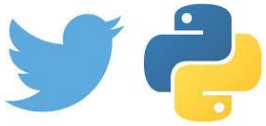
A
Project Report
On
"TWEETS SENTIMENT ANALYSIS & VISUALIZATION"
(CE351 – Software Group Project III)



Prepared by

Tarun Adavani	**Shivam Bhanvadia**	**Param Panara**
(16CE002)	**(16CE008)**	**(16CE056)**

Under the Supervision of
Prof. Vaishali Koria
Prof. Aniruddh Fataniya

Submitted to
Charotar University of Science & Technology (CHARUSAT)
Bachelor of Technology (B. Tech.)
in Computer Engineering (CE)
for 6th semester B. Tech
Submitted at

Accredited with Grade A by NAAC Accredited with Grade A by KCG

U & P U. PATEL DEPARTMENT OF COMPUTER ENGINEERING
(NBA Accredited)
Chandubhai S. Patel Institute of Technology (CSPIT) Faculty of Technology & Engineering (FTE), CHARUSAT
At: Changa, Dist.: Anand, Pin: 388421.

April, 2019

CERTIFICATE

This is to certify that the report entitled "Tweets Sentiment Analysis & visualization" is a bonafied work carried out by **Mr. Shivam Bhanvadia(16CE008)** under the guidance and supervision **Prof. Vaishali Koria, Prof. Aniruddh Fataniya** for the subject **Software Group Project-III (CE351)** of 6th Semester of Bachelor of Technology in **Computer Engineering** at Faculty of Technology & Engineering (C.S.P.I.T.) – CHARUSAT, Gujarat.

To the best of my knowledge and belief, this work embodies the work of candidate **Mr. Shivam Bhanvadia(16CE008)** 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.

Under the supervision of,

Prof. Vaishali Koria
Prof. Aniruddh Fataniya
Asst. Professor
U. & P U. Patel Dept. of Computer Engg.
C.S.P.I.T., CHARUSAT-Changa.

Dr. (Prof.) Amit Ganatra
Dean,
Faculty of Technology & Engineering
Head, U. & P U. Patel Department of Computer Engineering
C.S.P.I.T., CHARUSAT- Changa, Gujarat.

Chandubhai S Patel Institute of Technology (C.S.P.I.T.)

Faculty of Technology & Engineering, CHARUSAT

At: Changa, Ta. Petlad, Dist. Anand, PIN: 388 421. Gujarat

TABLE OF CONTENTS

ABSTRACT

In today's scenario, Sentiment Analysis or opinion mining is the computational study of people's opinions, sentiments, attitudes, and emotions expressed in written language. It is one of the most active research areas in natural language processing and text mining in recent years. Its popularity is mainly due to two reasons. First, it has a wide range of applications because opinions are central to almost all human activities and are key influencers of our behaviors. Whenever we need to make decision, we want to hear others' opinions. Second, it presents many challenging research problems, which had never been attempted before the year 2000. Part of the reason for the lack of study before was that there was little opinionated text in digital forms. It is thus no surprise that the inception and the rapid growth of the field coincide with those of the social media on the Web. In fact, the research has also spread outside of computer science to management sciences and social sciences due to its importance to business and society as a whole. In this talk, I will start with the discussion of the mainstream sentiment analysis research and then move on to describe some recent work on modelling comments, discussions, and debates, which represents another kind of analysis of sentiments and opinions.

ACKNOWLEDGEMENT

This work was done as a part of the Software Group Project-III (CE351) course undertaken by the authors at Faculty of Technology & Engineering (C.S.P.I.T.)-CHARUSAT, Gujarat during the period dated January 5,2019 to April 11, 2019.

Authors would like to express deep sense of gratitude towards our Head of the CE Department, Dr. Amit Ganatra and would like to acknowledge the constant guidance, encouragement and supervision rendered by Prof. Vaishali Koria, Prof. Aniruddh Fataniya.

The authors also express their sincere gratitude to CHARUSAT and access the library and computational facilities on campus.

We perceive this opportunity as a big milestone in career development. We will strive to use gained skills and knowledge in the best possible way, and will continue to work on their improvement, in order to attain desired career objectives.

We would like to thank our friends for their support. We owe regards to the faculty of department of computer at engineering from where we have learnt the basics of computer science and whose informal discussions and able guidance become light for us in entire duration of this work.

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

We are also thankful to https://github.com/, https://stackoverflow.com/,etc. for the huge support in the coding part of the project.

With sincere regards,

Tarun Adavani

Shivam Bhanvadia

Param Panara

LIST OF FIGURES

CHAPTER 1: INTRODUCTION

1.1 PROJECT OVERVIEW

Tweets are imported using Python and the data is cleaned by removing emoticons and URLs. Lexical Analysis is used to predict the sentiment of tweets and subsequently express the opinion graphically through pie chart, bar chart.

What is twitter sentiment analysis?

Twitter is an online news and social networking service that enables users to send and read short 140-character messages called "tweets". Registered users can read and post tweets, but those who are unregistered can only read them.

Hence Twitter is a public platform with a mine of public opinion of people all over the world and of all age categories.

As of October 2016, Twitter has more than 315 million monthly active users.

Twitter Sentiment Analysis is the process of determining the emotional tone behind a series of words, used to gain an understanding of the the attitudes, opinions and emotions expressed within an online mention.

1.2 PURPOSE

Why analytics?

- What is trending positively/negatively over a period of time and why?
- Who is being talked about, where, and why?
- What college is being talked about?
- What topics are being discussed the most?
- Who is being talked about most positively?
- What are the best sources for positive exposure?
- What is the geographic location of the comments?

Why visualization?

Data visualization is the presentation of data in a pictorial or graphical format. It enables decision makers to see analytics presented visually, so they can grasp difficult concepts or identify new patterns. With interactive visualization, you can take the concept a step further by using technology to drill down into charts and graphs for more detail, interactively changing what data you see and how it's processed. Tables, timelines, word clouds, histograms and pie charts can be used for visualization.

1.3 SCOPE

Applications:

The applications of sentiment analysis are broad and powerful. Shifts in sentiment on social media have been shown to correlate with shifts in the stock market.

For example, the Obama administration used sentiment analysis to gauge public opinion to policy announcements and campaign messages ahead of 2012 presidential election.

The ability to quickly understand consumer attitudes and react accordingly is something that Expedia Canada took advantage of when they noticed that there was a steady increase in negative feedback to the music used in one of their television adverts.

1.4 OBJECTIVE

Why twitter sentiment analysis?

The applications for sentiment analysis are endless. It is extremely useful in social media monitoring as it allows us to gain an overview of the wider public opinion behind certain

topics However, it is also practical for use in business analytics and situations in which text needs to be analyzed.

Sentiment analysis is in demand because of its efficiency. Thousands of text documents can be processed for sentiment in seconds, compared to the hours it would take a team of people to manually complete. Because it is so efficient (and accurate – Semantria has 80% accuracy for English content) many businesses are adopting text and sentiment analysis and incorporating it into their processes.

1.5 TECHNOLOGY AND LITERATURE REVIEW

Sentiment analysis has been handled as a Natural Language Processing task at many levels of granularity. Starting from being a document level classification task (Turney, 2002; Pang and Lee, 2004), it has been handled at the sentence level (Hu and Liu, 2004; Kim and Hovy, 2004) and more recently at the phrase level (Wilson et al., 2005; Agarwal et al., 2009). Microblog data like Twitter, on which users post real time reactions to and opinions about "everything", poses newer and different challenges. Some of the early and recent results on sentiment analysis of Twitter data are by Go et al. (2009), (Bermingham and Smeaton, 2010) and Pak and Paroubek (2010). Go et al. (2009) use distant learning to acquire sentiment data. They use tweets ending in positive emoticons like ":)" ":-)" as positive and negative emoticons like ":(" ":-(" as negative. They build models using Naive Bayes, MaxEnt and Support Vector Machines (SVM), and they report SVM outperforms other classifiers. In terms of feature space, they try a Unigram, Bigram model in conjunction with parts-of-speech (POS) features. They note that the unigram model outperforms all other models. Specifically, bigrams and POS features do not help.

Pak and Paroubek (2010) collect data following a similar distant learning paradigm. They perform a different classification task though: subjective versus objective. For subjective data they collect the tweets ending with emoticons in the same manner as Go et al. (2009). For objective data they crawl twitter accounts of popular newspapers like "New York Times", "Washington Posts" etc. They report that POS and bigrams both help

(contrary to results presented by Go et al. (2009)). Both these approaches, however, are primarily based on n-gram models. Moreover, the data they use for training and testing is collected by search queries and is therefore biased. In contrast, we present features that achieve a significant gain over a unigram baseline. In addition, we explore a different method of data representation and report significant improvement over the unigram models. Another contribution of this paper is that we report results on manually annotated data that does not suffer from any known biases. Our data is a random sample of streaming tweets unlike data collected by using specific queries. The size of our hand- labeled data allows us to perform cross-validation experiments and check for the variance in performance of the classifier across folds. Another significant effort for sentiment classification on Twitter data is by Barbosa and Feng (2010). They use polarity predictions from three websites as noisy labels to train a model and use 1000 manually labeled tweets for tuning and another 1000 manually labeled tweets for testing. They however do not mention how they collect their test data. They propose the use of syntax features of tweets like retweet, hashtags, link, punctuation and exclamation marks in conjunction with features like prior polarity of words and POS of words. We extend their approach by using real valued prior polarity, and by combining prior polarity with POS. Our results show that the features that enhance the performance of our classifiers the most are features that combine prior polarity of words with their parts of speech. The tweet syntax features help but only marginally. Gamon (2004) perform sentiment analysis on feedback data from Global Support Services survey. One aim of their paper is to analyze the role of linguistic features like POS tags. They perform extensive feature analysis and feature selection and demonstrate that abstract linguistic analysis features contribute to the classifier accuracy. In this paper we perform extensive feature analysis and show that the use of only 100 abstract linguistic features performs as well as a hard unigram baseline.

CHAPTER 2: SYSTEM ANALYSIS

2.1 USER CHARACTERISTICS

- **End Users**

  This product is designed mainly for END USERS. So, it is feasible for the end-users to directly analyze the tweets and will able to predict the sentiment of tweets and subsequently express the opinion graphically through bar chart, pie chart.

2.2 TOOLS & TECHNOLOGY

- Installation of Python
- Twitter Developer Account
- Twitter Authentication to access API
- Spyder IDE

CHAPTER 3: SYSTEM DESIGN

3.1 FLOW OF SYSTEM

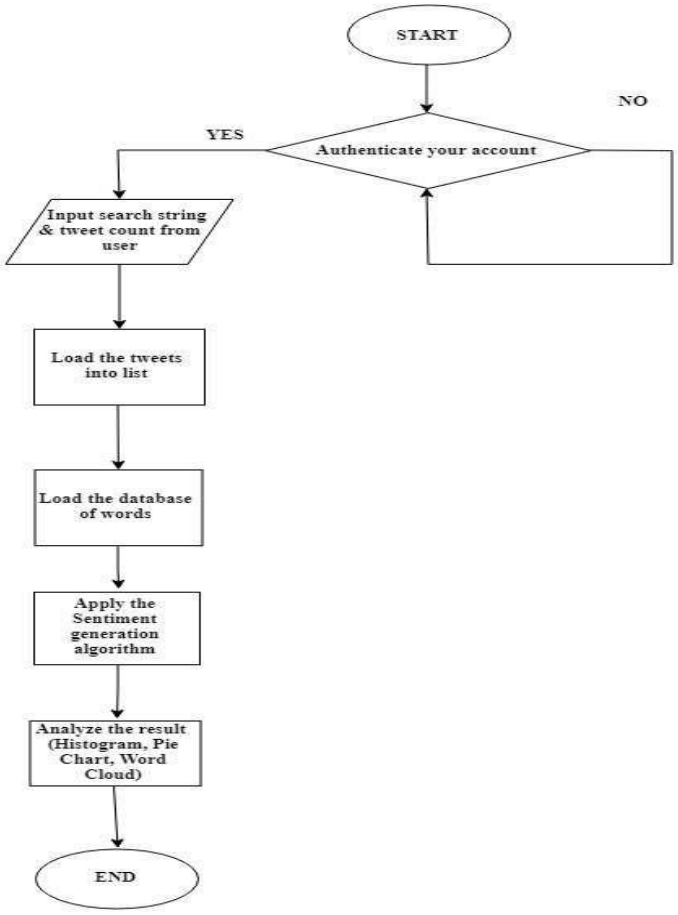


Fig 3.1: Flowchart of Sentiment Analysis of Twitter

3.2 MAJOR FUNCTIONALITY

- Sentence Level Sentiment Analysis in Twitter

3.3 PROJECT SCHEDULING

GANTT CHART

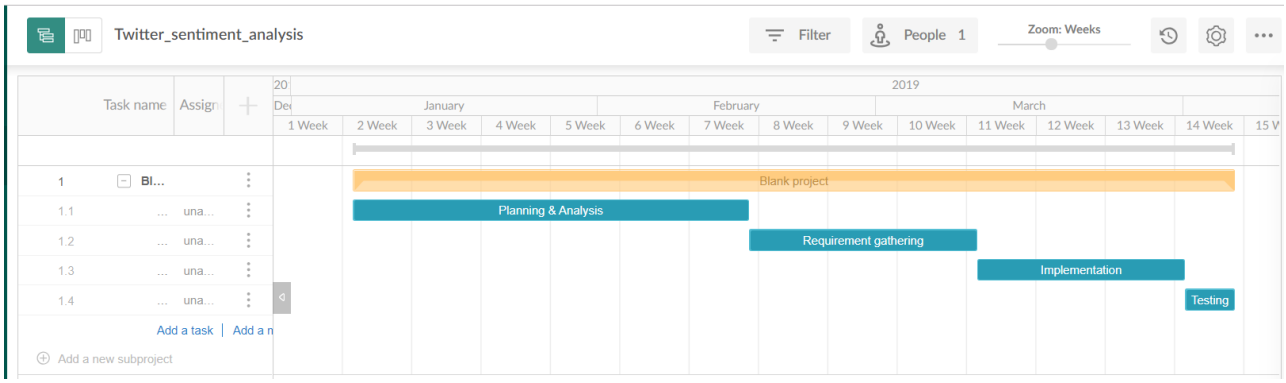


Fig 3.2: Gantt Chart

CHAPTER 4: IMPLEMENTATION

4.1 IMPLEMENTATION ENVIRONMENT

There can be various parameters for describing characteristics of the environment on which project implementation is done, one of the main parameters and its justification is mentioned below:

There can be various parameters for describing characteristics of the environment on which project implementation is done, one of the main parameters and its justification is mentioned below:

Single vs Multi-user

The software is single user, because only one user can use it at a time.

4.2 MODULE SPECIFICATION

FEATURES

1. Extraction of Tweets

(i) Create twitter application

(ii) tweepy - Provides an interface to the Twitter web API

(iii) OAuth - Python Interface for OAuth

(iv) Create twitter authenticated credential object (using key from step): It is done using consumer key, consumer secret, access token, access secret.

(v) During authentication, we are redirected to a URL automatically where we click on Authorize app as shown in the image below and enter the unique

2. Cleaning Tweets

The tweets are cleaned in python by removing:

- Extra punctuation
- Stop words (Most commonly used words in a language like *the*, *is*, *at*, *which*, and *on*.)
- Redundant Blank spaces
- Emoticons
- URLS

3. Algorithms used

- **Lexical Analysis:** By comparing uni-grams to the pre-loaded word database, the tweet is assigned sentiment score - positive, negative or neutral and overall score is calculated.

4. Calculating percentage

We are calculating percentage based on positive score and negative score. If positive score percentage = positive score / positive score + negative score.

5. Bar Chart : bar chart plot

A bar chart is a rectangular statistical graphic, which is divided into range to illustrate the sentiment of the hashtag. In a bar chart is proportional to the quantity it represents.

6. Pie Chart : pie chart plot

A pie chart is a circular statistical graphic, which is divided into slices to illustrate the sentiment of the hashtag. In a pie chart, the arc length of each slice (and consequently its central angle and area), is proportional to the quantity it represents.

7. Tweets

Here in tweets tab we are displaying recent tweets from twitter based on that search.

PACKAGES USED

- **Tweepy:** Provides an interface to the Twitter web API
- **Tweepy.OAuth:** Provides an interface to the OAuth 1.0 specification allowing users to authenticate via OAuth to the server of their choice.
- **TextBlob:** It is the python library for processing textual data.
- **re:** Tools for Splitting, Applying and Combining Data
- **matplotlib:** Lots of plots, various labeling, axis and color scaling functions.

4.3 SNAPSHOTS OF PROJECT

Accessing Twitter API

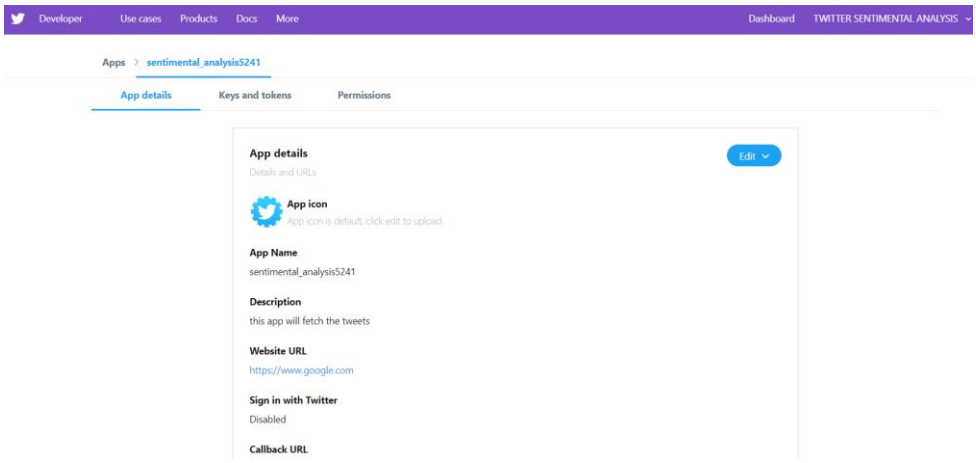


Fig 4.1: Authentication to use twitter

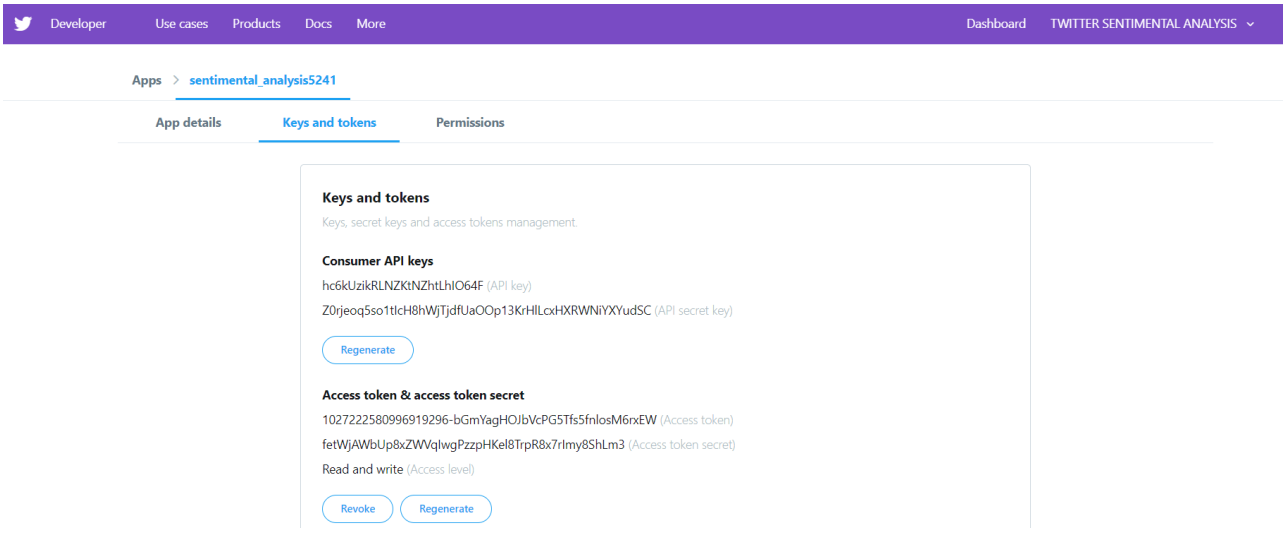


Fig 4.2: Keys and tokens Generated on Browser

```
what you want to search? :  modi
```

Fig 4.3: User Input to search

```
what you want to search? :  modi

No of total tweets:  31
no of positive tweets:  6
no of negative tweets:  5
no of neutral tweets:  20

Positive tweets percentage: 19.35483870967742 %
Negative tweets percentage: 16.129032258064516 %
Neutral tweets percentage: 64.51612903225806 %
```

Fig 4.4: Sentiment Score

```
Positive tweets:

RT siddiqui No one should be surprised at the Hindutva amp Anti
Muslim narrative of BJP in LS polls Modi Amit planned it for last
f
RT Thanks to MODI Ji One of Chowkidar enjoying on the Bank of
River in the smart city
RT Modi gave her toilet light connection gas connection ayushyman
health card n most important freedom from triple tal
Surya Great going Take India to a new height with Hon ble Modi ji
Jai Hind
RT Opposition hurled so many abuses and personal attacks against
PM Modi and yet he remain poised and calm Oppn panelists went
RT Congress will be responsible if Modi returns to power Arrogance
will doom this party Congress has a history of cheating
```

Fig 4.5: List of Positive tweets

```
Negative tweets:

When you know this very well that modi government is failure on
every front and against article 37
RT In 24 hours 1 75 year old Man killed by supporter of DMK
Congress because He was Modi fan 2 BJP leader shot dead in Odi
Ab 15lac mere account mai dalwa dena plz modi g ko bol do na
RT A 75 year old man Govindarajan killed by a supporter of the DMK
Congress alliance at Orthanadu in Tamil Nadu for support
RT Sick 75Yrs old Govindraj murdered by a DMK supporter for
praising PM Sh near Orthanadu village in Tanjore of
```

Fig 4.6: List of Negative tweets

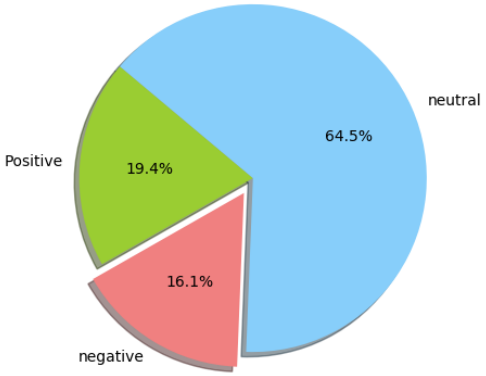


Fig 4.7: Pie chart (By Polarity)

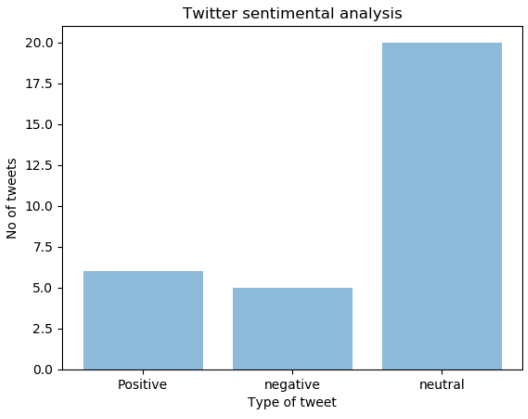


Fig 4.8: Bar Chart (By Polarity)

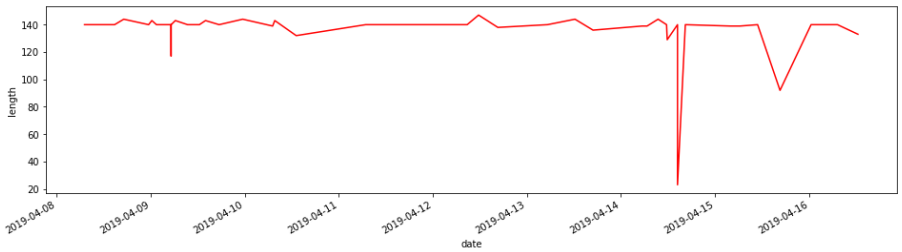


Fig 4.9: Length of tweets

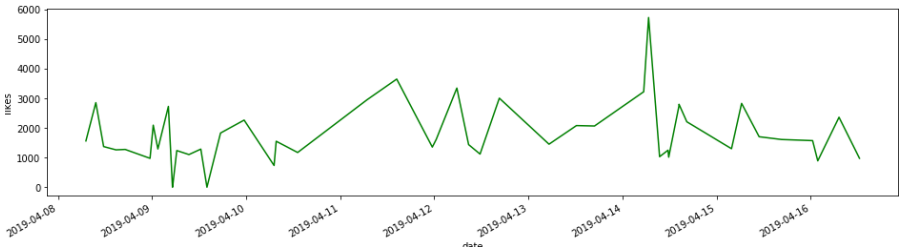


Fig 4.10: Like of tweets

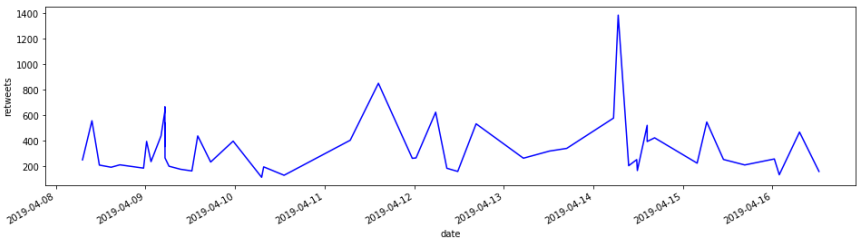


Fig 4.11: Retweet of tweets

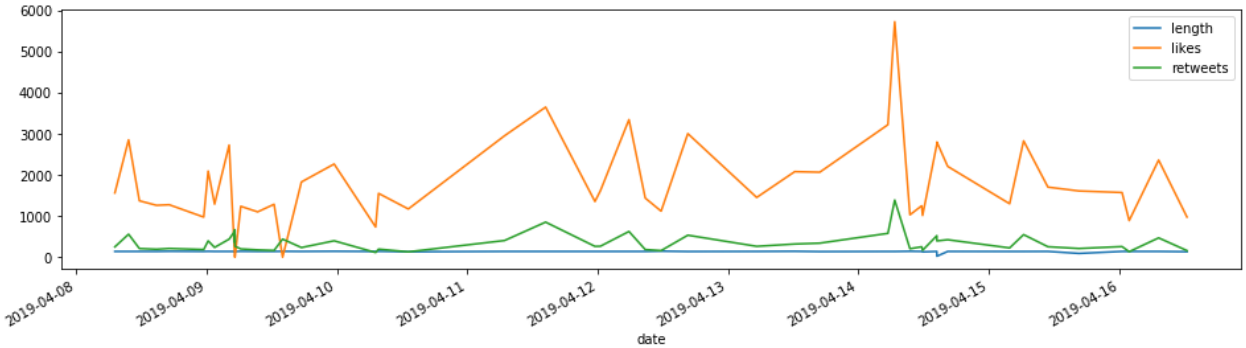


Fig 4.12: Length vs like vs retweet

CHAPTER 5: CONSTRAINTS AND FUTURE ENHANCEMENT

5.1 CONSTRAINTS:

- The Twitter Search API can get tweets up to a maximum of 7 days old.
- Not effective in detecting sarcasm.
- Cannot get 100% efficiency in analyzing sentiment of tweets.
- Can only retrieve a maximum of 100 tweets per query without authenticating via OAuth before receiving a 403 error or timeout.
- Giving a hash tag under the wrong category will still give results: No error message

5.2 FUTURE ENHANCEMENT:

- Add different language words to dataset.
- Star rating (Negative and Positive [According to percentage]) (BOX PLOT).
- Find no of mentions of n particular organizations (And analyze sentiment).
- Timeline of 7 days for emotions: anger, anticipation, disgust, fear, joy, sadness, surprise, trust.
- Word cloud to displayed and other visualization to be provided.
- GUI to be provided.
- Apply better Machine Learning Algorithms (Like Support Vector Machine, RNN etc.)

CHAPTER 6: CONCLUSION

6.1 SELF ANALYSIS OF PROJECT VIABILITIES

Twitter is a source of vast unstructured and noisy data sets that can be processed to locate interesting patterns and trends. Real time data analysis makes it possible for business organizations to keep track of their services and generates opportunities to promote, advertise and improve from time to time. Our heartfelt appreciation goes to Prof. Vaishali Koria and Prof. Aniruddh Fataniya with regards to his feedback across the course of project from the initial proposal up to the conclusion and for the valuable lessons learned along the way including collaboration within a group and the challenges involved in some large-scale software development efforts.

6.2 PROBLEM ENCOUNTERED AND THEIR SOLUTIONS

- **Lack of full knowledge about the technology**
  No clear ideas of how python Programming and data mining. After making python script how to make user friendly Interface. It can be achieved by tkinter GUI.
- **Optimization**
  How to write python script in efficient way. To solve this problem Practice is must.

6.3 SUMMARY OF PROJECT WORK

We presented results for sentiment analysis on Twitter. We use previously proposed state-of-the-art unigram model as our baseline and report for a 3-way positive versus negative versus neutral. We presented a comprehensive set of experiments for these task on manually annotated data that is a random sample of stream of tweets.

REFERENCES

PDF

- https://www.cs.uic.edu/~liub/FBS/sentiment-analysis.html

BLOG

- http://docs.tweepy.org/en/v3.5.0/api.html
- https://www.datacamp.com/community/tutorials/wordcloud-python
- https://github.com/vprusso/youtube_tutorials/blob/master/twitter_python/part_2_cursor_and_pagination/twitter_credentials.py

YOUTUBE

- https://www.youtube.com/watch?v=4Mf0h3HphEA&list=PLCg3kKKVIe7HbHliulxsDdJH2TePGDpnZ
- https://www.youtube.com/watch?v=wlnx-7cm4Gg
- https://www.youtube.com/watch?v=D8-snVfekto