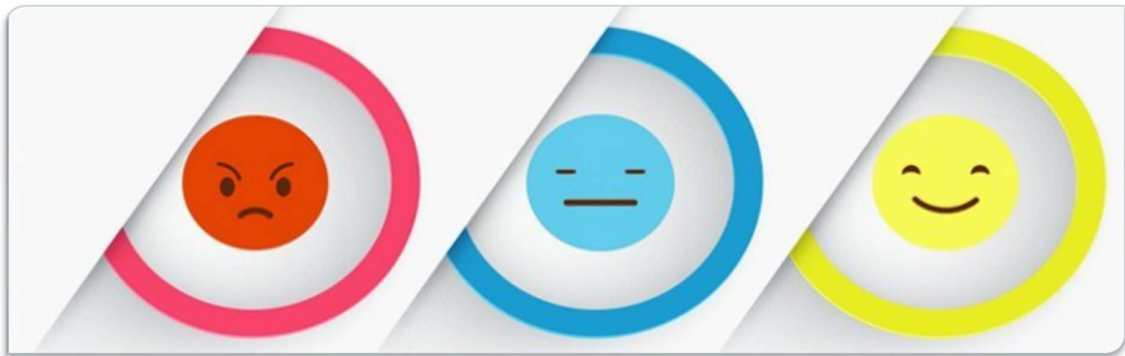


BERHAMPUR UNIVERSITY

BHANJA BIHAR – 760007

ODISHA



A Project report on

Sentiment analysis and image search using twitter

Under the guidance of

Prof. Manas Ranjan Patra

Presented by:

B Kewal

Roll No.: MCLE 2618

BERHAMPUR UNIVERSITY

This is to certify that the project work titled “Sentiment analysis and image search using twitter” submitted by Mr. B Kewal, Roll No.- MCLE 2618, Regd. No.-22349/2014, Student of “Berhampur University” has been examined by Berhampur University. It has been found fit and approved the award for the MCA degree.

Internal examiner

HOD

ACKNOWLEDGEMENT

It is a matter of great pleasure to present this progress report on development of “Sentiment Analysis and Image Search based on Twitter”. We are grateful to Dept. of Computer Science, Berhampur university for providing us with this great opportunity to develop a web application for the major project.

We are also grateful to our project supervisor Prof. Manas Ranjan Patra, Head of Department for his valuable advice and suggestion. We would also like to thanks all my teachers for providing us all the necessary resources that meets our project requirements.

We would like to convey our thanks to the teaching and non-teaching staffs of the Department of Computer Science, Berhampur University for their invaluable help and support throughout the period of the project hours. We will not miss to express our gratitude to all our friends and everyone who has been the part of this project by providing their comments and suggestions.

Group Members

N Smruti Ranjan Reddy

B Kewal

Nihal Kumar Nayak

DECLARATION

It is a great pleasure for me that our great personalities of my college, who helped a lot in completing and submitting the project. I am thankful to Berhampur University and especially to our project guide for allowing me to do this project entitled “**Sentiment Analysis and Image Search Based on Twitter**” in Python for fulfilment of “**MCA Final year**” Course. I am very grateful to Prof. Manas Ranjan Patra (H.O.D) for providing support and their kind co-operation for making this project successful.

At Last, from the bottom of my heart I am thankful to all of my faculties and my parents for their valuable inspiration during the preparation of my project work.

Name: B Kewal

Roll No.: MCLE 2618

Regd. No.: 22349/2014

ABSTRACT

Analysis of public information from social media could yield interesting results and insights into the world of public opinions about almost any product, service or personality. Social network data is one of the most effective and accurate indicators of public sentiment. The explosion of Web 2.0 has led to increased activity in Podcasting, Blogging, Tagging, Contributing to RSS, Social Bookmarking, and Social Networking. As a result, there has been an eruption of interest in people to mine these vast resources of data for opinions. Sentiment Analysis or Opinion Mining is the computational treatment of opinions, sentiments and subjectivity of text. In this project we will be discussing a methodology which allows utilization and interpretation of twitter data to determine public opinions.

Developing a program for sentiment analysis is an approach to be used to computationally measure customers' perceptions. This project reports on the design of a sentiment analysis, extracting and training a vast number of tweets. Results classify customers' perspective via tweets into positive, negative and neutral which is represented in a pie chart, bar diagram, etc using mat plot lib.

Image search is a specialized data search used to find images using web scrapping. To search for images, a user may provide query terms such as keyword, image file/link, or click on some image, and the system will return images "similar" to the query. The similarity used for search criteria could be meta tags, colour distribution in images, region/shape attributes, etc.

CONTENT

- ❖ Introduction
- ❖ Detailed description of the project
 - Twitter
 - Twitter API
 - Data mining
 - Opinion mining
 - Web scrapping
 - System overview
 - Python
 - Python modules
- ❖ System designing
- ❖ Screenshots
- ❖ Conclusion
- ❖ References

INTRODUCTION

What is sentiment analysis?

Sentiment Analysis is the process of ‘computationally’ determining whether a piece of writing is positive, negative or neutral. It’s also known as **opinion mining**, deriving the opinion or attitude of a speaker.

Why sentiment analysis?

- **Business:** In marketing field companies use it to develop their strategies, to understand customers’ feelings towards products or brand, how people respond to their campaigns or product launches and why consumers don’t buy some products.
- **Politics:** In political field, it is used to keep track of political view, to detect consistency and inconsistency between statements and actions at the government level. It can be used to predict election results as well!
- **Public Actions:** Sentiment analysis also is used to monitor and analyse social phenomena, for the spotting of potentially dangerous situations and determining the general mood of the blogosphere.

In the past few years, there has been a huge growth in the use of microblogging platforms such as Twitter. Spurred by that growth, companies and media organizations are increasingly seeking ways to mine Twitter for information about what people think and feel about their products and services. Companies such as Twitrratr (twitrratr.com), tweetfeel (www.tweetfeel.com), and Social Mention (www.socialmention.com) are just a few who advertise Twitter sentiment analysis as one of their services.

While there has been a fair amount of research on how sentiments are expressed in genres such as online reviews and news articles, how sentiments are expressed given the informal language and message-length constraints of microblogging has been much less studied. Features such as automatic part-of-speech tags and resources such as sentiment lexicons have proved useful for sentiment analysis in other domains, but will they also prove useful for sentiment analysis in Twitter?

Another challenge of microblogging is the incredible breadth of topic that is covered. It is not an exaggeration to say that people tweet about anything and

everything. Therefore, to be able to build systems to mine Twitter sentiment about any given topic, we need a method for quickly identifying data that can be used for training. In this project, we explore one method for building such data: using Twitter hashtags (e.g., #bestfeeling, #epicfail, #news) to identify positive, negative, and neutral tweets to use for training three-way sentiment classifiers. The online medium has become a significant way for people to express their opinions and with social media, there is an abundance of opinion information available. Using sentiment analysis, the polarity of opinions can be found, such as positive, negative, or neutral by analysing the text of the opinion. Sentiment analysis has been useful for companies to get their customer's opinions on their products predicting outcomes of elections, and getting opinions from movie reviews. The information gained from sentiment analysis is useful for companies making future decisions.

Many traditional approaches in sentiment analysis uses the bag of words method. The bag of words technique does not consider language morphology, and it could incorrectly classify two phrases of having the same meaning because it could have the same bag of words. The relationship between the collection of words is considered instead of the relationship between individual words. When determining the overall sentiment, the sentiment of each word is determined and combined using a function. Bag of words also ignores word order, which leads to phrases with negation in them to be incorrectly classified.

Sentiment analysis refers to the broad area of natural language processing which deals with the computational study of opinions, sentiments and emotions expressed in text. Sentiment Analysis (SA) or Opinion Mining (OM) aims at learning people's opinions, attitudes and emotions towards an entity. The entity can represent individuals, events or topics. An immense amount of research has been performed in the area of sentiment analysis. But most of them focused on classifying formal and larger pieces of text data like reviews. With the wide popularity of social networking and microblogging websites and an immense amount of data available from these resources, research projects on sentiment analysis have witnessed a gradual domain shift. The past few years have witnessed a huge growth in the use of microblogging platforms. Popular microblogging websites like Twitter have evolved to become a source of varied information. This diversity in the information owes to such microblogs being elevated as platforms where people post real time messages about their opinions on a wide variety of topics, discuss current affairs and share their experience on products and services they use in daily life. Stimulated by the growth of microblogging platforms, organizations are exploring ways to mine Twitter for information about how people are responding to their products and services. A fair amount of research has been carried out on how sentiments are expressed in formal text patterns such as product or movie reviews and news articles, but how

sentiments are expressed given the informal language and message-length constraints of microblogging has been less explored.

Twitter is an innovative microblogging service aired in 2006 with currently more than 550 million users. The user created status messages are termed tweets by this service. The public timeline of twitter service displays tweets of all users worldwide and is an extensive source of real-time information. The original concept behind microblogging was to provide personal status updates. But the current scenario surprisingly witnesses tweets covering everything under the world, ranging from current political affairs to personal experiences. Movie reviews, travel experiences, current events etc. add to the list. Tweets (and microblogs in general) are different from reviews in their basic structure. While reviews are characterized by formal text patterns and are summarized thoughts of authors, tweets are more casual and restricted to 140 characters of text. Tweets offer companies an additional avenue to gather feedback. Sentiment analysis to research products, movie reviews etc. aid customers in decision making before making a purchase or planning for a movie. Enterprises find this area useful to research public opinion of their company and products, or to analyse customer satisfaction. Organizations utilize this information to gather feedback about newly released products which supplements in improving further design. Different approaches which include Machine Learning (ML) techniques, sentiment lexicons, hybrid approaches etc. have been proved useful for sentiment analysis on formal texts. But their effectiveness for extracting sentiment in microblogging data will have to be explored. A careful investigation of tweets reveals that the 140-character length text restricts the vocabulary which imparts the sentiment. The hyperlinks often present in these tweets in turn restrict the vocabulary size. The varied domains discussed would surely impose hurdles for training. The frequency of misspellings and slang words in tweets (microblogs in general) is much higher than in other language resources which is another hurdle that needs to be overcome. On the other way around the tremendous volume of data available from microblogging websites on varied domains are incomparable with other data resources available. Microblogging language is characterized by expressive punctuations which convey a lot of sentiments. Bold lettered phrases, exclamations, question marks, quoted text etc. leave scope for sentiment extraction. The proposed work attempts a novel approach on twitter data by aggregating an adapted polarity lexicon which has learnt from product reviews of the domains under consideration, the tweet specific features and unigrams to build a classifier model using machine learning techniques.

Image Search using Twitter

Image search is a specialized data search used to find images using web scrapping. To search for images, a user may provide query terms such as keyword, image file/link, or click on some image, and the system will return images "similar" to the query. The similarity used for search criteria could be meta tags, colour distribution in images, region/shape attributes, etc.

A powerful social media web crawler/web scrapper that dumps images, tweets, captions, external links and hashtags from Instagram and Twitter in an organized form. It also shows the most relevant hashtags with their frequency of occurrence in the posts.

Web scraping, web harvesting, or web data extraction is data scraping used for extracting data from websites. Web scraping software may access the World Wide Web directly using the Hypertext Transfer Protocol, or through a web browser. While web scraping can be done manually by a software user, the term typically refers to automated processes implemented using a bot or web crawler. It is a form of copying, in which specific data is gathered and copied from the web, typically into a central local database or spreadsheet, for later retrieval or analysis.

Web scraping a web page involves fetching it and extracting from it. Fetching is the downloading of a page (which a browser does when you view the page). Therefore, web crawling is a main component of web scraping, to fetch pages for later processing. Once fetched, then extraction can take place. The content of a page may be parsed, searched, reformatted, its data copied into a spreadsheet, and so on. Web scrapers typically take something out of a page, to make use of it for another purpose somewhere else. An example would be to find and copy names and phone numbers, or companies and their URLs, to a list (contact scraping).

For scrapping Twitter, we need to setup a Twitter App. First of all, login from your Twitter account and go to Twitter Apps. Create a new app , go to Keys and access tokens and copy Consumer Key, Consumer Secret, Access Token and Access Token Secret.

Detailed Description of the project

Twitter

Twitter is an American microblogging and social networking service on which users post and interact with messages known as "tweets". Registered users can post, like, and retweet tweets, but unregistered users can only read them. Users access Twitter through its website interface, through Short Message Service (SMS) or its mobile-device application software. **Twitter, Inc.** is based in San Francisco, California, and has more than 25 offices around the world. Tweets were originally restricted to 140 characters, but was doubled to 280 for non-CJK languages in November 2017. Audio and video tweets remain limited to 140 seconds for most accounts.

Twitter was created by Jack Dorsey, Noah Glass, Biz Stone, and Evan Williams in March 2006 and launched in July of that year. By 2012, more than 100 million users posted 340 million tweets a day, and the service handled an average of 1.6 billion search queries per day. In 2013, it was one of the ten most-visited websites and has been described as "the SMS of the Internet. As of 2018, Twitter had more than 321 million monthly active users.

Tweets are publicly visible by default, but senders can restrict message delivery to only their followers. Users can tweet via the Twitter website, compatible external applications (such as for smartphones), or by Short Message Service (SMS) available in certain countries. Users may subscribe to other users' tweets—this is known as "following" and subscribers are known as "followers" or "tweeps", a portmanteau of Twitter and peeps. Individual tweets can be forwarded by other users to their own feed, a process known as a "retweet". Users can also "like" (formerly "favourite") individual tweets.

As a social network, Twitter revolves around the principle of followers. When you choose to follow another Twitter user, that user's tweets appear in reverse chronological order on your main Twitter page.

Users can group posts together by topic or type by use of hashtags – words or phrases prefixed with a "#" sign. Similarly, the "@" sign followed by a username is used for mentioning or replying to other users.

Twitter API

Twitter is what's happening in the world and what people are talking about right now. You can access Twitter via the web or your mobile device. To share information on Twitter as widely as possible, we also provide companies, developers, and users with programmatic access to Twitter data through our APIs (application programming interfaces). This article explains what Twitter's APIs are, what information is made available through them, and some of the protections Twitter has in place for their use.

At a high level, APIs are the way computer programs “talk” to each other so that they can request and deliver information. This is done by allowing a software application to call what's known as an endpoint: an address that corresponds with a specific type of information we provide (endpoints are generally unique like phone numbers). Twitter allows access to parts of our service via APIs to allow people to build software that integrates with Twitter, like a solution that helps a company respond to customer feedback on Twitter.

Twitter data is unique from data shared by most other social platforms because it reflects information that users choose to share publicly. Our API platform provides broad access to public Twitter data that users have chosen to share with the world. We also support APIs that allow users to manage their own non-public Twitter information (e.g., Direct Messages) and provide this information to developers whom they have authorized to do so.

Accessing Twitter Data

When someone wants to access our APIs, they are required to register an application. By default, applications can only access public information on Twitter. Certain endpoints, such as those responsible for sending or receiving Direct Messages, require additional permissions from you before they can access your information. These permissions are not granted by default; you choose on a per-application basis whether to provide this access, and can control all the applications authorized on your account.

The Twitter APIs include a wide range of endpoints, which fall into five primary groups:

Accounts and users: We allow developers to programmatically manage an account's profile and settings, mute or block users, manage users and followers, request information about an authorized account's activity, and more. These endpoints can help citizen services like the Commonwealth of Virginia's

Department of Emergency Management which provides information to residents about emergency responses and emergency alerts.

Tweets and replies: We make public Tweets and replies available to developers, and allow developers to post Tweets via our API. Developers can access Tweets by searching for specific keywords, or requesting a sample of Tweets from specific accounts.

These endpoints are used by NGOs like the UN to identify, understand and counter misinformation around public health initiatives. Similarly, we help researchers listen for early symptomatic indications of disease outbreaks and monitor their spread. A team at North-eastern University recently developed a new flu-tracking technique using Twitter data that forecasts outbreaks up to six weeks in advance — much earlier than many other models, without sacrificing accuracy.

Direct Messages: Our Direct Message endpoints provide access to the DM conversations of users who have explicitly granted permission to a specific application. We do not sell Direct Messages. Our DM APIs provide limited access to developers to create personalized experiences on Twitter, like Wendy's March Madness bracket builder. For accounts they own or manage, businesses can create these human- or chatbot-powered conversational experiences to communicate directly with customers for customer service, marketing, and brand engagement experiences.

Ads: We provide a suite of APIs to let developers, like Sprinklr, help businesses automatically create and manage ad campaigns on Twitter. Developers can use public Tweets to identify topics and interests, and provide businesses with tools for running advertising campaigns to reach the diverse audiences on Twitter.

Publisher tools and SDKs: We provide tools for software developers and publishers to embed Twitter timelines, share buttons, and other Twitter content on webpages. These tools allow brands to bring live, public conversations from Twitter into their web experience and make it easy for their customers to share information and articles from their sites.

Across all of our APIs and data products, we take our responsibility to protect our users' data seriously. We maintain strict policies and processes to assess how developers are using Twitter data, and restrict improper use of this data. When we learn that a developer violates our policies, we will take appropriate action, which can include suspension and termination of access to Twitter's APIs and data products.

Data Mining

Data mining is a process of discovering patterns in large data sets involving methods at the intersection of machine learning, statistics, and database systems. Data mining is an interdisciplinary subfield of computer science and statistics with an overall goal to extract information from a data set and transform the information into a comprehensible structure for further use. Data mining is the analysis step of the "knowledge discovery in databases" process. Aside from the raw analysis step, it also involves database and data management aspects, data pre-processing, model and inference considerations, interestingness metrics, complexity considerations, post-processing of discovered structures, visualization, and online updating.

The term "data mining" is a misnomer, because the goal is the extraction of patterns and knowledge from large amounts of data, not the extraction (*mining*) of data itself. It also is a buzzword and is frequently applied to any form of large-scale data or information processing (collection, extraction, warehousing, analysis, and statistics) as well as any application of computer decision support system, including artificial intelligence (e.g., machine learning) and business intelligence.

The actual data mining task is the semi-automatic or automatic analysis of large quantities of data to extract previously unknown, interesting patterns such as groups of data records (cluster analysis), unusual records (anomaly detection), and dependencies (association rule mining, sequential pattern mining). This usually involves using database techniques such as spatial indices. These patterns can then be seen as a kind of summary of the input data, and may be used in further analysis

Opinion Mining

Opinion Mining (also known as Sentiment Analysis or **emotion AI**) 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 widely applied to voice of the customer materials such as reviews and survey responses, online and social media, and healthcare materials for applications that range from marketing to customer service to clinical medicine.

A basic task in opinion mining is classifying the *polarity* of a given text at the document, sentence, or feature/aspect level—whether the expressed opinion in a document, a sentence or an entity feature/aspect is positive, negative, or neutral. Advanced, "beyond polarity" sentiment classification looks, for instance, at emotional states such as "angry", "sad", and "happy".

Web Scrapping

A powerful social media web crawler/web scrapper that dumps images, tweets, captions, external links and hashtags from Instagram and Twitter in an organized form. It also shows the most relevant hashtags with their frequency of occurrence in the posts.

Web scraping, web harvesting, or web data extraction is data scraping used for extracting data from websites. Web scraping software may access the World Wide Web directly using the Hypertext Transfer Protocol, or through a web browser. While web scraping can be done manually by a software user, the term typically refers to automated processes implemented using a bot or web crawler. It is a form of copying, in which specific data is gathered and copied from the web, typically into a central local database or spreadsheet, for later retrieval or analysis.

Web scraping a web page involves fetching it and extracting from it. Fetching is the downloading of a page (which a browser does when you view the page). Therefore, web crawling is a main component of web scraping, to fetch pages for later processing. Once fetched, then extraction can take place. The content of a page may be parsed, searched, reformatted, its data copied into a spreadsheet, and so on. Web scrapers typically take something out of a page, to make use of it for another purpose somewhere else. An example would be to find and copy names and phone numbers, or companies and their URLs, to a list (contact scraping).

For scrapping Twitter, we need to setup a Twitter App. First of all, login from your Twitter account and go to Twitter Apps. Create a new app , go to Keys and access tokens and copy Consumer Key, Consumer Secret, Access Token and Access Token Secret.

Web Scraping (also termed Screen Scraping, Web Data Extraction, Web Harvesting etc.) is a technique employed to extract large amounts of data from websites whereby the data is extracted and saved to a local file in your computer or to a database in table (spreadsheet) format.

Data displayed by most websites can only be viewed using a web browser. They do not offer the functionality to save a copy of this data for personal use. The only option then is to manually copy and paste the data - a very tedious job which can take many hours or sometimes days to complete. Web Scraping is the technique of automating this process, so that instead of manually copying the data from websites, the Web Scraping software will perform the same task within a fraction of the time.

Web pages are built using text-based mark-up languages (HTML and XHTML), and frequently contain a wealth of useful data in text form. However, most web pages are designed for human end-users and not for ease of automated use. As a result, specialized tools and software have been developed to facilitate the scraping of web pages.

Uses of Web Scrapping

1. Extract product details including price, images etc. from eCommerce websites for populating other websites, competition monitoring etc.
2. Extract business contact details including name, address, email, phone, website etc. from Yellow Pages, Google Maps etc. for marketing and lead generation.
3. Extract property details as well as agent contact details from real estate websites.
4. Price Comparison: Services such as Parse Hub use web scraping to collect data from online shopping websites and use it to compare the prices of products.
5. Email address gathering: Many companies that use email as a medium for marketing, use web scraping to collect email ID and then send bulk emails.
6. Social Media Scraping: Web scraping is used to collect data from Social Media websites such as Twitter to find out what's trending.
7. Research and Development: Web scraping is used to collect a large set of data (Statistics, General Information, Temperature, etc.) from websites, which are analysed and used to carry out Surveys or for R&D.
8. Job listings: Details regarding job openings, interviews are collected from different websites and then listed in one place so that it is easily accessible to the user.

Statement of the problem

The problem at hand consists of two subtasks:

- ☐ Phrase Level Sentiment Analysis in Twitter:

Given a message containing a marked instance of a word or a phrase, determine whether that instance is positive, negative or neutral in that context.

- ☐ Sentence Level Sentiment Analysis in Twitter:

Given a message, decide whether the message is of positive, negative, or neutral sentiment. For messages conveying both a positive and negative sentiment, whichever is the stronger sentiment should be chosen.

Objectives

The objectives of this project are:

- ☐ To implement an algorithm for automatic classification of text into positive, negative and neutral

- ☐ Sentiment Analysis to determine the attitude of the mass is positive, negative or neutral towards the subject of interest

- ☐ Graphical representation of the sentiment in form of Pie-Chart, Bar Diagram, etc.

Scope of project

This project will be helpful to the companies, political parties as well as to the common people. It will be helpful to political party for reviewing about the program that they are going to do or the program that they have performed. Similarly, companies also can get review about their new product on newly released hardware's or software's. Also, the movie maker can take review on the currently running movie. By analysing the tweets analyser can get result on how positive or negative or neutral are peoples about it.

System Overview

This proposal entitled “Sentiment analysis and image search based on Twitter” is a web application which is used to analyse the tweets. We will be performing sentiment analysis in tweets and determine where it is positive, negative or neutral. This web application can be used by any organization office to review their works or by political leaders or by any others company to review about their products or brands.

System Features

The main feature of our web application is that it helps to determine the opinion about the peoples on products, government work, politics or any other by analysing the tweets. Our system is capable of training the new tweets taking reference to previously trained tweets. The computed or analysed data will be represented in various diagram such as Pie chart, Bar graph, etc.

Python

Python is an interpreted, high-level, general-purpose programming language. Created by Guido van Rossum and first released in 1991, Python's design philosophy emphasizes code readability with its notable use of significant whitespace. Its language constructs and object-oriented approach aim to help programmers write clear, logical code for small and large-scale projects.

Python is dynamically typed and garbage-collected. It supports multiple programming paradigms, including structured object-oriented, and functional programming. Python is often described as a "batteries included" language due to its comprehensive standard library.

Python is an interpreted, object-oriented, high-level programming language with dynamic semantics. Its high-level built in data structures, combined with dynamic typing and dynamic binding, make it very attractive for Rapid Application Development, as well as for use as a scripting or glue language to connect existing components together. Python's simple, easy to learn syntax emphasizes readability and therefore reduces the cost of program maintenance. Python supports modules and packages, which encourages program modularity and code reuse. The Python interpreter and the extensive standard library are available in source or binary form without charge for all major platforms, and can be freely distributed.

Often, programmers fall in love with Python because of the increased productivity it provides. Since there is no compilation step, the edit-test-debug cycle is incredibly fast. Debugging Python programs is easy: a bug or bad input will never cause a segmentation fault. Instead, when the interpreter discovers an error, it raises an exception. When the program doesn't catch the exception, the interpreter prints a stack trace. A source level debugger allows inspection of local and global variables, evaluation of arbitrary expressions, setting breakpoints, stepping through the code a line at a time, and so on. The debugger is written in Python itself, testifying to Python's introspective power. On the other hand, often the quickest way to debug a program is to add a few print statements to the source: the fast edit-test-debug cycle makes this simple approach very effective.

Python provides number of modules and libraries which makes it easier to develop an application in a simple way. In this project we have used different python modules like tweepy, oauthhandler, textblob, matplotlib, tkinter, etc...

Tweepy:

Tweepy supports accessing Twitter via Basic Authentication and the newer method, OAuth. Twitter has stopped accepting Basic Authentication so OAuth is now the only way to use the Twitter API.

The main difference between Basic and OAuth authentication are the consumer and access keys. With Basic Authentication, it was possible to provide a username and password and access the API, but since 2010 when the Twitter started requiring OAuth, the process is a bit more complicated. An app has to be created at dev.twitter.com.

OAuth is a bit more complicated initially than Basic Auth, since it requires more effort, but the benefits it offers are very lucrative:

- Tweets can be customized to have a string which identifies the app which was used.
- It doesn't reveal user password, making it more secure.
- It's easier to manage the permissions, for example a set of tokens and keys can be generated that only allows reading from the timelines, so in case someone obtains those credentials, he/she won't be able to write or send direct messages, minimizing the risk.
- The application doesn't rely on a password, so even if the user changes it, the application will still work.

After logging in to the portal, and going to "Applications", a new application can be created which will provide the needed data for communicating with Twitter API.

Tweepy provides access to the well documented Twitter API. With tweepy, it's possible to get any object and use any method that the official Twitter API offers. For example, a User object has its documentation at <https://dev.twitter.com/docs/platform-objects/users> and following those guidelines, tweepy can get the appropriate information.

TextBlob:

TextBlob is a Python (2 and 3) library for processing textual data. It provides a simple API for diving into common natural language processing (NLP) tasks such as part-of-speech tagging, noun phrase extraction, sentiment analysis, classification, translation, and more.

It does have a lot of features like:

- Noun phrase extraction
- Part-of-speech tagging
- Sentiment analysis
- Classification (Naive Bayes, Decision Tree)
- Tokenization (splitting text into words and sentences)
- Word and phrase frequencies
- Parsing
- **n**-grams
- Word inflection (pluralization and singularization) and lemmatization
- Spelling correction
- Add new models or languages through extensions
- WordNet integration

Matplotlib:

Matplotlib is a comprehensive library for creating static, animated, and interactive visualizations in Python. Matplotlib produces publication-quality figures in a variety of hardcopy formats and interactive environments across platforms. Matplotlib can be used in Python scripts, the Python and IPython shell, web application servers, and various graphical user interface toolkits. It is a cross-platform library for making 2D plots from data in arrays.

Matplotlib is a plotting library for the Python programming language and its numerical mathematics extension NumPy. It provides an object-oriented API for embedding plots into applications using general-purpose GUI toolkits like Tkinter, wxPython. Using this we can build multiple graphical plots like bar graph, pie chart, histogram, line plot, box plot, etc...

Tkinter:

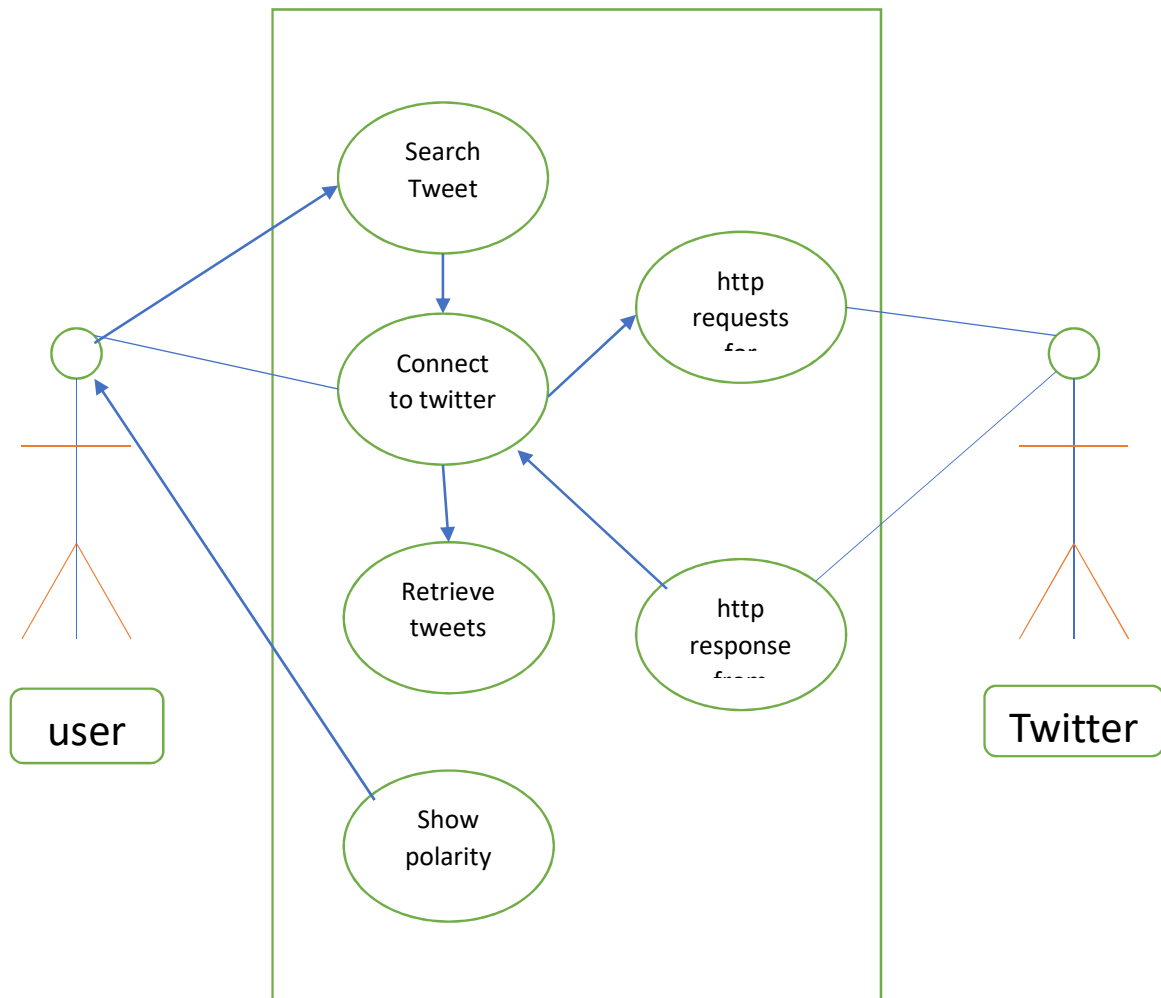
Tkinter is the standard GUI library for Python. Python when combined with Tkinter provides a fast and easy way to create GUI applications. Tkinter provides a powerful object-oriented interface to the Tk GUI toolkit.

As with most other modern Tk bindings, Tkinter is implemented as a Python wrapper around a complete Tcl interpreter embedded in the Python interpreter. Tkinter calls are translated into Tcl commands which are fed to this embedded interpreter, thus making it possible to mix Python and Tcl in a single application.

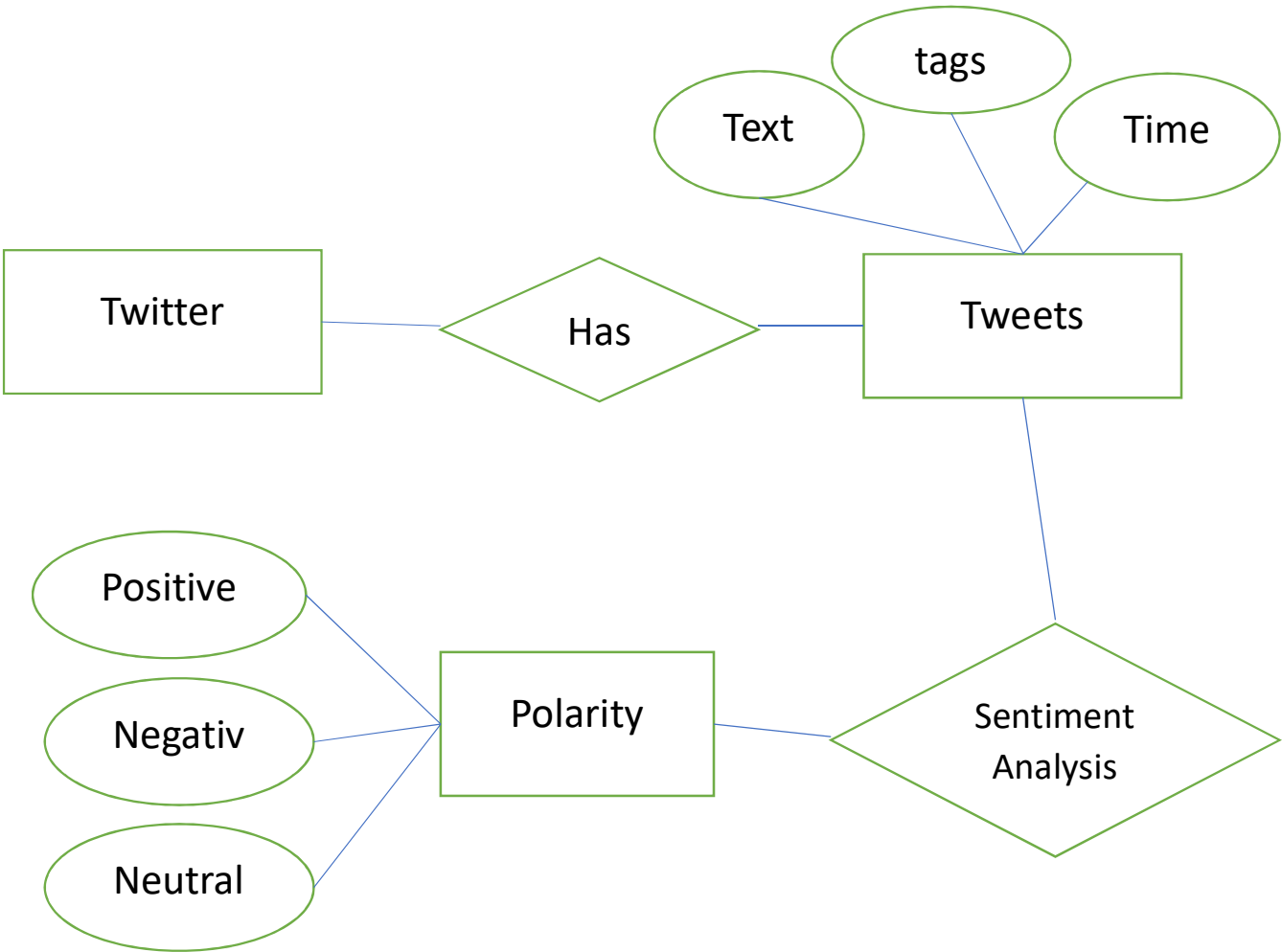
Tkinter provides various controls, such as buttons, labels and text boxes used in a GUI application. These controls are commonly called widgets. All Tkinter widgets have access to specific geometry management methods, which have the purpose of organizing widgets throughout the parent widget area.

System Designing

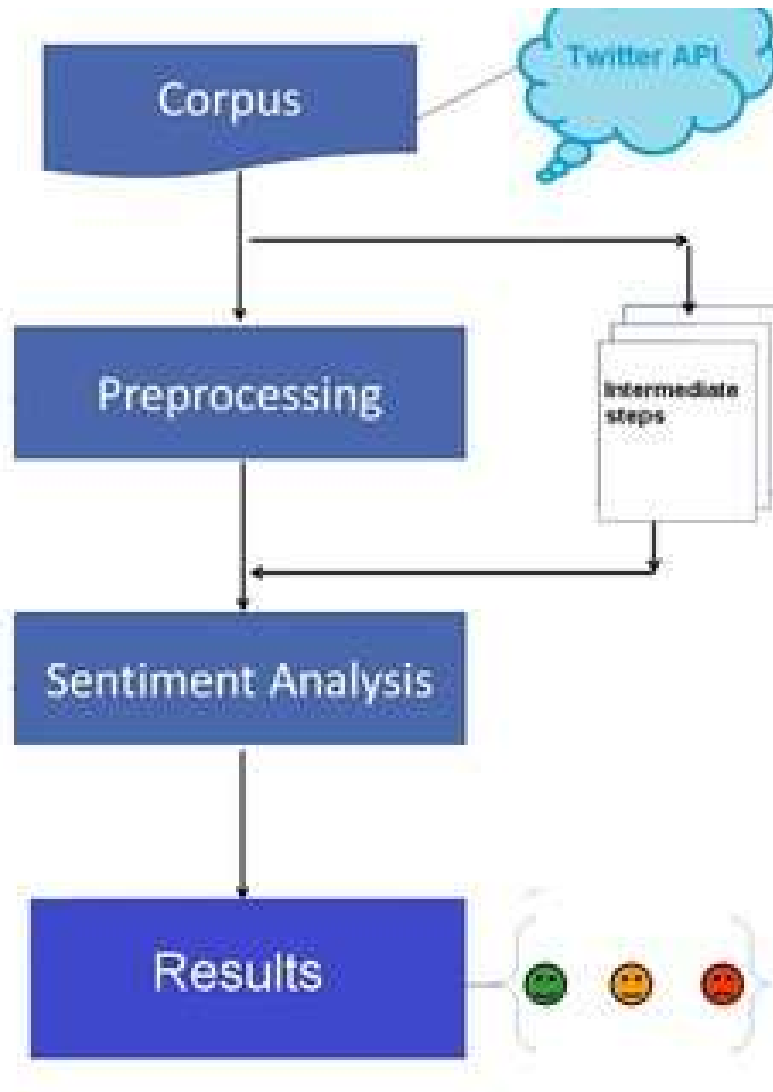
1. Use Case Diagram



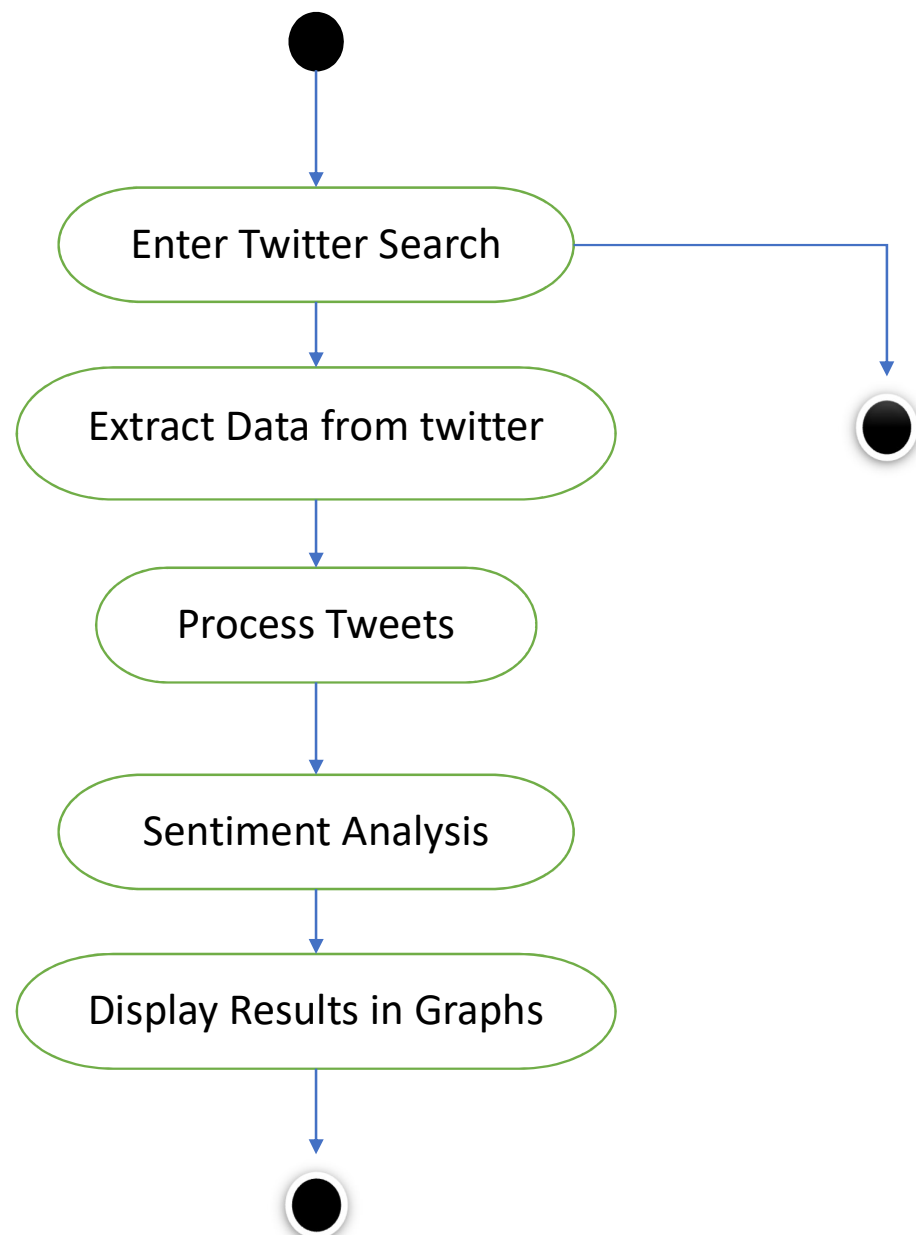
2. Entity Relationship Diagram



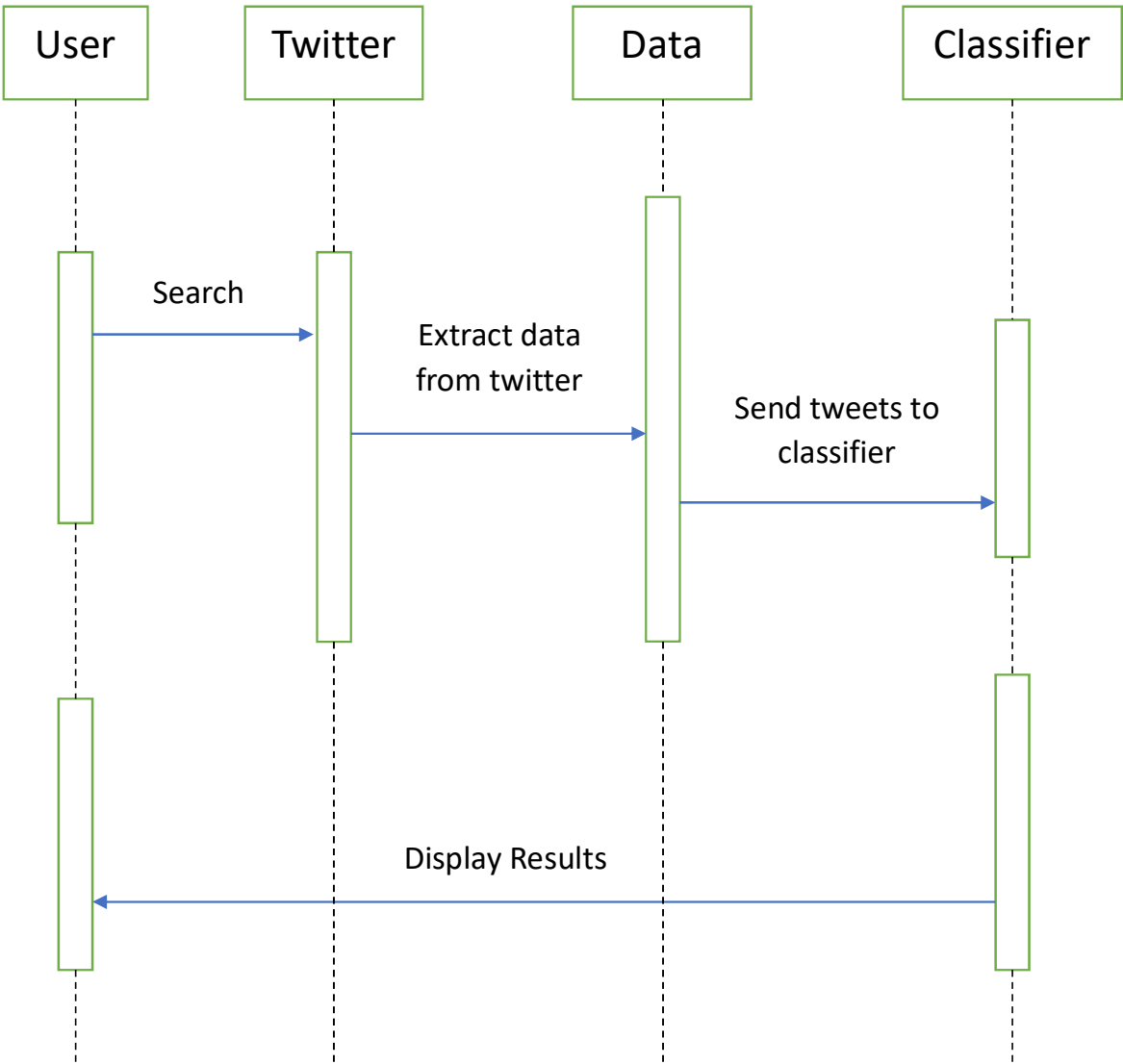
3. Class Diagram



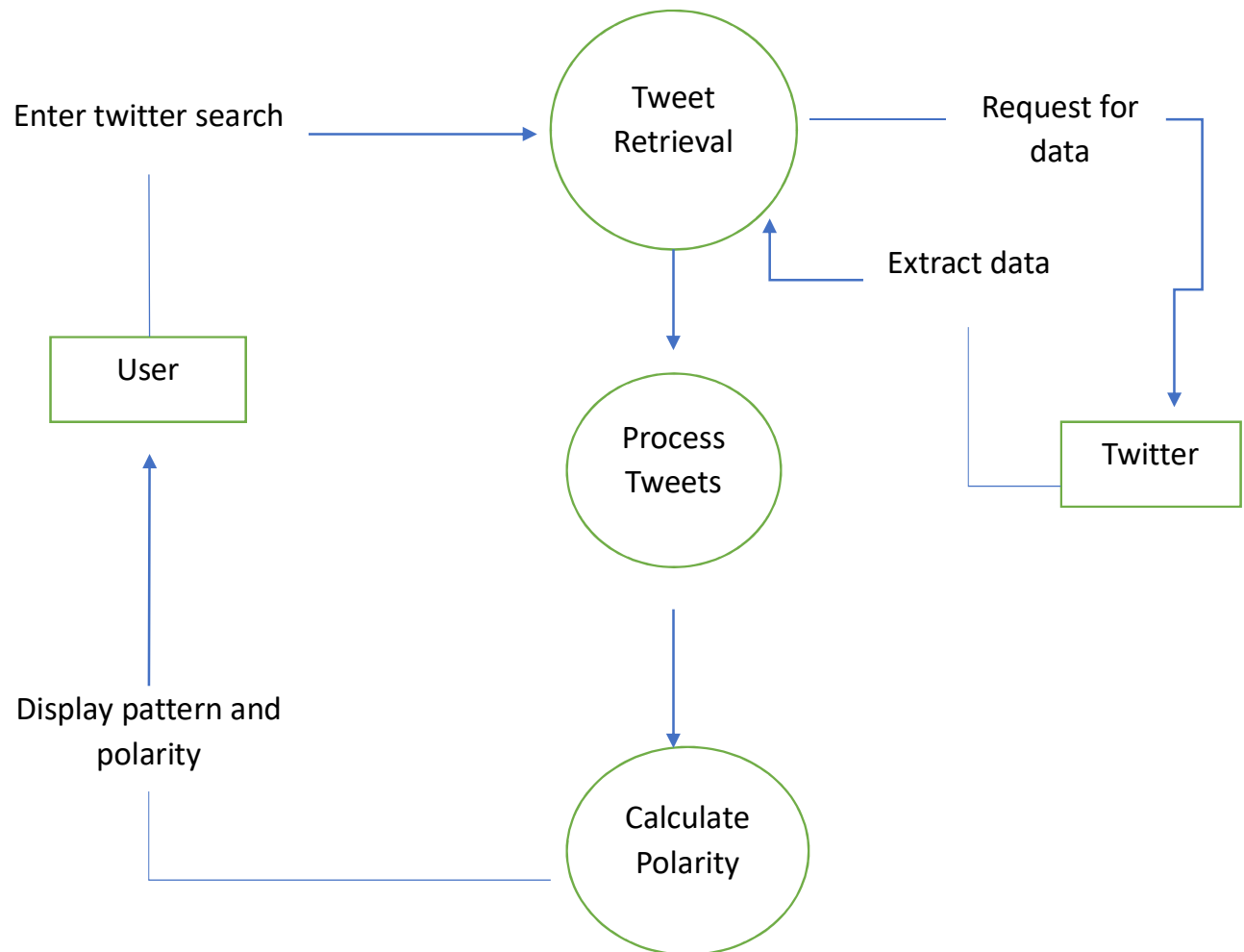
4. Activity Diagram



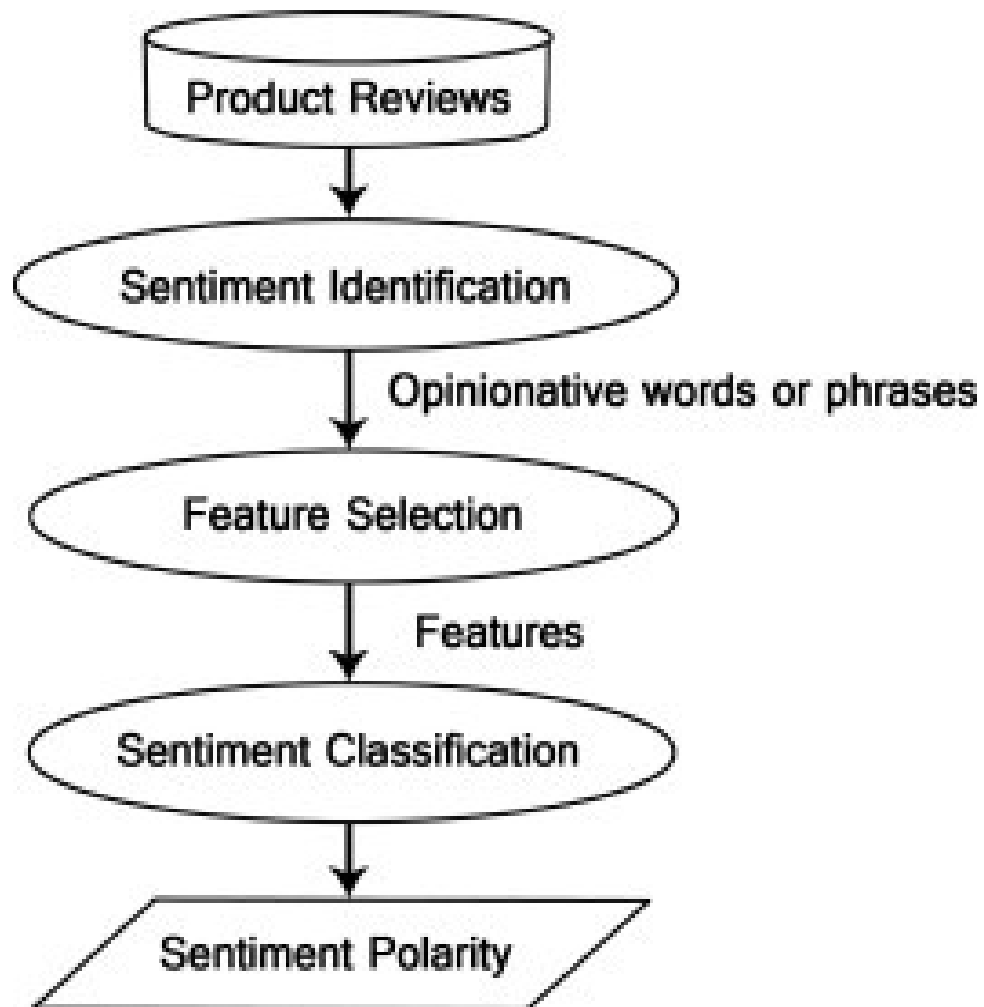
5. Sequence Diagram



6. Data Flow Diagram



7. Flow Diagram



Project Screenshots

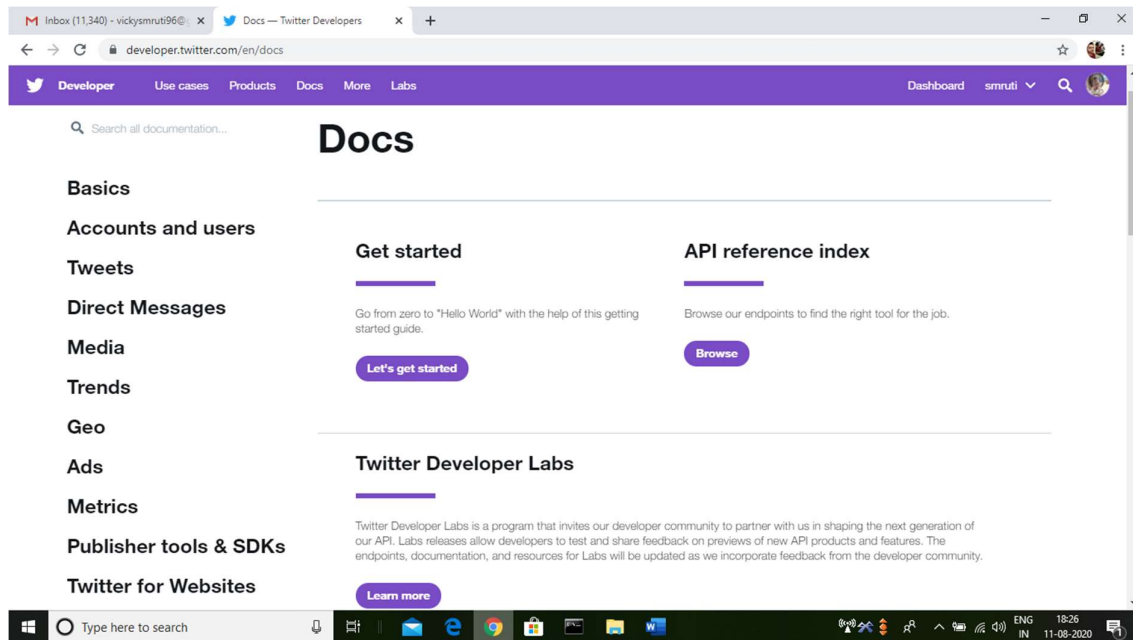


Fig. 1 - Twitter API

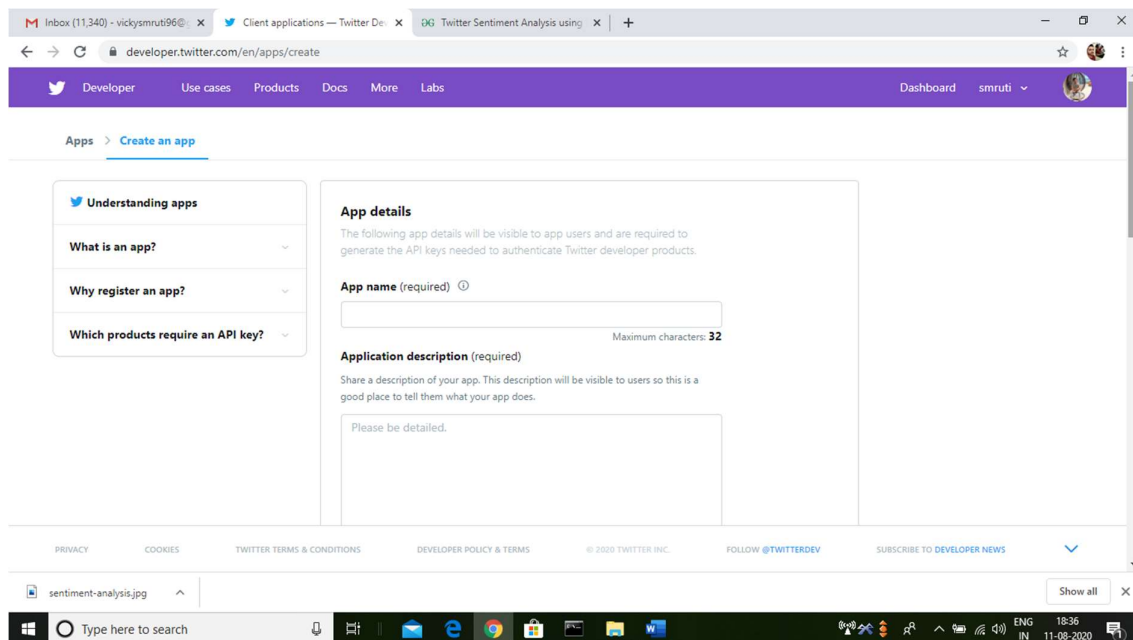


Fig. 2 – Create an app using twitter API

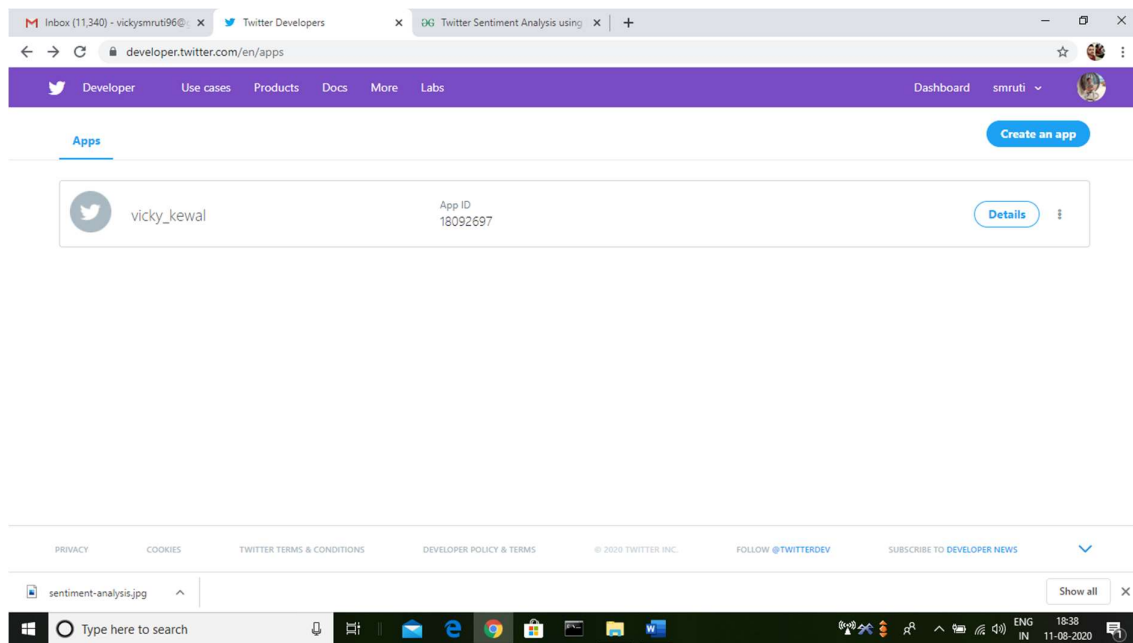


Fig. 3 – Application created

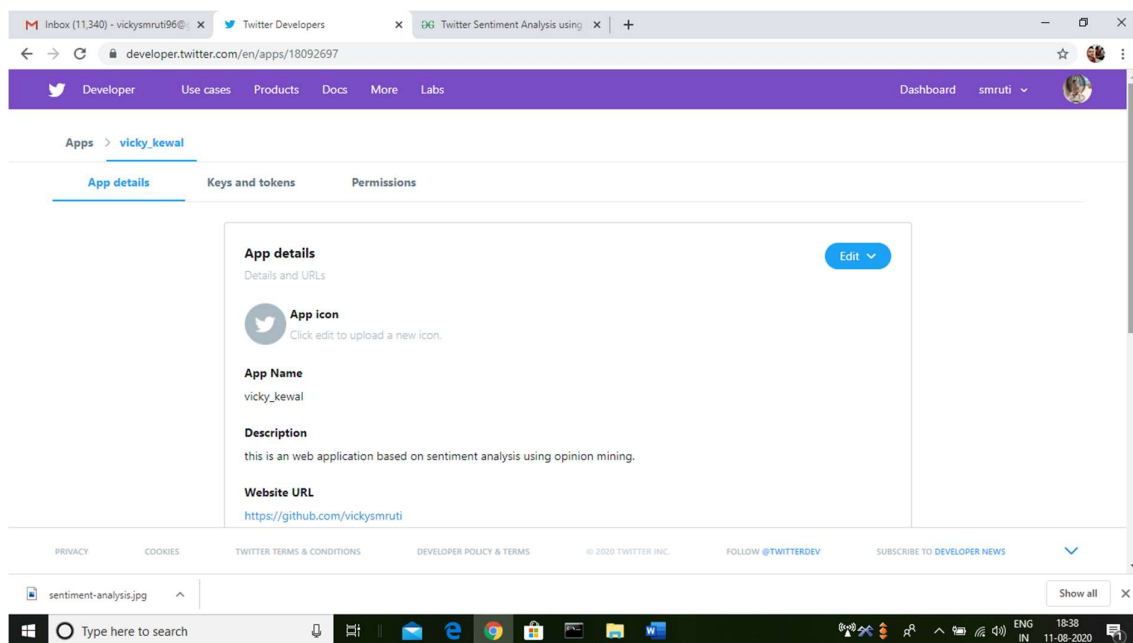


Fig. 4 – Details of the application

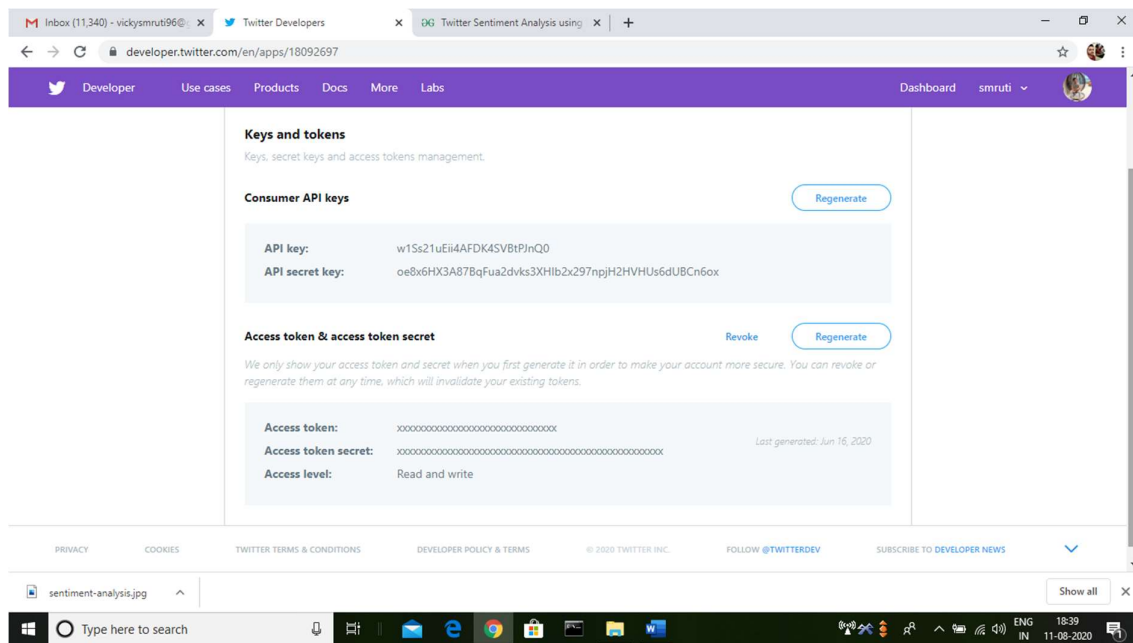


Fig. 5 – Keys and access tokens

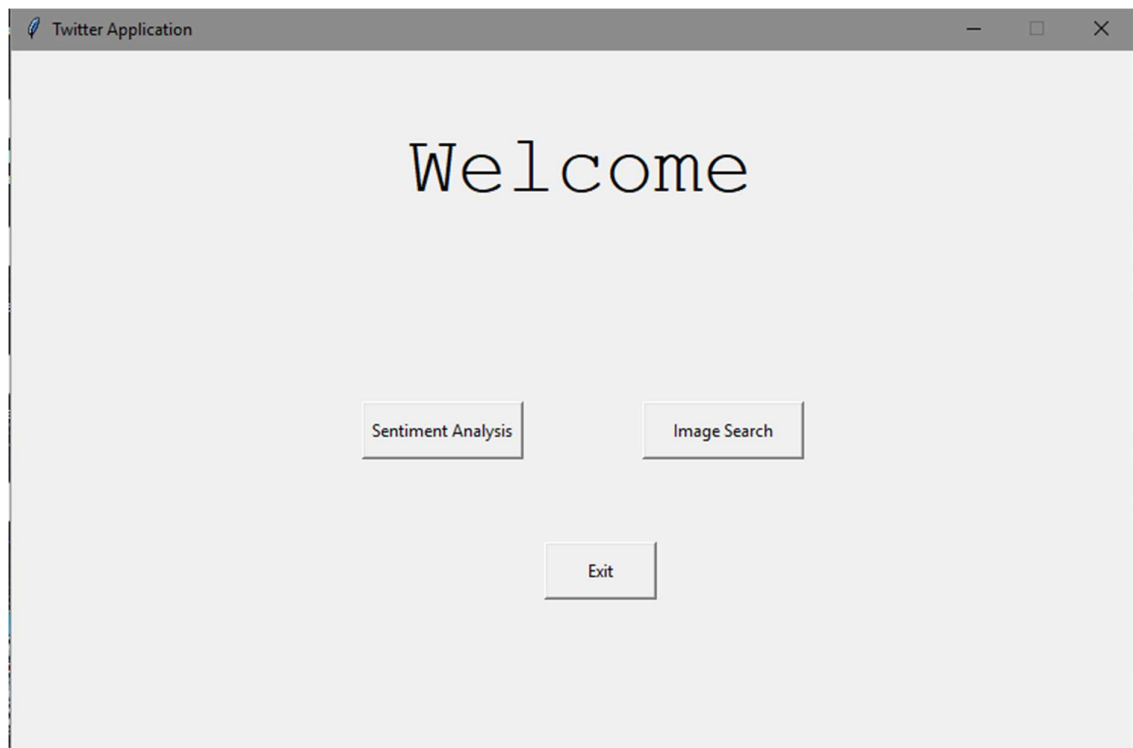


Fig. 6 – Welcome window

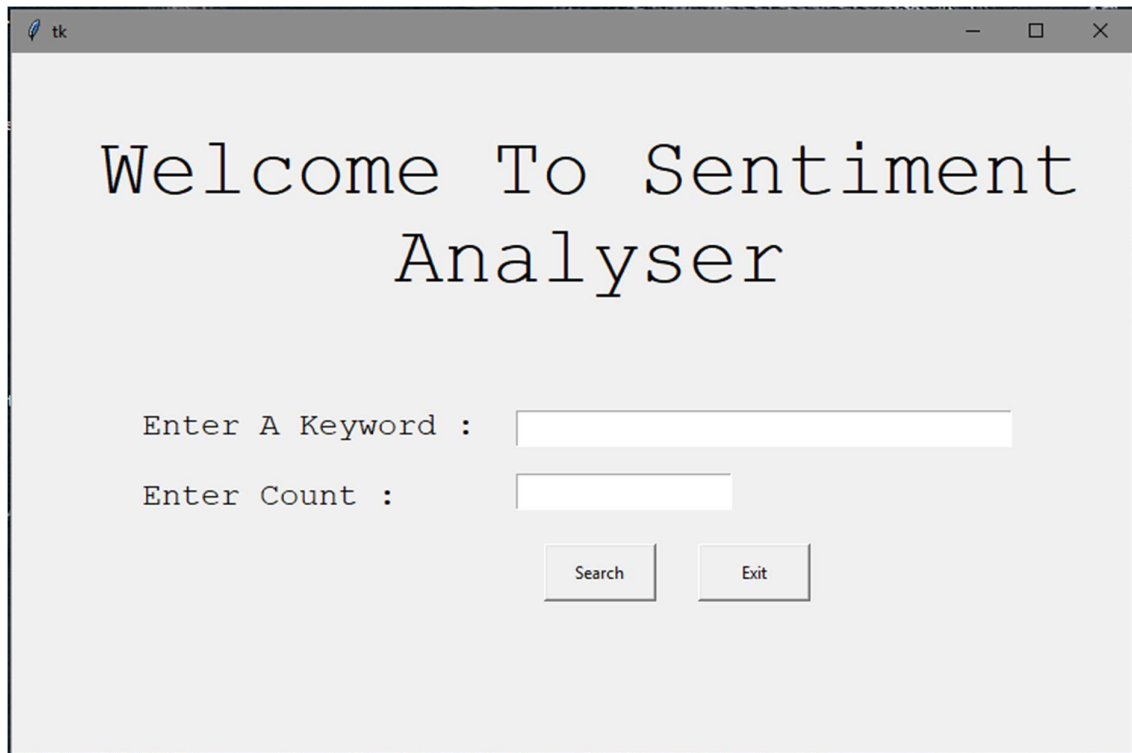


Fig. 7 – Sentiment analysis window

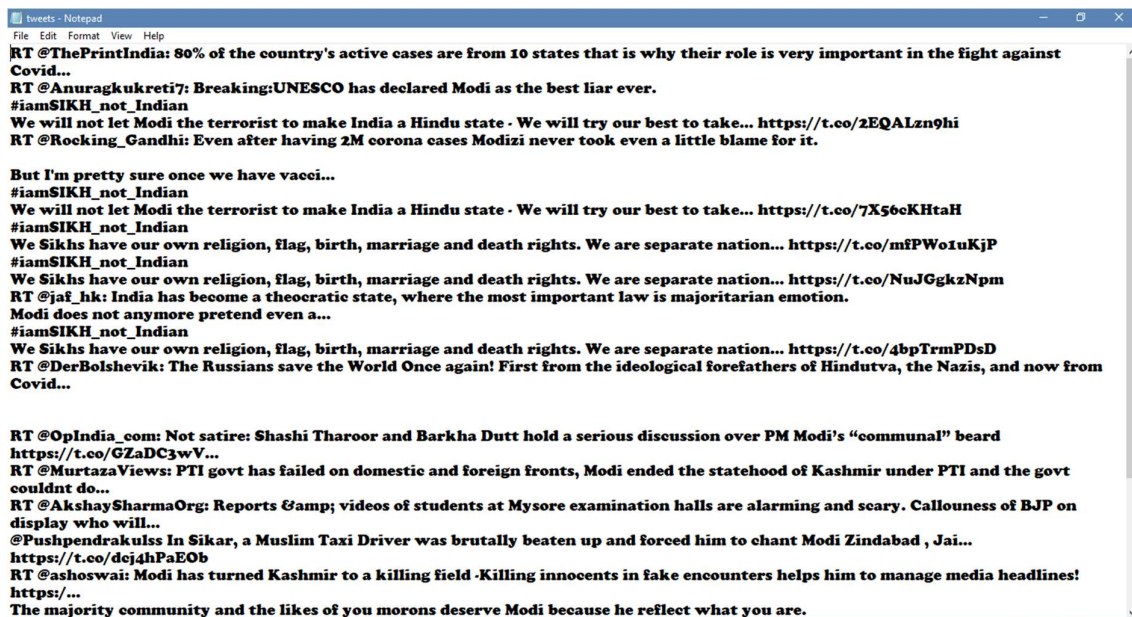


Fig. 8 – Tweets window

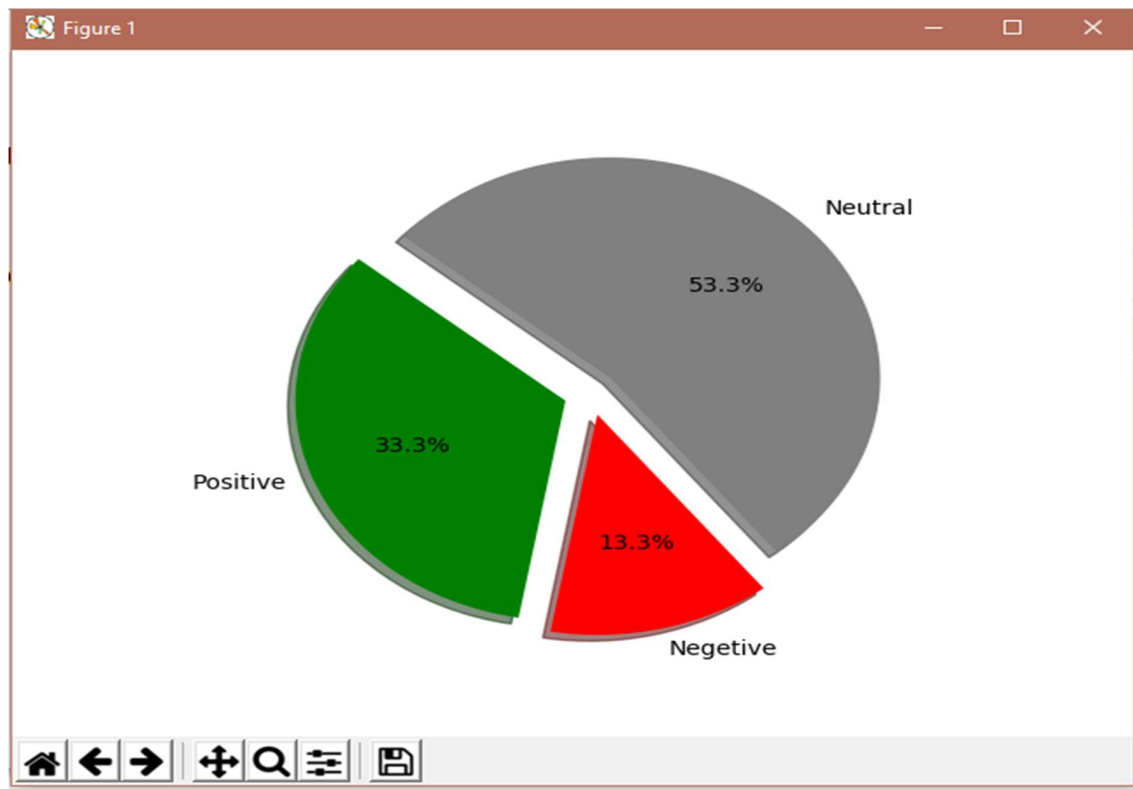


Fig. 9 – Graph based on polarity

A Tkinter window titled "Tag Based Image Search" is shown. It has a light blue background. The title is in a large, black, serif font. Below the title are two input fields: "Enter Image Keyword :" and "Enter Limit :". At the bottom, there are two buttons: "Search" and "Exit".

Fig. 10 – Image search window

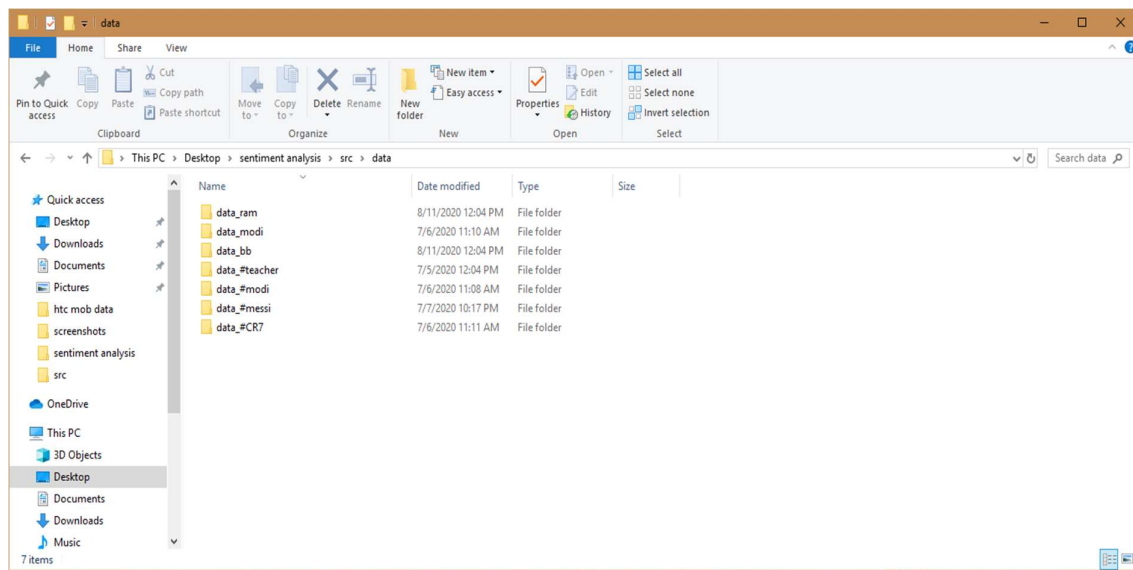


Fig. 11 – Image search folder view

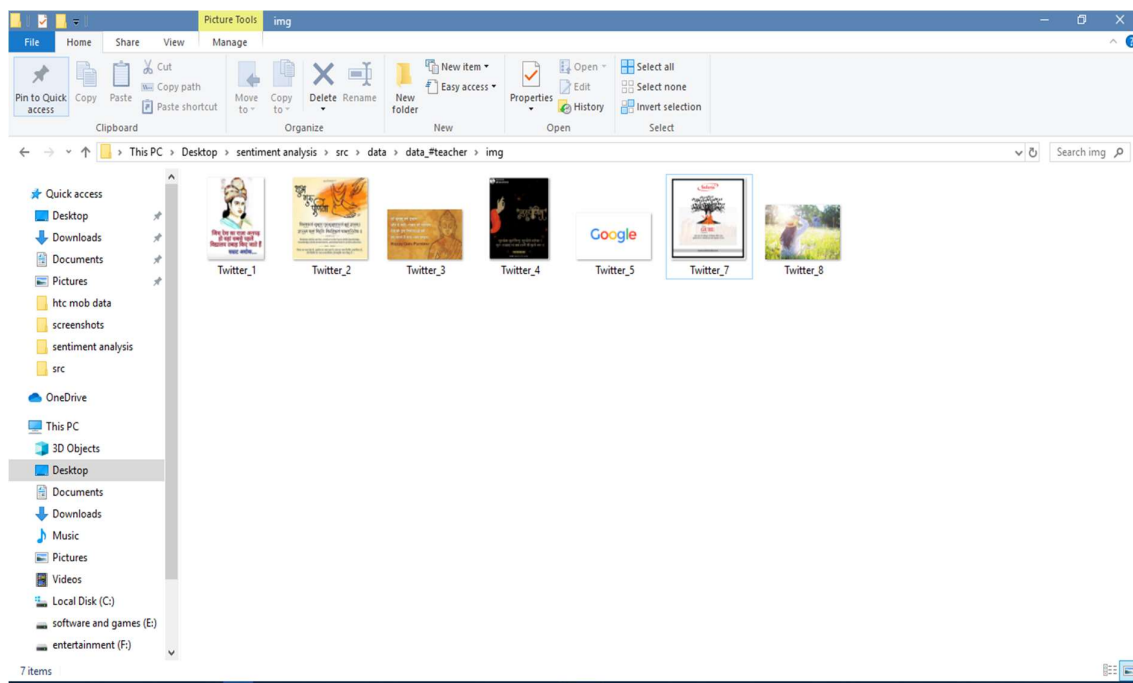


Fig. 12 – Image search inside view(images)

Conclusion

Nowadays, sentiment analysis or opinion mining is a hot topic in machine learning. We are still far to detect the sentiments of a corpus of texts very accurately because of the complexity in the English language and even more if we consider other languages such as Hindi.

The task of sentiment analysis, especially in the domain of micro-blogging, is still in the developing stage and far from complete. So, we propose a couple of ideas which we feel are worth exploring in the future and may result in further improved performance.

Right now, we have worked with only the very simplest unigram models; we can improve those models by adding extra information like closeness of the word with a negation word. We could specify a window prior to the word (a window could for example be of 2 or 3 words) under consideration and the effect of negation may be incorporated into the model if it lies within that window. The closer the negation word is to the unigram word whose prior polarity is to be calculated, the more it should affect the polarity. For example, if the negation is right next to the word, it may simply reverse the polarity of that word and farther the negation is from the word the more minimized its effect should be.

Apart from this, we are currently only focusing on unigrams and the effect of bigrams and trigrams may be explored. As reported in the literature review section when bigrams are used along with unigrams this usually enhances performance. However for bigrams and trigrams to be an effective feature we need a much more labelled data set than our meagre 9,000 tweets.

Right now, we are exploring Parts of Speech separate from the unigram models, we can try to incorporate POS information within our unigram models in future. So, say instead of calculating a single probability for each word like $P(\text{word} | \text{obj})$ we could instead have multiple probabilities for each according to the Part of Speech the word belongs to. For example we may have $P(\text{word} | \text{obj}, \text{verb})$, $P(\text{word} | \text{obj}, \text{noun})$ and $P(\text{word} | \text{obj}, \text{adjective})$. Pang et al. [5] used a somewhat similar approach and claims that appending POS information for every unigram result in no significant change in performance (with Naive Bayes performing slightly better and SVM having a slight decrease in performance), while there is a significant decrease in accuracy if only adjective unigrams are used as features. However, these results are for classification of reviews and may be verified for sentiment analysis on micro blogging websites like Twitter.

In this research we are focussing on general sentiment analysis. There is potential of work in the field of sentiment analysis with partially known context. For example, we noticed that users generally use our website for specific types of keywords which can be divided into a couple of distinct classes, namely: politics/politicians, celebrities, products/brands, sports/sportsmen, media/movies/music. So, we can attempt to perform separate sentiment analysis on tweets that only belong to one of these classes (i.e. the training data would not be general but specific to one of these categories) and compare the results we get if we apply general sentiment analysis on it instead.

Image search: As the Internet has grown astronomically and businesses have become increasingly dependent on data, it is now a compulsion to have access to the latest data on every given subject.

Data has become the basis of all decision-making processes whether it's a business or a non-profit organization. Therefore, web scraping has found its applications in every endeavour of note in contemporary times.

It is also becoming increasingly clear that those who will make creative and advanced use of web scraping tool will race ahead of others and gain a competitive advantage. So, leverage web scraping and boost your prospects in your chosen area of endeavour!

The future of web data scraping is indeed bright and for a simple reason that it will become even more essential for one and all as time passes by. The web has transformed into a gigantic mechanism absorbing data in the past 2 years. Statistics vouch for that. 90 percent of data has piled up within the last 2 years makes it crystal clear that contemporary business strategies are data-driven. The ever-increasing use of mobile and hand-held devices backed by faster internet speeds and cheaper access to the internet, all are contributing to increasing popularity of Web Scraping Companies in India.

Reference

1. <https://developer.twitter.com/en/docs>
2. <https://help.twitter.com/en/rules-and-policies/twitter-api>
3. <https://www.geeksforgeeks.org/>
4. <https://www.geeksforgeeks.org/twitter-sentiment-analysis-using-python/>
5. <https://en.wikipedia.org/wiki/>
6. https://www.tutorialspoint.com/python/python_gui_programming.htm
7. <http://www.dataisolutions.com/future-of-web-scraping/>
8. [https://textblob.readthedocs.io/en/dev/#:~:text=TextBlob%20is%20a%20Python%20\(2,classifi,cation%2C%20translation%2C%20and%20more.](https://textblob.readthedocs.io/en/dev/#:~:text=TextBlob%20is%20a%20Python%20(2,classifi,cation%2C%20translation%2C%20and%20more.)
9. <https://pypi.org/project/matplotlib/>

