SENTIMENT ANALYSIS OF TRENDING HASHTAGS IN TWITTER USING WEB SCRAPING

A Project Report Submitted By

AYSHA HARIS A

SUBMITTED IN PARTIAL FULFILLMENT OF THE
REQUIREMENTS FOR THE DEGREE OF
MASTER OF SCIENCE (M.SC. COMPUTER SCIENCE)
(University of Calicut)
AT

DEPARTMENT OF COMPUTER SCIENCE
FAROOK COLLEGE (AUTONOMOUS)
FAROOK COLLEGE P.O., CALICUT, KERALA, INDIA
MARCH 2020

ACKNOWLEDGEMENT

Foremost,I am grateful for the blessings that ALMIGHTY GOD has showered upon me over the years, this project would not have completed without them.I wish to express my sincere gratitude to the head of the institute Dr.NASEER for providing me the necessary facilities to carry out my work. I respect and thank our head of the department Dr.V KABEER for all the support and guidance which made me complete the project duly.I am extremely thankful to him for providing the support amidst his busy schedule. I heartily thank my project coordinator,Dr.ABDUL HALEEM , for his guidance,motivations, suggestions and sincere effort during this project work. I owe my deep gratitude to my project guide Mr.SAMEER V.V , who took keen interest on my project work and guided me all along, till the completion of my project work by providing all the necessary information for developing a good system. I am thankful to and fortunate enough to get constant encouragement, support and guidance from all teaching staffs of Computer Science Department who helped me in successfully completing my project work. Also, I would like to extend my sincere esteems to all my fellow classmates,my parents ,my husband for their timely support.

AYSHA HARIS A

DECLARATION

I hereby declare that this submission is my own work and that, to the best of my knowledge and belief, it contains no material previously published or written by another person or material which has been accepted for the award of any other degree or diploma of the university or other institute of higher learning, except where due acknowledgement has been made in the text.

Dated: March 2020

Signature of the Author: _____

AYSHA HARIS A

DEPARTMENT OF COMPUTER SCIENCE
FAROOK COLLEGE (AUTONOMOUS)

Date: **March 2020**

This is to certify that the project report entitled SENTIMENT ANALYSIS OF TRENDING HASHTAGS IN TWITTER USING WEB SCRAPING submitted by AYSHA HARIS A to the University of Calicut for the award of the Master of Science (M.Sc.) in Computer Science is a bonafide record of the project work carried out by her under my supervision and guidance. The content of the report, in full or parts have not been submitted to any other institute or University for the award of any other degree or diploma.

Research Supervisor: _____
Mr. SAMEER V.V

Head of the Department: _____
Dr. V. Kabeer

This is to certify that the candidate was examined by us in the Project Viva Voce Examination held on ..... and her Register Number is FKASMCS004.

Examiners:

Examiner 1: _____

Examiner 2: _____

Table of Contents

Table of Contents **v**

List of Tables **vii**

List of Figures **viii**

Abstract **ix**

1 Introduction **1**
1.1 Project Category 2
1.2 project profile 2
1.2.1 Tools/platforms 2

2 Problem analysis and Scope of the System **3**
2.1 Problem Definition 3
2.2 Need identification 3
2.3 Problem Domain 4
2.4 Objectives 4
2.5 Methodology 5
2.6 Scope 5
2.7 Motivation 5

3 Analysis and Software Specification **6**
3.1 Literature review 6
3.2 Analysis and observation 7
3.3 Feasibility study 8
3.3.1 Economic Feasibility 8
3.3.2 Operational Feasibility 8
3.3.3 Technical Feasibility 8

3.4 Functional and Non Functional Specifications 9
3.4.1 Functional requirements 9
3.4.2 Non functional requirements 9
3.5 Technical Requirements 9

4 Design **11**
4.1 Modules of the system 11
4.1.1 Data collection 11
4.1.2 Tweet management 12
4.1.3 Data analysis 12
4.1.4 Visualisation 13
4.2 Data Flow Diagrams 13
4.3 User interface 15
4.4 Database and Design 17
4.4.1 List of Entities and Attributes 17
4.4.2 Structure of Tables 17

5 Implementation **21**
5.1 Tools/scripts for implementation 21
5.2 Coding 22
5.2.1 Extracting Tweets 22
5.2.2 Classifiying Tweets 28
5.2.3 Displaying Tweets with Sentiments 36
5.3 Important screenshots 38

6 Testing **42**
6.1 Introduction 42
6.2 Testing Methodologies adopted 42
6.2.1 Unit Testing 42
6.2.2 Integration Testing 43
6.2.3 System Testing 44

7 Conclusion **45**

References **46**

Bibliography **46**

List of Tables

4.1 List of entities and attributes. 18
4.2 Login Table. 19
4.3 Signup table. 19
4.4 Twitter Table. 20

List of Figures

4.1 Level 0 13
4.2 Level 1 14
4.3 Level 1(User management) 14
4.4 Level 2 14
4.5 Login page 15
4.6 Home page 15
4.7 Get Tweets 16
4.8 Result page 16

5.1 Login page 38
5.2 Home page 39
5.3 Tweet select page 39
5.4 Result page 40
5.5 Result page 40
5.6 Result Graph 41

Abstract

Social media has become an exceptional medium of communication among us that people like to express their views on any burning public issue through social media like Twitter facebook etc. As the tweet size in Twitter is limited ,the person expresses their views precisely in a few words . So it gets conveyed very clearly without ambiguity.Various open problems exist in this area of opinion mining where people's text or comments are analysed for its sentiments .People interact through messages called 'tweets'. Registered users can post, like, tweet, retweet whereas unregistered users can only read them.People commonly express their viewpoints regarding anything in this platform. Millions express their stand through tweets regarding the issues happening n the current world. Sentiment analysis is a part of natural language processing where data analysis is done to extract meaningful results on large amounts of data.In the proposed system ,for getting data in large amount,Web scraping technology is used instead of Twitter API used.web scraping is a technology which allows extraction of data in large amounts from a web page in a structured format.The extracted tweets are analysed to know its sentiments and are categorised in to positive negative and neutral comments and are visually represented through a pie chart.

Chapter 1

Introduction

Twitter is a microblogging web platform which helps people to stay connected through frequent short messages or entries on a blog.This platforms where millions of messages or data are being stored are used by researchers for their machine learning,natural language processing studies which requires a large data set.Twitter being a public platform,it creates approximately one billion user generated content accross the globe[1].People expresses their views precisely in a few words as the tweets size is limited.Data researchers have started using Twitter for data collection due is its usability and preciseness. In this study,Twitter is used to extract large amount of data,tweets which are now trending is used for analysis of the sentiments or public emotion of people .Twitter provides an API which allows to extract its data by granting access to tokens uniquely provided to users.But the amount of extraction of data through Twitter has some limitation .Extraction of older tweets are not possible through twitter API which poses some problem to the study where older data is to retrieved through Twitter.Web scraping technology bypassses the limitation by allowing access to older tweets in large number[2].The extracted tweets undergoes the classification based on the words used in them as positive,negative and neutral comments.

1.1 Project Category

The area of project named SENTIMENT ANALYSIS OF TRENDING HASHTAGS IN TWITTER USING WEB SCRAPING is data mining and falls under the category of research.

1.2 project profile

1.2.1 Tools/platforms

Hardware requirements

- processor-intel i3 or above
- Hard disk-320 GB
- RAM-4GB

Software requirements

- OS:windows 8 or above
- Language :Python 3.7
- Database:SQLyog Database
- Python intrepreter: Pycharm framework

Chapter 2

Problem analysis and Scope of the System

2.1 Problem Definition

The proposed study includes sentiment analysis which is a part of natural language processing. Sentiment analysis is a process of finding the sentiments of a text based on the words used in it. A text is classified into positive negative and neutral by checking the polarity score of each text. The texts or tweets are treated as negative positive neutral by checking the words used in it. Tweets are the main data set needed for the project.Data extraction is through web scraping.Web scraping technology is used to extract large amounts of data from websites in a structured format. The proposed study comes under the area of data mining or text mining and falls under the research category.

2.2 Need identification

People's opinions are important especially in a secular democratic country . So in a social platform where the people are allowed to express their public opinion about

the current burning public issues has to be analyzed to know the trend and mindset of the people. So this area where people fearlessly express their views has to be analyzed.Twitter has become an exceptional medium of communication among us that we like to express our views on any burning public issue through it. As everything happening around the globe is discussed in this platform, people try to express their stands through tweets .so that explains the relevance of analyzing people's opinion about this act and getting a clear picture of the stands of the people towards the current issues happening in and out of the country. This work will lead to overall average sentiments of people towards the trending hashtags related to Covid19,the racism happening in America ,Citizenship ammendment act in India etc.. The above points explain the need of the project and the social relevance of this work.

2.3 Problem Domain

The data set needed for the study are taken from Twitter.The tweets regarding the topic are scraped from the Twitter as soon as the user selects the hashtag.The tweets are then classified into positive negative and neutral using naive bayes classification and decision tree.And the result is displayed in a graphical pie chart.

2.4 Objectives

The main objective of this work is to retrieve tweets related to the topic in large numbers and do a sentiment analysis of each of the tweets.The data extraction is using web scraping. To know the sentiments of each tweet and classify it into neutral ,positive and negative comments. And to know the overall sentiment of the tweets in a visual representation.

2.5 Methodology

The methodologies adopted for the data mining applications are mostly based on the natural language processing techniques offered by programming language like Python.The data essential for the study is taken from Twitter through webscraping techniques.The tweets are classifies using naive bayes and decision tree classifiers.

2.6 Scope

This system analyzes the tweets related to different topics like covid19 which are now trending in the internet. Twitter is a large platform where everything happening around us is discussed . Here in this project the limiting its scope to only the selected topic of discussion.Tweets are sometimes written in one's native language such as Hindi Tamil or Malayalam.Here we limit this study to only tweets written in English. Tweets can sometimes be sarcastic and can contain idioms. The system may not deliver expected output for sarcastic comments.

2.7 Motivation

Twitter being a microblogging platform is equally used by celebreties and common people to pen down their opinions. New trends with hashtags even lead to public campaign , marches and protests.This user generated data is used by several researchers to do their machine learning and natural language processing works.Moreover the public issue like caa nrc ,anti racist movement are more discussed online.This leads to the development of this project work.

Chapter 3

Analysis and Software Specification

3.1 Literature review

Internet penetration in India went from a lower percentage to a very significant percentage in recent years . 420 million users are online via their mobile phones. People highly depend upon the social media for news updates and business updates. Twitter appears as a public domain for people to write down their views. Much research on the usage of Twitter and analyzing the tweets during the elections has taken place in the past. Sentiment analysis has been a very important research topic for the past many years . It helps to know the public opinion on a particular product ,public issue etc. To know the public trends during election and predicting the outcome of the election by analyzing tweets has been a part of study by many researchers. A study by [1] shows the empirical significance of Twitter during the Indian elections in 2014 and 2019. It also did the quantitative estimations of the two major parties to predict the outcome.Mainly three collection strategy was used by the above study.They were candidate based collection,election day collection,hashtag based collection .When GST was first implemented in India,people wrote their stand on GST through tweets.Opinion mining on peoples view on GST was carried out

by [3].Thousands of tweets were taken through Twitter streaming API .Nathional research council(NRC) lexicons were used for opinion mining which expressed 8 emotions like joy,trust,anticipation,fear etc.Twitter data were taken for forecasting future marketing outcomes by [4].Collecting data for research and studies requires large amount of dataset.Twitter allows the extraction of public streams of information by granting access to unique set of tokens by which recent tweets of not more than a month old can be easily retrieved through Twitter API.Well known works [1][5] have used keywords for querying the Twitter API endpoint and retrieving formatted objects containing tweets.In this study the tweets from the beginning of this year is to be taken which is not possible through Twitter API.Limitation occurs when historical tweets like these are to be retrieved which are only available within a maximum range of three weeks by using Twitter API. A study [2] shows the restriction in using Twitter API for collecting tweets and how it can be bypassed using webscraping technique which allows querying chronological tweets without any restrictions.

3.2 Analysis and observation

Analysis of the tweets for the election prediction, natural disaster rescue ,and other calamities have been done by several researchers .From the analysis of the previous studies done in this area,most of the researchers used Twitter API to collect tweets for their work. One can collect tweets which are of a maximum of 3 weeks old with Twitter API by creating a developer account in twitter.But in this study the tweets which popped from the beginning of this year has to be taken. For collecting old tweets a study by [4] used a technology called web scraping. A web scraping technology allows one to extract data in large amounts.Some of the issues happened this year are heavily

discussed and communicated online . People express their viewpoints regarding this issue highly rely on social media platforms to express their stand towards different topics.

3.3 Feasibility study

Feasibility study is a preliminary study which investigates the resource requirements ,costs ,benefits of the proposed system .It takes into account various constraints within which the system is to be operated and implemented.The main objective of this study is to determine the operational,economical and technical feasibility .

3.3.1 Economic Feasibility

Economic feasibility attempts to weigh the cost of developing and implementing the new system. This is a web based application and requires no capital.

3.3.2 Operational Feasibility

The system will be useful only if it meets the required user specification and operates accordingly to the user needs.The study checks if the system will operate when deployed.The proposed system will operate without failure as it is a web based application and a person with basic operating knowledge can use the system.

3.3.3 Technical Feasibility

A number of issues have to be addressed while checking the system for its technical feasibility. Before commencing the proposed system we have to look for the availability of the resources needed for the project. This system is technically feasible since

all the required tools are easily available as open source free tools. The framework Pycharm is user friendly and easy to work with.

3.4 Functional and Non Functional Specifications

3.4.1 Functional requirements

Functional requirements of the system are features that must be included in the current system that will satisfy the business needs of the client or user. The proposed system should be able to collect Data. The proposed system should be able to analyze the tweets. The proposed system should be able to visually display the polarity of the tweets .

3.4.2 Non functional requirements

Nonfunctional requirement defines the quality attribute of a software system which represents a set of quality standards used to judge the operation of the system. They are essentially based on performance throughput and services. Based on these the nonfunctional requirements are as follows System is user friendly .It is accurate for non sarcastic comments and has a large Data set availability for Research oriented studies.

3.5 Technical Requirements

The hardware and software requirements needed for the study are discussed in this section. Hardware requirements

- processor-intel i3 or above
- Hard disk-320 GB

- RAM-4GB Software requirements
- OS:windows 8 or above Windows 8 is a personal computer operating system that was produced by Microsoft as part of the Windows NT family of operating systems. The operating system was released to manufacturing on August 1, 2012, with general availability on October 26, 2012
- Language :Python 3.7 Python is a high level programming language created by Guido Vann Roussum And its first release was on 1991.Python provides several builtin modules and libraries which helps in the study .Especially it has libraries which helps in Data mining .
- Front End: Pycharm Framework Pycharm is a famous framework used for doing projects written in Python. PyCharm is an integrated development environment used in computer programming, specifically for the Python language. It is developed by the Czech company JetBrains. It has 2 versions for download . A community version and a Free version.
- Database:SQLyog Database used to store the data is SQLyog. SQLyog is a GUI tool for the RDBMS MySQL. It is developed by Webyog, Inc., based in Bangalore, India, and Santa Clara, California.

Chapter 4

Design

Software Design is a mechanism to convert user requirements to some implementable form which can be delivered. It deals with representing the client's requirement, as described in SRS (Software Requirement Specification) document, which is made using programming language. The software design phase is the first step in SDLC (Software Design Life Cycle), which moves the concentration from the problem domain to the solution domain. In software design, we consider the system to be a set of components or modules having clearly defined or preset behaviors and boundaries.

4.1 Modules of the system

4.1.1 Data collection

Data collection includes retrieving of tweets related to the topic.Data is collected through web scraping techniques. The tweets are filtered using hashtags and are collected from specific dates onwards. The number of tweets are also given to limit the data set.The user can select from a list of hashtags related to the different topics and the tweets related to the selected tags are scraped from the Twitter,which requires a live internet connection. Web scraping techniques allows you to collect more old

tweets from twitter allowing one to make a larger Data set which otherwise is not possible using twitter API method.

4.1.2 Tweet management

This module includes collecting or retrieving thousands of tweets related to the topic through web scraping technology. The structured data obtained from twitter is stored in a database.The scraped tweets contain several attributes containing userid,username,text,text html,video url ,number of likes,retweets,image url etc. Some of these attributes of the scraped tweets are irrelevant for the study. The attributes which are essential for our study are taken exclusively for sentiment analysis.

4.1.3 Data analysis

The collected data is analysed to know its sentiments.The tweets collected from Twitter contains many attributes like user id ,username,text,replyto,imageurl, etc .Some of these attributes are irrelevant for the study.They are stored in a Tweets table temporarily in a database.Only the relevant attributes like username userid text are retrieved by the user.The tweets also contains words which has to be removed. They are stop words,punctuations etc .All the characters, words and phrases that are less significant or carry less weightage in sentiment analysis are removed.The analysed texts are classified into positive negative and neutral comments on the basis of polarity score of each text. The texts having a score greater than zero is considered to be a positive comment. The texts of polarity score less than zero are treated as negative and texts of score zero as neutral comment.

4.1.4 Visualisation

The data collected from twitter through web scraping technology is made to analyse to know its sentiments to categorise it into positive,negative and neutral comments.Each person who login to the system is allowed to select from a drop down list of options in trending hashtags .The tweets related to the selected hashtags are fetched from the Twitter through web scraping.And view results option for all the category of tweets gives the visual representation of the sentiments in pie chart.Visualisation of tweets pictures the overall opinion of the people.The tweets classified according to the sentiments are visualised through a visualisation tool.

4.2 Data Flow Diagrams

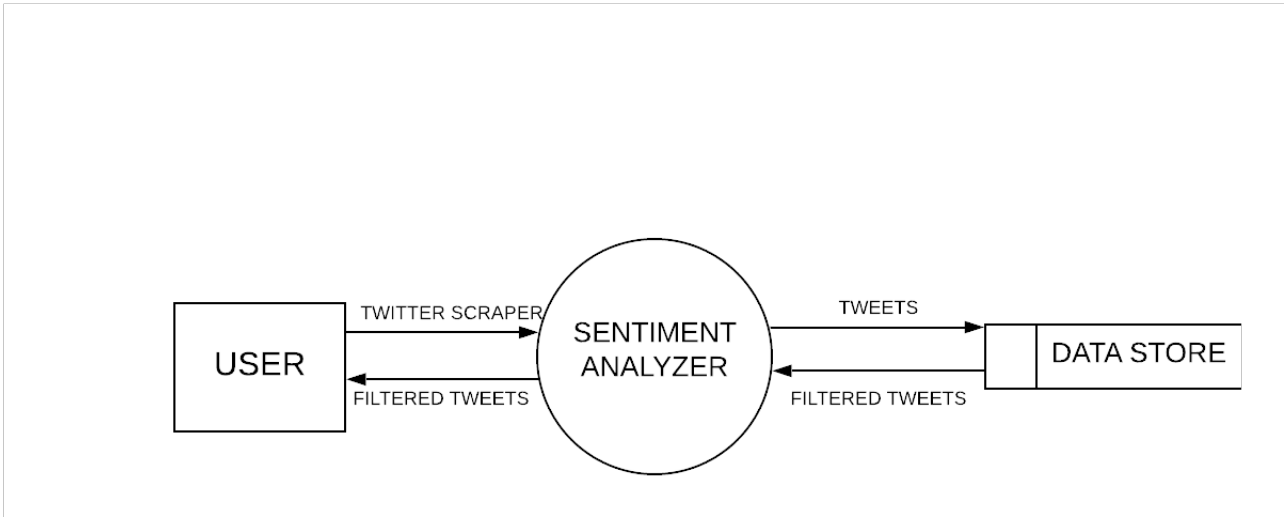


Figure 4.1: Level 0

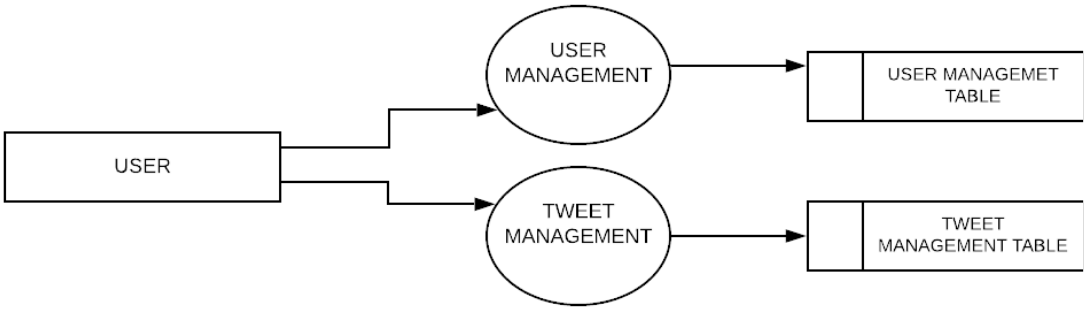


Figure 4.2: Level 1

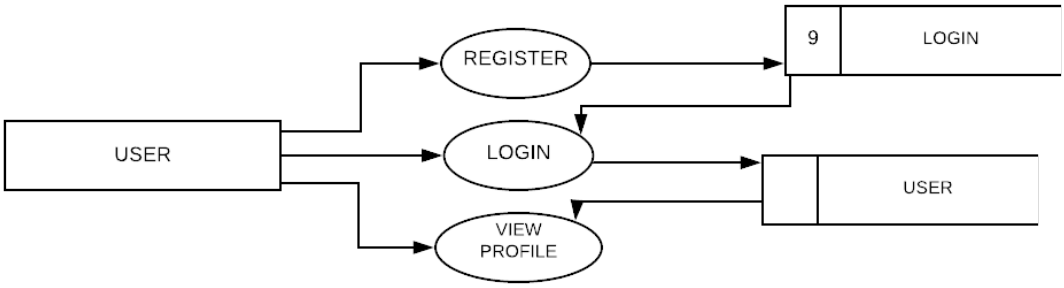


Figure 4.3: Level 1(User management)

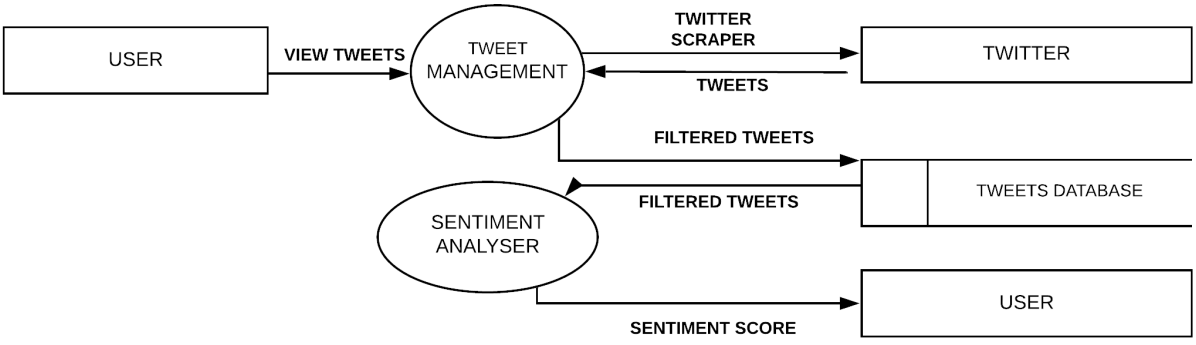


Figure 4.4: Level 2

4.3 User interface

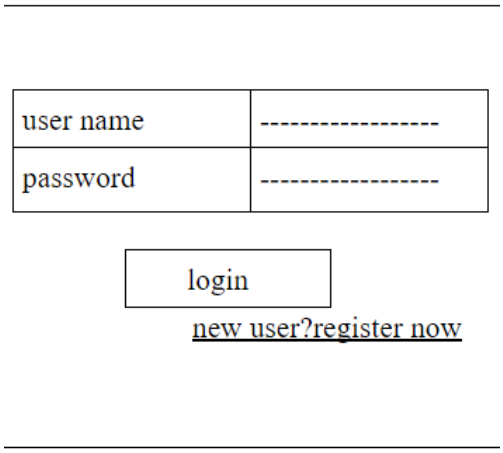


Figure 4.5: Login page

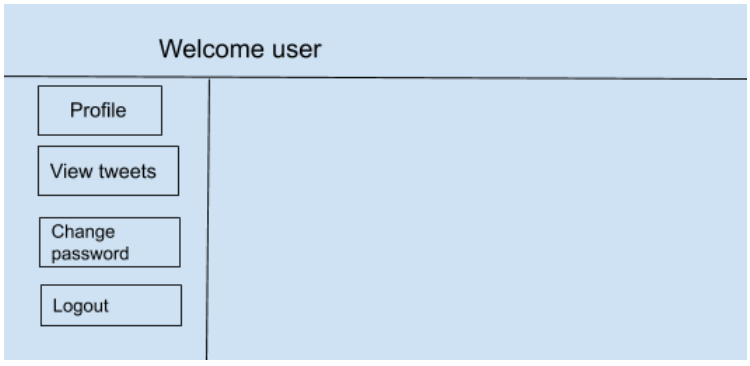


Figure 4.6: Home page

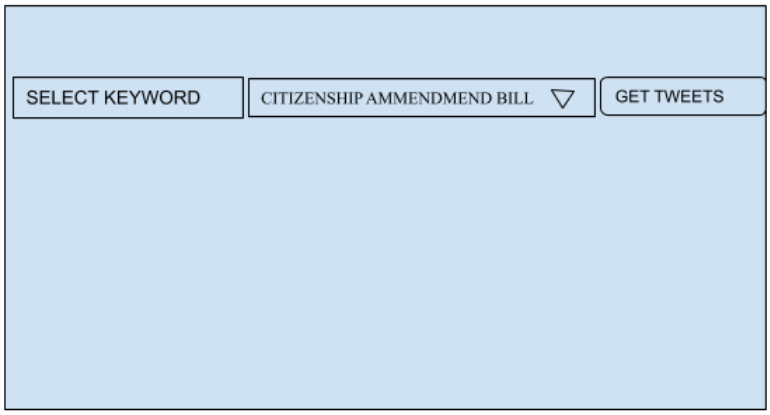


Figure 4.7: Get Tweets

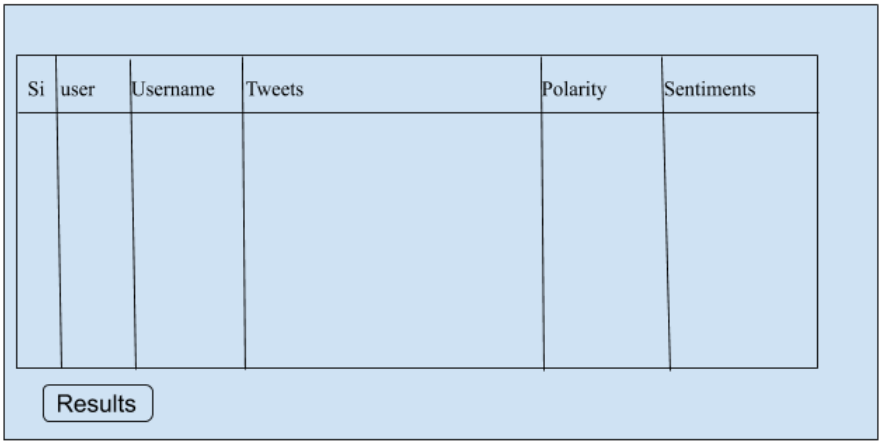


Figure 4.8: Result page

4.4 Database and Design

Database design run in parallel with the application design. As information is collected it has to be stored and retrieved.So in building a new system ,The structure and contents of the database are relevant.Database design is the organization of data according to a database model and data stored in a database is related to one other.

4.4.1 List of Entities and Attributes

4.4.2 Structure of Tables

Tables	Attributes
LoginTable	Name Loginid Username password usertype
Sign up Table	Signid Name age Dob Gender Email District Phone Locality Password Pincode Phone
Twitter Table	Index Screename Username Userid Tweetid Tweeturl Timestamp Imestzmpepoch Text Texthtml links Hashtags Hasmedia imgurls vediourl likes Retweets Replies isreplyto parenttweetid Replytousers sentiments

Table 4.1: List of entities and attributes.

Name	Type	Null	Default	Extra
Loginid	int	no	no	Auto increment
username	varchar	no	no	
password	int	no	no	
usertype	varchar	no	no	

Table 4.2: Login Table.

Name	Type	Null	Default	Extra
signid	int	no	no	Auto increment
name	varchar	no	no	
age	int	no	no	
dob	date	no	no	
gender	varchar	no	no	
email	varchar	no	no	
phone	bigint	no	no	
District	varchar	no	no	
Locality	varchar	no	no	
pincode	int	no	no	
loginid	int	no	no	

Table 4.3: Signup table.

Name	Type	Null	Default	Extra
index	int	no	no	Auto increment
screename	varchar	no	no	
username	varchar	no	no	
userid	varchar	no	no	
tweetid	varchar	no	no	
tweeturl	varchar	no	no	
timestamp	varchar	no	no	
imestamp	varchar	no	no	
text	varchar	no	no	
texthtml	varchar	no	no	
links	varchar	no	no	
hashtags	varchar	no	no	
hasmedia	varchar	no	no	
imgurls	varchar	no	no	
vediourl	varchar	no	no	
likes	varchar	no	no	
retweets	varchar	no	no	
replies	varchar	no	no	
isreplyto	varchar	no	no	
parentweetid	varchar	no	no	
geolocation	varchar	no	no	
sentiments	varchar	no	no	

Table 4.4: Twitter Table.

Chapter 5

Implementation

The implementation phase is the process of converting a system specification into an executable system. It yields the lowest-level system elements in the system hierarchy.This phase is the realization of the design as a program where the software system is developed as an executable form.

5.1 Tools/scripts for implementation

- python 3.8.1

Python is an interpreted, high-level, general-purpose programming language.Python is meant to be an easily readable language.python is having a simple easy to learn syntax.Python also supports modules and packages therefore nehances modularity and code reuse. Its formatting is visually uncluttered, and it often uses English keywords where other languages use punctuation. Unlike many other languages, it does not use curly brackets to delimit blocks, and semicolons after statements are optional.Python intrepreters and standard libraries are available freely.

- SQLyog

SQLyog is a GUI tool for the RDBMS MySQL. It is developed by Webyog, Inc., based in Bangalore, India, and Santa Clara, California.Session restore,Query builder,Autocomplete and SQL fomatting are some of its stunning features.SQLyog is the most powerful manager, admin and GUI tool for MySQL, combining the features of MySQL Query Browser, Administrator, phpMyAdmin and other MySQL Front Ends and MySQL GUI tools in a single intuitive interface. SQLyog is a fast, easy to use and compact graphical tool for managing MySQL databases.Backup is made easy with compressed backup and scheduling option.Complex Queries can be typed easily with Query builder option.Session restore automatically saves all the previous functions.

5.2 Coding

Only core sections of the source code are discussed in this section

5.2.1 Extracting Tweets

```
from twitterscraper import query_tweets
import datetime as dt
import pandas as pd
from DBConnection import Db
db=Db()
begin_date = dt.date(2020, 1, 10)
end_date = dt.date(2020, 2, 29)

limit = 500
lang = 'english'
```

```
class scratch:
    def scr(self,keyword):
        qrydel = "delete from tweets"
        db.delete(qrydel)

        # tweets = query_tweets("delhi riots", limit=limit,
        begindate=begin_date, enddate=end_date,  lang=lang)
        tweets = query_tweets(keyword, limit=limit, begindate=begin_date
         // enddate=end_date, lang=lang)

        df = pd.DataFrame(t.__dict__ for t in tweets)

        k = 1

        for i in df.values:
            screen_name = (i[0])
            if screen_name is None:
                screen_name=""
            print("1",screen_name)

               username = (i[1])
              if username is None:
                username=""
            print("2",username)

             user_id = (i[2])
            if user_id is None:
                user_id=""
            print("3",user_id)
```

```
 tweet_id = (i[3])
if tweet_id is None:
    tweet_id=""
print("4",tweet_id)

  tweet_url = (i[4])
if tweet_url is None:
    tweet_url=""
print("5",tweet_url)

 timestamp = (i[5])
if timestamp is None:
    timestamp=""
print("6",timestamp)

imestamp_epoch = (i[6])
if imestamp_epoch is None:
    imestamp_epoch=""
print("7",imestamp_epoch)

text = (i[7])
if text is None:
    text=""
print("8", text)

text_html = (i[8])
if text_html is None:
    text_html=""
```

```
print("9", text_html)

links = (i[9])
if links is None:
    links=""
links = ""
print("10", links)

hashtags = i[10]
if hashtags is None:
    hashtags=""
hashtags=""
print("11",hashtags )

has_media = (i[11])
if has_media is None:
    has_media=""
# print("12", has_media)

img_urls = (i[12])
if img_urls is None:
    img_urls=""
img_url = ""
print("13", img_urls)

vedio_url = (i[13])
if vedio_url is None:
    vedio_url=""
vedio_url = ""
```

```
print("14", vedio_url)

likes = (i[14])
if likes is None:
    likes=""
print("15", likes)

retweets = (i[15])
if retweets is None:
    retweets=""
print("16", retweets)

replies = (i[16])
if replies is None:
    replies=""
print("17", replies)

is_replied = (i[17])
if is_replied is None:
    is_replied=""
print("18", is_replied)

is_reply_to = (i[18])
if is_reply_to is None:
    is_reply_to=""
print("19", is_reply_to)

parent_tweet_id = (i[19])
```

```
            if parent_tweet_id is None:
                parent_tweet_id=""
            print("20", parent_tweet_id)

            reply_to_users = (i[20])
            if reply_to_users is None:
                reply_to_users=""
            reply_to_users=""
            print("21", reply_to_users)

            print(type(screen_name))

 text_html=text_html.replace
 ('<p class="TweetTextSize js-tweet-text tweet-text"
 data-aria-label-part="0" lang="en">',"")
  text_html=text_html.replace('</p>',"")
 text_html=text_html.replace("'","")
 text=text.replace("'","")
 text=text.replace("/","")
 username=username.replace("'","")
text_html=""


qry ="insert into  tweets  values
(null,'"+screen_name+"','"+username+"',
'"+str(user_id)+"','"+str(tweet_id)+"',
'"+str(tweet_url)+"','"+str(timestamp)+"',
'"+str(imestamp_epoch)+"','"+str(text)+"',
'"+str(text_html)+"','"+str(links)+"',
```

```
'"+str(hashtags)+"','"+str(has_media)+"'
,'"+""+"','"+""+"','"+str(likes)+"'
,'"+str(retweets)+"','"+str(replies)+"',
'"+str(is_replied)+"','"+""+"','"+str(parent_tweet_id)+"'
,'00','00','00')"

            print(qry)
            db.insert(qry)
```

5.2.2 Classifiying Tweets

```
import nltk

from textblob.compat import basestring
from textblob.decorators import cached_property
from textblob.exceptions import FormatError
from textblob.tokenizers import word_tokenize
from textblob.utils import strip_punc, is_filelike
import textblob.formats as formats

### Basic feature extractors ###


def _get_words_from_dataset(dataset):
    def tokenize(words):
        if isinstance(words, basestring):
            return word_tokenize(words, include_punc=False)
        else:
```

```
            return words
    all_words = chain.from_iterable
    (tokenize(words) for words, _ in dataset)
    return set(all_words)

def _get_document_tokens(document):
    if isinstance(document, basestring):
        tokens = set((strip_punc(w, all=False)
                    for w in word_tokenize
                    (document, include_punc=False)))
    else:
        tokens = set(strip_punc(w, all=False) for w in document)
    return tokens

def basic_extractor(document, train_set):
    try:
        el_zero = next(iter(train_set))
         # Infer input from first element.
    except StopIteration:
        return {}
    if isinstance(el_zero, basestring):
        word_features = [w for w in chain([el_zero], train_set)]
    else:
        try:
            assert(isinstance(el_zero[0], basestring))
            word_features =
            _get_words_from_dataset(chain([el_zero], train_set))
        except Exception:
            raise ValueError('train_set is probably malformed.')
```

```
    tokens = _get_document_tokens(document)
features = dict(((u'contains({0})
'.format(word), (word in tokens))
 for word in word_features))
    return features


def contains_extractor(document):
    tokens = _get_document_tokens(document)
features = dict((u'contains({0}
)'.format(w), True) for w in tokens)
    return features


class BaseClassifier(object):

    def __init__(self, train_set,
    feature_extractor=basic_extractor, format=None, **kwargs):
        self.format_kwargs = kwargs
        self.feature_extractor = feature_extractor
        if is_filelike(train_set):
            self.train_set = self._read_data(train_set, format)
        else: # train_set is a list of tuples
            self.train_set = train_set
        self._word_set = _get_words_from_dataset
        (self.train_set) # Keep a hidden set of unique words.
        self.train_features = None
```

```
    def _read_data(self, dataset, format=None):
        if not format:
            format_class = formats.detect(dataset)
            if not format_class:
                raise FormatError
('Could not automatically detect format for the given '
  'data source.')
        else:
            registry = formats.get_registry()
            if format not in registry.keys():
                raise ValueError("'{0}'
format not supported.".format(format))
            format_class = registry[format]
        return format_class(dataset,
         **self.format_kwargs).to_iterable()


class NLTKClassifier(BaseClassifier):

        class MyClassifier(NLTKClassifier):
 nltk_class = nltk.classify.svm.SvmClassifier
nltk_class = None

    def __init__(self, train_set,
                 feature_extractor=
                 basic_extractor, format=None, **kwargs):
        super(NLTKClassifier,
        self).__init__(train_set, feature_extractor, format, **kwargs)
        self.train_features =
```

```
        [(self.extract_features(d), c) for d, c in self.train_set]

    def __repr__(self):
        class_name = self.__class__.__name__
        return "<{cls}
trained on {n} instances>".format
         (cls=class_name,
           n=len(self.train_set))



    def labels(self):
        return self.classifier.labels()

    def classify(self, text):

        text_features = self.extract_features(text)
        return self.classifier.classify(text_features)

    def accuracy(self, test_set, format=None):

        if is_filelike(test_set):
            test_data = self._read_data(test_set, format)
        else:  # test_set is a list of tuples
            test_data = test_set
        test_features =
        [(self.extract_features(d), c) for d, c in test_data]
        return nltk.classify.accuracy(self.classifier, test_features)

    def update(self, new_data, *args, **kwargs):
```

```
        self.train_set += new_data
        self._word_set.update(_get_words_from_dataset(new_data))
        self.train_features = [(self.extract_features(d), c)
                               for d, c in self.train_set]
        try:
            self.classifier = self.nltk_class.train(self.train_features,
                                                    *args, **kwargs)
        except AttributeError:  # Descendant has not defined nltk_class
            raise ValueError("NLTKClassifier must have a nltk_class"
                             " variable that is not None.")
        return True


class NaiveBayesClassifier(NLTKClassifier):

    nltk_class = nltk.classify.NaiveBayesClassifier

    def prob_classify(self, text):
        text_features = self.extract_features(text)
        return self.classifier.prob_classify(text_features)

    def informative_features(self, *args, **kwargs):
        return self.classifier.most_informative_features(*args, **kwargs

    def show_informative_features(self, *args, **kwargs):
        return self.classifier.show_most_informative_features(*args, **k


class DecisionTreeClassifier(NLTKClassifier):
```

```
    nltk_class = nltk.classify.decisiontree.DecisionTreeClassifier

    def pretty_format(self, *args, **kwargs):
        return self.classifier.pretty_format(*args, **kwargs)

    # Backwards-compat
    pprint = pretty_format

    def pseudocode(self, *args, **kwargs):
        return self.classifier.pseudocode(*args, **kwargs)


class PositiveNaiveBayesClassifier(NLTKClassifier):

    nltk_class = nltk.classify.PositiveNaiveBayesClassifier

    def __init__(self, positive_set, unlabeled_set,
                 feature_extractor=contains_extractor,
                 positive_prob_prior=0.5, **kwargs):
        self.feature_extractor = feature_extractor
        self.positive_set = positive_set
        self.unlabeled_set = unlabeled_set
        self.positive_features = [self.extract_features(d)
                                  for d in self.positive_set]
        self.unlabeled_features = [self.extract_features(d)
                                  for d in self.unlabeled_set]
        self.positive_prob_prior = positive_prob_prior
```

```
    def __repr__(self):
        class_name = self.__class__.__name__
        return "<{cls} trained on {n_pos} labeled and {n_unlabeled}
unlabeled instances>"\
                        .format(cls=class_name, n_pos=len(self.positive_
                                n_unlabeled=len(self.unlabeled_set))

    # Override
    def train(self, *args, **kwargs):
        self.classifier = self.nltk_class.train(self.positive_features,
                                                self.unlabeled_features,
                                                self.positive_prob_prior
        return self.classifier

    def update(self, new_positive_data=None,
               new_unlabeled_data=None, positive_prob_prior=0.5,
               *args, **kwargs):
        self.positive_prob_prior = positive_prob_prior
        if new_positive_data:
            self.positive_set += new_positive_data
            self.positive_features += [self.extract_features(d)
                                       for d in new_positive_data]
        if new_unlabeled_data:
            self.unlabeled_set += new_unlabeled_data
            self.unlabeled_features += [self.extract_features(d)
                                        for d in new_unlabeled_data]
        self.classifier = self.nltk_class.train(self.positive_features,
                                                self.unlabeled_features,
                                                self.positive_prob_prior
```

```
                                                    *args, **kwargs)
        return True


class MaxEntClassifier(NLTKClassifier):
    __doc__ = nltk.classify.maxent.MaxentClassifier.__doc__
    nltk_class = nltk.classify.maxent.MaxentClassifier

    def prob_classify(self, text):
        feats = self.extract_features(text)
        return self.classifier.prob_classify(feats)
```

5.2.3 Displaying Tweets with Sentiments

```
@app.route("/viewtweetspost",methods=["post"])
def viewtweetspost():

    db=Db()
    qrydel = "delete from tweets"
    db.delete(qrydel)




    keyw=request.form["List"]
    obj=scratch()
    obj.scr(keyw)
    print("keyword=",keyw)
```

```
qryt= "select * from tweets"
res = db.select(qryt)


ress=[]

po=0
ne=0
ntr=0



for i in res:
    # print("gg=",i)

    ty = i['text']

    polarity = TextBlob(ty).sentiment.polarity
    scr=polarity
    if scr > 0:
        po = po + 1
    elif scr == 0:
        ntr = ntr + 1
    else:
        ne = ne + 1


    # print("Polarity = ", polarity)
    ress.append(polarity)
```

```
        return render_template('viewtweets.html',res=res,
          pol=ress,po=po,ntr=ntr,ne=ne)
ert(qry)
```

5.3 Important screenshots

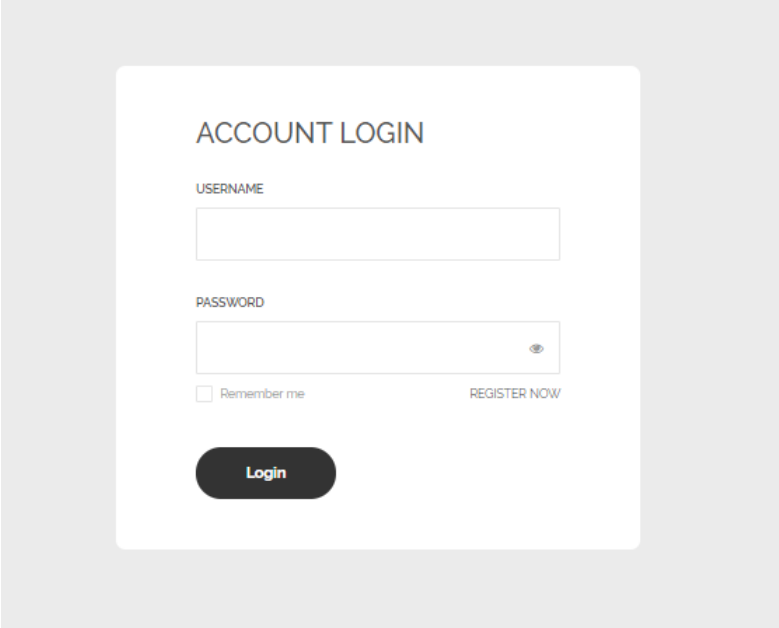


Figure 5.1: Login page

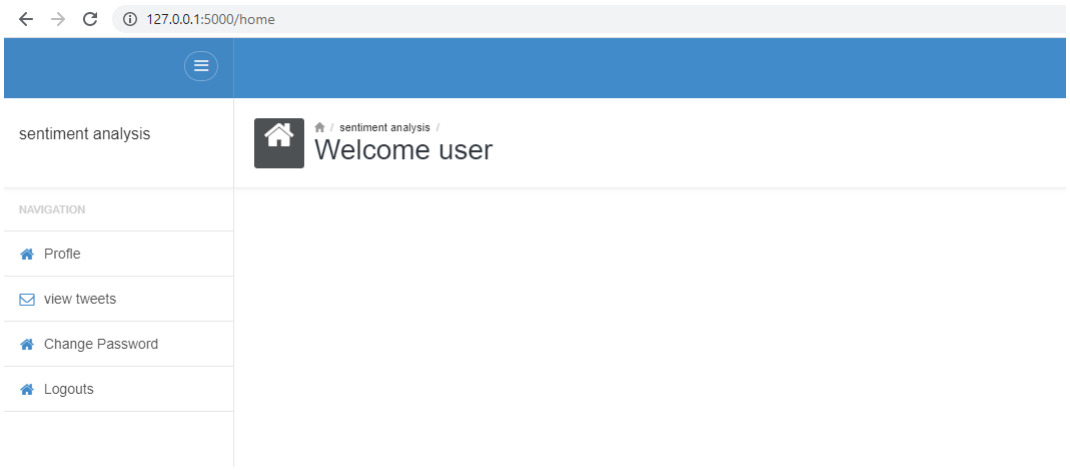


Figure 5.2: Home page

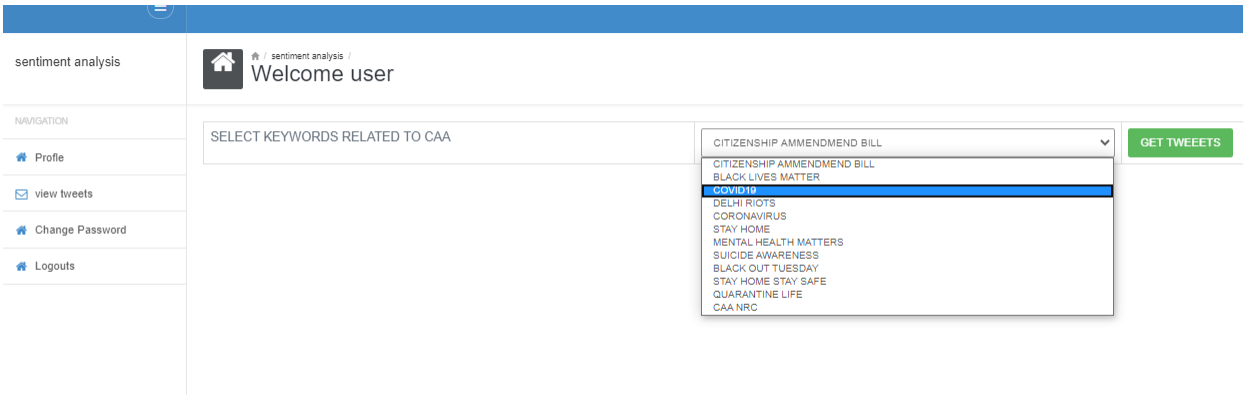
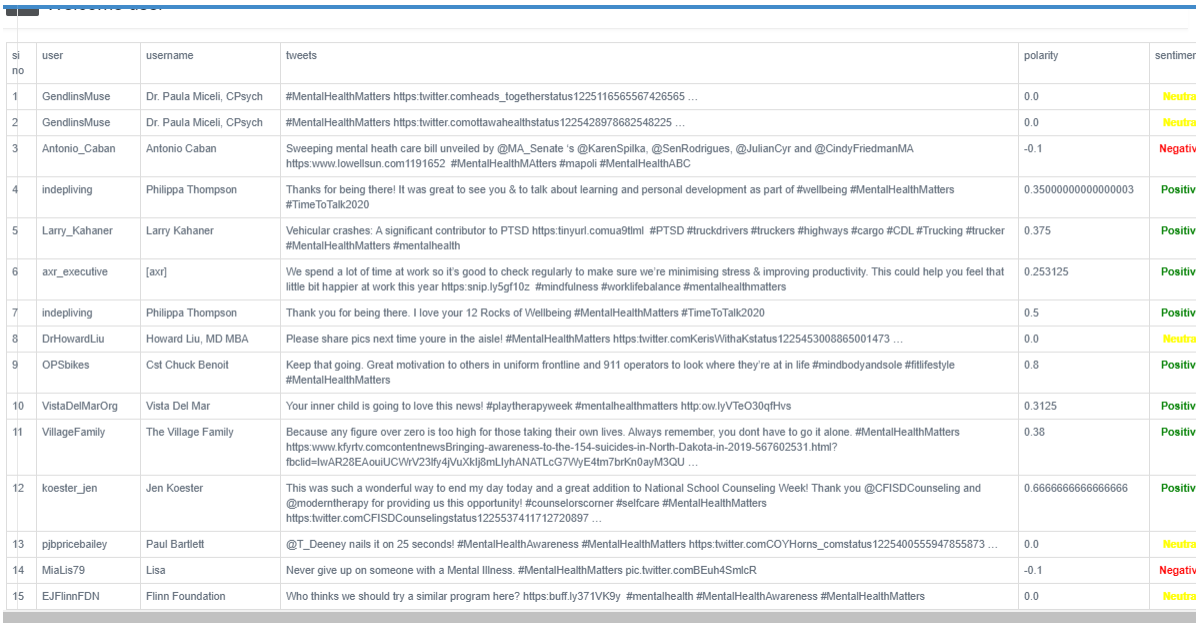
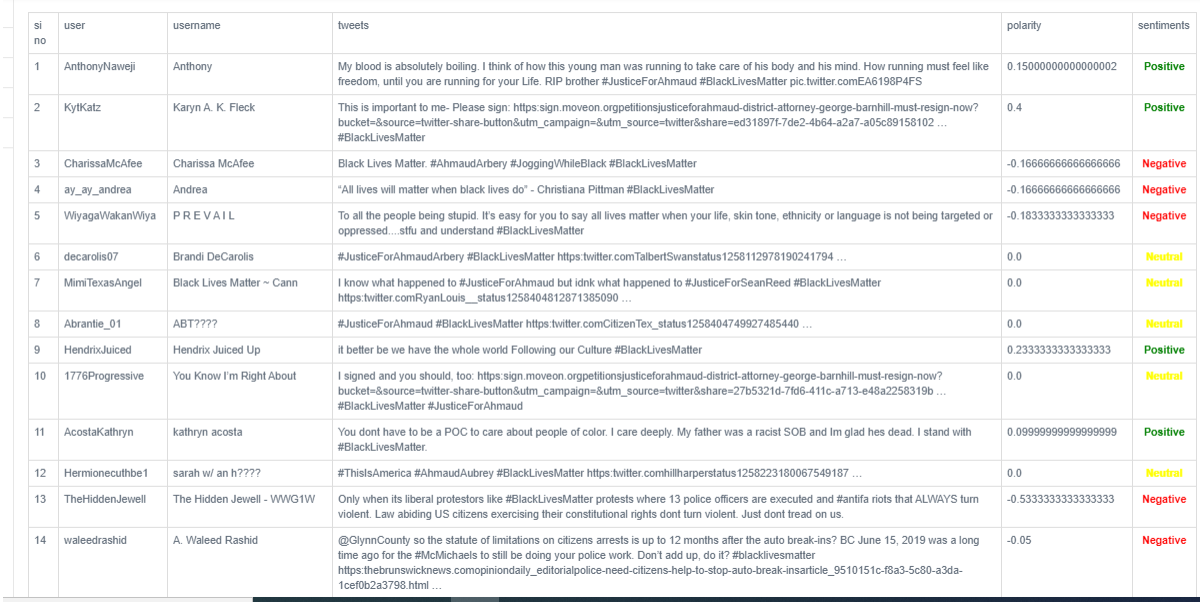


Figure 5.3: Tweet select page



si no	user	username	tweets	polarity	sentimer
1	GendlinsMuse	Dr. Paula Miceli, CPsych	#MentalHealthMatters https:twitter.comheads_togetherstatus1225116565567426565 …	0.0	Neutra
2	GendlinsMuse	Dr. Paula Miceli, CPsych	#MentalHealthMatters https:twitter.comottawahealthstatus1225428978682548225 …	0.0	Neutra
3	Antonio_Caban	Antonio Caban	Sweeping mental heath care bill unveiled by @MA_Senate 's @KarenSpilka, @SenRodrigues, @JulianCyr and @CindyFriedmanMA https:www.lowellsun.com1191652 #MentalHealthMAtters #mapoli #MentalHealthABC	-0.1	Negativ
4	indepliving	Philippa Thompson	Thanks for being there! It was great to see you & to talk about learning and personal development as part of #wellbeing #MentalHealthMatters #TimeToTalk2020	0.35000000000000003	Positiv
5	Larry_Kahaner	Larry Kahaner	Vehicular crashes: A significant contributor to PTSD https:tinyurl.comua9tlml #PTSD #truckdrivers #truckers #highways #cargo #CDL #Trucking #trucker #MentalHealthMatters #mentalhealth	0.375	Positiv
6	axr_executive	[axr]	We spend a lot of time at work so it's good to check regularly to make sure we're minimising stress & improving productivity. This could help you feel that little bit happier at work this year https:snip.ly5gf10z #mindfulness #worklifebalance #mentalhealthmatters	0.253125	Positiv
7	indepliving	Philippa Thompson	Thank you for being there. I love your 12 Rocks of Wellbeing #MentalHealthMatters #TimeToTalk2020	0.5	Positiv
8	DrHowardLiu	Howard Liu, MD MBA	Please share pics next time youre in the aisle! #MentalHealthMatters https:twitter.comKerisWithaKstatus1225453008865001473 …	0.0	Neutra
9	OPSbikes	Cst Chuck Benoit	Keep that going. Great motivation to others in uniform frontline and 911 operators to look where they're at in life #mindbodyandsole #fitlifestyle #MentalHealthMatters	0.8	Positiv
10	VistaDelMarOrg	Vista Del Mar	Your inner child is going to love this news! #playtherapyweek #mentalhealthmatters http:ow.lyVTeO30qfHvs	0.3125	Positiv
11	VillageFamily	The Village Family	Because any figure over zero is too high for those taking their own lives. Always remember, you dont have to go it alone. #MentalHealthMatters https:www.kfyrtv.comcontentnewsBringing-awareness-to-the-154-suicides-in-North-Dakota-in-2019-567602531.html?fbclid=IwAR28EAouiUCWrV23lfy4jVuXkIj8mLIyhANATLcG7WyE4tm7brKn0ayM3QU …	0.38	Positiv
12	koester_jen	Jen Koester	This was such a wonderful way to end my day today and a great addition to National School Counseling Week! Thank you @CFISDCounseling and @moderntherapy for providing us this opportunity! #counselorscorner #selfcare #MentalHealthMatters https:twitter.comCFISDCounselingstatus1225537411712720897 …	0.6666666666666666	Positiv
13	pjbpricebailey	Paul Bartlett	@T_Deeney nails it on 25 seconds! #MentalHealthAwareness #MentalHealthMatters https:twitter.comCOYHorns_comstatus1225400555947855873 …	0.0	Neutra
14	MiaLis79	Lisa	Never give up on someone with a Mental Illness. #MentalHealthMatters pic.twitter.comBEuh4SmlcR	-0.1	Negativ
15	EJFlinnFDN	Flinn Foundation	Who thinks we should try a similar program here? https:buff.ly371VK9y #mentalhealth #MentalHealthAwareness #MentalHealthMatters	0.0	Neutra

Figure 5.4: Result page



si no	user	username	tweets	polarity	sentiments
1	AnthonyNaweji	Anthony	My blood is absolutely boiling. I think of how this young man was running to take care of his body and his mind. How running must feel like freedom, until you are running for your Life. RIP brother #JusticeForAhmaud #BlackLivesMatter pic.twitter.comEA6198P4FS	0.15000000000000002	Positive
2	KytKatz	Karyn A. K. Fleck	This is important to me- Please sign: https:sign.moveon.orgpetitionsjusticeforahmaud-district-attorney-george-barnhill-must-resign-now?bucket=&source=twitter-share-button&utm_campaign=&utm_source=twitter&share=ed31897f-7de2-4b64-a2a7-a05c89158102 … #BlackLivesMatter	0.4	Positive
3	CharissaMcAfee	Charissa McAfee	Black Lives Matter. #AhmaudArbery #JoggingWhileBlack #BlackLivesMatter	-0.16666666666666666	Negative
4	ay_ay_andrea	Andrea	"All lives will matter when black lives do" - Christiana Pittman #BlackLivesMatter	-0.16666666666666666	Negative
5	WiyagaWakanWiya	P R E V A I L	To all the people being stupid. It's easy for you to say all lives matter when your life, skin tone, ethnicity or language is not being targeted or oppressed....stfu and understand #BlackLivesMatter	-0.1833333333333333	Negative
6	decarolis07	Brandi DeCarolis	#JusticeForAhmaudArbery #BlackLivesMatter https:twitter.comTalbertSwanstatus1258112978190241794 …	0.0	Neutral
7	MimiTexasAngel	Black Lives Matter ~ Cann	I know what happened to #JusticeForAhmaud but idnk what happened to #JusticeForSeanReed #BlackLivesMatter https:twitter.comRyanLouis__status1258404812871385090 …	0.0	Neutral
8	Abrantie_01	ABT????	#JusticeForAhmaud #BlackLivesMatter https:twitter.comCitizenTex_status1258404749927485440 …	0.0	Neutral
9	HendrixJuiced	Hendrix Juiced Up	it better be we have the whole world Following our Culture #BlackLivesMatter	0.2333333333333333	Positive
10	1776Progressive	You Know I'm Right About	I signed and you should, too: https:sign.moveon.orgpetitionsjusticeforahmaud-district-attorney-george-barnhill-must-resign-now?bucket=&source=twitter-share-button&utm_campaign=&utm_source=twitter&share=27b5321d-7fd6-411c-a713-e48a2258319b … #BlackLivesMatter #JusticeForAhmaud	0.0	Neutral
11	AcostaKathryn	kathryn acosta	You dont have to be a POC to care about people of color. I care deeply. My father was a racist SOB and Im glad hes dead. I stand with #BlackLivesMatter.	0.09999999999999999	Positive
12	Hermionecuthbe1	sarah w/ an h????	#ThisIsAmerica #AhmaudAubrey #BlackLivesMatter https:twitter.comhillharperstatus1258223180067549187 …	0.0	Neutral
13	TheHiddenJewell	The Hidden Jewell - WWG1W	Only when its liberal protestors like #BlackLivesMatter protests where 13 police officers are executed and #antifa riots that ALWAYS turn violent. Law abiding US citizens exercising their constitutional rights dont turn violent. Just dont tread on us.	-0.5333333333333333	Negative
14	waleedrashid	A. Waleed Rashid	@GlynnCounty so the statute of limitations on citizens arrests is up to 12 months after the auto break-ins? BC June 15, 2019 was a long time ago for the #McMichaels to still be doing your police work. Don't add up, do it? #blacklivesmatter https:thebrunswicknews.comopiniondaily_editorialpolice-need-citizens-help-to-stop-auto-break-insarticle_9510151c-f8a3-5c80-a3da-1cef0b2a3798.html …	-0.05	Negative

Figure 5.5: Result page

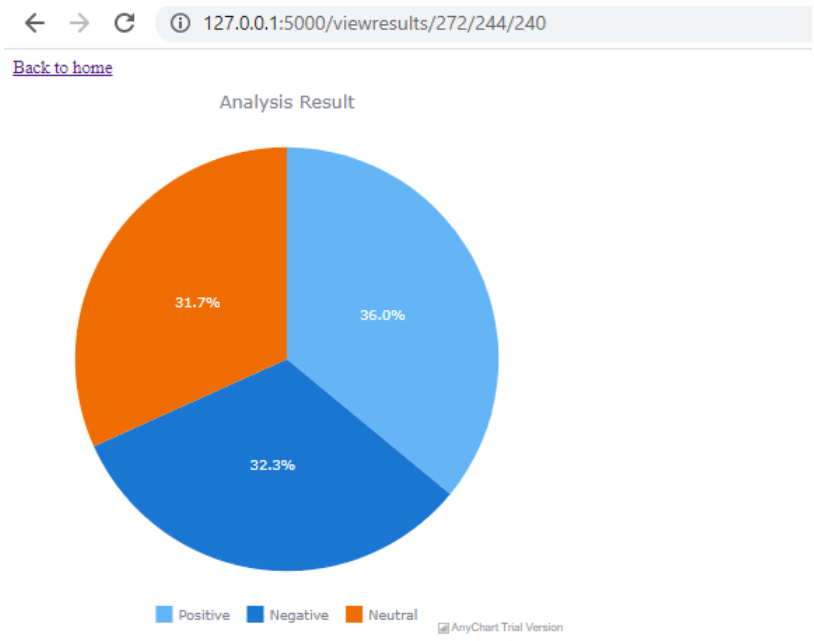


Figure 5.6: Result Graph

Chapter 6

Testing

6.1 Introduction

Software Testing is a critical process done to assure the quality of the software. It can be stated as the activity carried out to check whether the actual results match the expected results and to ensure that the software system is Defect free.Software testing aims at finding the missing requirements of the client's actual requirements and helps to identify errors.Software testing also ensures application is bug free and works effectively and efficiently with handling all the exceptional and boundary cases. The process of software testing aims not only at finding faults in the existing software but also at finding measures to improve the software in terms of efficiency, accuracy and usability.

6.2 Testing Methodologies adopted

6.2.1 Unit Testing

Unit testing focuses on verifying the individual units of a software.It validates the smallest module of the code which can be isolated.The local data structures are examined to ensure that the data stored temporarily maintains the integrity during

the execution.The Boundary conditions are tested to check whether the desired results are obtained.In SENTIMENT ANALYSIS OF CAA NRC TWEETS USING WEB SCRAPING modules are tested one by one.

Test case 01

- Test case ID: UT01
- Test Case Description: Trying to Login without entering Login details
- Expected Result: Not Able to Login

6.2.2 Integration Testing

Integration testing is a level of software testing where the individual units are combined and tested as a group.It helps to uncover the errors associated with the module s when integrated together.The major concerns of integration testing are developing an incremental strategy that will limit the complexity of entire actions among the components as they are added to the system.

Test case 01

- Test case ID: IT01
- Test Case Description: Details of the user.
- Test Strategy:Input Data Filled in Two forms.

  1.Complete Information Filled

  2.Incomplete Information with missing Field.

- Expected result:

1. If the Information filled by the user is complete the user is directed to login page where he /she can login with the username and password created.

2. If the information filled by the user is incomplete and have missing fields in between,A pop up message shows in the respective textbox where user accidentally forgot to fill the information.

6.2.3 System Testing

System testing the level of testing where the system as a whole is tested and evaluted.The main objective of this testing is to check the system's compliance with the specified requirements.

Test case 01

- Test case ID: ST01
- Test Case Description: Checks the Working of the system
- Test Strategy:Requirement specifications are compared with the working of the system.

  1.Partially meets the system requirements.

  2.Requirements fully met.

- Expected result:

  1.If the Requirements specified are fully met the The system works and delivers the output as desired.

  2.If the Requirements are partially fulfilled ,system may fail to deliver the output as needed.

Chapter 7

Conclusion

The project named SENTIMENT ANALYSIS OF TRENDING HASHTAGS IN TWITTER USING WEB SCRAPING mainly emphasize on extracting the tweets related to different topics which are now trending in the internet and finding its emotions through sentiment analysis.Sentiment analysis or opinion mining uses natural language processing techniques where user generated informations are taken as dataset.The tweets are fetched or scraped from Twitter when user selects the hashtag listed . Web Scraping technology is used to extract the tweets in large amount instead of the Twitter API method which does not allow one to collect tweets which are older than three weeks. The collected tweets are undergone a classification based on Naive bayes and decision tree to know the sentiments of each tweets. Polarity score of each Tweets are displayed to the user and are graphically represented through a pie chart.The system is developed with the insight of future modifiations which can be easily incoperated.Hence the maintanence of the sysetm would require less effort.

Bibliography

[1] A. S. Tulasi, K. Gupta, O. Gurjar, S. S. Buggana, P. Mehan, A. B. Buduru, and P. Kumaraguru, "Catching up with trends: The changing landscape of political discussions on twitter in 2014 and 2019," *ArXiv*, vol. abs/1909.07144, 2019.

[2] A. Hernandez-Suarez, G. Sanchez-Perez, K. Toscano-Medina, V. Martinez-Hernandez, V. Sanchez, and H. Perez-Meana, "A web scraping methodology for bypassing twitter api restrictions," *arXiv preprint arXiv:1803.09875*, 2018.

[3] D. Uniyal and A. Rai, "Citizens' emotion on gst: A spatio-temporal analysis over twitter data," *arXiv preprint arXiv:1906.08693*, 2019.

[4] S. Asur and B. A. Huberman, "Predicting the future with social media," in *2010 IEEE/WIC/ACM international conference on web intelligence and intelligent agent technology*, vol. 1. IEEE, 2010, pp. 492–499.

[5] B. O'Connor, R. Balasubramanyan, B. R. Routledge, and N. A. Smith, "From tweets to polls: Linking text sentiment to public opinion time series," in *Fourth international AAAI conference on weblogs and social media*, 2010.

[6] S. Ahmed, K. Jaidka, and J. Cho, "The 2014 indian elections on twitter: A comparison of campaign strategies of political parties," *Telematics Informatics*, vol. 33, pp. 1071–1087, 2016.

[7] M. Korakakis, E. Spyrou, and P. Mylonas, "A survey on political event analysis in twitter," *2017 12th International Workshop on Semantic and Social Media Adaptation and Personalization (SMAP)*, pp. 14–19, 2017.