Sentimental Analysis on Twitter Data

A Major Project Report Submitted

in Partial Fulfillment of the Requirements

For the Degree of

**BACHELORS OF TECHNOLOGY**

In

**COMPUTER SCIENCE & ENGINEERING**

**Submitted by**

**Panwar Shivani 190180101038**

**Priya Joshi 190180101042**

**Shivani Sharma 190180101051**

**Yogita 190180101063**

**Under the Guidance of**

**Ms. Swati Verma**

**Assistant Professor**



DEPARTMENT OF COMPUTER SCIENCE ENGINEERING

Bipin Tripathi Kumaon Institute of Technology, Dwarahat

(Affiliated to Uttarakhand Technical University Dehradun)

**JUNE-2023**

**Declaration**

We, Panwar Shivani, Priya Joshi, Shivani Sharma, Yogita students of Bachelor of Technology in COMPUTER SCIENCE AND ENGINEERING, BIPIN TRIPATHI KUMAON INSTITUTE OF TECHNOLOGY, DWARAHAT, hereby declare that the project entitled “Sentimental Analysis on Twitter Data” submitted for the degree of BACHELOR OF TECHNOLOGY under the supervision of Ms. Swati Verma, Assistant Professor, Department of Computer Science and Engineering, Bipin Tripathi Kumaon Institute of Technology is the outcome of my own work, is bonafide and correct to the best of our knowledge and this work has been carried out taking care of Engineering Ethics.

Date-………………………...

Panwar Shivani (190180101038)

Priya Joshi (190180101042)

Shivani Sharma (190180101051)

Yogita (190180101063)

Department of Computer Science Engineering

Bipin Tripathi Kumaon Institue of Technology, Dwarahat

**Certificate**

This is to certify that the project work entitled as “SENTIMENTAL ANALYSIS ON TWITTER DATA” is being Submitted by Panwar Shivani (190180101038), Priya Joshi (190180101042), Shivani Sharma (190180101051), Yogita (190180101063) in the partial fulfillment for the award of the Degree of Bachelor of Technology in “COMPUTER SCIENCE AND ENGINNERING” in the academic during 2019-2023.

To the best of knowledge, the matter embodied in the project has not been submitted to any other university or institute for the award of any degree.

Under the esteemed Guidance of Head of the Department

Ms. Swati Verma Dr. Ajit Singh

CSE Department Professor & Head

**Acknowledgement**

We are pleased to present this Project report entitled SENTIMENTAL ANALYSIS ON TWITTER DATA. It is indeed a great pleasure and a moment of immense satisfaction for us to express our sense of profound gratitude and indebtedness towards our guide Ms. Swati Verma whose enthusiasm are the source of inspiration for me. We are extremely thankful for the guidance and untiring attention, which he bestowed on us right from the beginning. His valuable and timely suggestions at crucial stages and above all his constant encouragement have made it possible for us to achieve this work. We would like to thank the entire Teaching staff who are directly or indirectly involved in the various data collection and software assistance to bring forward this semester report. Last but not the least, we would like to thank all my B.Tech. colleagues for their co-operation and useful suggestion and all those who have directly or indirectly helped us in completion of this project work.

Sincerely,

Panwar Shivani

Priya Joshi

Shivani Sharma

Yogita

**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 machine learning algorithms.

Table of Contents

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

Certificate………………………………………………………………………..…………….3

Acknowledgement……………………………………………………..……………………...4

Abstract………………………………………………………………………………………..5

Chapter 1: Introduction………..………………………………………………………...…….8

[1.1 Introduction to Sentimental Analysis](#_Toc137330232)………………..……………………………….8

[1.2 Introduction to Python](#_Toc137330233) 9

[1.3 Introduction to NLTK 9](#_Toc137330234)

[1.4 Introduction to Machine Learning 10](#_Toc137330235)

1.4.1 Naïve-Bayes (NB) Classifier …………………………………... ……...…10

1.4.2 MultinomialNB Classifier …………………………………………..…….11

1.4.3 BernoulliNB Classifier ………………………………………....................11

1.4.4 Logistic Regression Classifier ……………………………………………12

1.4.5 Decision Tree Classifier…………………….. ………….………………. 12

1.5 Goal of Project…………….………………………………………………………… 13

1.6 Need of Sentimental Analysis ………………………………..………………………14

1.6.1 Industry Evolution …………………………..……………………………14

1.6.2 Research Demand ……………………………….…………………….....15

1.6.3 Decision Making …………………………….………………………….. 15

1.6.4 Understanding Contextual ……………………………………………… 15

1.6.5 Internet Marketing ……………………………………………………… 16

1.7 Applications of Sentiment Analysis …………………………….............................. 16

1.7.1 Word of Mouth (WOM) ………………………………………..…….… 16

1.7.2 Voice of Voters ……………………………………………………….… 17

1.7.3 Online Commerce ………………………………………………………. 17

1.7.4 Voice of the Market (VOM) ………………………………..…..………. 17

1.7.5 Brand Reputation Management (BRM)……………………………………..17

1.7.6 Government …………………….………………………..…..…………….. 18

Chapter 2: Literature Review …………………………………………..………….………. 19

Chapter 3: Problem Statement …………………………………...………………………... 23

3.1 Objectives ……………………………………………………………………... 23

3.2 Methodology ………………………………………………….……………..... 24

Chapter 4: Implementation ……………………………………….………………………… 25

4.1 Proposed Architecture………………………………………………………….. 25

4.2 Data Collection ………………………………………………………………… 26

4.3 Data Pre-Processing ………………………………………………………….... 26

4.4 Classification ……………………………….………………………………….. 28

4.4.1 Feature Extraction………………………………………………………….29

Chapter 5: Results and Analysis …………………………………………….…..………….30

Chapter 6: Conclusion and Future Scope ……………………………….……………….…33

6.1 Conclusion …………………………………………………………………..………... 33

6.2 Future Scope ………………………………………………………….......................... 34

Chapter 7: References ………………… …………………………………………..……...35

# 

# Chapter -1: 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 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 Sentimental 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. Sentimental Analysis can be done at document, phrase

and sentence level. In document level, summary of whole level, summary of the entire

document is taken first and then it is analyzed 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

another 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.

### 1.2 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.

**1.3 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.

**1.4 Introduction to Machine Learning**

Machine Learning enables a machine to automatically learn from data, improve performance

from experiences, and predict things without being explicitly programmed. Machine Learning

is said as a subset of artificial intelligence that is mainly concerned with the development of

algorithms which allow a computer to learn from the data and past experience on their own.

A Machine Learning system learns from historical data, builds the prediction models, and

Whenever it receives new data, predicts the output for it. The accuracy of predicted output

Depends upon the amount of data, as the huge amount of data helps to build a better model

which predicts the output more accurately.

Different machine learning classifiers which we are going to use to build our classifier are:

* Naïve-Bayes Classifier
* Multinomial Classifier
* Bernoulli NB Classifier
* Logistic Regression Classifier
* Decision Tree Classifier

**1.4.1 Naïve-Bayes (NB) Classifier**

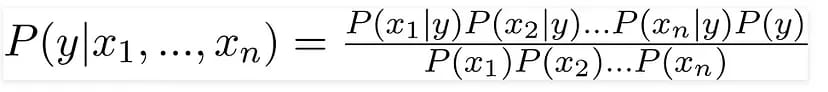
Naïve-Bayes classifiers are probabilistic classifiers which come under supervised machine

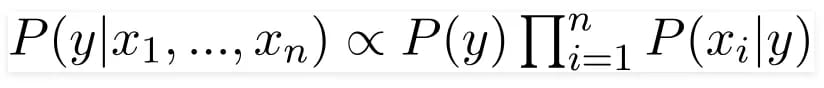
learning techniques. These classifiers are based on applying Bayes’ theorem with strong (naïve)

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





**1.4.2 MultinomialNB Classifier**

The Multinomial Naive Bayes algorithm is a Bayesian learning approach popular in Natural

Language Processing (NLP). The program guesses the tag of a text, such as an email or a

newspaper story, using the Bayes theorem. It calculates each tag's likelihood for a given sample

and outputs the tag with the greatest chance. The Naive Bayes classifier is made up of a number

of algorithms that all have one thing in common: each feature being classed is unrelated to any

another feature. A feature's existence or absence has no bearing on the inclusion or exclusion

of another feature.

**1.4.3 BernoulliNB Classifier**

BernoulliNB also implements NB algorithm for training and classification. It uses NB

for multivariate Bernoulli distribution of data; i.e., there can be many features but

each and every one is assumed to have a binary value or Boolean (true or false)

variable. Hence, every class requires samples which have to be represented in binary

value variables; also, if any other kind of data is given then BernoulliNB can binaries

its input.

The BernoulliNB decision rule is explained as:



which is different from MultinomialNB’s rule, this rule directly punishes any unavailability of

feature which behaves like a feature of class, where as in multinomial it simply ignores if there

is any non-occuring feature.

**1.4.4 Logistic Regression Classifier**

Despite its name Logistic regression, is not a regression model but a linear model for

classification. This model is also known by other names as Maximum-Entropy (MaxEnt)

classification or log-linear classifier. A logistic function is used in this model, were probability

describe the outcome of single trial. The logistic regression can be implemented from Scikit-

learn library of Python in which there is a class named Logistic Regression. This

implementation fits a OvR (one-vs-rest) multiclass regression with an optional L1 or L2

regularization.

**1.4.5 Decision Tree Classifier**

A decision tree is a type of supervised machine learning used to categorize or make predictions

based on how a previous set of questions were answered. The model is a form of supervised

learning, meaning that the model is trained and tested on a set of data that contains the desired

categorization. The decision tree may not always provide a clear-cut answer or decision.

Instead, it may present options so the data scientist can make an informed decision on their

own. Decision trees imitate human thinking, so it’s generally easy for data scientists to

understand and interpret the results.

**1.5 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 post 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 number 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 different

Tweets. Public opinions are mined from Twitter and then classified into sentiments, whether

positive or negative by using supervised machine learning classifiers. These results will let us

know about the reviews and opinions of people. 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.

To many companies and organizations, a customer’s perception of a product or service is

extremely valuable information. From the knowledge gained from an analysis such as this a

company can identify issues with their products, spot trends before their competitors, create

improved communications with their target audience, and gain valuable insight into how

effective their marketing campaigns were. Through this knowledge companies gain valuable

feedback which allows them to further develop the next generation of their product.

In the context of the sentiment analysis being carried out for this application, the results will

allow user to gain insight into how topics are being perceived by the public. This is very

valuable information as public is uploading their expectations, opinions and views on different

events. This really revolutionizes the feedback process. An application such as this has the

potential to analyze the sentiment in real time giving the users immediate feedback on how a

organization is being help in the eyes of its audience.

**1.6 Need of Sentimental Analysis**

**1.6.1 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, 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.

**1.6.2 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.

**1.6.3 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 will 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.

**1.6.4 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.

**1.6.5 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.

**1.7 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:

**1.7.1 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.

**1.7.2 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.

**1.7.3 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 suggestions. These details and reviews have been collected by company and conversion of

data into the geographical form with updates of the recent online commerce websites who use

these current techniques.

**1.7.4 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.

**1.7.5 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.

**1.7.6 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. 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-2: Literature Survey

Many research has been done on the subject of sentiment analysis in past. Latest research in

this area is to perform sentiment analysis on data generated by user from many social

networking websites like Facebook, Twitter, Amazon, etc. Mostly research on sentiment

analysis depend on machine learning algorithms, whose main focus is to find whether given

text is in favor or against and to identify polarity of text. In this chapter we will provide

insight of some of the research work which helps us to understand the topic deep.

**P. Pang, L. Lee, S. Vaithyanathan et al [8]**

They were the first to work on sentiment analysis. Their main aim was to classify text by

overall sentiment, not just by topic e.g., classifying movie review either positive or negative.

They apply machine learning algorithm on movie review database which results that these

algorithms out-perform human produced algorithms. The machine learning algorithms they

use are Naïve-Bayes, maximum entropy, and support vector machines. They also conclude by

examining various factors that classification of sentiment is very challenging. They show

supervised machine learning algorithms are the base for sentiment analysis.

**P. Pang, L. Lee et al [9]**

By collecting large amount of data has always been a key to find out what people is thinking

or expecting. With the emergence in the field of social media, availability of data which is

full of opinion resources is very high. Other resources such as blogs, review sites, messages,

etc. are helping us to know what people can do and their opinion about the topic. The sudden

increase of work in the field of data mining and sentiment extraction deals with the

computational power to solve the problem of opinion mining or subjectivity in text. Hence

various new systems are created based on different languages and commands that can deal

directly with opinion mining as the first-class object and direct response or live research also

becoming the area of interest. They take a survey which covers that methodology and

approaches that are used in direct response of opinion mining are more helpful than others.

Their focus is on functions that can solve new challenges rising in sentiment analysis

applications. They also compared these new techniques to already present traditional analysis

which is based on facts.

**E. Loper, S. Bird et al [10]**

Natural Language Toolkit (NLTK) is a library which consists of many program modules,

large set of structured files, various tutorials, problem sets, many statistics functions, ready-

to-use machine learning classifiers, computational linguistics courseware, etc. The main

purpose of NLTK is to carry out natural language processing, i.e., to perform analysis on

human language data. NLTK provides corpora which are used for training classifiers.

Developers create new components and replace them with existing component, more than

structured programs are created and more sophisticated results are given by dataset.

**H. Wang, D. Can, F. Bar, S. Narayana et al [11]**

They were the researchers who proposed a system for real time analysis of public responses

for 2012 presidential elections in U.S. They collect the responses from Twitter, a micro

blogging platform. Twitter is one the social network site where people share their views,

thoughts and opinions on any trending topic. People responses on Twitter for election

candidates in U.S. created a large amount of data, which helps to create a sentiment for each

candidate and also created a prediction of whom winning.

A relation is created between sentiments that arise from people response on twitter with the

complete election events. They also explore how sentiment analysis affects these public

events. They also show this live sentiment analysis is very fast as compared to traditional

content analysis which takes many days or up to some weeks to complete. The system they

demonstrated analyzes sentiment of entire Twitter data about the election, candidates,

promotions, etc. and delivering results at a continuous rate. It offers media, politicians and

researchers a new way which is timely effective which is completely based on public opinion.

**L. Jiang, M. Yu, M. Zhou, X. Liu, T. Zhao et al [13]**

Twitter sentiment analysis was growing at faster rate as amount of data is increasing. They

created a system which focuses on target dependent classification. It is based onTwitter in

which a query is given first; they classify the tweets as positive, negative or neutral

sentiments with respect to that query that contain sentiment as positive, negative or neutral. In

their research, query sentiment serves as target. The target independent strategy is always

adopted to solve these problems with the help of state-of-the-art approaches, which may

sometime assign immaterial sentiments to target.

Also, when state-of-the-art approaches are used for classification, they only take tweet into

consideration. These approaches ignore related tweet, as they classify based on current tweet.

However, because tweets have property to be short and mostly ambiguous, considering

current tweet only for sentiment analysis is not enough. They propose a system to improve

target-dependent Twitter sentiment classification by:

1) Integrating target-dependent features, and

2) Taking related tweets into consideration.

**C. Tan, L. Lee, J. Tang, L. Jiang, M. Zhou, P. Li, et al [14]**

They show that information that can be used to improve user-level sentiment analysis. There

base of research is social relationships, i.e., users that are connected in any social platform

will somehow hold similar opinions, thoughts; therefore, relationship information can

supplement what they extract from user’s viewpoint. They use Twitter as there source of

experimental data and they use semi-supervised machine learning framework to carry out

analysis. They propose systems that are persuade either from the network of Twitter

followers or from the network formed by users in Twitter in which users referring to each

other using “@username”. According to them, these semi-supervised learning results shows

that by including this social-network information leads to statistically significant

improvement in performance of sentiment analysis classification over the performance based

on the approach of SVM (Support Vector Machines) that have only access to textual features.

# Chapter-3: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.

Today, if any one wants to purchase a product or to give vote or to watch a movie, etc. then

that person will first want to know what are other people reviews, reactions and opinions

about that product or candidate or movie on social media websites like Twitter, Facebook,

etc. So, there is a need of system that can automatically generate sentiment analysis from this

huge amount of data.

## 3.1 Objective

The main objective of this project work is to perform the sentiment analysis on tweets which

are extracted from Twitter. Thus to achieve this objective we build a classifier based on

machine learning and perform live sentiment analysis on data collected.

**3.2 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.

∑ Comparing results of each classifier and plotting a graph that show the trend of

positive and negative sentiment of different people.

# Chapter-4: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

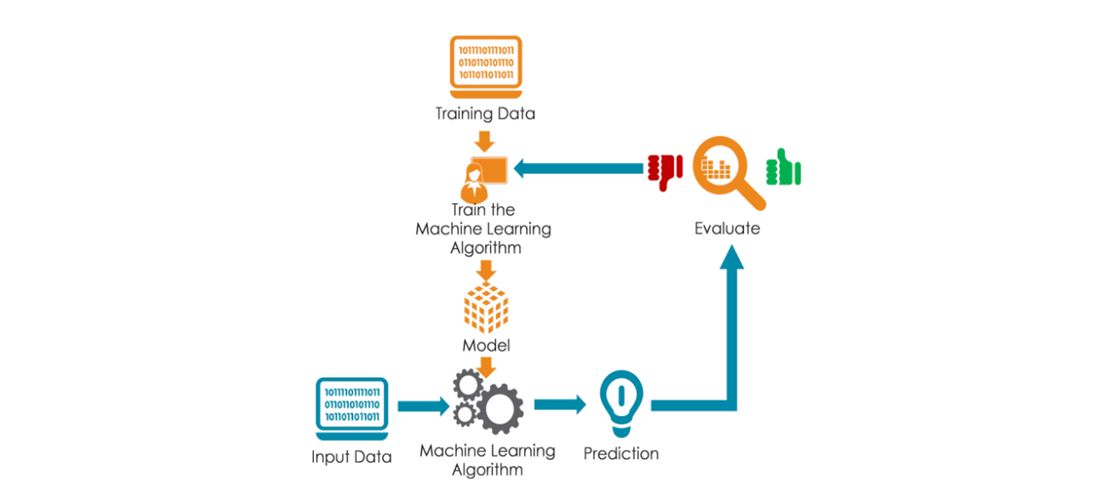
our proposed architecture.

## 4.1 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 below figure:



## 4.2 Data Collection

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 my collecting and annotating them manually which

very long and fastidious.

Additionally to find a way of getting a corpus of tweets, we need to take of having abalanced

data set, 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 1578612 tweets in English coming from two

sources: Kaggle and Sentiment140. It is composed of four columns that are Item ID,

Sentiment, Sentiment Source and Sentiment Text. 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 Sentiment Text columns containing the tweets in a raw format.

## 4.3 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 preprocessing 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.

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’s name with a word ‘AT\_USER' and add it to stop words

∑ 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

Below Table shows the various types of contents that are included in tweets and also the actions

performed on these contents.

|  |  |
| --- | --- |
| **CONTENT** | **ACTION** |
| Punctuation(! ? , . “” ; :) | Removed |
| Uppercase characters | Lowercase all content |
| URLs and web links | Remove URLs or replaced with “URL” and then added in stop words |
| Number | Removed |
| Words not starting with alphabets | Removed |
| All Words | Stemmed all word  (Converted into simple form) |
| Stop Words | Removed |
| White spaces | Removed |

Once our data is cleaned and ready for processing our next step is to classify this cleaned data

into different classes.

**4.4 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 scikitlearn’.

In order to build our classifier, we use seven in-build classifiers which come in Scikitlearn

library, which are:

∑ Naïve-Bayes Classifier

∑ MultinomialNB Classifier

∑ BernoulliNB Classifier

∑ Logistic Regression Classifier

∑Decision Tree Classifier

**4.4.1 Feature Extraction**

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

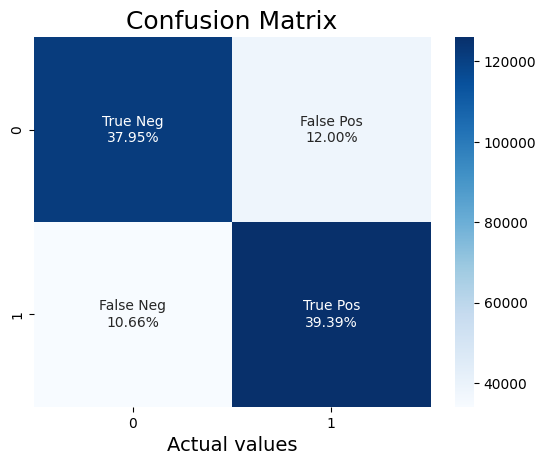
helps to indicate the polarity of the sentences.

# Chapter-5:Result And Analysis

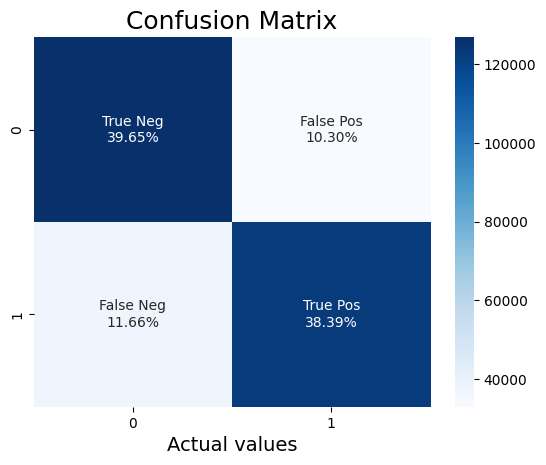
# RESULTS:

Accuracy of Algorithms used in our model:

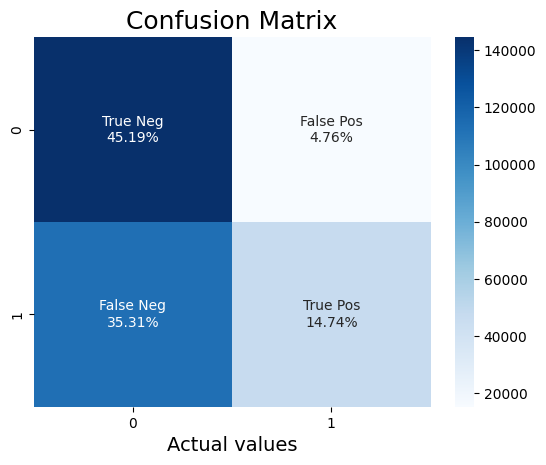
1. **Bernoulli NB** : 76.64%



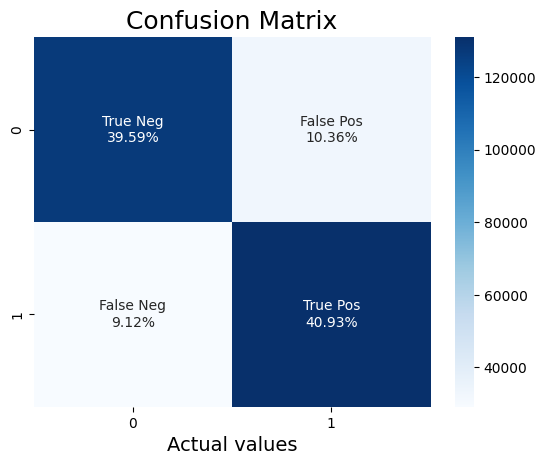
1. **Multinomial NB** : 78.85%



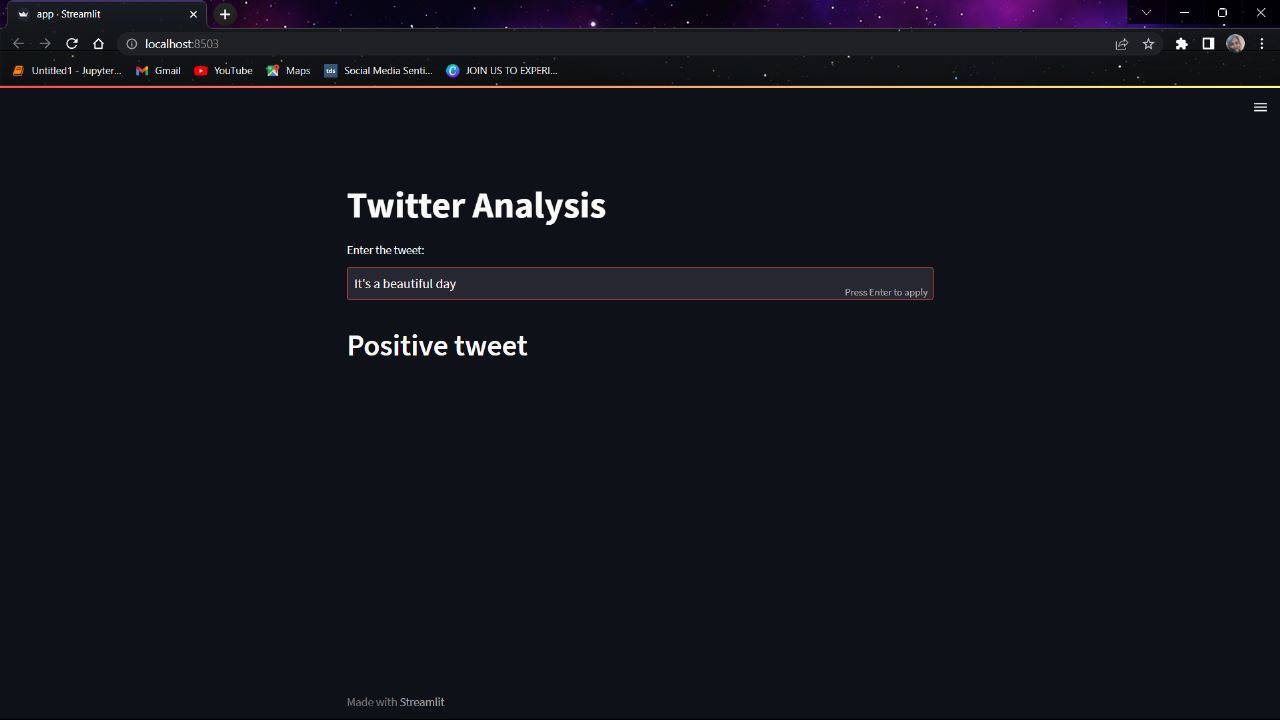
1. **Decision Tree Classifier** : 75.58%



1. **Logistic Regression**: 79.80%



**Front-end:**



# 

# Chapter-6: Conclusion and Future Scope

## 6.1 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. In conclusion, sentiment

analysis on Twitter data using machine learning algorithms can provide valuable insights into

public opinion and sentiment towards various topics. Each algorithm has its advantages and

limitations, and the choice of algorithm depends on the specific task and data at hand.

Logistic Regression showed maximum accuracy of 82% among all the algorithms we applied.

By understanding the strengths and weaknesses of each algorithm, we can make informed

decisions when analyzing Twitter data and extract meaningful insights that can inform

decision-making in various fields such as marketing, politics, and social sciences.

## 

## 6.2 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.

∑ 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 more local.

∑ We have not considered neutral tweets in this project, which limited the scope of our project.

# 

# Chapter-7: References

[1] H. Zang, “The optimality of Naïve-Bayes”, Proc. FLAIRS, 2004

[2] C.D. Manning, P. Raghavan and H. Schütze, “Introduction to Information Retrieval”, Cambridge University Press, pp. 234-265, 2008

[3] A. McCallum and K. Nigam, “A comparison of event models for Naive Bayes text

classification”, Proc. AAAI/ICML-98 Workshop on Learning for Text Categorization, pp.

41-48, 1998

[4] M. Schmidt, N. L. Roux and F. Bach, “Minimizing finite Sums with the Stochastic

Average Gradient”, 2002

[5] ‘Y. LeCun, L. Bottou, G. Orr and K. Muller, “Efficient BackProp”, Proc. In Neural

Networks: Tricks of the trade 1998.

[6] T. Wu, C. Lin and R. Weng, “Probality estimates for multi-class classification by pairwise

coupling”, Proc. JMLR-5, pp. 975-1005, 2004

[7] “Support Vector Machines” [Online],http://scikitlearn.org/stable/modules/svm.html#svm-

classification, Accessed Jan 2016

[8] P. Pang, L. Lee and S. Vaithyanathan, “Thumbs up? sentiment classification using

machine learning techniques”, Proc. ACL-02 conference on Empirical methods in natural

language processing, vol.10, pp. 79-86, 2002

[9] P. Pang and L. Lee, “Opinion Mining and Sentiment Analysis. Foundation and Trends in

Information Retrieval”, vol. 2(1-2), pp.1-135, 2008

[10] E. Loper and S. Bird, “NLTK: the Natural Language Toolkit”, Proc. ACL-02 Workshop

on Effective tools and methodologies for teaching natural language processing and

computational linguistics ,vol. 1,pp. 63-70, 2002

[11] H. Wang, D. Can, F. Bar and S. Narayana, “A system for real-time Twitter sentiment

analysis of 2012 U.S.presidential election cycle”, Proc. ACL 2012 System Demostration, pp.

115-120, 2012

[12] O. Almatrafi, S. Parack and B. Chavan, “Application of location-based sentiment

analysis using Twitter for identifying trends towards Indian general elections 2014”. Proc.

The 9th International Conference on Ubiquitous Information Management and

Communication,2015

[13] L. Jiang, M. Yu, M. Zhou, X. Liu and T. Zhao, “Target-dependent twitter sentiment

classification”, Proc. The 49th Annual Meeting of the Association for Computational

Linguistics: Human Language Technologies, vol. 1, pp. 151-160, 2011

[14] C. Tan, L. Lee, J. Tang, L. Jiang, M. Zhou and P. Li, “User-level sentiment analysis

incorporating social networks”, Proc. The 17th ACM SIGKDD international conference on

Knowledge discovery and data mining, pp. 1397-1405, 2011

[15] A. Pak and P. Paroubek, “Twitter as a Corpus for Sentiment Analysis and Opinion

Mining”, vol. 10, pp. 1320-1326, 2010

[16] B. Sun and TY. V. Ng, “Analyzing Sentimental influence of Posts on Social Networks”,

Proc. The 2014 IEEE 18th International Conference on Computer Supported Cooperative

Work in Design, 2014

[17] A. Go, R. Bhayani and L. Huang, “Twitter sentiment classification using distant

supervision”, CS224N Project Report, Stanford, vol.1-12, 2009

[18] A. Barhan and A. Shakhomirov, “Methods for Sentiment Analysis of Twitter Messages”,

Proc.12th Conference of FRUCT Association, 2012

[19] T. Mitchell, “Machine Learning”, McGraw Hill, 1997

[20] F. Jensen, “An Introduction to Bayesian Networks”, Springer, 1996

[21] T. C. Peng and C. C. Shih, “An Unsupervised Snippet-based Sentiment Classification Method for Chinese Unknown Phrases without using Reference Word Pairs”.IEEE/WIC/ACM Int. Conf. on Web Intelligence and Intelligent Agent Technology, vol. 3, pp. 243-248, 2010

[22] R. Feldman, “Techniques and applications for sentiment analysis”, Proc. ACM, pp. 56-

82, 2009

[23] N. Cristianini and J. Shawe-Taylor, “ An Introduction to Support Vector Machines and

Other Kernel-based Learning Methods”, Cambridge University Press, March 2000