. Because of the usage of Twitter, it is a perfect source of data to determine the current overall opinion about anything.

Data

To gather the data many options are possible. In some previous paper researches, they built a program to collect automatically a corpus of tweets based on two classes, “positive” and

“negative”, by querying Twitter with two type of emoticons:

- Happy emoticons, such as “:)”, “:P”, “:)” etc.
- Sad emoticons, such as “:(”, “:’(”, “=(“.

Others make their own dataset of tweets by collecting and annotating them manually which is very long and fastidious. Additionally to find a way of getting a corpus of tweets, we need to take care of having a balanced dataset, meaning we should have an equal number of positive and negative tweets, but it needs also to be large enough. Indeed, more the data we have, more we can train our classifier and more the accuracy will be. After many researches, I found a dataset of 3010 tweets in English coming from Kaggle. It is composed of four columns that are *tweet_id*, *sentiment*, *author* and *tweet_data*. We are only interested by the *Sentiment* column corresponding to our label class taking a binary value, 0 if the tweet is negative, 1 if the tweet is positive and the *tweet_data* column containing the tweets in a raw format.

[illegible]

Example of twitter posts annotated with their corresponding sentiment, 0 if it is negative, 1 if it is positive.

- The presence of **acronyms** "bf" or more complicated "APL". Does it mean apple?

Apple (the company)? In this context we have "friend" after so we could think that he refers to his smartphone and so Apple, but what about if the word "friend" was not here?

- The presence of **sequences of repeated characters** such as "Juuuuuuuuuuuuuuuuuusssst", "hmmmm". In general, when we repeat several characters in a word, it is to emphasize it, to increase its impact.

- The presence of **emoticons**, "O", "T_T", ":|" and much more, give insights about user's moods.

- **Spelling mistakes** and "urban grammar" like "im gunna" or "mi".

- The presence of **nouns** such as "TV", "New Moon".

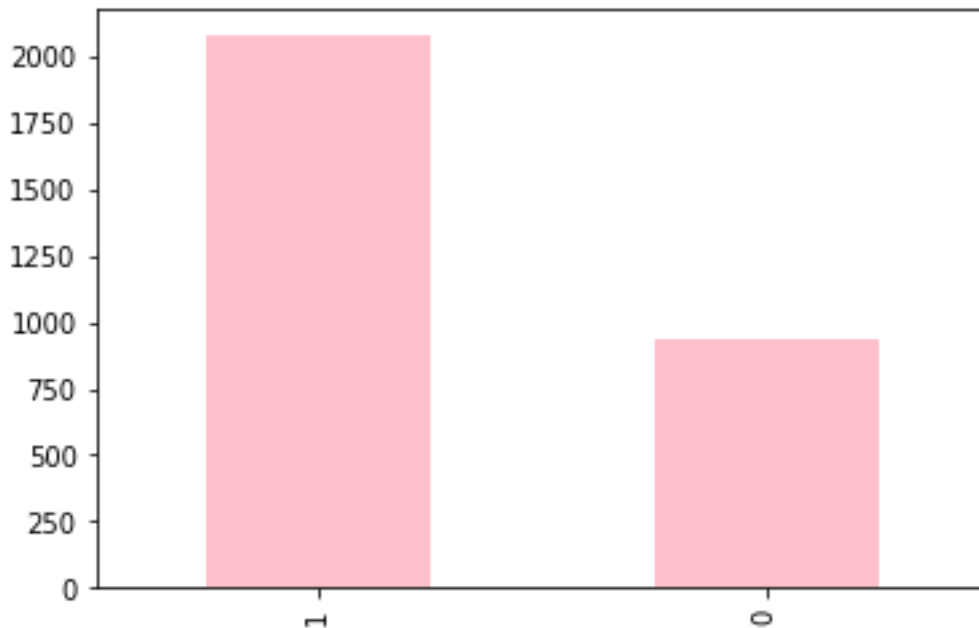
Furthermore, we can also add,

- People also indicate their moods, emotions, states, between two *such as*, *I cries*, *hummin*, *sigh*.

- The negation, "can't", "cannot", "don't", "haven't" that we need to handle like: "I don't like chocolate", "like" in this case is negative.

We could also be interested by the grammar structure of the tweets, or if a tweet is subjective/objective and so on. As you can see, it is **extremely complex** to deal with languages and even more when we want to analyse text typed by users on the Internet because people don't take care of making sentences that are grammatically correct and use a ton of acronyms and words that are more or less English in our case.

We can visualize a bit more the dataset by making a chart of how many positive and negative tweets does it contains,



Histogram of the tweets according to their sentiment

We have exactly positive 2088 tweets and 922 negative tweets which signify that the dataset is well balanced.

There is also no duplicates.

Finally, let's recall the Twitter terminology since we are going to have to deal with in the tweets:

- **Hashtag:** A hashtag is any word or phrase immediately preceded by the # symbol. When you click on a hashtag, you'll see other Tweets containing the same keyword or topic.
- **@username:** A username is how you're identified on Twitter, and is always preceded immediately by the @ symbol. For instance, Katy Perry is @katyperry.
- **MT:** Similar to RT (Retweet), an abbreviation for "Modified Tweet." Placed before the Retweeted text when users manually retweet a message with modifications, for example shortening a Tweet.
- **Retweet:** RT, A Tweet that you forward to your followers is known as a Retweet. Often used to pass along news or other valuable discoveries on Twitter, Retweets always retain original attribution.
- **Emoticons:** Composed using punctuation and letters, they are used to express emotions concisely, ";) :) ...".

Chapter 3

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.

Chapter 4

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 analysing 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. One of the interesting problems which can be taken up is applying this method in the multilingual country like the India where content of the generating mixture of the different languages (e.g., Bengali English) is a very common practice.

Chapter 5

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 Twitter data, such that we can track the negative comments so that it can't hurt public emotions.

Thus to achieve this objective we build a classifier based on supervised learning and perform sentiment analysis on data collected from Kaggle which contain 3010 tweets data.

Methodology

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

is used:

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

- Σ Collection of related data from Kaggle.

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

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

- Σ Training and testing of build classifier using datasets.

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

Chapter 6

Implementation

Data collection is not a simple task, as it may seem. Various decisions have to be made for collecting data. For our thesis 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 classifier Once our classifier is ready and trained we are going to follow the steps shown in Process to classify tweets using build classifier

Step-1 First we are going to read tweets which are collected by Kaggle used in our build classifier with the help of library in python.

Step-2 Then we pre-process these tweets, so that they can be fit for mining and feature extraction.

Step-3 After pre-processing we pass this data in our trained classifier, which then classify them into positive or negative class based on trained results.

Data Collection

Twitter Data

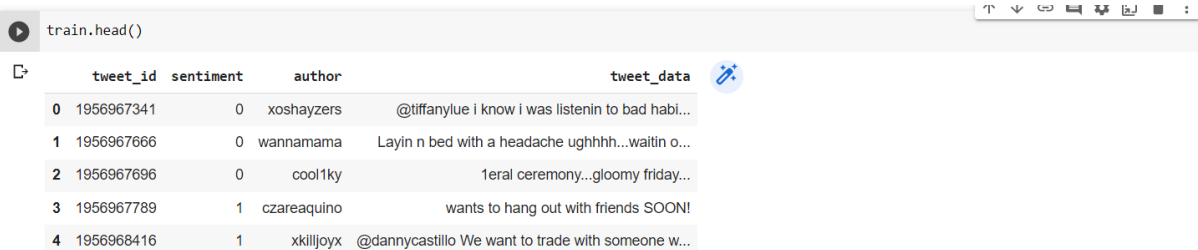
For the project of Data Science data is very important we need to extract data which is best fit for our model. It is good to have large data set so our model is test well. But large data need time to execute so I choose data from Kaggle which contain 3010 tweets data on Kaggle data is come

directly by extracting data from twitter API. Data contain four columns that are tweet_id, sentiment, author and tweet_data.

Training Data

First we need to see the data so that we can easily work on it. So we start with shape of data we got (3010,4) which means 3010 rows and 4 columns.

Then we see the head which is top 5 rows of the data below show image is what we got



```
train.head()
```

	tweet_id	sentiment	author	tweet_data
0	1956967341	0	xoshayzers	@tiffanylue i know i was llistenin to bad habi...
1	1956967666	0	wannamama	Layin n bed with a headache ughhhh...waitin o...
2	1956967696	0	cool1ky	1eral ceremony...gloomy friday...
3	1956967789	1	czareaquino	wants to hang out with friends SOON!
4	1956968416	1	xkilljoyx	@dannycastillo We want to trade with someone w...

Then we check for the null value in the data so we got null value in tweet_data so the null value need to be remove so by the help of dropna method I remove the null value only one row contain null value.

So by the checking of head we see that sentiment column has two kind of values 0 and 1, 0 for negative values and 1 for positive values so now we need to see the positive and negative data separately.

```
# checking out the negative comments from the train set
```

```
train[train['sentiment'] == 0].head(10)
```

	tweet_id	sentiment	author	tweet_data
0	1956967341	0	xoshayzers	@tiffanylue i know i was listenin to bad habi...
1	1956967666	0	wannamama	Layin n bed with a headache ughhhh...waitin o...
2	1956967696	0	cool1ky	1eral ceremony...gloomy friday...
6	1956968487	0	ShansBee	I should be sleep, but im not! thinking about ...
8	1956969035	0	nic0lepaula	@charviray Charlene my 1. I miss you
9	1956969172	0	Ingenue_Em	@kelcouch I'm sorry at least it's Friday?
12	1956970047	0	Danied32	Ugh! I have to beat this stupid song to get to...
13	1956970424	0	Samm_xo	@BrodyJenner if u watch the hills in london u ...
15	1956971077	0	Sim_34	The storm is here and the electricity is gone
17	1956971206	0	brokenangel1982	So sleepy again and it's not even that late. I...

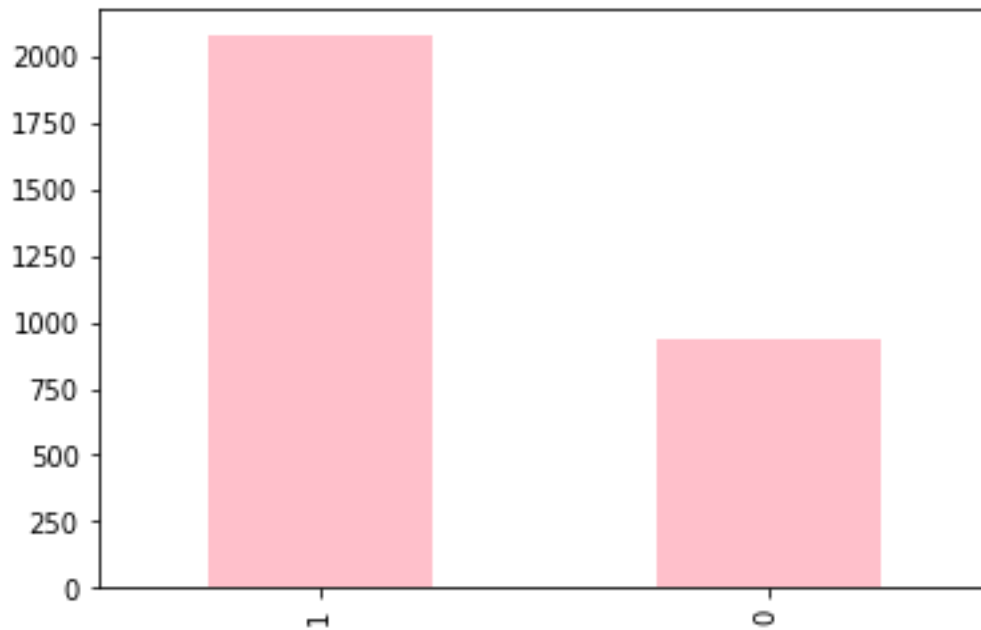
Negative Comments from the data set

Positive Comments from data set

```
[9] # checking out the postive comments from the train set
```

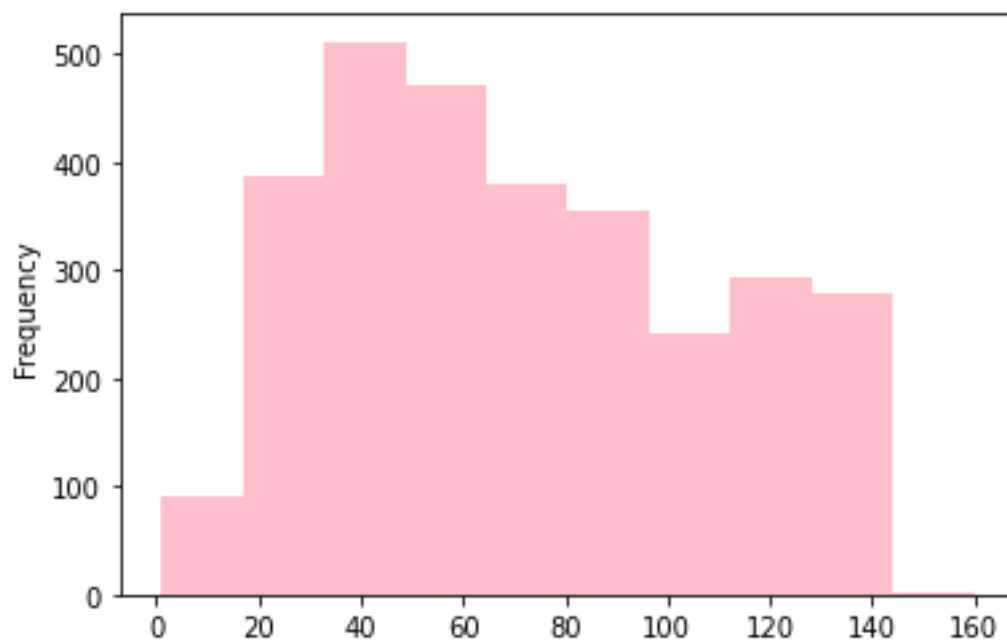
```
train[train['sentiment'] == 1].head(10)
```

	tweet_id	sentiment	author	tweet_data
3	1956967789	1	czareaquino	wants to hang out with friends SOON!
4	1956968416	1	xkilljoyx	@dannycastillo We want to trade with someone w...
5	1956968477	1	xxxPEACHESxxx	Re-pinging @ghostidah14: why didn't you go to...
7	1956968636	1	mcsleazy	Hmmm. http://www.djhero.com/ is down
10	1956969456	1	feinyheiny	cant fall asleep
11	1956969531	1	dudeitsmanda	Choked on her retainers
14	1956970860	1	okiepeanut93	Got the news
16	1956971170	1	poppygallico	@annarosekerr agreed
18	1956971473	1	LCJ82	@PerezHilton lady gaga tweeted about not being...
20	1956971981	1	andreagauster	@raaaaaaek oh too bad! I hope it gets better. ...



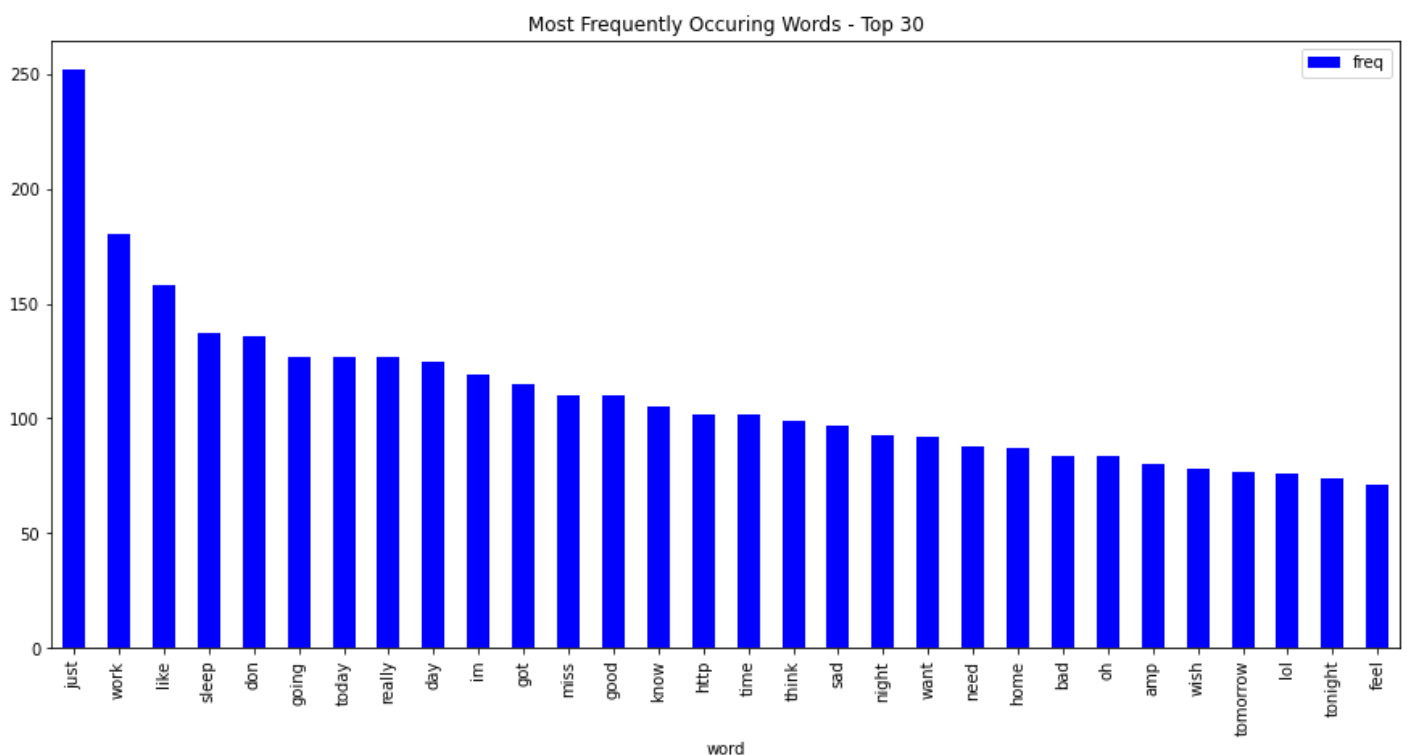
We have total 3010 data set in which 1 is null value data and 938 negative comments and 2073 positive comments above plot shows the same.

Then distribution of the tweets on the basis of the length below shown graph shows the frequency of length.



While seen above graph it is clear that most of the tweets are of the length between 20 words to 90 words and no tweet exceed to 150 words. Then describe positive tweets and negative tweets separately.

Now I decided to find top 30 most used words in this tweets data for this data I used countvectorizer with fit_transform method fit_transform():This fit_transform() method is basically the combination of fit method and transform method, it is equivalent to **fit().transform()**. **fit()** method, where we use the required formula and perform the calculation on the feature values of input data and fit this calculation to the transformer. transform() method, where we apply the calculations that we have calculated in fit() to every data point in feature F. **CountVectorizer** is a great tool provided by the scikit-learn library in Python. It is used to transform a given text into a vector on the basis of the frequency (count) of each word that occurs in the entire text. This is helpful when we have multiple such texts, and we wish to convert each word in each text into vectors (for using in further text analysis).then arrange the frequency of words in descending order and make graph of top 30 graph show below.

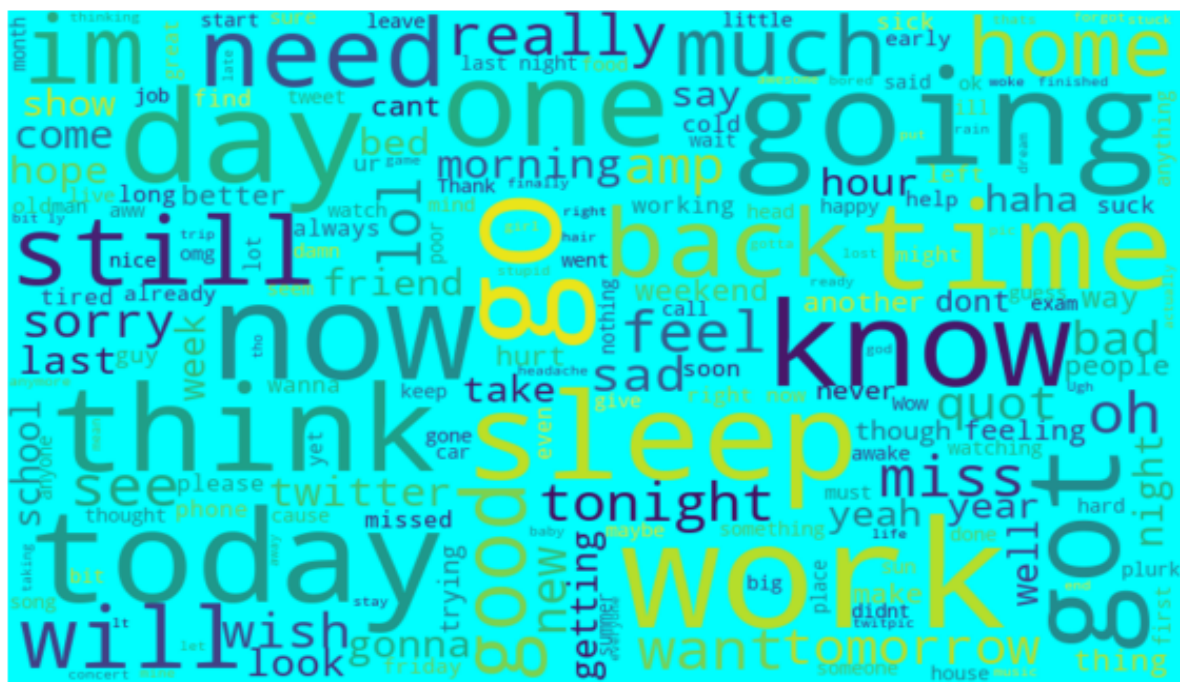


Now I create a word cloud from these words below show image is of this word cloud.



Now make the wordcloud for separate negative and neutral comments.

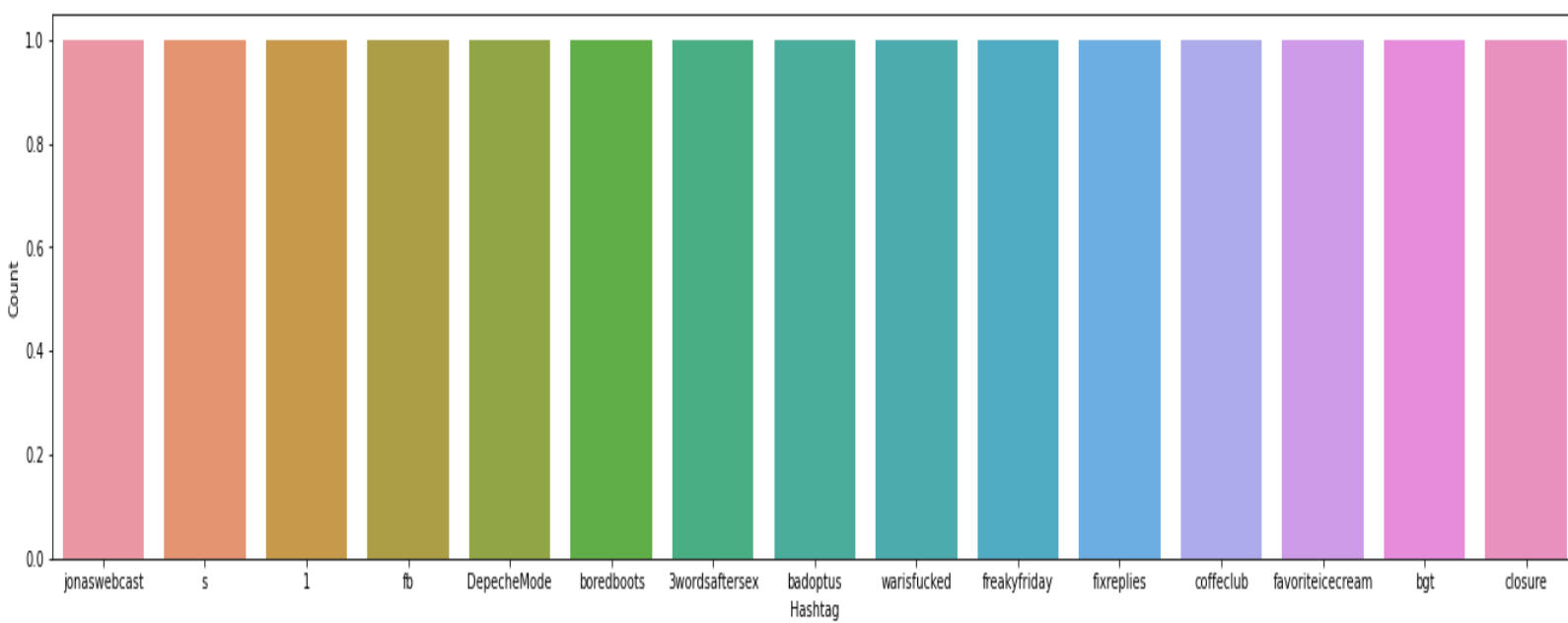
The Neutral Words



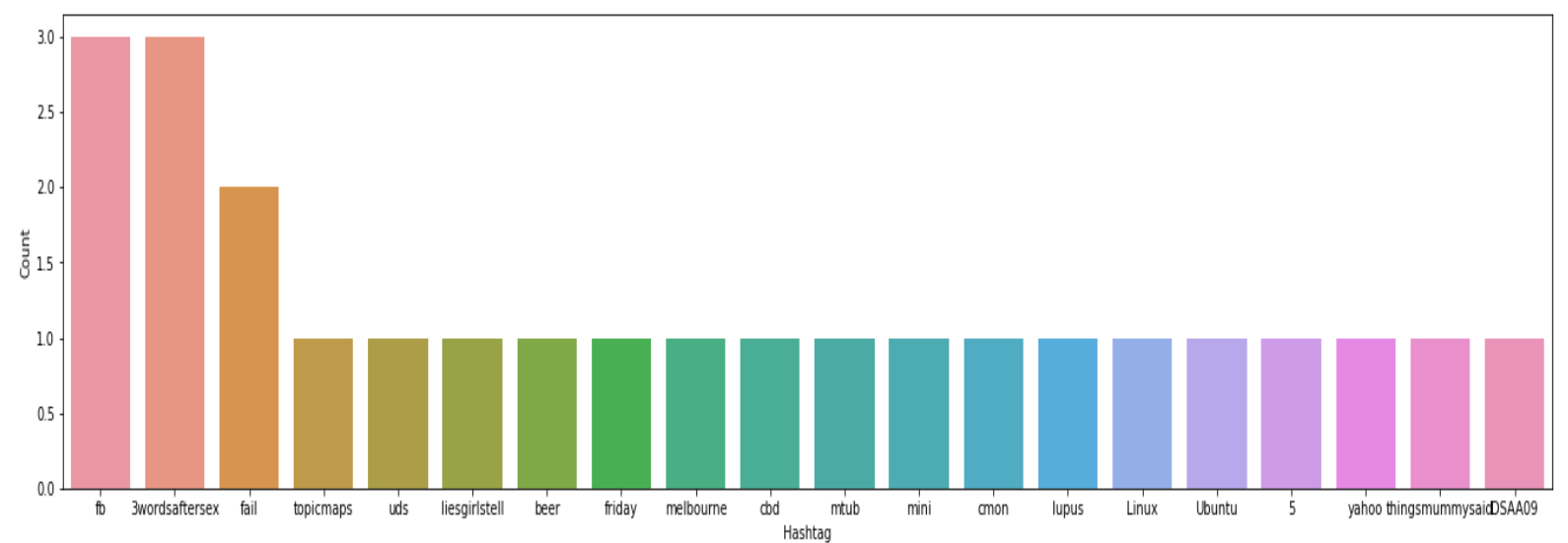
The Negative Words



Now hashtag extracting from the tweets while extracting hashtag we separate the neutral tweets to the negative tweets the further make the graph for top 20 negative hashtag and as well as neutral once.



Top 20 Negative hashtag

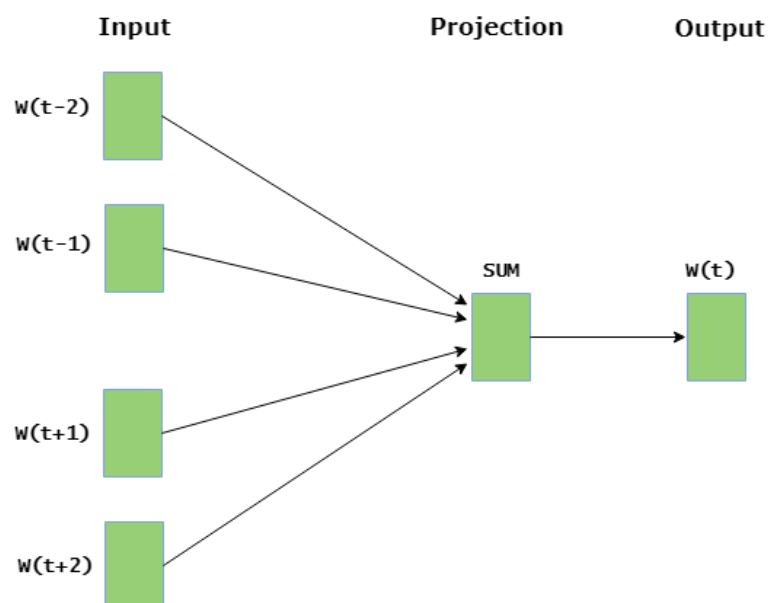


Top 20 Neutral hashtag

Now after data visualization and data cleaning need to create vectorize the data I use for this word2vec this process is known as word embedding.

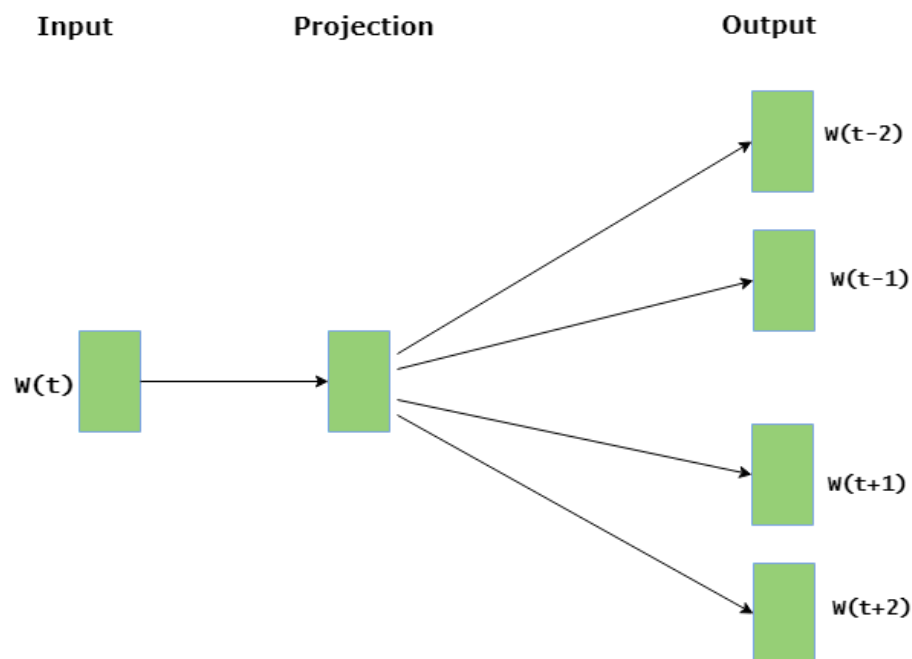
Word Embedding is a language modelling technique used for mapping words to vectors of real numbers. It represents words or phrases in vector space with several dimensions. Word embeddings can be generated using various methods like neural networks, co-occurrence matrix, probabilistic models, etc. **Word2Vec** consists of models for generating word embedding. These models are shallow two-layer neural networks having one input layer, one hidden layer, and one output layer. Word2Vec utilizes two architectures:

CBOW (Continuous Bag of Words): CBOW model predicts the current word given context words within a specific window. The input layer contains the context words and the output layer contains the current word. The hidden layer contains the number of dimensions in which we want to represent the current word present at the output layer.



Skip Gram: Skip gram predicts the surrounding context words

within specific window given current word. The input layer contains the current word and the output layer contains the context words. The hidden layer contains the number of dimensions in which we want to represent current word present at the input layer.



Then I go to label the tweets data so to label each tweet in data it required iterator to each tweet. So, for this I use LabelSentence which classify in two classification words and tags. Then after label the data then I remove the unwanted patterns in the data.

After removal of all the unwanted patterns in the data it time to do the **stemming process** with the help of porterStemmer. Stemming is the process of producing morphological variants of a root/base word. Stemming programs are commonly referred to as stemming algorithms or stemmers. A stemming algorithm reduces the words “chocolates”, “chocolatey”, “Choco” to the root word, “chocolate” and “retrieval”, “retrieved”, “retrieves” reduce to the stem “retrieve”.

Errors in Stemming:

There are mainly two errors in stemming

– **Overstemming** and **Understemming**. Overstemming occurs when two words are stemmed to same root that are of different stems. Under-stemming occurs when two words are stemmed to same root that are not of different stems.

Applications of stemming are:

Stemming is used in information retrieval systems like search engines.

It is used to determine domain vocabularies in domain analysis.

Stemming is desirable as it may reduce redundancy as most of the time the word stem and their inflected/derived words mean the same.

So now we have reduced data of 2712 tweets and we are preform spiting of the data 25% data for testing and 75% of data for train data set I do this with the help of train_test_split method of sklearn library. After splitting 2034 tweets is in training set and 678 testing set.

To **standardize** a dataset means to scale all of the values in the dataset such that the mean value is 0 and the standard deviation is 1.

We use the following formula to standardize the values in a dataset:

$$x_{\text{new}} = (x_i - x) / s$$

where:

x_i : The i^{th} value in the dataset

x : The sample mean

s : The sample standard deviation

then in the end using different classifier to create model and test the accuracy by the help of confusion matrix and f1_score matrix. f1_score matrix is giving most authential result.

First classifier I use is Rendomforest classifier before going toward the result first know about RendomForest classifier.

Random Forest Algorithm

Random Forest is a popular machine learning algorithm that belongs to the supervised learning technique. It can be used for both Classification and Regression problems in ML. It is based on the concept of **ensemble learning**, which is a process of *combining multiple classifiers to solve a complex problem and to improve the performance of the model*.

As the name suggests, ***"Random Forest is a classifier that contains a number of decision trees on various subsets of the given dataset and takes the average to improve the predictive accuracy of that dataset."*** Instead of relying on one decision tree, the random forest takes

the prediction from each tree and based on the majority votes of predictions, and it predicts the final output.

The greater number of trees in the forest leads to higher accuracy and prevents the problem of overfitting.

So by this classifier I got the training accuracy of 99.01% and validation accuracy of 64.75% and f1 score 76.266% the most accurate one is f1 Score. THEN second classifier is Logical Regression Classifier.

Logistic Regression

Logistic regression is a supervised learning classification algorithm used to predict the probability of a target variable. The nature of target or dependent variable is dichotomous, which means there would be only two possible classes.

In simple words, the dependent variable is binary in nature having data coded as either 1 (stands for success/yes) or 0 (stands for failure/no).

Mathematically, a logistic regression model predicts $P(Y=1)$ as a function of X . It is one of the simplest ML algorithms that can be used for various classification problems such as spam detection, Diabetes prediction, cancer detection etc.

So by logical Regression I got training accuracy of 98.43% and validation accuracy of 57.08% and f1 Score is 67.63% logical regression not perform that well as random forest do.

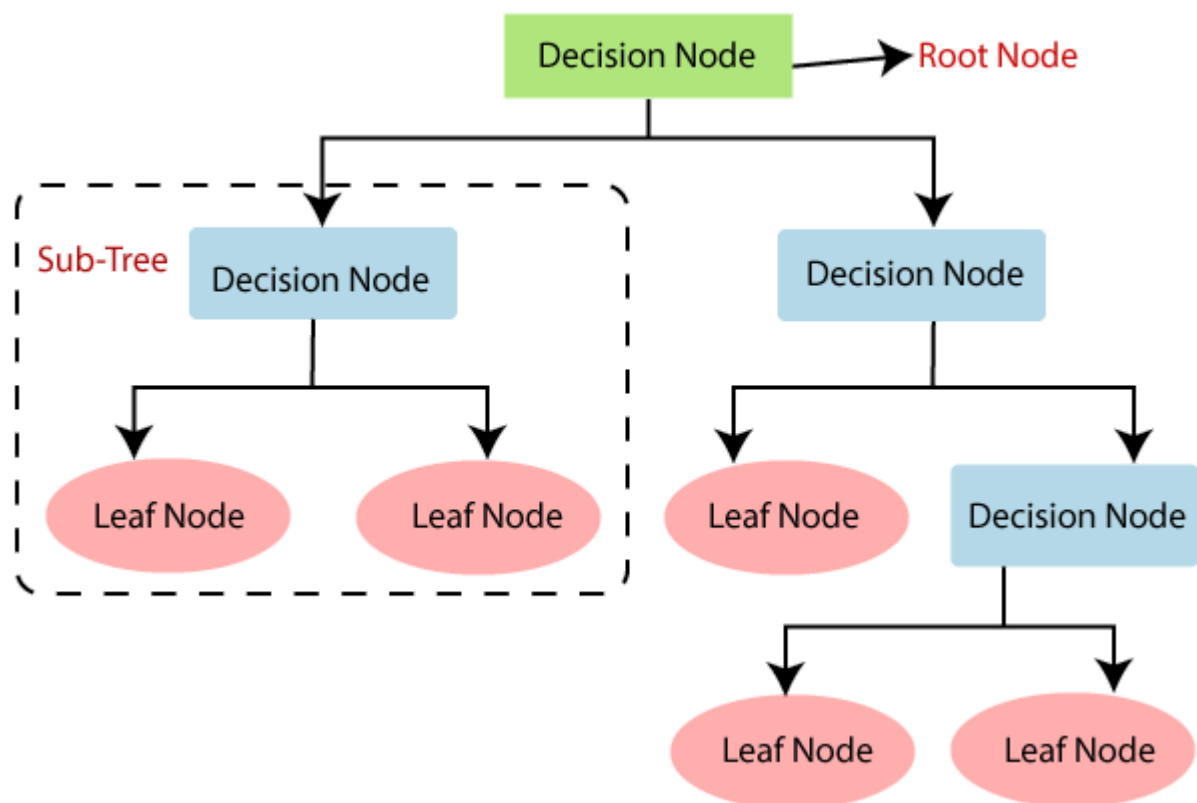
Our third classifier is Decision Tree Classifier.

Decision Tree Classification Algorithm

- o Decision Tree is a **Supervised learning technique** that can be used for both classification and Regression problems, but mostly it is preferred for solving Classification problems. It is a tree-structured classifier, where **internal nodes represent the features of a dataset, branches represent the decision rules and each leaf node represents the outcome.**
- o In a Decision tree, there are two nodes, which are the **Decision Node** and **Leaf Node**. Decision nodes are used to make any decision and have multiple branches, whereas Leaf nodes are the output of those decisions and do not contain any further branches.

- o The decisions or the test are performed on the basis of features of the given dataset.
- o ***It is a graphical representation for getting all the possible solutions to a problem/decision based on given conditions.***
- o It is called a decision tree because, similar to a tree, it starts with the root node, which expands on further branches and constructs a tree-like structure.
- o In order to build a tree, we use the **CART algorithm**, which stands for **Classification and Regression Tree algorithm**. A decision tree simply asks a question, and based on the answer (Yes/No), it further split the tree into subtrees.

Below diagram explains the general structure of a decision tree:



By Decision tree classifier I got training accuracy is 99.01% and validation accuracy of 59.23% f1 Score accuracy of 69.06% Decision Tree classifier is performing better then Logical Regression but not as good as Random Forest classifier.

Our fourth classifier is SVC(Support Vector Classifier)

Support Vector Machines

Support vector machines (SVMs) are a set of supervised learning methods used for [classification](#), [regression](#) and [outliers detection](#).

The advantages of support vector machines are:

Effective in high dimensional spaces.

Still effective in cases where number of dimensions is greater than the number of samples.

Uses a subset of training points in the decision function (called support vectors), so it is also memory efficient.

Versatile: different [kernel functions](#) can be specified for the decision function. Common kernels are provided, but it is also possible to specify custom kernels.

The disadvantages of support vector machines include:

If the number of features is much greater than the number of samples, avoid over-fitting in choosing [kernel functions](#) and regularization term is crucial.

SVMs do not directly provide probability estimates, these are calculated using an expensive five-fold cross-validation.

Support Vectors Classifier tries to find the best hyperplane to separate the different classes by maximizing the distance between sample points and the hyperplane.

Support Vector Classifier used in classifying model making the result are surprisingly very good I got training accuracy of 86.23% and validation accuracy is 67.99% and f1 Score accuracy is 80.88% which so far best even more then random forest classifier.

Fifth and last classifier is xgbclassifier first know what is xgbclassifier. XGBoost is one of the most popular machine learning algorithm these days. Regardless of the type of prediction task at hand; regression or classification.

XGBoost is well known to provide better solutions than other machine learning algorithms. In fact, since its inception, it has become the "state-of-the-art" machine learning algorithm to deal with structured data.

But what makes XGBoost so popular?

Speed and performance : Originally written in C++, it is comparatively faster than other ensemble classifiers.

Core algorithm is parallelizable : Because the core XGBoost algorithm is parallelizable it can harness the power of multi-core computers. It is also parallelizable onto GPU's and across networks of computers making it feasible to train on very large datasets as well.

Consistently outperforms other algorithm methods : It has shown better performance on a variety of machine learning benchmark datasets.

Wide variety of tuning parameters : XGBoost internally has parameters for cross-validation, regularization, user-defined objective functions, missing values, tree parameters, scikit-learn compatible API etc.

XGBoost (Extreme Gradient Boosting) belongs to a family of boosting algorithms and uses the gradient boosting (GBM) framework at its core. It is an optimized distributed gradient boosting library.

By this classifier I got great accuracy slightly similar to that of SVC I got training accuracy of 71.73% and validation accuracy of 68.44% and f1 score is 80.91% accuracy.

Chapter 7

Conclusion and Future Scope

Conclusion

Nowadays, sentiment analysis or opinion mining is a hot topic in machine learning. We are still far to detect the sentiments of a corpus of texts very accurately because of the complexity in the English language and even more if we consider other languages such as Chinese.

In this project we tried to show the basic way of classifying tweets into positive or negative category using Naive Bayes as baseline and how language models are related to the Naive Bayes and can produce better results. We could further improve our classifier by trying to extract more features from the tweets, trying different kinds of features, tuning the parameters of the naïve Bayes classifier, or trying another classifier all together.

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 compute the frequency of each term in tweet. Using machine learning supervised approach help to obtain the results.

Future Scope

In future work , we aim to handle emoticons , dive deep into emotional analysis to further detect idiomatic statements .We will also explore richer linguistic analysis such as parsing and semantic analysis.

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

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

Σ A web-based application can be made for our work in future.

37

Σ We can improve our system that can deal with sentences of multiple

meanings.

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

Σ We can start work on multi languages like Hindi, Spanish, and Arabic to provide sentiment analysis to more local.