A Project Report on

Twitter Sentiment Analysis

Submitted in partial fulfillment of the requirements

in

Computer Engineering

by

Atharva Kulkarni (17102070)
Manas Mohite (17102069)
Parin Dodhiya (17102036)

Under the Guidance of

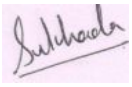
Prof. S.S.Aloni

Department of Computer Engineering
A.P. Shah Institute of Technology
G.B.Road,Kasarvadavli, Thane(W), Mumbai-400615
UNIVERSITY OF MUMBAI

Academic Year 2020-2021

Approval Sheet

This Project Report entitled ***"Twitter sentimental analysis"*** Submitted by ***"Atharva Kulkarni"(17102070), "Manas Mohite"(17102069) ,"Parin Dodhiya"(17102036)*** is approved for the partial fulfillment of the requirement in ***Computer Engineering*** from ***University of Mumbai*** .



(Prof S.S.Aloni)

Guide

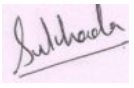
Prof. S.H.Malave

Head, Computer Engineering Department

Place:A.P.Shah Institute of Technology, Thane
Date:16th december 2020

CERTIFICATE

This is to certify that the project entitled ***"Twitter sentiment analysis"*** submitted by ***"Atharva Kulkarni"(17102070), "Manas Mohite"(17102069) ,"Parin Dodhiya"(17102036)*** for the partial fulfillment of the requirement for award of a degree ***Bachelor of Engineering*** in Computer Engineering .,to the University of Mumbai,is a bonafide work carried out during the academic year 2020-2021.



Project coordinator

(Prof
S.S.Aloni)
Guide

Prof. Sachin Malve
Head, Computer Engineering Department

Dr. Uttam D. Kolekar
Principle

External Examiner

Place:A.P.Shah Institute of Technology, Thane
Date: 16th december, 2020

Declaration

We declare that this written submission represents our ideas in our own words and where others' ideas or words have been included, We have adequately cited and referenced the original sources. We also declare that We have adhered to all principles of academic honesty and integrity and have not misrepresented or fabricated or falsified any idea/data/fact/source in our submission. We understand that any violation of the above will be cause for disciplinary action by the Institute and can also evoke penal action from the sources which have thus not been properly cited or from whom proper permission has not been taken when needed.

_____

(Signature)

_____

(Atharva Kulkarni,17102070)
(Manas Mohite , 17102069)
(Parin Dodhiya , 17102036)

Date: 16th december,2020

Contents

Chapter 1

Project Concept and Initiation

1.1 Abstract

Twitter for example is a platform in which users send, read posts known as `tweets' and interact with different communities

Users share their daily lives, post their opinions on everything such as brands and places. Companies can benefit from this massive platform by collecting data related to opinions on them.

Data in Twitter is highly unstructured which makes it difficult to analyze. The aim of this paper is to present a model that can perform sentiment analysis of real data collected from Twitter.

1.2 Objectives

The aim is to find the opinionative data and classify it according to its polarity, i.e. positive, negative or neutral feedback, known as sentiment classification and then analysing it which is known as sentiment analysis

To implement an algorithm for automatic classification of text into positive, negative or neutral Graphical representation of the sentiment in form of Pie-Chart

1.3 Literature Review

A. **TA. Pak, and P. Paroubek**, “Twitter as a corpus for sentiment analysis and opinion mining,” pp. 1320–1326, May 2010. Recent research based on sentiment analysis says that the analysis of opinion utilizes simultaneous learning. Pak and Paroubek in utilized tweets which end with emoticons like ”:)” ”:-)” as positive, and ”:(” ”:-(” as negative. They accumulated models including Max Entropy, Support Vector Machines (SVM) and Naive Bayes and concluded that SVM performed the best amongst various others, attaining more precision which led SVM to be the best performer of all the classifiers.

B. **Min Song, Meen Chul Kim , Yoo Kyung Jeong** ,”Analyzing the Political Landscape of 2012 Korean Presidential Election in Twitter ”,Intelligent Systems, IEEE (Volume:29 , Issue: 2) ,2014 IEEE. Song, Kim and Jeong initially collected the tweets. Their collection and mining techniques processed large datasets in real time. To demonstrate the usefulness of their approaches, they specifically focused on topical trend analysis and network analysis to examine presidential issues embedded in Twitter data.

1.4 Problem Definition

1) The problem in sentiment analysis is classifying the polarity of a given text at the document,sentence, or feature/aspect level

2) Whether the expressed opinion in a document, sentence or an entity feature/aspect is positive,negative, or neutral

1.5 Scope

The scope for Phase I is Limited to gathering the raw tweets, pre-processing the raw tweets and preparing a clean dataset for training as well as testing purpose. The training of the model will be then done. A front end will be useful for the end user to use the software. Data Visualization and Front End to Back End connection will be reserved for Phase II.

1.6 Technology Stack

- Python 3.7
- Tweepy
- TextBlob
- HTML, CSS, PHP
- Google Collab

1.7 Benefits For Environment And Society

This project helps the entrepreneurs understand how their product is being received by the public and accordingly act to modify or upgrade their product or service as per the need of the customer.

This project is not only beneficial for the companies to get a review of their product but also for the common users to search for their favourite topics and interests on twitter and know other users' opinions.

Chapter 2

2.1 Proposed System

- **Retrieval of Data**: Public Twitter data is mined using the existing Twitter APIs for data extraction. Tweets would be selected based on a few chosen keywords pertaining to the domain of our concern, i.e. product reviews.
- **Preprocessing**: In this stage, the data is put through a preprocessing stage in which we remove identifying information such as Twitter handles, timestamps of the message and embedded links and videos. Such information is largely irrelevant and may cause false results to be given by our system.
- **Data Training**: With a large data set available, we can train the algorithm to have a high accuracy.
- **Data Visualization**: Classify data into positive, negative and neutral tweets and visualize it in the form of Pie Charts and Graphs.

2.2 Design (Flow of Modules)

The proposed flow of modules can be classified into four modules: Gathering the data, processing the data for unwanted characters and preparing data set, applying machine level algorithms to training and testing data for generating outputs, and lastly creating a frontend for data visualization. Gathering the tweets is a basic module for accessing the tweets using the Twitter Search API. The next step is to divide the main dataset into training and testing dataset. Then feature engineering the data to suit our project.

1) Gather data

2) Obtain the train,test dataset 3)Process the data

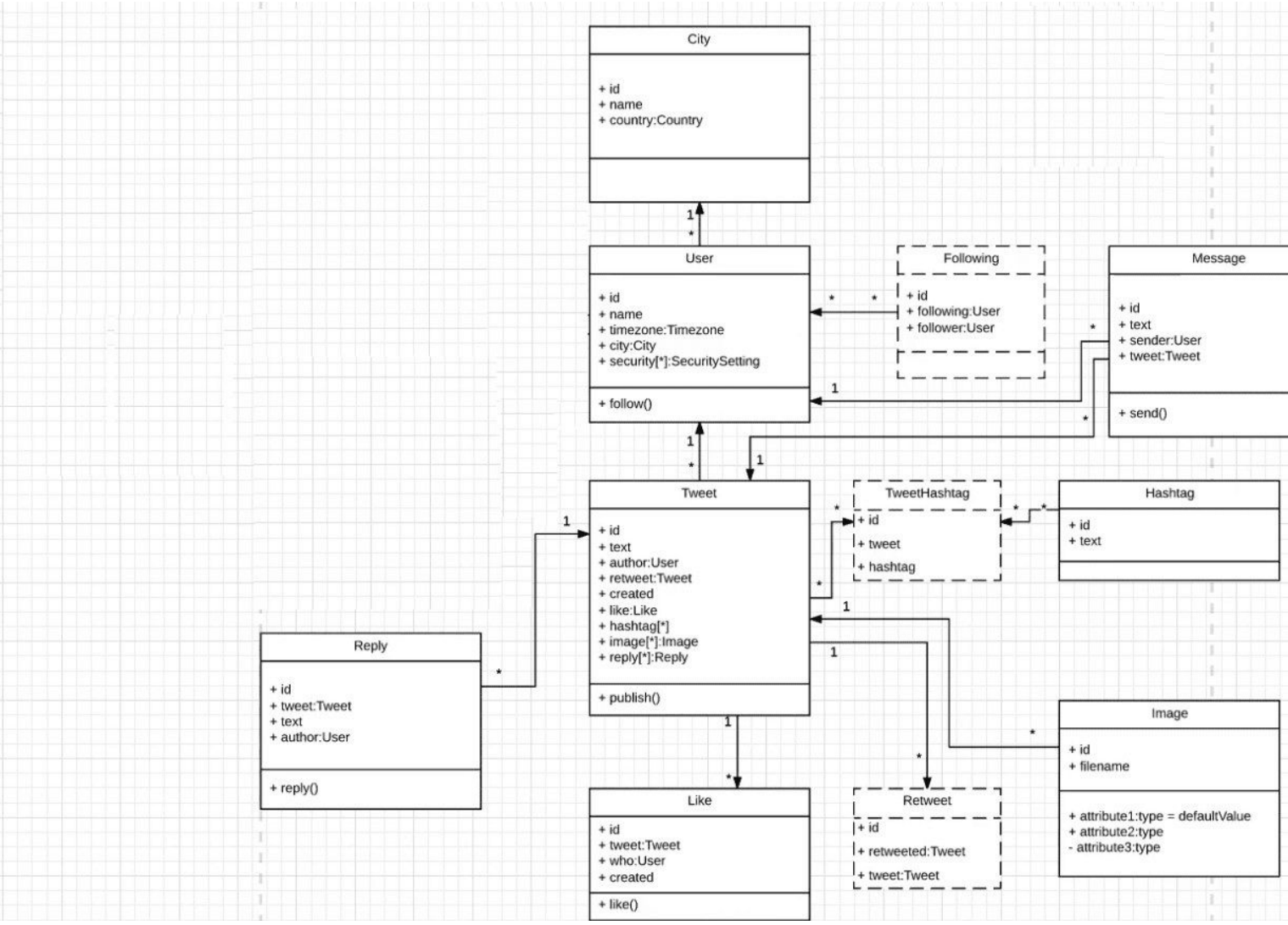
   i. Removing the special characters from the tweets. ii.Converting all the sentences into lower case. iii.Apply actual algorithm on the training data

4) Create a front end for the end user to search tweets

5) Connect front and back end for data visualization.

2.3 Figures and Tables-

2.3.1 Class Diagram



2.3.2 Gantt Chart-

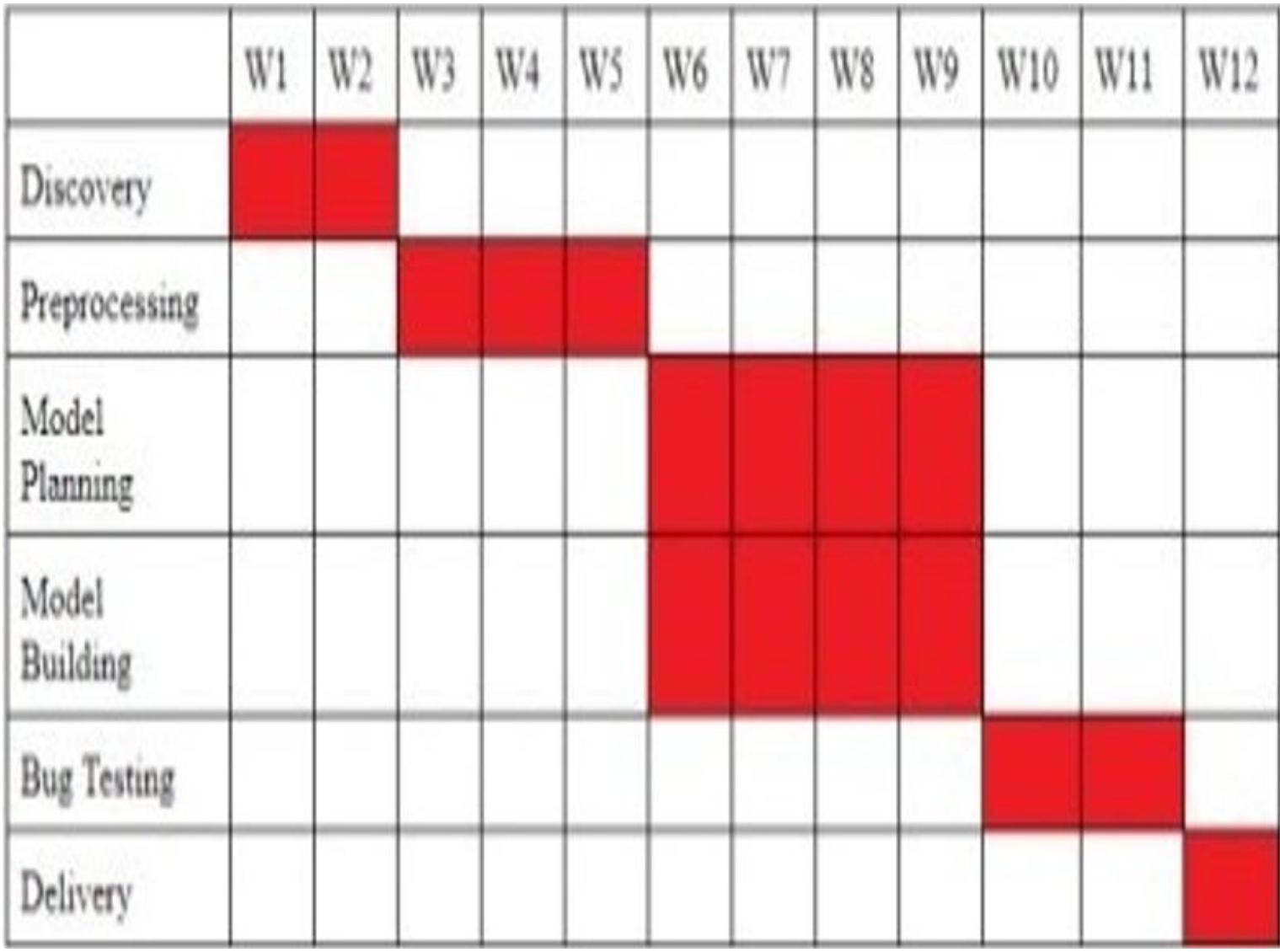


Fig 2.3.2 Gantt Chart

2.4 Modules-

2.4.1 Gathering the tweets -

Twitter provides the REST search api for searching tweets from Twitter's search index. This is different than using the streaming filter API, in that the later is real time and starts giving you results from the point of query, while the Search API will give you results from past, up to as far back as the search index goes (usually last 7 days).To start with the API Rate Limit page details the limits of various Twitter APIs, and as per the page the limit for the Search API is 180 Requests per 15 mins window for per-user authentication. The twitter Search API is limited 180 Requests/15 mins limit, and per request you can ask for maximum 100 tweets, giving you a grand total limit of 18,000 tweets/15 mins, If you download 18K tweets before 15 mins, you won't be able to get any more results until your 15 minute window expires and you search again.

2.4.2 Preprocessing the Tweets-

After gathering the tweets, we need to remove any unnecessary qualities in the data which would make the trained model a poor generalizer. Any redundant characters or words can make the training model less accurate to recognize the actual words. Text preprocessing involves many things like removing emojis, properly formatting the text to remove extra spaces or any other information in the text that we don't believe would add information to our model. We also have to make sure that the information we pass the model is in a format that computers can understand. After this pre-processing step, our data should be ready to use for a machine learning classification task.

2.4.3 Model Training And Testing-

This stage is a continuation from the previous modules that manages the tweets for unwanted characters using text preprocessing for a machine learning(ML) sentiment analysis task. In this stage we'll split our data into training and test sets. After training the model, we will then use it to classify sentiment on unseen twitter data that has been preprocessed in the same manner as the training data.

2.4.4 Front End and Data Visualization-

The objective is to make a front end for the end user to search for a tweet. The end result of the front end is to visualize the data into positive, negative and neutral tweets. The end user should be able to analyze the tweets with the help of the graphical representation of data.

3. Planning-

1) To develop data visualization on the front end.

2) Collect a more vast training data to obtain more accuracy for the model.

3) Make the front end cleaner and more practical.