A
Project Report
On

"Tweeks"
(Emotion & Sentimental Analysis)

Prepared by

Archan Bhalani(19DCE004)

Dhruv Maradiya(19DCE066)

Femil Mori(19DCE075)

Under the guidance of

Prof. Urvashi Chaudhari

Prof. Sanjay Patel

A Report Submitted to

Charotar University of Science and Technology

for Partial Fulfillment of the Requirements for the

6th Semester Software Group Project-I (CE359)

Submitted at

CE

DEPSTAR

At: Changa, Dist.: Anand – 388421

April 2022

CERTIFICATE

This is to certify that the report entitled "**Tweeks**" is a bonafied work carried out by **Femil Mori(19DCE075), Dhruv Maradiya(19DCE066), Archan Bhalani(19DCE004)** under the guidance and supervision of **Assistant Prof. Sanjay Patel and Prof. Urvashi Chaudhari** for the subject CE359- **Software Group Project-IV** (CE/CSE/IT) of 6th Semester of Bachelor of Technology in **DEPSTAR** at Faculty of Technology & Engineering – CHARUSAT, Gujarat.

To the best of my knowledge and belief, this work embodies the work of candidate himself, has duly been completed, and fulfills the requirement of the ordinance relating to the B.Tech. Degree of the University and is up to the standard in respect of content, presentation and language for being referred to the examiner.

Sanjay Patel
Teaching cum Research Assistant
Branch - CE
DEPSTAR, Changa, Gujarat.

Urvashi Chaudhari
Teaching cum Research Assistant
Branch - CE
DEPSTAR, Changa, Gujarat.

Dr. Amit Ganatra
Principal, DEPSTAR
Dean, FTE
CHARUSAT, Changa, Gujarat.

Dr Dweepna Garg
I/C HOD & Assistant Professor
Branch – CE
DEPSTAR, Changa, Gujarat.

ABSTRACT

Online social networks have emerged as new platform that provide an arena for people to share their views and perspectives on different issues and subjects with their friends, family, relatives, etc. We can share our thoughts, mental state, moments, stand on specific social, national, international issues through text, photos, audio and video messages and posts. Indeed, despite the availability of other forms of communication, text is still one of the most common ways of communication in a social network. The target of the work described in this paper is to detect and analyze sentiment and emotion expressed by people from text in their twitter posts and use them for generating recommendations. We collected tweets and replies on few specific topics and created a dataset with text, user, emotion, sentiment information, etc. We used the dataset to detect sentiment and emotion from tweets and their replies and measured the influence scores of users based on various user-based and tweet-based parameters. Finally, we used the latter information to generate generalized and personalized recommendations for users based on their twitter activity. The method we used in this paper includes some interesting novelties such as,

(i) including replies to tweets in the dataset and measurements.

(ii) introducing agreement score, sentiment score and emotion score of replies in influence score calculation.

(iii) generating general and personalized recommendation containing list of users who agreed on the same topic and expressed similar emotions and sentiments towards that particular topic.

ACKNOWLEDGEMENT

We thank **Prof. Urvashi Chaudhari and Prof. Sanjay Patel** who have been the great inspiration and who have provided sufficient background knowledge and understanding of this subject.

Our humble prostration goes to our H.O.D. Dweepna Garg, for the support during the whole session of study and development. It is because of them, that we were prompted to do hard work, adopting new technologies which have aided us to complete this project successfully.

we would also like to thank our mentor Prof. Minal Patel Sir for his guidelines throughout the development phase of the project.

They altogether provided me favorable environment, and without them it would not have been possible to achieve my goal.

Thanks,

FEMIL MORI
DHRUV MARADIYA
ARCHAN BHALANI

INDEX

CHAPTER - 1 INTRODUCTION

INTRODUCTION

Social media sentiment analysis has turn out to be a distinguished area of study and experimentation in current years..Twitter a micro-blogging site, has lion's share in social media info. Most research has been confined to classify tweets into positive,negative categories ignoring sarcasm.Human emotions are extremely diverse and cannot be restricted to certain metrics alone.Polarity analysis gives limited information on the actual intent of message delivered by author and just positive or negative classes are not sufficient to understand nuances of underlying tone of a sentence.This brings the need to take one step above sentiment analysis leading to emotion analysis. In this paper we throw light on methods we have used to derive sentiment analysis considering sarcasm and how we have accomplished emotion analysis of user opinions.

A supervised learning techniqueprovides labels to classifier to make it understand the insights among various features.Once the classifiergetsfamiliarized with train data it can perform classification on unseen test data.We have chosen Naive Bayes and Support Vector Machine classification algorithms to carry out sentiment and emotional analysis respectively.

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

What is Sentiment Analysis?

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

Sentiment analysis is contextual mining of text which identifies and extracts subjective information in source material, and helping a business to understand the social sentiment of their brand, product or service while monitoring online conversations. However, analysis of social media streams is usually restricted to just basic sentiment analysis and count based metrics. This is akin to just scratching the surface and missing out on those high value insights that are waiting to be discovered

What is Emotional Analysis?

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

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

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

PROBLEM STATEMENT

Generating statistical information regarding emotions, sentiments out of analysis of user's opinions from tweets, which can be used as an inference to understand how users feel thereby improving users experiences regarding. Despite the availability of software to extract data regarding a person's sentiment on a specific product or service, organizations and other data workers still face issues regarding the data extraction.

With the rapid growth of the World Wide Web, people are using social media such as Twitter which generates big volumes of opinion texts in the form of tweets which is available for the sentiment analysis. This translates to a huge volume of information from a human viewpoint which make it difficult to extract a sentence, read them, analyse tweet by tweet, summarize them and organize them into an understandable format in a timely manner

CHAPTER – 2 SYSTEM REQUIREMENTS

SYSTEM REQUIREMENTS

1. SYSTEM CONFIGURATION:

This project can run on commodity hardware.We ran entire project on an Intel I5 processor with 8 GB Ram , 2 GB Nvidia Graphic Processor , It also has 2 cores which runs at 1.7 GHz , 2.1 GHz respectively. First part of the project just takes very little amount of time that depends on the size of data set upon which classifier is working upon.Second part Emotional analysis takes some time around 5-10 mins to produce results because of its large volume data set.

2. Software Requirements:

Following are the software and modules that needs to be installed for successful execution of the project.They are:

1. Anaconda
2. Spyder
3. Jupiter Note Book
4. Nltk
5. Scikit-learn
6. Matplotlib
7. Tweepy
8. Pandas
9. Numpy
10. TextBlob
11. VaderSentiment
12. Csv

3. Hardware Requirements:

Following are the hardware requirements necessary for faster execution of the code.

1.A minimum of Intel Core I3 processor

2.A minimum of 4 GB Ram

3.Cpu with atleast 2 cores of clock speeds > 1.5GHz

CHAPTER – 3 METHODOLOGY

METHODOLOGY

The sentiment analysis of Twitter data is an emerging field that needs much more attention.We use Tweepy an API to stream live tweets from Twitter.User based on his interest chooses a keyword and tweets containing that keyword are collected and stored into a csv file.Then we make it a labeled dataset using textblob and setting the sentiment fields accordingly.Thus our train data set without preprocessing is ready.Next we perform preprocessing to clean,remove unwanted text,characters out of the tweets.

Then we train our classifier by fitting the train data to the classifier ,there after prediction of results over unseen test data set is made which there after provides us with the accuracy with which the classifier had predicted the outcomes.There after we present our results in a pictorial manner which is the best way to showcase results because of its easiness to understand information out of it.

PROPOSED SYSTEM

1) Extraction Of Data:

Tweets based on a keyword of user's choice of interest have been collected using a famous twitter API known as Tweepy and stored into a csv file.This data set collected for sentiment analysis have tweets based on a keyword e.g.,cybertruck. Tweets mimicking various emotionsas a dataset downloaded from kaggle is used for emotional analysis.Since both the machines are trained using supervised learning and work on different parameters different data sets have been considered.

In order to extract the opinion first of all data is selected and extracted from twitter in the form of tweets. After selecting the data set of the tweets, these tweets were cleaned from emoticons, unnecessary punctuation marks and a database was created to store this data in a specific transformed structure. In this structure, all the transformed tweets are in lowercase alphabets and are divided into different parts of tweets in the specific field. The details about the steps adopted for the transformation of information are described in next subsections

Fig 1 Extraction of Data

2) Preprocessing Of Data:

Following are the Preprocessing steps that have been carried out

3) Removing Html tags and urls:

Html tags and urls often have minimum sentiments thus they are removed from tweets.Using regular expressions.

4) Conversion to lowercase:

To maintain uniformity all the tweets are converted to lowercase .This will benefit to avert inconsistency in data.Python provides a function called lower() to convert sentences to lower case.

5) Tokenization:

Tokenization is the process of converting text into tokens before transforming it into vectors. It is also easier to filter out unnecessary tokens. For example, a document into paragraphs or sentences into words. In this case we are tokenising the reviews into words.

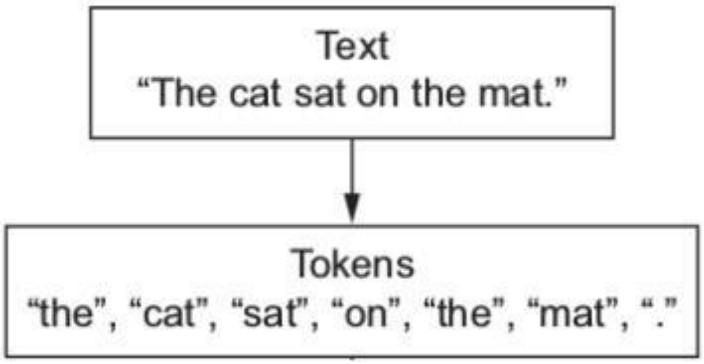


Fig 2 Tokenization

6) Removing punctuations and special symbols:

Apart from the considered set of emoticons punctuations and symbols like &,\,; are removed.

7) Stop words removal:

Stop words are the most commonly occuring words which are not relevant in the context of the data and do not contribute any deeper meaning to the phrase. In this case contain no sentiment. NLTK provide a library used for this.

"This is a sample sentence, showing off the stop words filtration."

['This', 'is', 'a', 'sample', 'sentence', ',', 'showing', 'off', 'the', 'stop', 'words', 'filtration', '.']

8) After stop words removal:

['This', 'sample', 'sentence', ',', 'showing', 'stop', 'words', 'filtration', '.']

9) Stemming and Lemmatization:

Sentences are always narrated in tenses,singular and plural forms making most words accompany with -ing,-ed,es and ies. Therefore,extracting the root word will suffice to identify sentiment behind the text.

Base forms are the skeleton for grammar stemming and lemmatization reduces inflectional forms and derivational forms to common base forms .
Example: Cats is reduced to cat ,ponies is reduced to poni.

Stemming is a crude way of reducing terms to their root, by just defining rules of chopping off some characters at the end of the word, and hopefully, gets good results most of the time. *The goal of both stemming and lemmatization is to reduce inflectional forms and sometimes derivationally related forms of a word to a common base form.*With that being said, stemming/lemmatizing helps us reduce the number of overall terms to certain “root” terms.

Rule			Example		
SSES	→	SS	caresses	→	caress
IES	→	I	ponies	→	poni
SS	→	SS	caress	→	caress
S	→		cats	→	cat

SYSTEM ARCHITECTURE

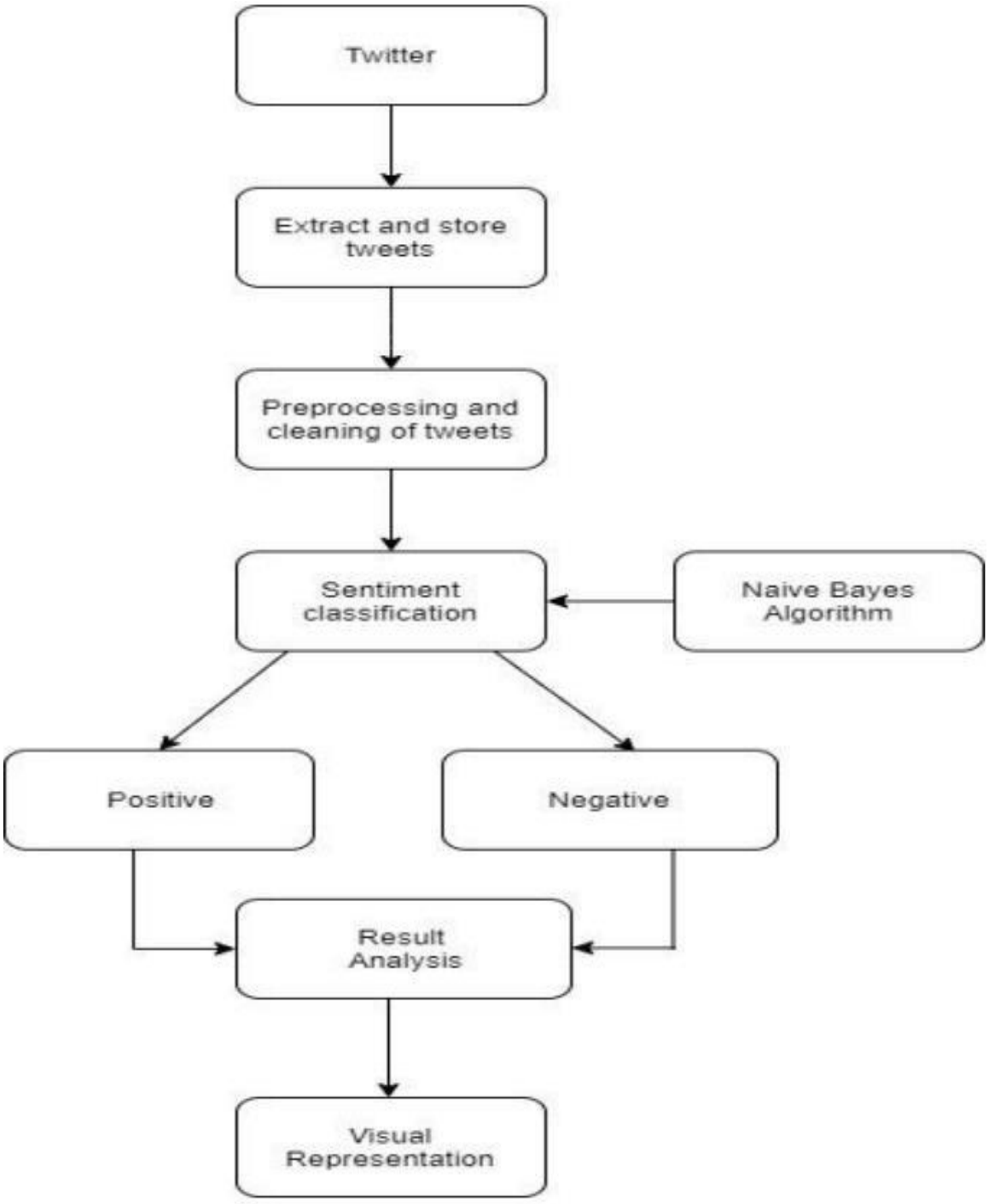


Fig 3. System Representation

STRUCTURE CHART

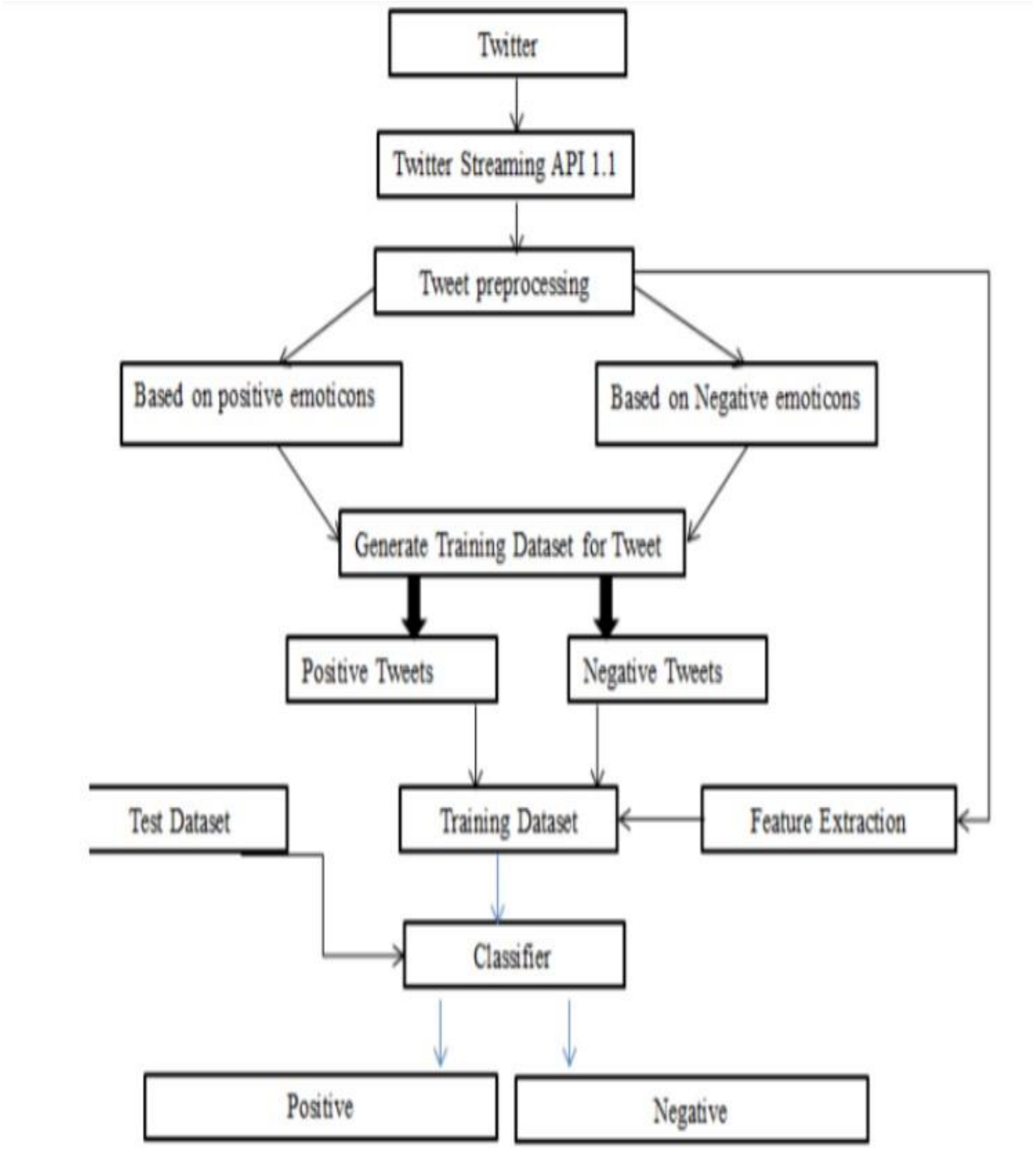


Fig 4. Structure chart

UML DIAGRAM

A UML diagram is a partial graphical representation (view) of a model of a system under design, implementation, or already in existence. UML diagram contains graphical elements (symbols) - UML nodes connected with edges (also known as paths or flows) - that represent elements in the UML model of the designed system. The UML model of the system might also contain other documentation such as use cases written as templated texts.

The kind of the diagram is defined by the primary graphical symbols shown on the diagram. For example, a diagram where the primary symbols in the contents area are classes is class diagram. A diagram which shows use cases and actors is use case diagram. A sequence diagram shows sequence of message exchanges between lifelines.

UML specification does not preclude mixing of different kinds of diagrams, e.g. to combine structural and behavioral elements to show a state machine nested inside a use case. Consequently, the boundaries between the various kinds of diagrams are not strictly enforced. At the same time, some UML Tools do restrict set of available graphical elements which could be used when working on specific type of diagram.

UML specification defines two major kinds of UML diagram: structure diagrams and behavior diagrams. Structure diagrams show the static structure of the system and its parts on different abstraction and implementation levels and how they are related to each other. The elements in a structure diagram represent the meaningful concepts of a system, and may include abstract, real world and implementation concepts.

Behavior diagrams show the dynamic behavior of the objects in a system, which can be described as a series of changes to the system over time

1) USE CASE DIAGRAM

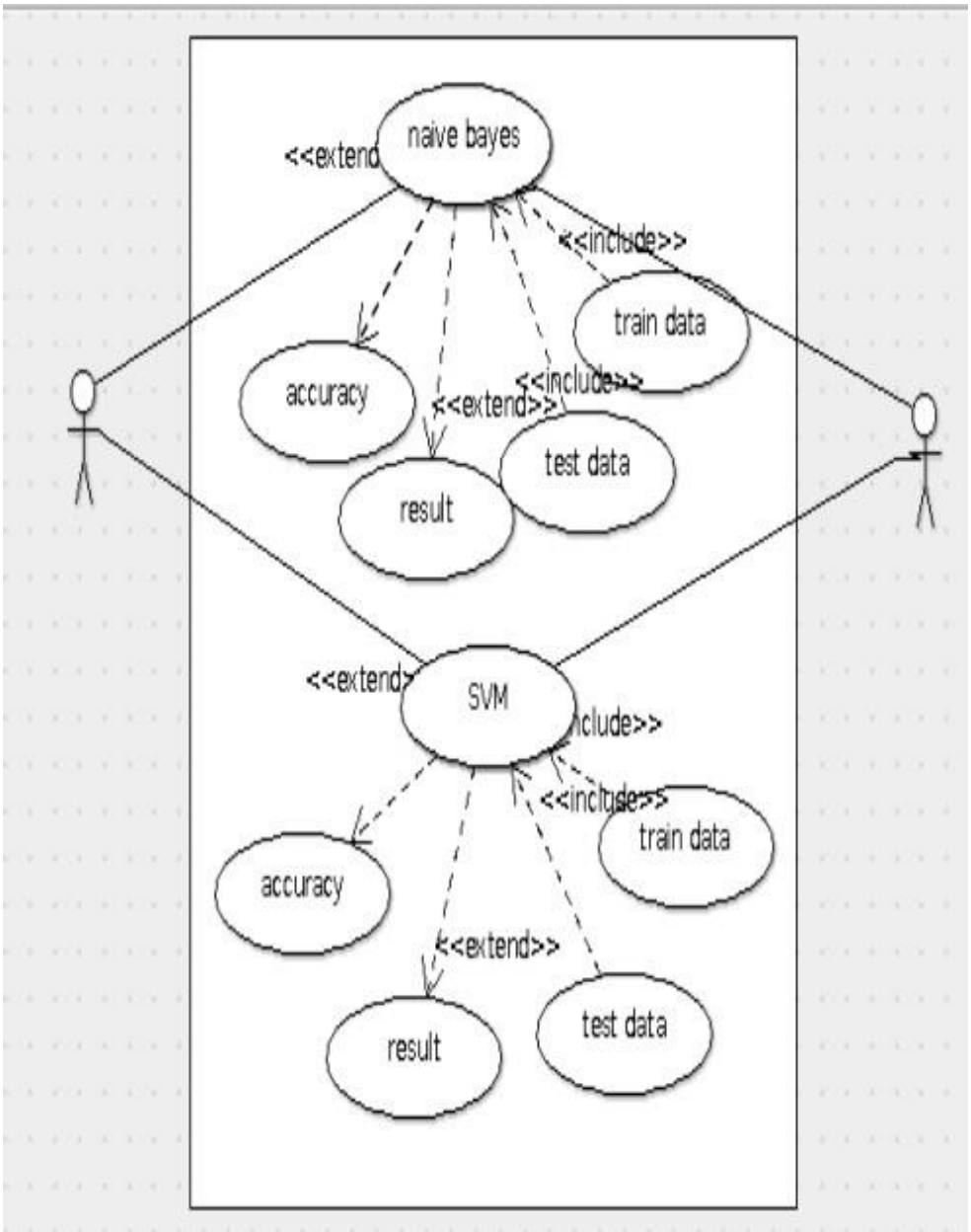


Fig 5. Use Case Diagram

2) SEQUENCE DIAGRAM

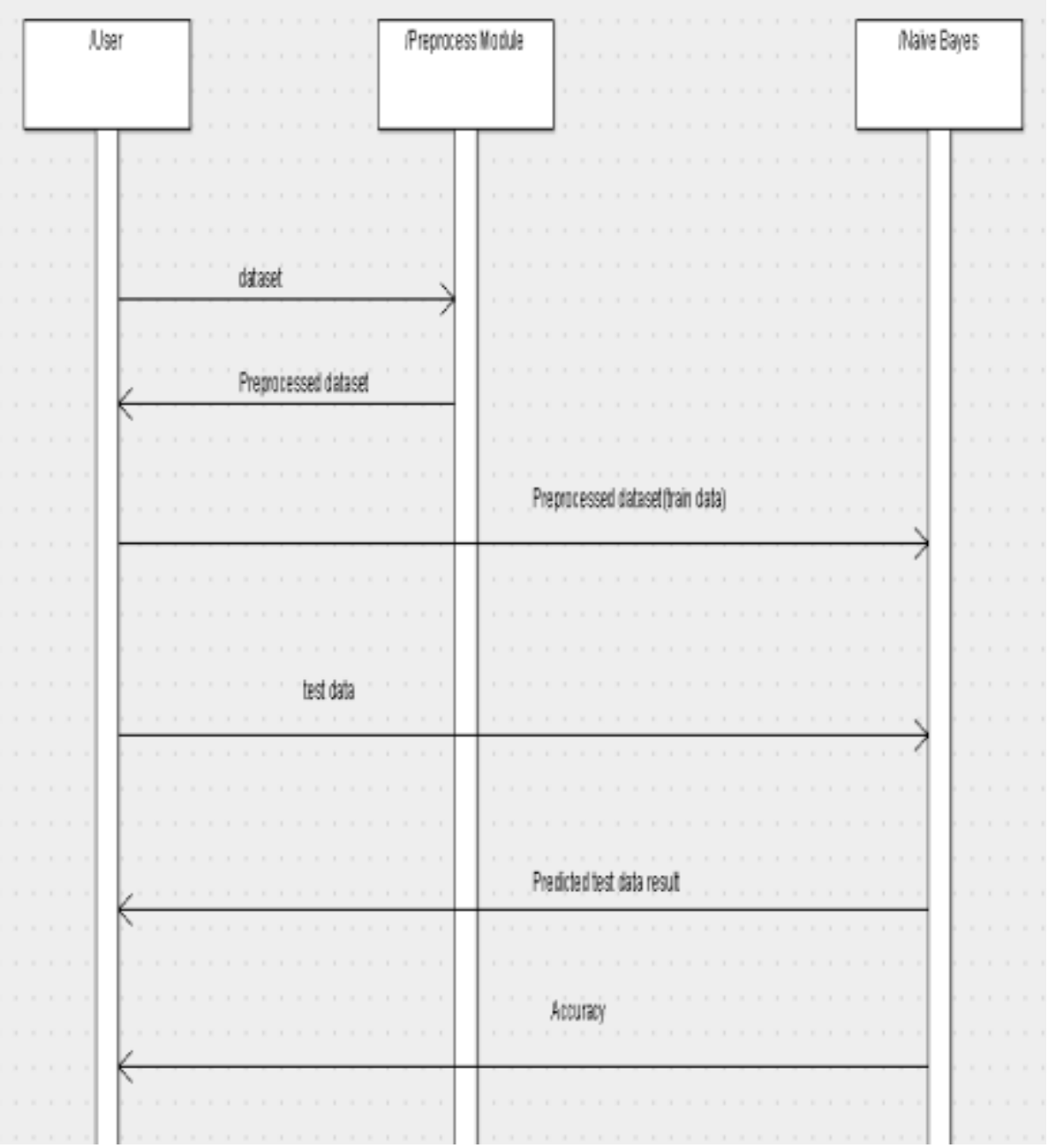


Fig 6. Sequence Diagram of Sentiment Analysis(I)

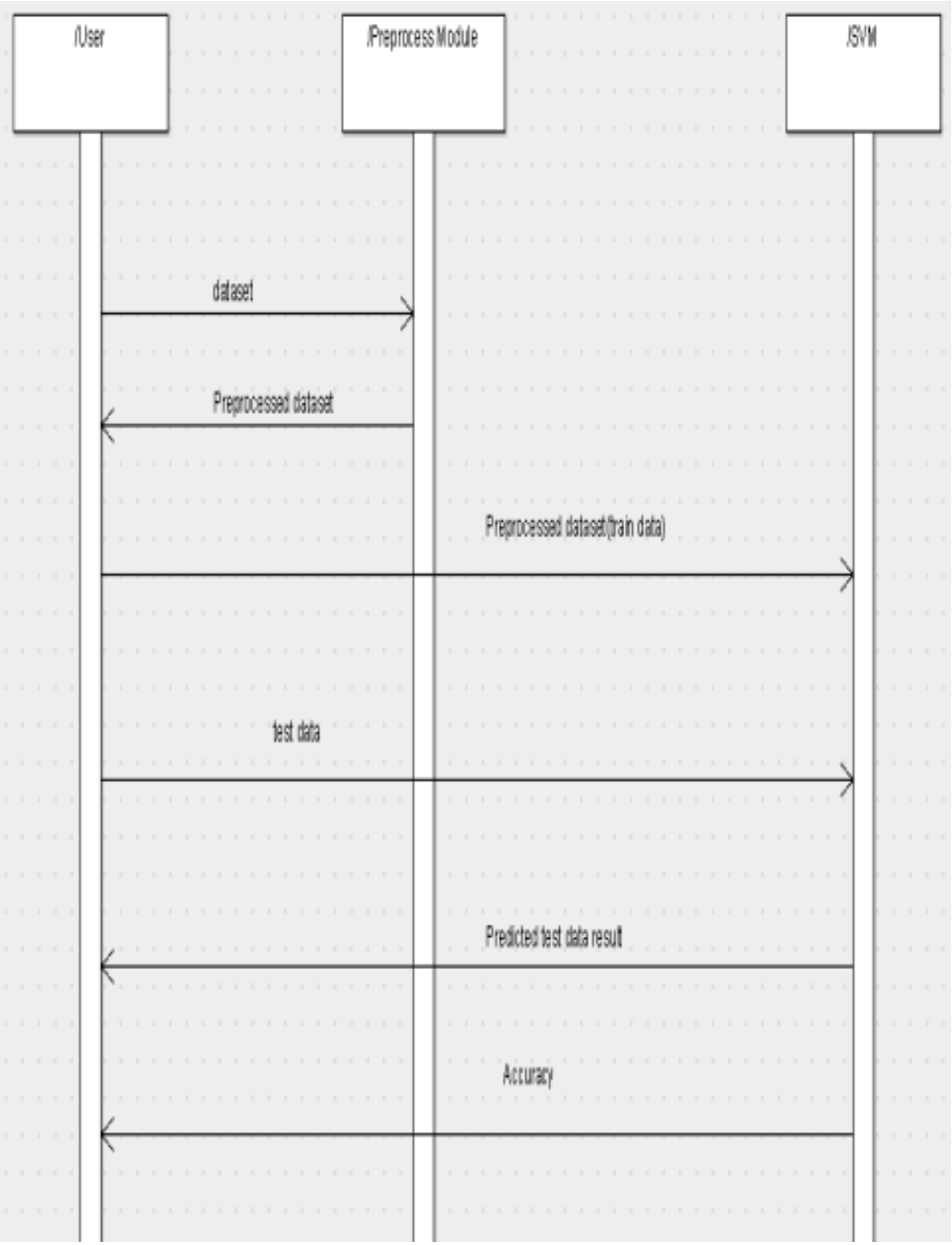


Fig 7. Sequence Diagram of Emotional Analysis(II)

3) FLOW CHART DIAGRAM

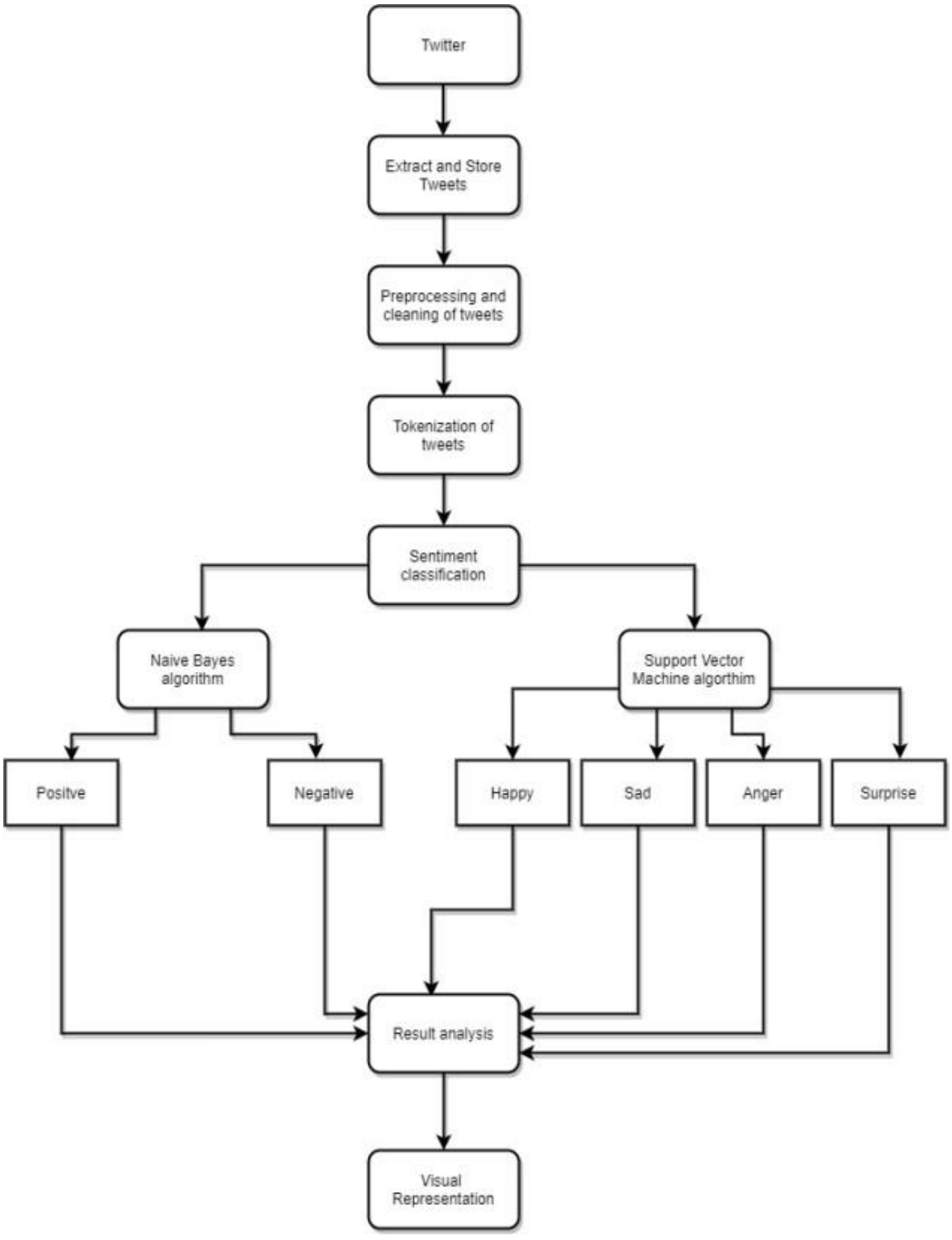


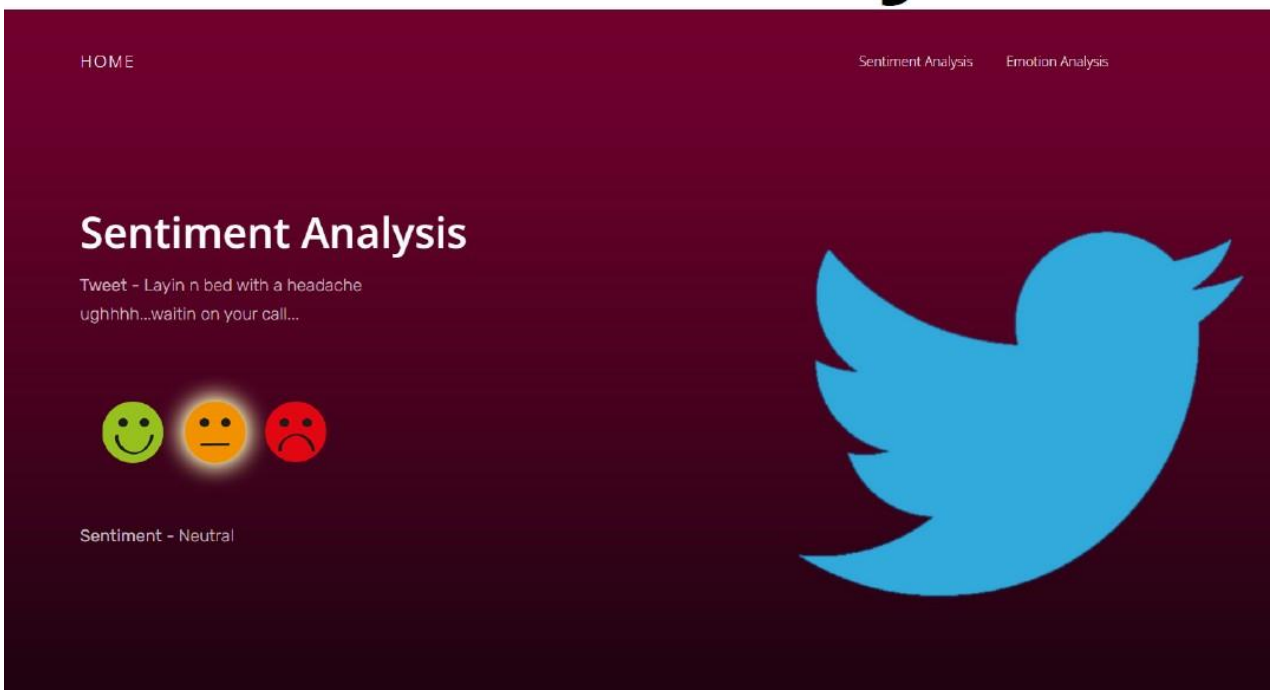
Fig 8. Flow Chart Diagram

CHAPTER - 4 IMPLEMENTATION SCREENSHOT & ROADMAP

IMPLEMENTATION SCREENSHOTS

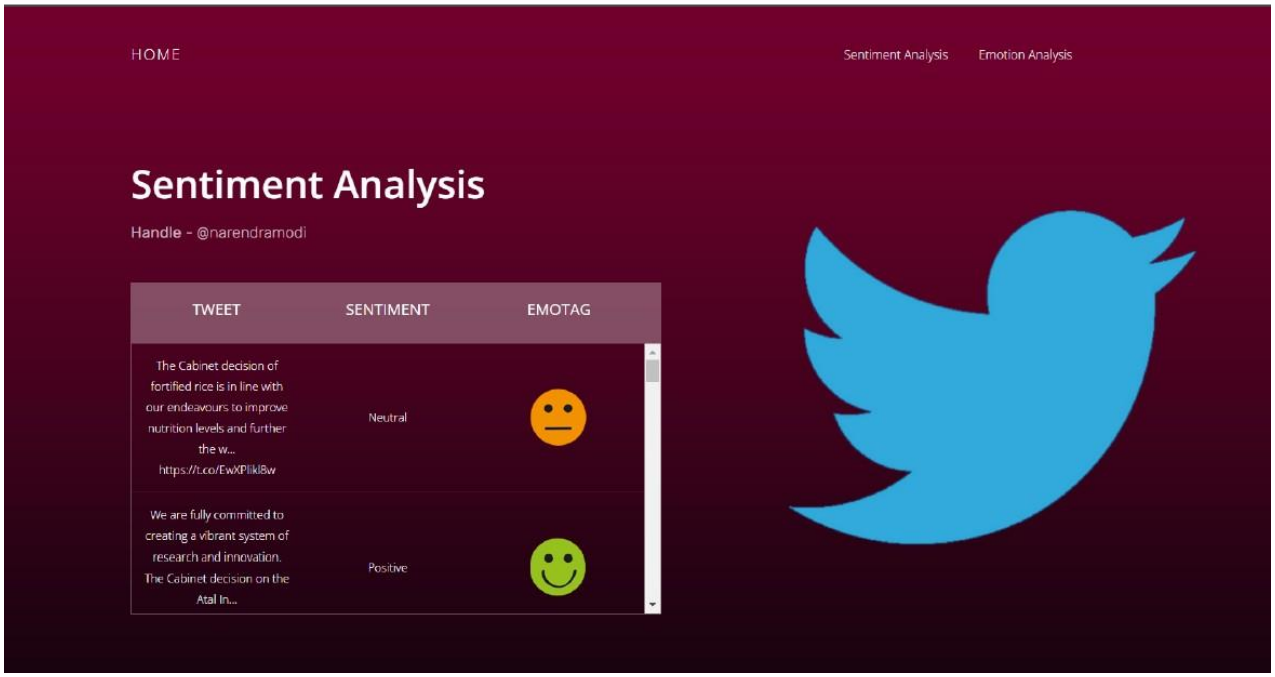
Home page

Sentiment Analysis



Emotion Analysis

Tweets of user




ROAD MAP
Project planning and selecting definition
1
UI designing and further planning
3
Debugging
5
2
Requirement Gathering and research
4
Implementing actual code
6
Final built project

EXPECTED OUTCOMES

We choose “Sentiment and Emotion Analysis of twitter tweets” as our project because we wanted to build a data mining app which will be able to create interaction between tweet data and data mining. We wanted to build an web app which will be able to collect twitter data for example tweets and analyze them. After the analysis the application will provide us the sorted overview of sentimental and emotional situation of targeted people.

So finally we decided to build a data mining application using R for sentiment analysis. Where we will be able to search tweets using twitter API or by writing them ourselves, for a particular topic, of a particular account or random data. User will also be able to see trending hot topics or keywords on their search topic. We will also build a web application for non-technical users. So that they will be able to analyze data themselves from twitter.

ROLES AND RESPONSIBILITIES

- Dhruv Maradiya(19DCE066): -
  Backend

- Femil Mori(19DCE075): -
  UI/UX & project manager

- Archan Bhalani (19DCE004): -
  Fronten

CONCLUSION

CONCLUSION

Nowadays, sentiment analysis or opinion mining is a hot topic in machine learning. We are still far to detect the sentiments of s corpus of texts very accurately because of the complexity in the English language and even more if we consider other languages such as Chinese. In this project we tried to show the basic way of classifying tweets into positive or negative category using Naive Bayes as baseline and how language models are related to the Naive Bayes and can produce better results. We could further improve our classifier by trying to extract more features from the tweets, trying different kinds of features, tuning the parameters of the naïve Bayes classifier, or trying another classifier all together.

FUTURE WORK

FUTURE WORK

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

BIBLIOGRAPHY

- Alexander Pak, Patrick Paroubek. 2010, Twitter as a Corpus for Sentiment Analysis and Opinion Mining.
- Alec Go, Richa Bhayani, Lei Huang. Twitter Sentiment Classification using Distant Supervision
- Jin Bai, JianYun Nie. Using Language Models for Text Classification.
- Apoorv Agarwal, Boyi Xie, Ilia Vovsha, Owen Rambow, Rebecca Passonneau. Sentiment Analysis of Twitter Data.
- Fuchun Peng. 2003, Augmenting Naive Bayes Classifiers with Statistical Language Models
- Ravinder Ahujaa, Aakarsha Chuga, Shruti Kohlia,Shaurya Guptaa,Pratyush Ahujaa- The Impact of Features Extraction on the Sentiment Analysis Noida, India.
- Richa Mathur, Devesh Bandil, Vibhakar Pathak-Analyzing Sentiment of Twitter Data using Machine Learning Algorithm
- Soumaya Chaffar and Diana Using a Heterogeneous Dataset forEmotion Analysis in Inkpen, School of Information Technology and Engineering, University of Ottawa ,Ottawa, ON, Canada.
- Introduction to Machine Learning Alex Smola and S.V.N. Vishwanathan Yahoo! Labs Santa Clara –and– Departments of Statistics and Computer Science Purdue University –and– College of Engineering and Computer Science Australian National University
- Machine Learning Tom M. Mitchell
- An Idiot's guide to Support vector machines (SVMs) R. Berwick, Village Idiot
- https://www.svm-tutorial.com/2014/10/svm-linear-kernel-good-text-classification
- https://towardsdatascience.com/naive-bayes-classifier-81d512f50a7c
- https://www.ijitee.org/wp-content/uploads/papers/v8i8/H6330068819.pdf