“Sentimental Analysis on Twitter tweets “

A Socially Relevant Project- II Report submitted to

JAWAHARLAL NEHRU TECHNOLOGICAL UNVERSITY ANANTAPUR.

In Partial Fulfillment of the Requirements for the Award of thedegree of

BACHELOR OF TECHNOLOGY
IN
COMPUTER SCIENCE AND ENGINEERING
BY

G .Naga Eshitha	**19121A0565**
C .Anusha	**19121A0538**
Abdul Bari Shaik	**19121A0502**
G .Hemanth Kumar	**19121A0558**
B .Nagendra Naik	**19121A0524**

Under the Guidance of

Dr. K. Reddy Madhavi

Professor
Dept of CSE, SVEC

Department of Computer Science and Engineering(Affiliated
SREE VIDYANIKETHAN ENGINEERING COLLEGE
to JNTUA, Anantapuramu)
Sree Sainath Nagar, Tirupati – 517 102
2019-2023

SREE VIDYANIKETHAN ENGINEERING COLLEGE

(Affiliated to Jawaharlal Nehru Technological University Anantapur) Sree Sainath Nagar,
A. Rangampet, Tirupati – 517 102, Chittoor Dist., A.P.

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

CERTIFICATE

This is to certify that the Project Work entitled

“Sentimental Analysis on twitter tweets “

is the bonafide work done by

G .Naga Eshitha	**19121A0565**
C .Anusha	**19121A0538**
Abdul Bari Shaik	**19121A0502**
G .Hemanth Kumar	**19121A0558**
B .Nagendra Naik	**19121A0524**

In the Department of Computer Science and Engineering, Sree Vidyanikethan Engineering College, A. Rangampet. is affiliated to JNTUA, Anantapuramu in partialfulfillment of the requirements for the award of Bachelor of Technology in Computer Science and Engineering during 2019-2023.

This is work has been carried out under my guidance and supervision.

The results embodied in this Project report have not been submitted in any Universityor Organization for the award of any degree or diploma.

Internal Guide

Dr. K. Reddy Madhavi
Professor
Dept of CSE
Sree Vidyanikethan Engineering College

Head

Dr. B. Narendra Kumar Rao
Prof & Head
Dept of CSE
Sree Vidyanikethan Engineering CollegeTirupathi
Tirupathi

INTERNAL EXAMINER

EXTERNAL EXAMINER

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

VISION AND MISSION

VISION

To become a Centre of Excellence in Computer Science and Engineering by imparting high quality education through teaching, training and research.

MISSION

The Department of Computer Science and Engineering is established to provide undergraduate and graduate education in the field of Computer Science and Engineering to students with diverse background in foundations of software and hardware through a broad curriculum and strongly focused on developing advanced knowledge to become future leaders.

Create knowledge of advanced concepts, innovative technologies and develop research aptitude for contributing to the needs of industry and society.

Develop professional and soft skills for improved knowledge and employability of students.

Encourage students to engage in life-long learning to create awareness of the contemporary developments in computer science and engineering to become outstanding professionals.

Develop attitude for ethical and social responsibilities in professional practice at regional, National and International levels.

Program Educational Objectives (PEO's)

1. Pursuing higher studies in Computer Science and Engineering and related disciplines

2. Employed in reputed Computer and I.T organizations and Government or have established startup companies.

3. Able to demonstrate effective communication, engage in team work, exhibit leadership skills, ethical attitude, and achieve professional advancement through continuing education.

Program Specific Outcomes (PSO’s)

PSO1: Use mathematical methodologies to model real-world problems, employ modern tools and platforms for efficient design and development of computer-based systems.

PSO2: Apply adaptive algorithms and methodologies to develop intelligent systems for solving problems from inter-disciplinary domains.

PSO3: Apply suitable models, tools and techniques to perform data analytics for effective decision making.

PSO4: Design and deploy networked systems using standards and principles, evaluate security measures for complex networks, apply procedures and tools to solve networking issues.

Program Outcomes (PO’s)

1. Apply the knowledge of mathematics, science, engineering fundamentals, and an engineering specialization to the solution of complex engineering problems (**Engineering knowledge**).

2. Identify, formulate, review research literature, and analyze complex engineering problems reaching substantiated conclusions using first principles of mathematics, natural sciences, andengineering sciences (**Problem analysis**).

3. Design solutions for complex engineering problems and design system components or processes that meet the specified needs withappropriate consideration for the public health and safety, and the cultural, societal, and environmental considerations(**Design/development of solutions**).

4. Use research-based knowledge and research methods including design of experiments, analysis and interpretation of data, and synthesis of the information to provide valid conclusions (**Conduct investigations of complex problems**).

5. Create, select, and apply appropriate techniques, resources, and modern engineering and IT tools including prediction and modeling to complex engineering activities with an understanding of the limitations (**Modern tool usage**)

6. Apply reasoning informed by the contextual knowledge to assess societal, health, safety, legal and cultural issues and the consequent responsibilities relevant to the professional engineering practice (**The engineer and society**)

7. Understand the impact of the professional engineering solutions in societal and environmental contexts, and demonstrate the

knowledge of, and need for sustainable development (**Environment and sustainability**).

8. Apply ethical principles and commit to professional ethics and responsibilities and norms of the engineering practice (**Ethics**).

9. Function effectively as an individual, and as a member or leader in diverse teams, and in multidisciplinary settings (**Individual and team work**).

10. Communicate effectively on complex engineering activities with the engineering community and with society at large, such as, being able to comprehend and write effective reports and design documentation, make effective presentations, and give and receive clear instructions (**Communication**).

11. Demonstrate knowledge and understanding of the engineering and management principles and apply these to one’s own work, asa member and leader in a team, to manage projects and in multidisciplinary environments (**Project management and finance**).

12. Recognize the need for, and have the preparation and ability to engage in independent and life-long learning in the broadest context of technological change (**Life-long learning**)

Course Outcomes

CO1. Create/Design engineering systems or processes to solve complex societal problems using appropriate tools and techniques following relevant standards, codes, policies, regulations and latest developments.

CO2. Consider environment, sustainability, economics and project management in addressing societal problems.

CO3. Perform individually or in a team besides communicating effectively in written, oral and graphical forms on socially relevan

CO-PO-PSO Mapping

Course Outcome	Program Outcomes												Program Specific Outcomes			
	PO1	PO2	PO3	PO4	PO5	PO6	PO7	PO8	PO9	PO10	PO11	PO12	PSO1	PSO2	PSO3	PSO4
CO1	3	3	3	3	3	3		3				3	3	3	3	3
CO2							3				3		3	3	3	3
CO3									3	3						

(Note: 3-High, 2-Medium, 1-Low)

DECLARATION

We hereby declare that this project report titled **“SENTIMENTAL ANALYSIS ON TWITTER TWEETS”** is a genuine project work carried out by us, in **B. Tech *(Computer Science and Engineering)*** degree course of **Jawaharlal Nehru Technological University Anantapur** and has not been submitted to any other course or University for the award of any degree by us.

Signature of the student

1.
2.
3.
4.
5.

ACKNOWLEDGEMENT

We are extremely thankful to our beloved Chairman and founder **Dr. M. Mohan Babu** who took keen interest to provide us the infrastructural facilities for carrying out the project work.

We are highly indebted to **Dr. B.M. Satish**, Principal of Sree Vidyanikethan Engineering College for his valuable support andguidance in all academic matters.

We are very much obliged to **Dr. B. Narendra Kumar Rao,** Professor & Head, Department of CSE, for providing us the guidance and encouragement in completion of this project.

We would like to express our indebtedness to the project coordinator, **Ms. K. Ghamya**, Assistant Professor, Department of CSE for his valuable guidance during the course of project work.

We would like to express our deep sense of gratitude to, **Dr. K. Reddy Madhavi**, Professor, Department of CSE, for the constant support and invaluable guidance provided for the successful completion of the project.

We are also thankful to all the faculty members of CSE Department, who have cooperated in carrying out our project. We would like to thank our parents and friends who have extended their help and encouragement either directly or indirectly in completion of our project work.

ABSTRACT

Technology today has become a momentous driving vehicle for communication world-wide. Social media platforms like twitter, facebook, instagram are the most important arenas for expressing views on transformations happening in and around the world everyday. Twitter is a rich origin of info for mining of user opinions. This paper reflects the idea of taking user opinions into consideration performing sentiment, emotion analysis and establishing conclusions on interested topics using Machine Learning algorithms. Naive Bayes and Support Vectors Machines in Machine Learning are tuned-up using supervised learning to obtain outputs for sentiment emotion analysis respectively. Sentiment analysis desires to obtain sentiment polarity (positive or negative) and Emotion analysis intent to obtain emotion (eg.., empty , sadness , anger etc..,) from user data. Such analysis essentially serves a gateway for consumer needs and generates growth opportunities in businesses.

TABLE OF CONTENTS

LIST OF FIGURES

1. 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.

A supervised learning technique provides labels to classifier to make it understand the insights among various features. Once the classifier gets familiarized 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) 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.

Generally people discuss a lot of things daily but it is difficult to get insights just by reading through each of their opinions so there should be a way that helps us to get insights of users opinions in an unbiased manner, So this model helps in drawing out Sentiment, Emotions of users, classify them and finally present them to us. Sentiment analysis is the prediction of emotions in a word, sentence or corpus of documents.

It is intended to serve as an application to understand the attitudes, opinions and emotions expressed within an online mention. The intention is to gain an overview of the wider public opinion behind certain topics.

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.

1.1 MOTIVATION

Businesses primarily run over customers satisfaction, customer reviews about their products. Shifts in sentiment on social media have been shown to correlate with shifts in stock markets. Identifying customer grievances thereby resolving them leads to customer satisfaction as well as trustworthiness of an organization. Hence there is a necessity of an unbiased automated system to classify customer reviews regarding any problem.

In today’s environment where we’re justifiably suffering from data overload (although this does not mean better or deeper insights), companies might have mountains of customer feedback collected; but for mere humans, it’s still impossible to analyze it manually without any sort of error or bias.

Oftentimes, companies with the best intentions find themselves in an insights vacuum. You know you need insights to inform your decision making and you know that you’re lacking them, but don’t know how best to get them.

Sentiment analysis provides some answers into what the most important issues are, from the perspective of customers, at least. Because sentiment analysis can be automated, decisions can be made based on a significant amount of data rather than plain intuition that isn’t always right.

1.2 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 user’s 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.

1.3 OBJECTIVES

- To implement automatic classification of text into positive, negative and neutral and achieve 80% accuracy.
- Sentiment Analysis is used for identifying the polarity of sentiments expressed by users.
- The output is represented in the form of pie-chart which is given based on the polarity.

1.4 SOCIETAL APPLICATIONS

- Customer support ticket analysis
- Social media monitoring
- Brand monitoring and reputation
- Product analysis
- Competitive research
- Market research and insights into industry trends
- Workforce analytics/employee engagement monitoring
- Listen to the voice of customer

1.5 LIMITATIONS

- Sentiment Neutral Statements with Negative or Positive Reputational Impact

- Incorrectly Targeted Sentiment

- Review Language is Dissimilar to Social Media and News Language

- problems recognizing things like sarcasm and irony, negations, jokes, and exaggerations

2. LITERATURE SURVEY

2.1 INTRODUCTION

"What other people think" has always been an important piece of information for most of us during the decision-making process. The Internet and the Web have now (among other things) made it possible to find out about the opinions and experiences of those in the vast pool of people that are neither our personal acquaintances nor well-known professional critics — that is, people we have never heard of. And conversely, more and more people are making their opinions available to strangers via the Internet. The interest that individual users show in online opinions about products and services, and the potential influence such opinions wield, is something that is driving force for this area of interest. And there are many challenges involved in this process which needs to be walked all over inorder to attain proper outcomes out of them.

In this survey we analyzed basic methodology that usually happens in this process and measures that are to be taken to overcome the challenges being faced.

2.2 ANALYZING THE EXISTING METHODS

2.2.1 Using a Heterogeneous Dataset for Sentimental Analysis in Text :

A supervised machine learning approach was adopted to recognize basic emotions using a emotion-annotated dataset which combines news headlines, fairy tales and blogs. For this purpose, different features sets, such as bags of words, and N-grams, were used. The Support Vector Machines classifier (SVM) performed significantly better than other classifiers, and it generalized well on unseen examples. Five data sets were considered to compare among various approaches. In bag of words Each sentence in the dataset was represented by a feature vector composed of Boolean attributes for each word that occurs in the sentence. If a word occurs in a given sentence, its corresponding attribute is set to 1; otherwise it is set to 0. In N grams approach they are defined as sequences of words of length n. N-grams can be used for catching syntactic patterns in text and may include important text features such as negations, e.g., “not happy”. Negation is an important feature for the analysis of emotion in text because it can totally change the expressed emotion of a sentence. The author concludes some research studies in sentiment analysis claimed that N-grams features improve performance beyond the BOW approach.

2.2.2 Analyzing Sentiment of Twitter Data using Machine Learning Algorithm :

Author here classifies the process of Sentiment analysis as follows : Tweets posted on twitter are freely available through a set of APIs of twitter. At first, we collected a corpus of positive, negative, neutral and irrelevant tweets from twitter API. Then pre-processing done by removing stop words, negations, URL, full stop, commas etc. to reduce noise from tweets and to prepare our data for sentiment classification. After that, we apply machine learning algorithms to our dataset and compare their results. Results helps to identify which machine learning algorithm is best suited for classification of SA. The stages involved in this process are:

1.Data Collection : obtain training data of

twitter

2.2.Pre-processing Setup: removing

unrelated contents

3.Sentiment Classifier : various machine learning

algorithms are used4.Evaluation: produces result.

And each step is further classified like in the pre-processing step some sub- steps like stemming, stop word extractor are also included. The efficiency of an classifier usually depends on the pre-processing step

2.2.3 The Impact of Features Extraction on the SentimentAnalysis:

In this project analysing the impact and importance of extracting thefeatures and how they play a crucial role in the performance of the classifier and its outcome. Here six pre-processing techniques are used and then features are extracted from the pre-processed data. There are many feature extraction techniques such as Bag of words, n-grams ,TF-IDF etc.., Here we analysed the impact of two features TF-IDF word level and N-Gram on a Twitter data set and found that performance of classifier is 3-4% in high when TF-IDF feature is chosen than N-Gram and analysis is done using many classification algorithms like Naïve Bayes, Support Vector Machines etc.., Eventually emphasizes on the importance of feature selection process which affects the Sentiment Evaluation result.

2.2.4 Methods for Sentiment Analysis:

In this paper, various approaches to sentiment analysis have been examined and analysed,techniques such as Streaming API SVM etc., discussed. These techniques all have different strengths and weaknesses.

1. Sentiment Analysis on Twitter using streaming API SVM:

It uses NLP where it helps in tokenization, stemming, classification, tagging, parsing and sentiment reasoning Its basic feature is to convert unstructured data into structured data. It uses Naive Bayes for classification which requires number of linear parameters.

To find out the sentiment an automated system must be developed "Support Vector Machine" can be used for this method. SVM is machine that takes the input and store them in a vector then using SentiWordNet it scores it decides the sentiment. It also classifies the opinion in overall way by positive, negative or Neutral.

There are many more techniques but these are the most familiar ones , performs more efficiently. And selection of both features and techniques affect the final outcome. So proper analysis must be done to get intended , as accurate results as possible.

3. 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.

3.1 PROPOSED SYSTEM

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 emotions as 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.

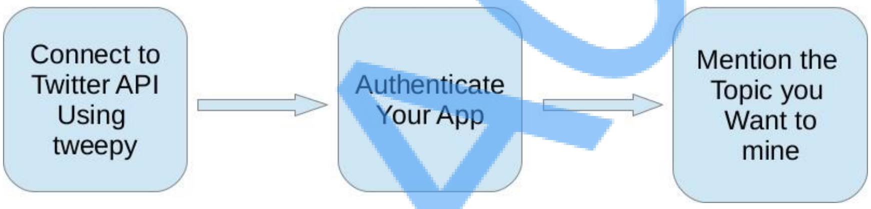

Connect to Twitter API Using tweepy
Authenticate Your App
Mention the Topic you Want to mine

Fig 3.1.1 Extraction of data

Preprocessing Of Data:

Following are the Preprocessing steps that have been carried out:

Removing Html tags and urls:

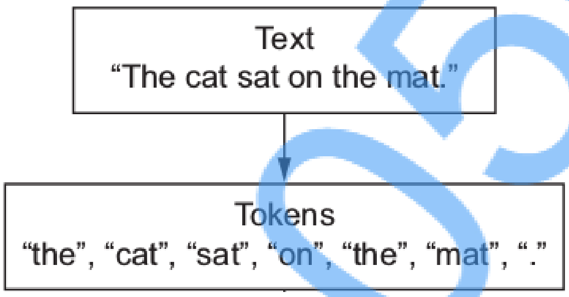
Html tags and urls often have minimum sentiments thus they are removed from tweets. Using regular expressions.

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.

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.


Text
“The cat sat on the mat.”
Tokens
“the”, “cat”, “sat”, “on”, “the”, “mat”, “.”

Removing punctuations and special symbols:

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

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', '.']

After stop words removal:

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

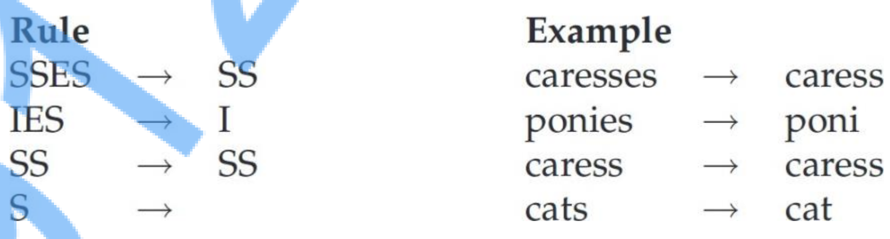
Stemming and Lemmatization:

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

Feature Extraction:

Text data demands a special measure before you train the model. Words after tokenization are encoded as integers or floating point values for feeding input to machine learning algorithm. This practice is described as vectorization or feature extraction. Scikit-learn library offers TF-IDF vectorizer to convert text to word frequency vectors.

Fitting Data to Classifier and predicting test data:

Train data is fitted to a suitable classifier upon feature extraction ,then once the classifier is trained enough then we predict the results of the test data using the classifier, then compare the original value to the value returned by the classifier.

Result Analysis:

Here the accuracy of different classifiers are shown among which the best classifier with highest accuracy percent is the chosen. Some factors such as f- score, mean, variance etc., also accounts for consideration of the classifiers.

Visual Representation:

Our final results are plotted as pie charts which contains different fields such as positive, negative, neutral in case of sentiment analysis. Pictorial representation is the best way to convey information without much efforts. Thus it is chosen.

3.1.1 SYSTEM ARCHITECTURE

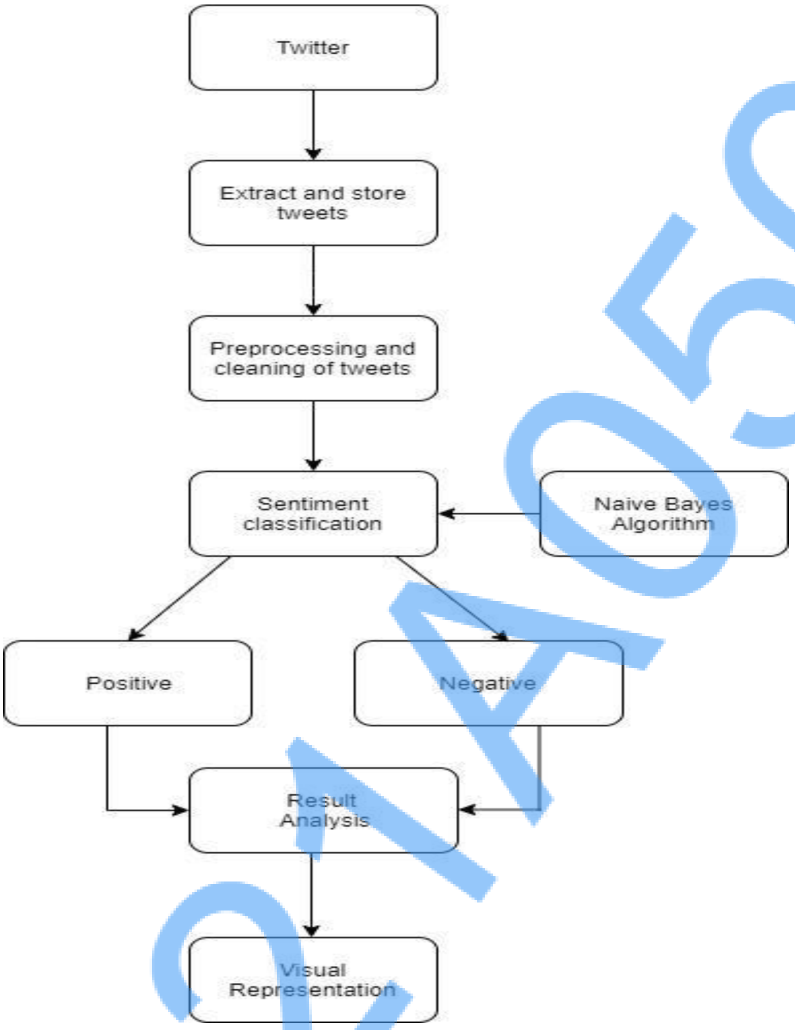


Fig 3.1.1.1 Architecture diagram for sentiment analysis using Naive Bayes

Naive Bayes Algorithm :

Naive Bayes algorithm which is based on well known Bayes theorem which is mathematically represented as

$$\boldsymbol{P\left(\frac{A}{B}\right) = \frac{P\left(\frac{B}{A}\right)P(A)}{P(B)}}$$

Where ,A and B are events and P(A/B) is the likeliness of happening of event A given that event B is true and has happened, which is known to be as posterior probability .

P(A) is the likeliness of happening of an event A being true,Which is known to be as prior probability.

P(B/A) is the likeliness of happening of an event B given A was true ,Which is known to be as Likelihood.

P(B) is the likeliness of happening of an event B, Which is known

to be as Evidence . Bayes theorem can now be applied on data sets

in following way

$$P\left(\frac{y}{X}\right) = \frac{P\left(\frac{X}{y}\right)P(y)}{P(X)}$$

Where y is a class variable and X is feature vector.

This is a classification method that relies on Bayes' Theorem with strong (naive) independence assumptions between the features. A Naive Bayes classifier expects that the closeness of a specific feature (element) in a class is disconnected to the closeness of some other elements. For instance, an organic fruit might be considered to be an apple if its color is red, its shape is round and it measures approximately three inches in breadth. Regardless of whether these features are dependent upon one another or upon the presence of other features, a Naïve Bayes classifier would consider these properties independent due to the likelihood that this natural fruit is an apple. Alongside effortlessness, the Naive Bayes is known to out-perform even exceedingly modern order strategies.

The Naive Bayes is widely used in the task of classifying texts into multiple classes and was recently utilized for sentiment analysis classification.

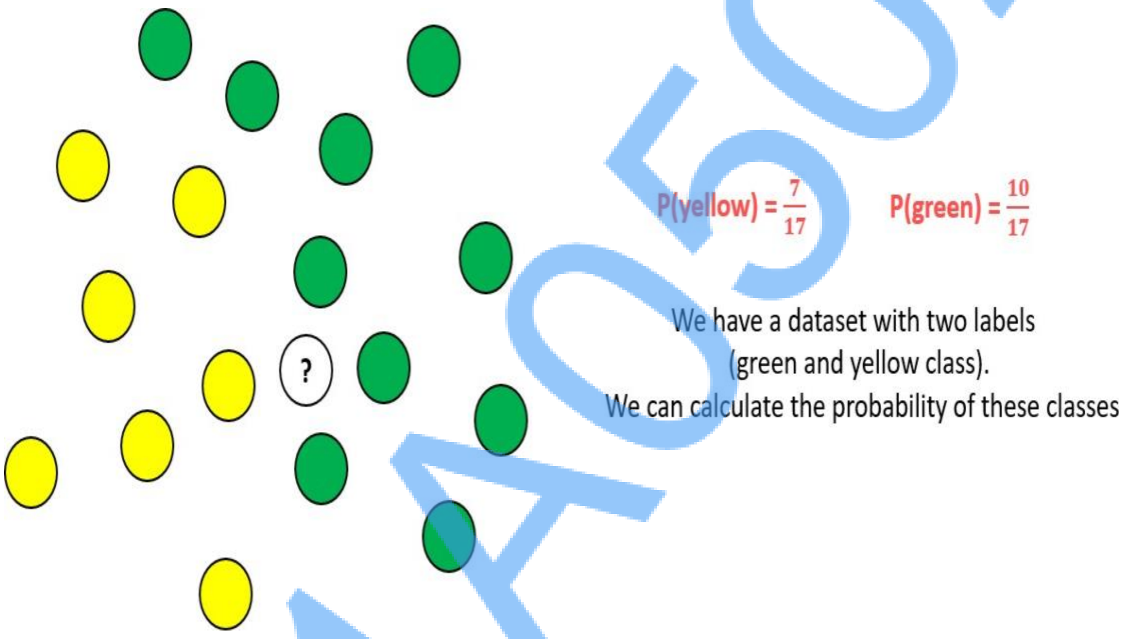

?
P(yellow) = $\frac{7}{17}$
P(green) = $\frac{10}{17}$
We have a dataset with two labels
(green and yellow class).
We can calculate the probability of these classes

Fig3.1.1.2 Naive Bayes Classifier

$$P\left(\frac{A}{B}\right) = \frac{P\left(\frac{B}{A}\right)P(A)}{P(B)}$$

Support Vector Machine:

Support Vector Machines is a supervised machine learning algorithm , adopted conventionally for classification as well as regression problems. SVMs for classification, work by figuring out the right hyperplane among the classes. After being trained by a labeled data set , SVM outputs an optimal hyperplane that categorizes new examples. Classification by SVMs for different data sets is governed by tuning parameters namely kernel , regularization, gamma and margin. When data is 2 dimensional Support vector classifier is a line, if it is 3D SVC forms a plane instead of a line. When data is more than 4D then classifier is a hyperplane. For highly distributed data Maximal margin and support vector classifier fail and hence SVMs are used. For linearly separable patterns optimal hyperplane is formed and for non-linearly separable patterns transformation of original data into a new space is performed determined by kernel function. The trouble of discovering an optimal hyperplane is an optimization problem and can be worked out using optimization techniques (eg. Lagrange). To classify tweets into different emotion classes a linear kernel has been utilized. Linear kernel is preferable for text classification problems because text has lot of features, linear kernel is faster and less parameters are optimized. When SVM is trained with a linear kernel only C regularization parameter need to be optimized whereas for other kernels you need to optimize gamma parameter also.

Support Vector Machines (SVM) is a machine learning model proposed by V. N. Vapnik. The basic idea of SVM is to find an optimal hyperplane to separate two classes with the largest margin from pre-classified data. By applying appropriate transformations to the data space before computing the separating hyperplane, SVM can be extended to cases where the margin between two classes is non-linear

Linearly Separable Case

If the training data are linearly separable, then there exists a pair (w, b) such that WT Xi + b ≥ 1, for all Xi ∈ P (1) WT Xi + b ≤ -1, for all Xi ∈ N

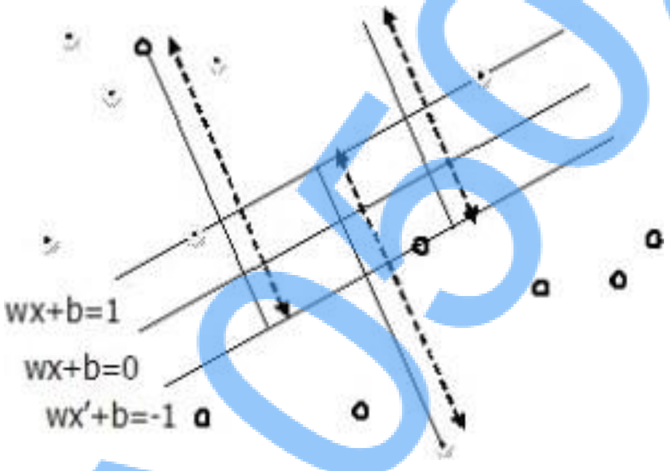


Fig 3.1.3 SVM Classifier(I)

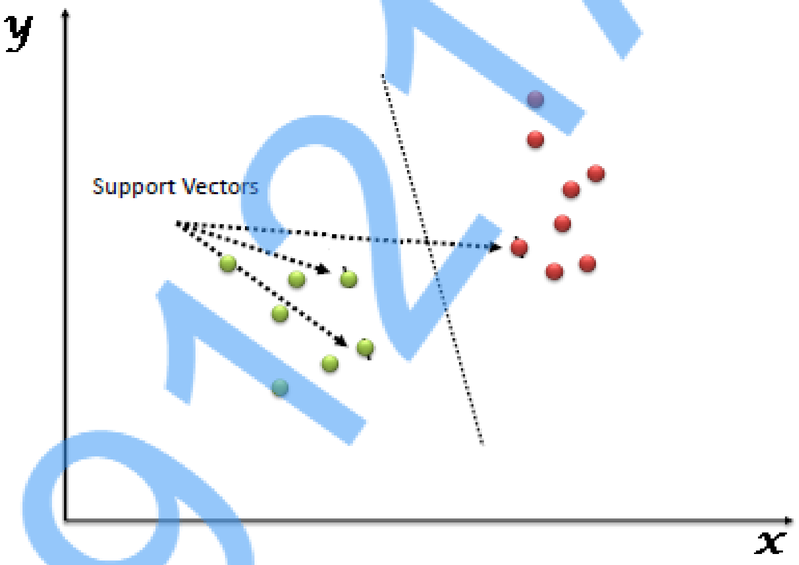


Fig 3.1.4 SVM Classifier (II)

4.1 STRUCTURE CHART

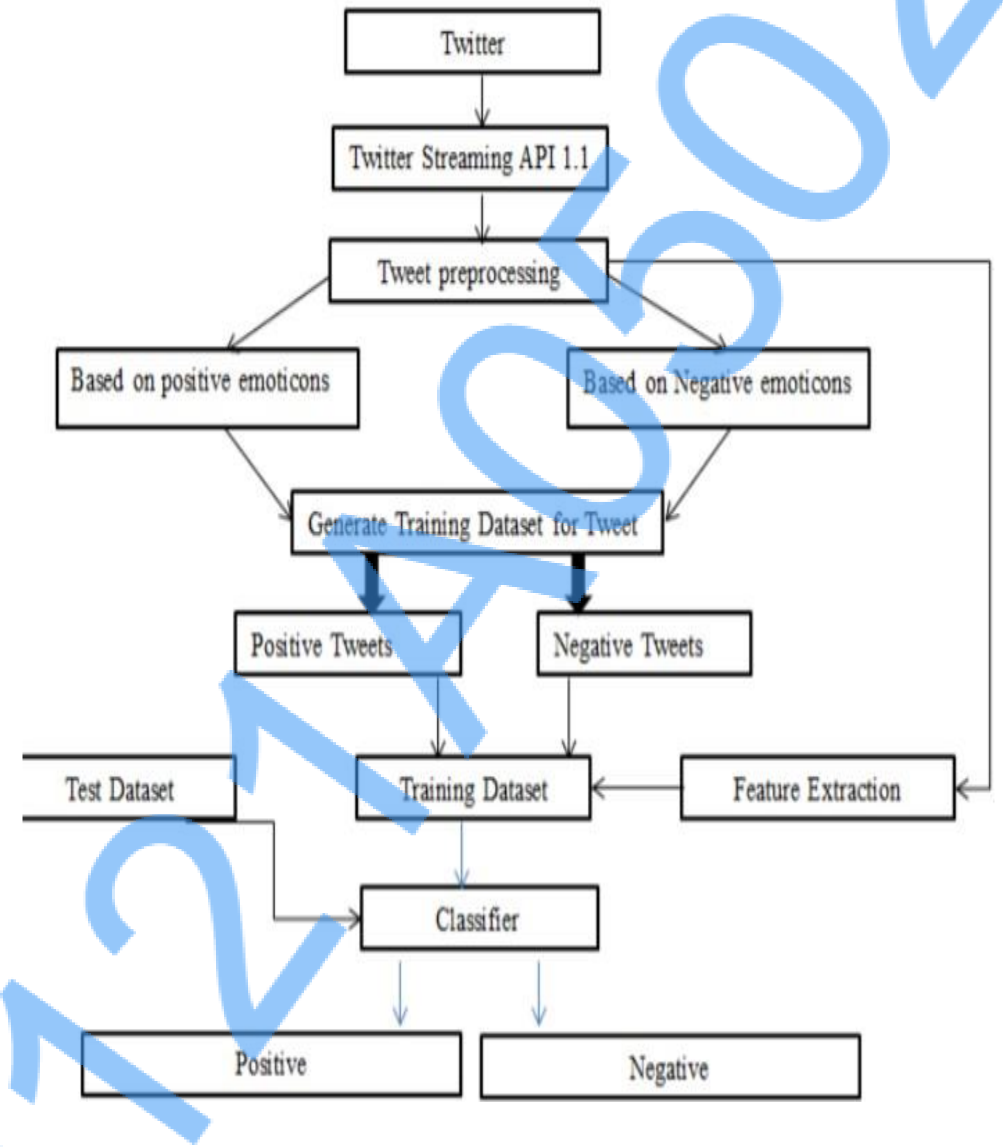


Fig 4.1.1 Structure Chart

4.2 UML DIAGRAMS

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**.

4.2.1 USE CASE DIAGRAM

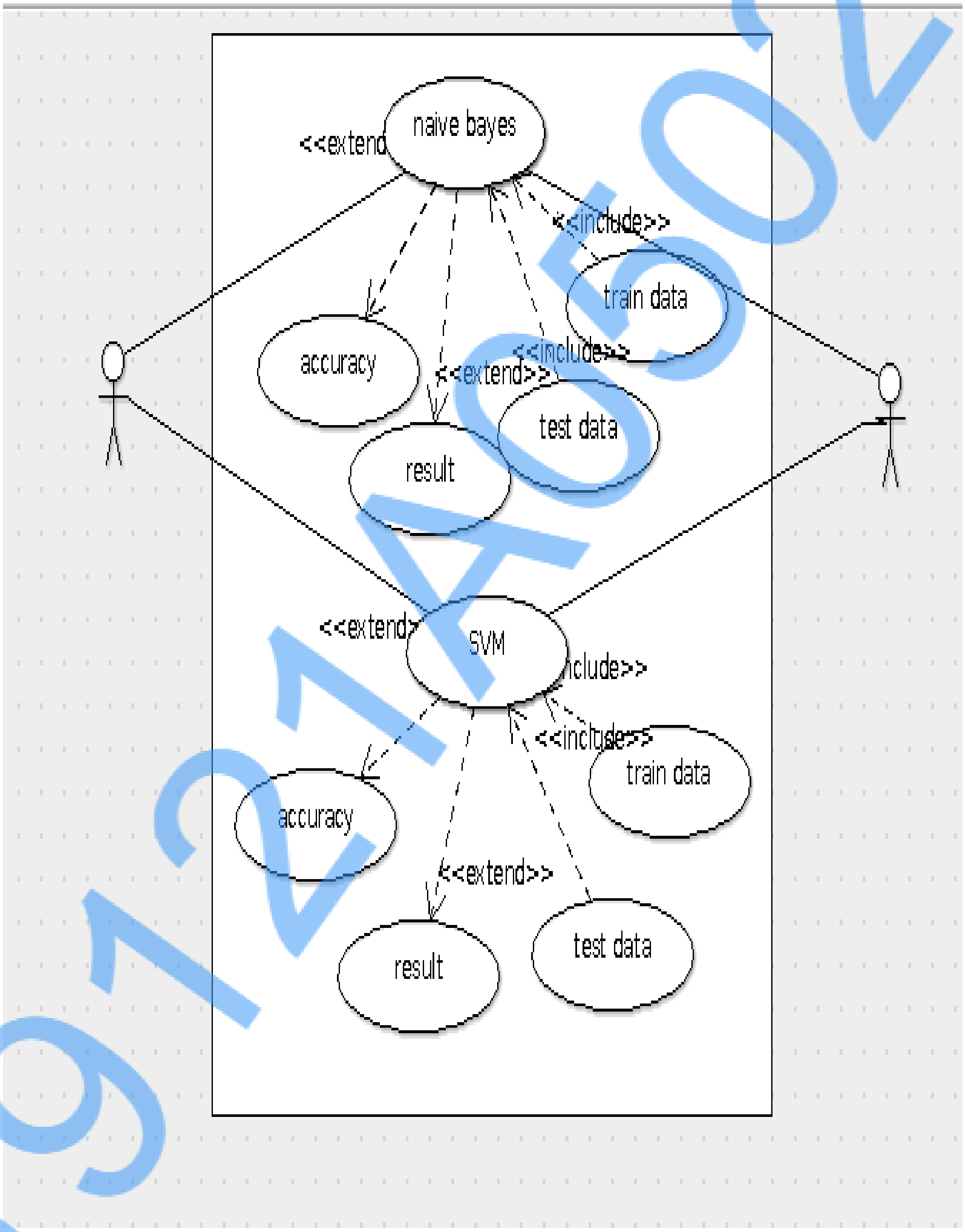


Fig 4.2.1.1 Use Case Diagram

4.2.2 CLASS DIAGRAM

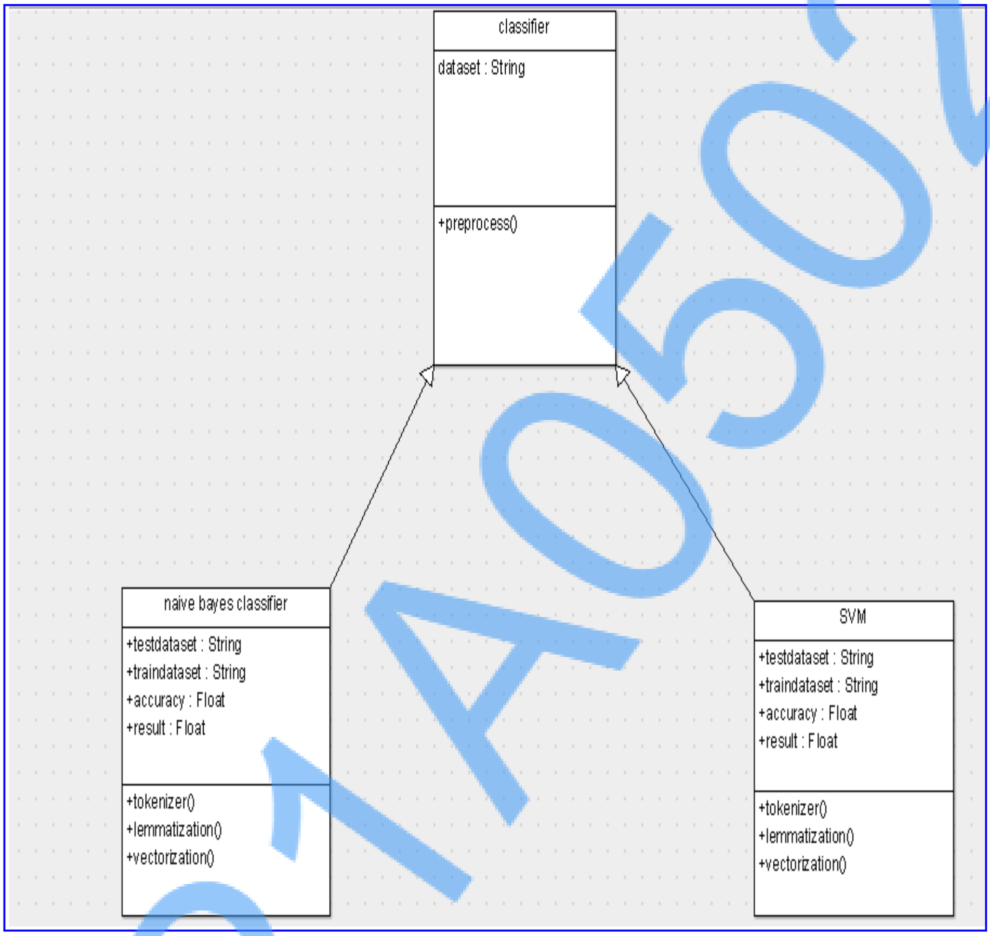


Fig 4.2.2.1 Class Diagram

4.2.3 SEQUENCE DIAGRAM

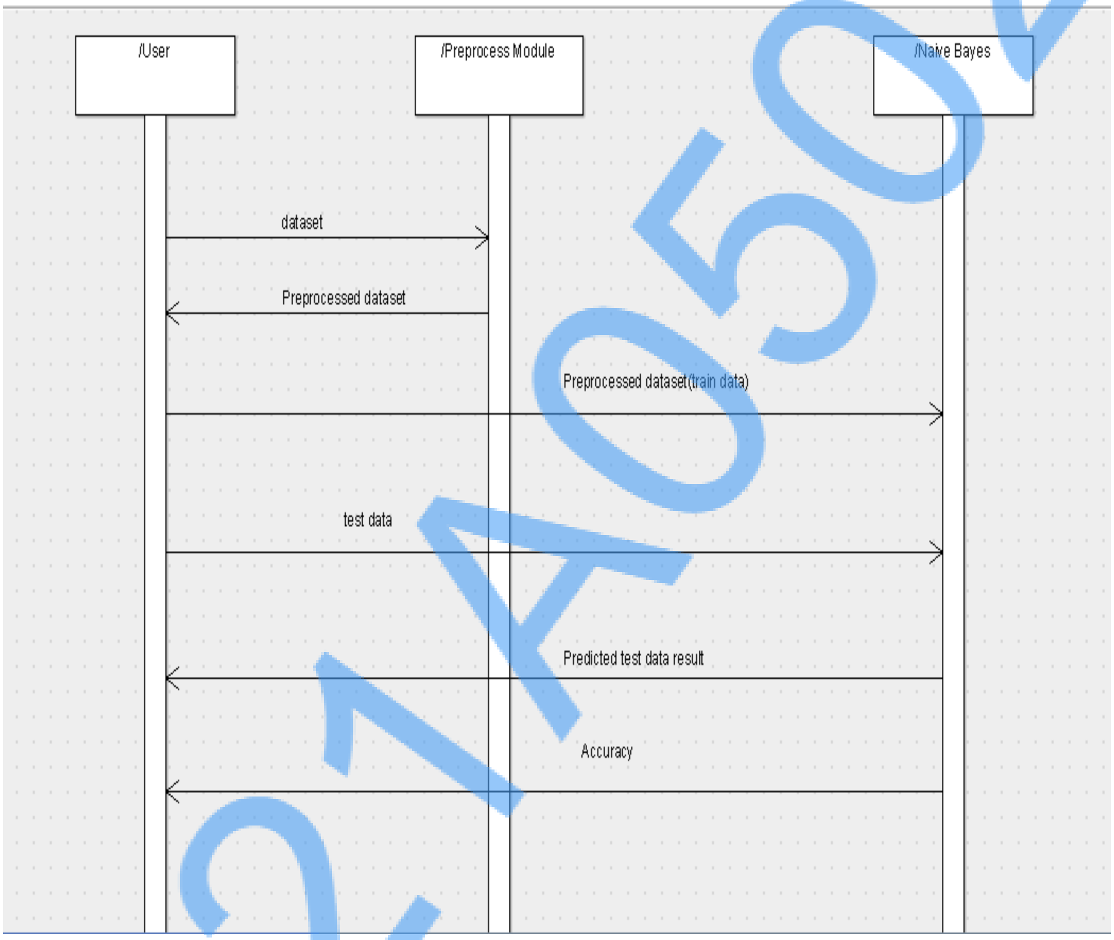


Fig 4.2.3.1 Sequence Diagram of Sentiment Analysis(I)

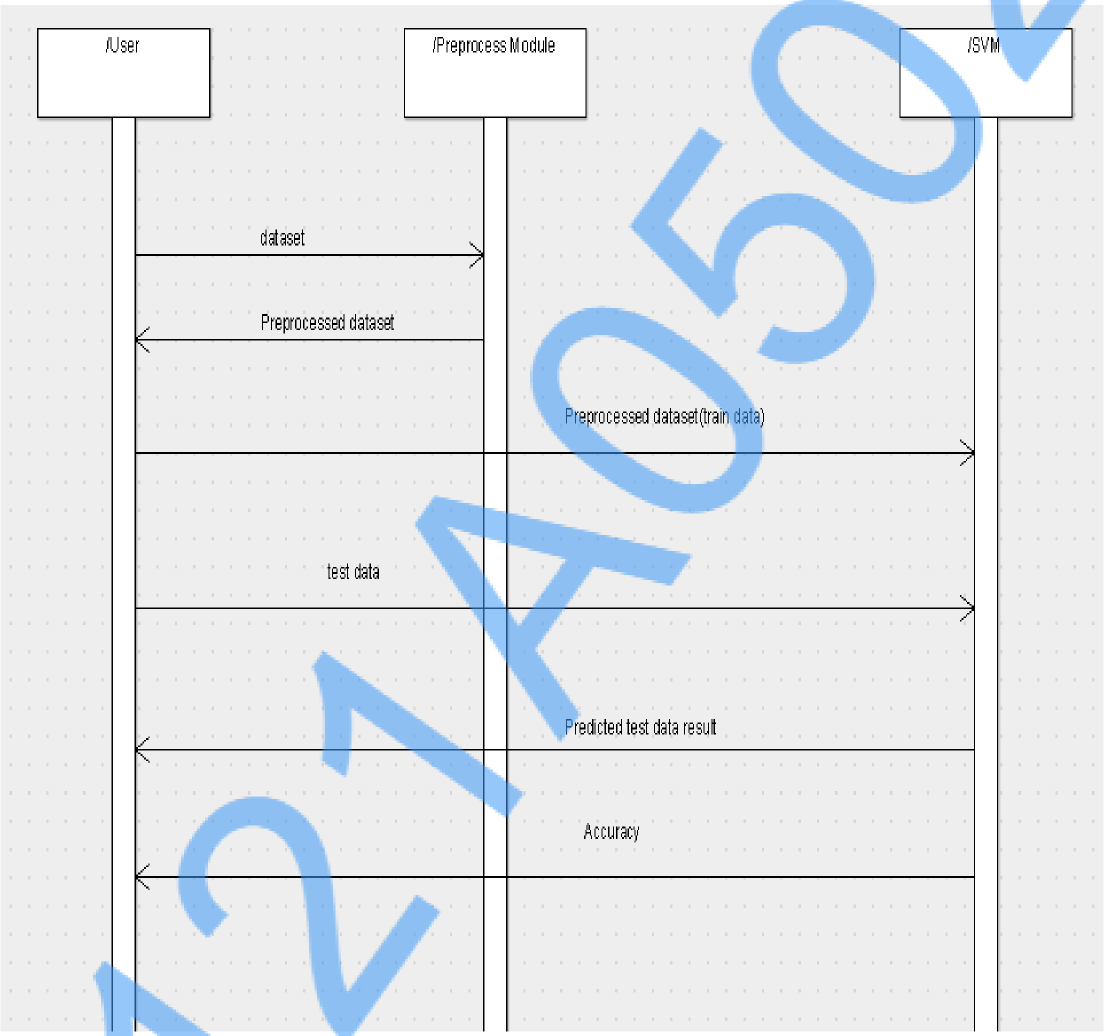


Fig 4.2.3.2 Sequence Diagram of Sentimental Analysis(II)

4.2.4 ACTIVITY DIAGRAM

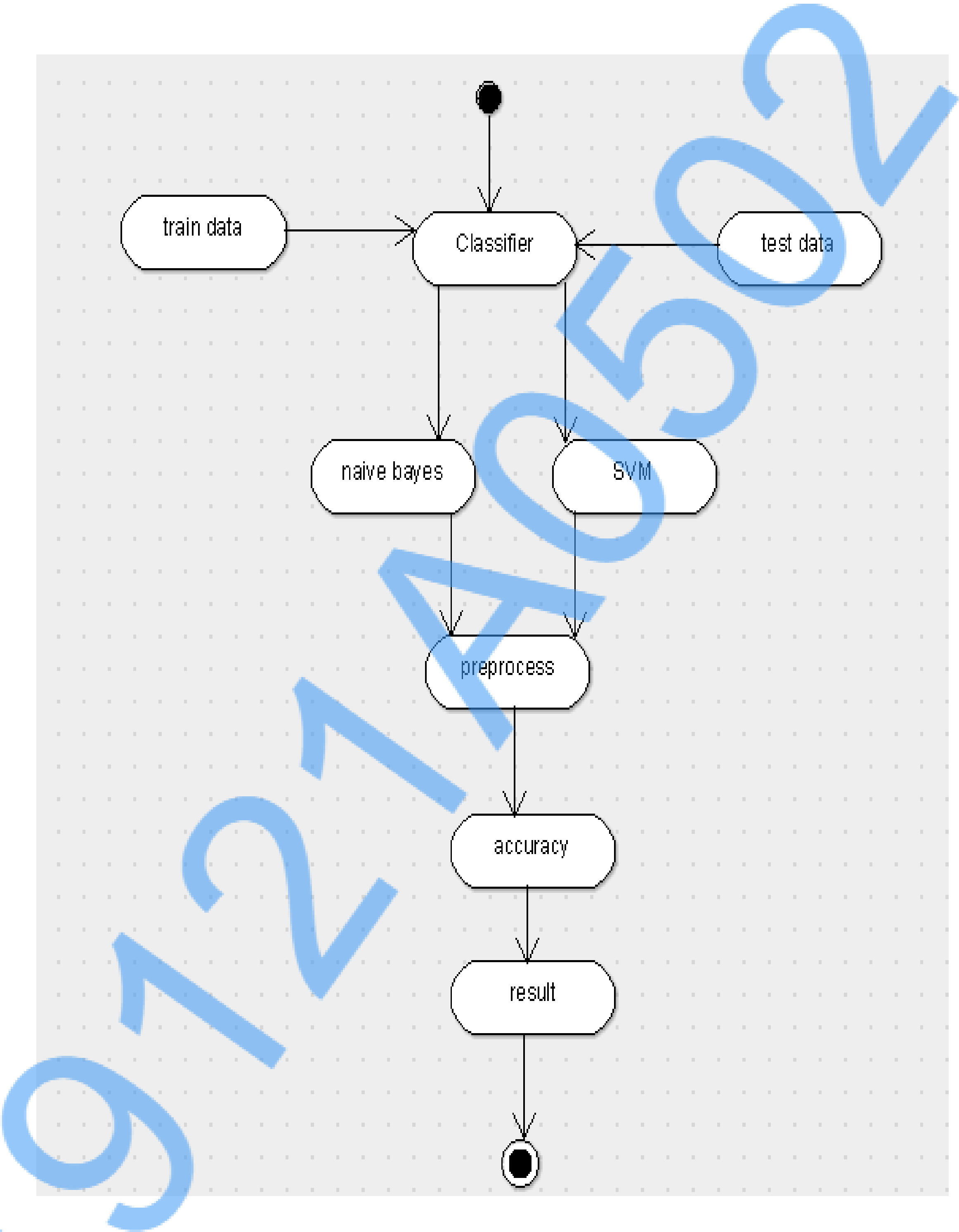


Fig 4.2.4.1 Activity Diagram

4.2.5 COLLABORATION DIAGRAM

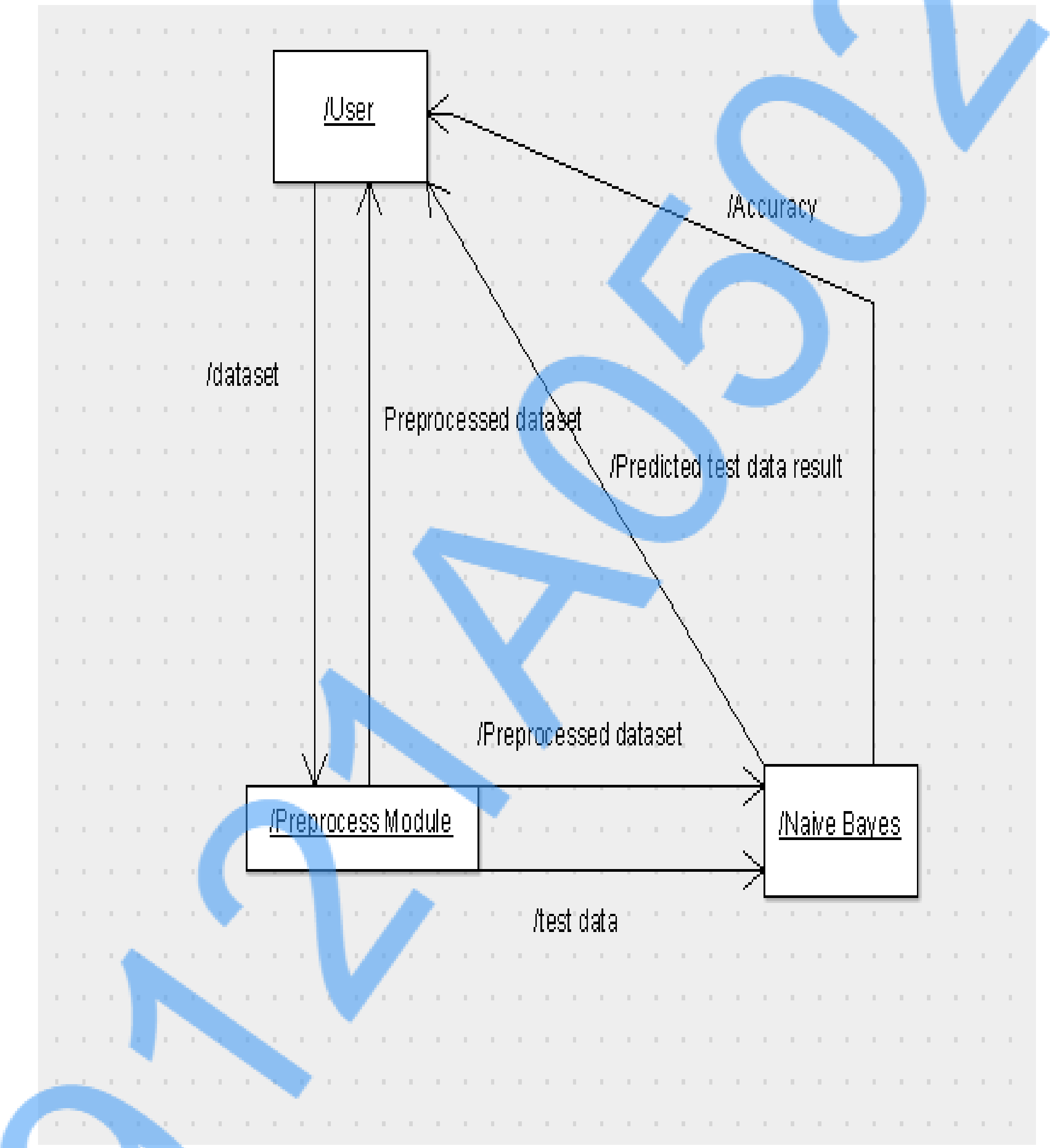


Fig 4.2.5.1 Collaboration Diagram

4.2.6 COMPONENT DIAGRAM

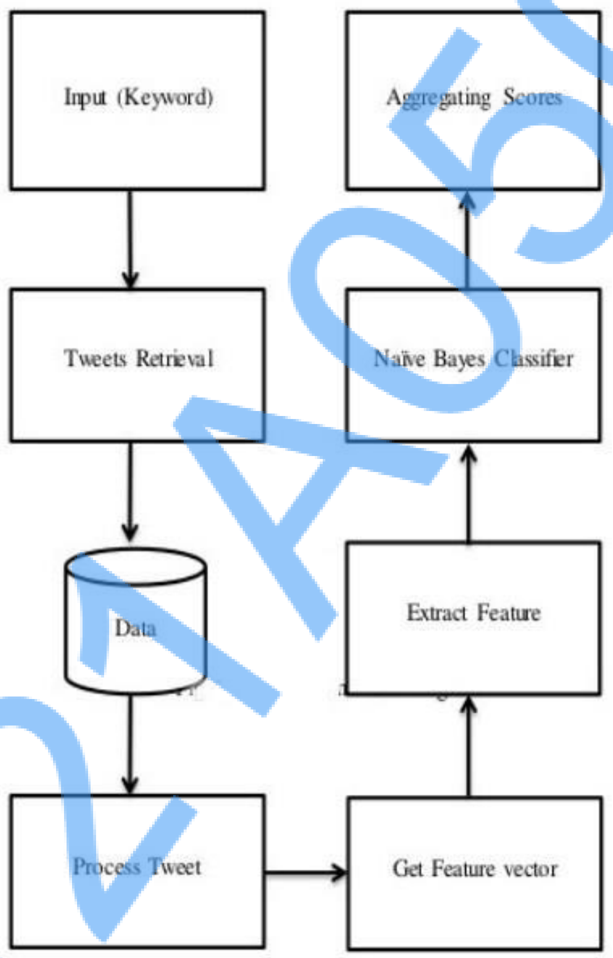


Fig 4.2.6.1 Component Diagram

5. EXPERIMENTAL ANALYSIS AND RESULTS

5.1 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.

5.1.1 SOFTWARE REQUIRMENTS

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

1. Jupiter Note Book or Google colab

2. NLTK

3. Scikit-learn

4. Matplotlib

5. Tweepy

6. Pandas

7. Numpy

10.TextBlob

11.VaderSentiment

12.Csv

13.Re(Regular Expressions)

14.Windows

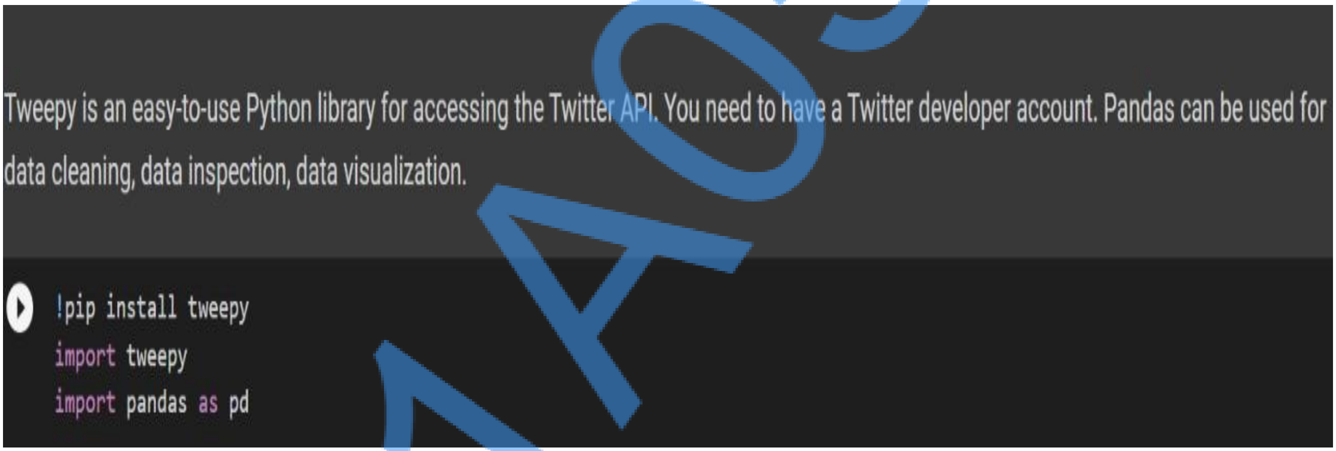
5.1.2 HARDWARE REQUIREMENTS

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

1.A minimum of Intel Core I3 processor

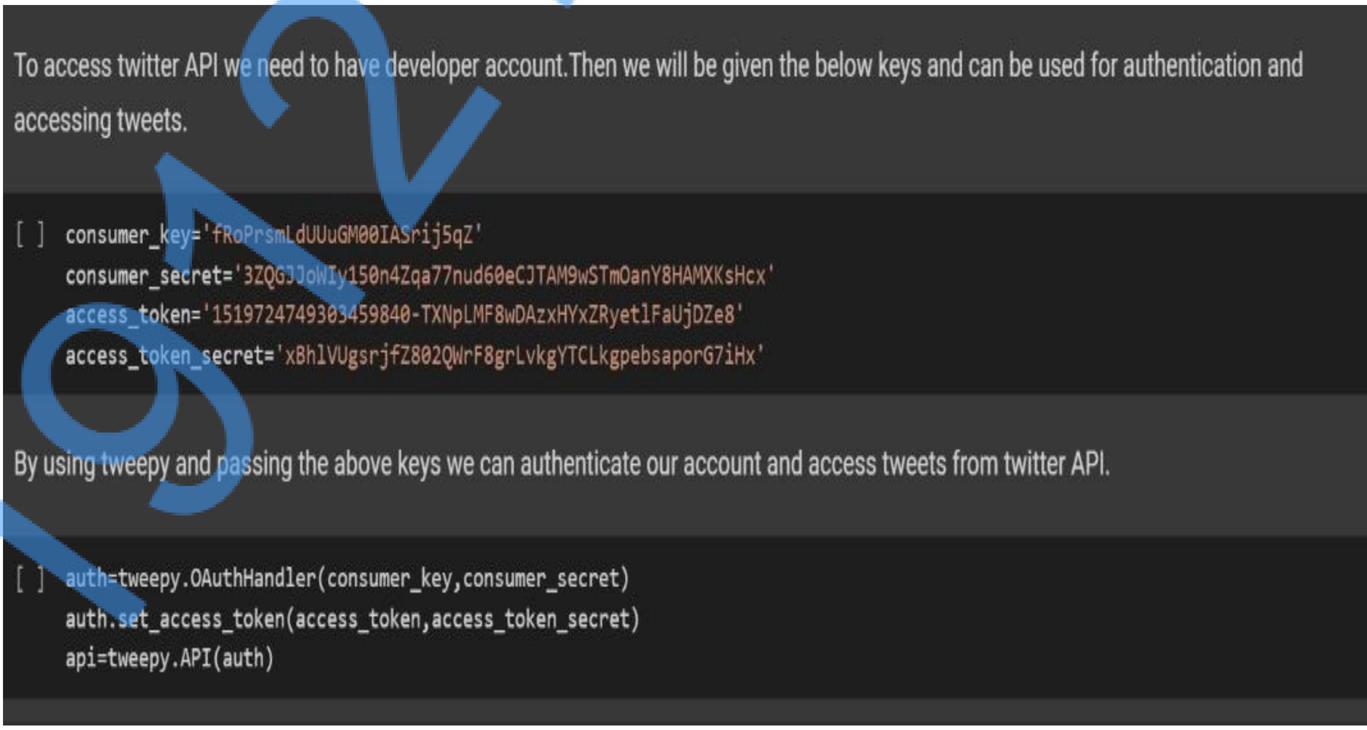
2.A minimum of 4 GB Ram

5.2 ANALYSIS OF INPUT AND OUTPUT



Tweepy is an easy-to-use Python library for accessing the Twitter API. You need to have a Twitter developer account. Pandas can be used for data cleaning, data inspection, data visualization.

```
!pip install tweepy
import tweepy
import pandas as pd
```

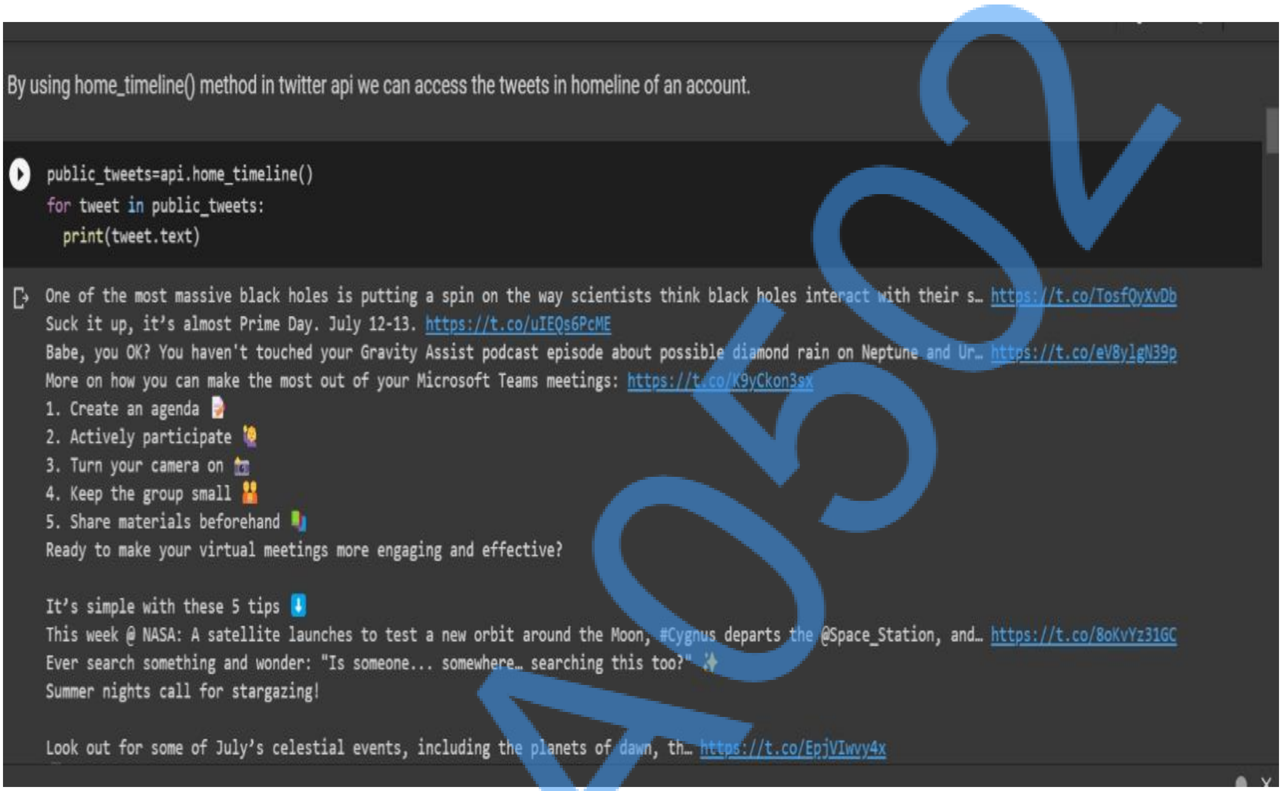


To access twitter API we need to have developer account.Then we will be given the below keys and can be used for authentication and accessing tweets.

```
[ ]  consumer_key='fRoPrsmLdUUuGM00IASrij5qZ'
     consumer_secret='3ZQGJJoWIy150n4Zqa77nud60eCJTAM9wSTmOanY8HAMXKsHcx'
     access_token='1519724749303459840-TXNpLMF8wDAzxHYxZRyetlFaUjDZe8'
     access_token_secret='xBhlVUgsrjfZ802QWrF8grLvkgYTCLkgpebsaporG7iHx'
```

By using tweepy and passing the above keys we can authenticate our account and access tweets from twitter API.

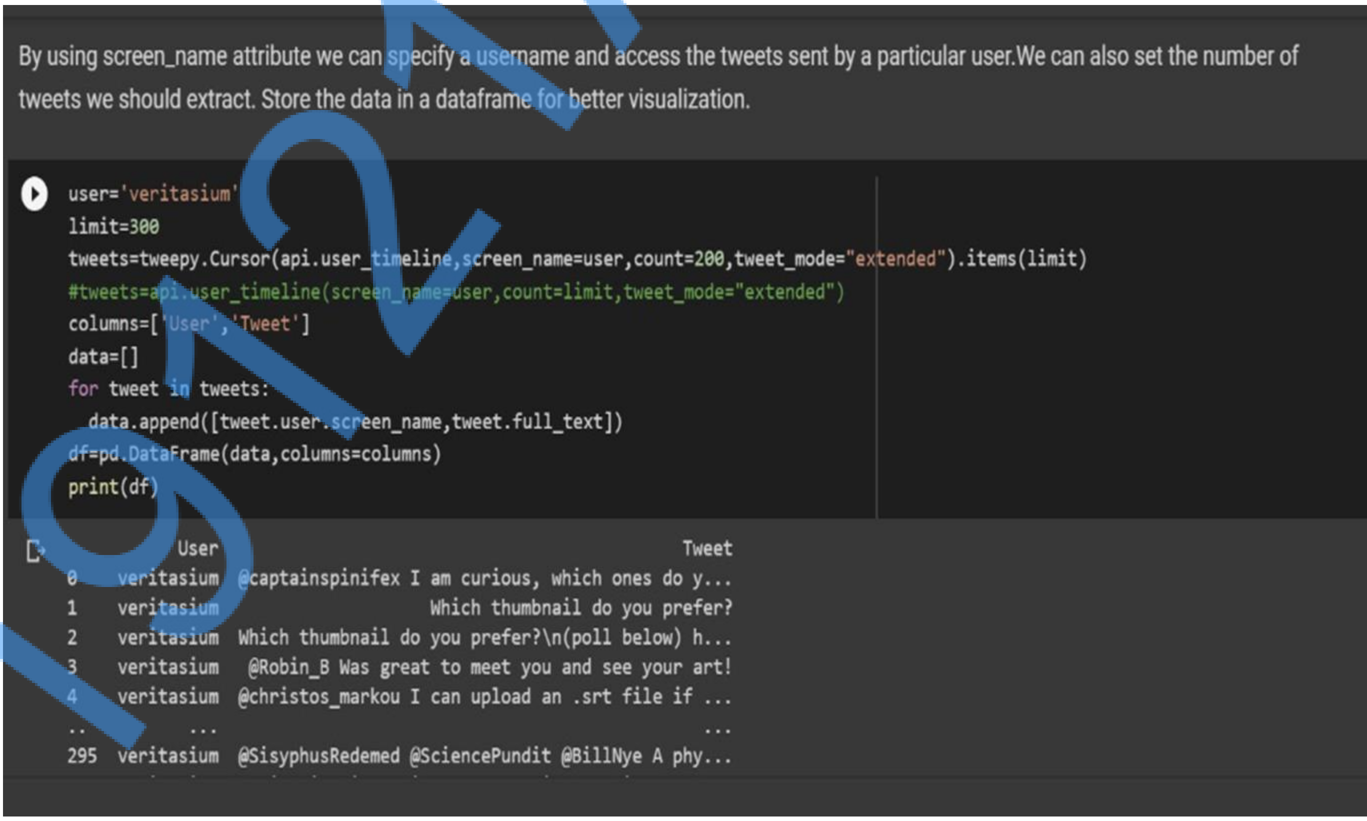
```
[ ]  auth=tweepy.OAuthHandler(consumer_key,consumer_secret)
     auth.set_access_token(access_token,access_token_secret)
     api=tweepy.API(auth)
```



By using home_timeline() method in twitter api we can access the tweets in homeline of an account.

```
public_tweets=api.home_timeline()
for tweet in public_tweets:
  print(tweet.text)
```

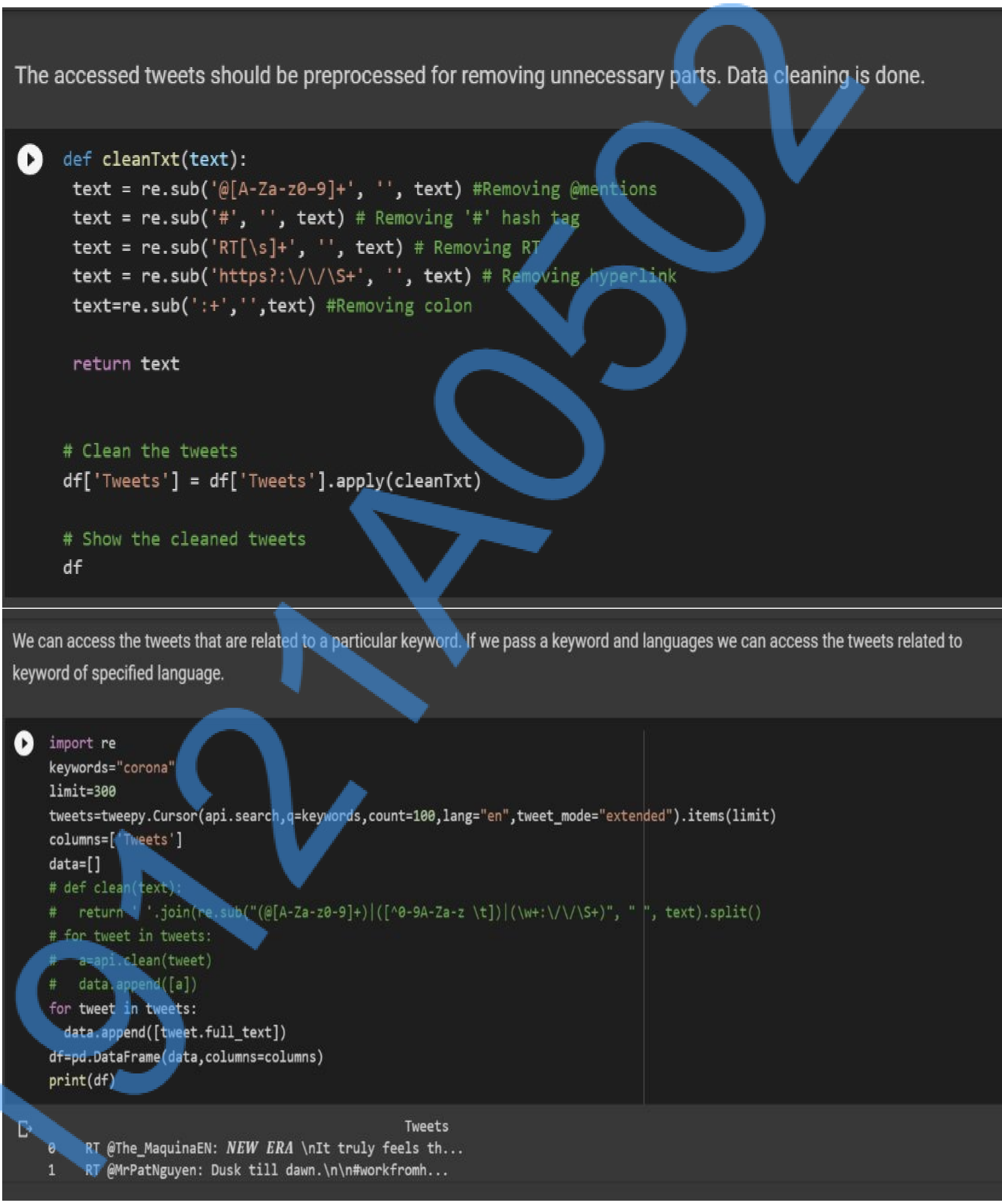
```
One of the most massive black holes is putting a spin on the way scientists think black holes interact with their s… https://t.co/TosfQyXvDb
Suck it up, it’s almost Prime Day. July 12-13. https://t.co/uIEQs6PcME
Babe, you OK? You haven't touched your Gravity Assist podcast episode about possible diamond rain on Neptune and Ur… https://t.co/eV8ylgN39p
More on how you can make the most out of your Microsoft Teams meetings: https://t.co/K9yCkon3sx
1. Create an agenda 📝
2. Actively participate 🙋
3. Turn your camera on 📷
4. Keep the group small 👪
5. Share materials beforehand 📚
Ready to make your virtual meetings more engaging and effective?

It’s simple with these 5 tips ⬇️
This week @ NASA: A satellite launches to test a new orbit around the Moon, #Cygnus departs the @Space_Station, and… https://t.co/8oKvYz31GC
Ever search something and wonder: "Is someone... somewhere… searching this too?" ✨
Summer nights call for stargazing!

Look out for some of July’s celestial events, including the planets of dawn, th… https://t.co/EpjVIwvy4x
```



By using screen_name attribute we can specify a username and access the tweets sent by a particular user.We can also set the number of tweets we should extract. Store the data in a dataframe for better visualization.

```
user='veritasium'
limit=300
tweets=tweepy.Cursor(api.user_timeline,screen_name=user,count=200,tweet_mode="extended").items(limit)
#tweets=api.user_timeline(screen_name=user,count=limit,tweet_mode="extended")
columns=['User','Tweet']
data=[]
for tweet in tweets:
  data.append([tweet.user.screen_name,tweet.full_text])
df=pd.DataFrame(data,columns=columns)
print(df)
```

```
           User                                              Tweet
0    veritasium  @captainspinifex I am curious, which ones do y...
1    veritasium                        Which thumbnail do you prefer?
2    veritasium  Which thumbnail do you prefer?\n(poll below) h...
3    veritasium   @Robin_B Was great to meet you and see your art!
4    veritasium  @christos_markou I can upload an .srt file if ...
..          ...                                                ...
295  veritasium  @SisyphusRedemed @SciencePundit @BillNye A phy...
```



The accessed tweets should be preprocessed for removing unnecessary parts. Data cleaning is done.

```
def cleanTxt(text):
  text = re.sub('@[A-Za-z0-9]+', '', text) #Removing @mentions
  text = re.sub('#', '', text) # Removing '#' hash tag
  text = re.sub('RT[\s]+', '', text) # Removing RT
  text = re.sub('https?:\/\/\S+', '', text) # Removing hyperlink
  text=re.sub(':+','',text) #Removing colon

  return text


# Clean the tweets
df['Tweets'] = df['Tweets'].apply(cleanTxt)

# Show the cleaned tweets
df
```

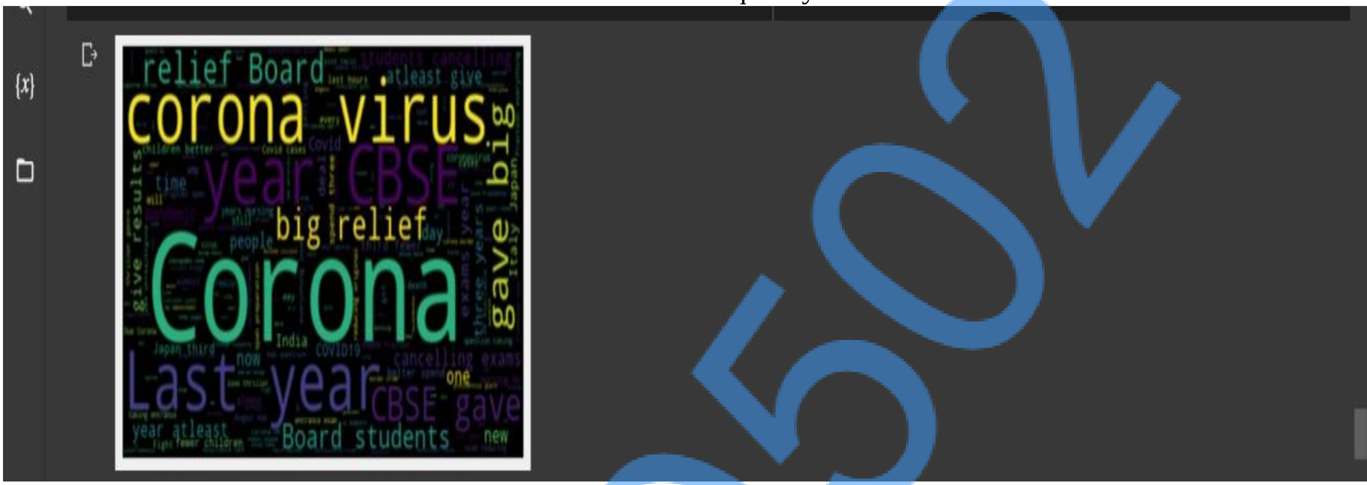
We can access the tweets that are related to a particular keyword. If we pass a keyword and languages we can access the tweets related to keyword of specified language.

```
import re
keywords="corona"
limit=300
tweets=tweepy.Cursor(api.search,q=keywords,count=100,lang="en",tweet_mode="extended").items(limit)
columns=['Tweets']
data=[]
# def clean(text):
#   return ' '.join(re.sub("(@[A-Za-z0-9]+)|([^0-9A-Za-z \t])|(\w+:\/\/\S+)", " ", text).split()
# for tweet in tweets:
#   a=api.clean(tweet)
#   data.append([a])
for tweet in tweets:
  data.append([tweet.full_text])
df=pd.DataFrame(data,columns=columns)
print(df)
```

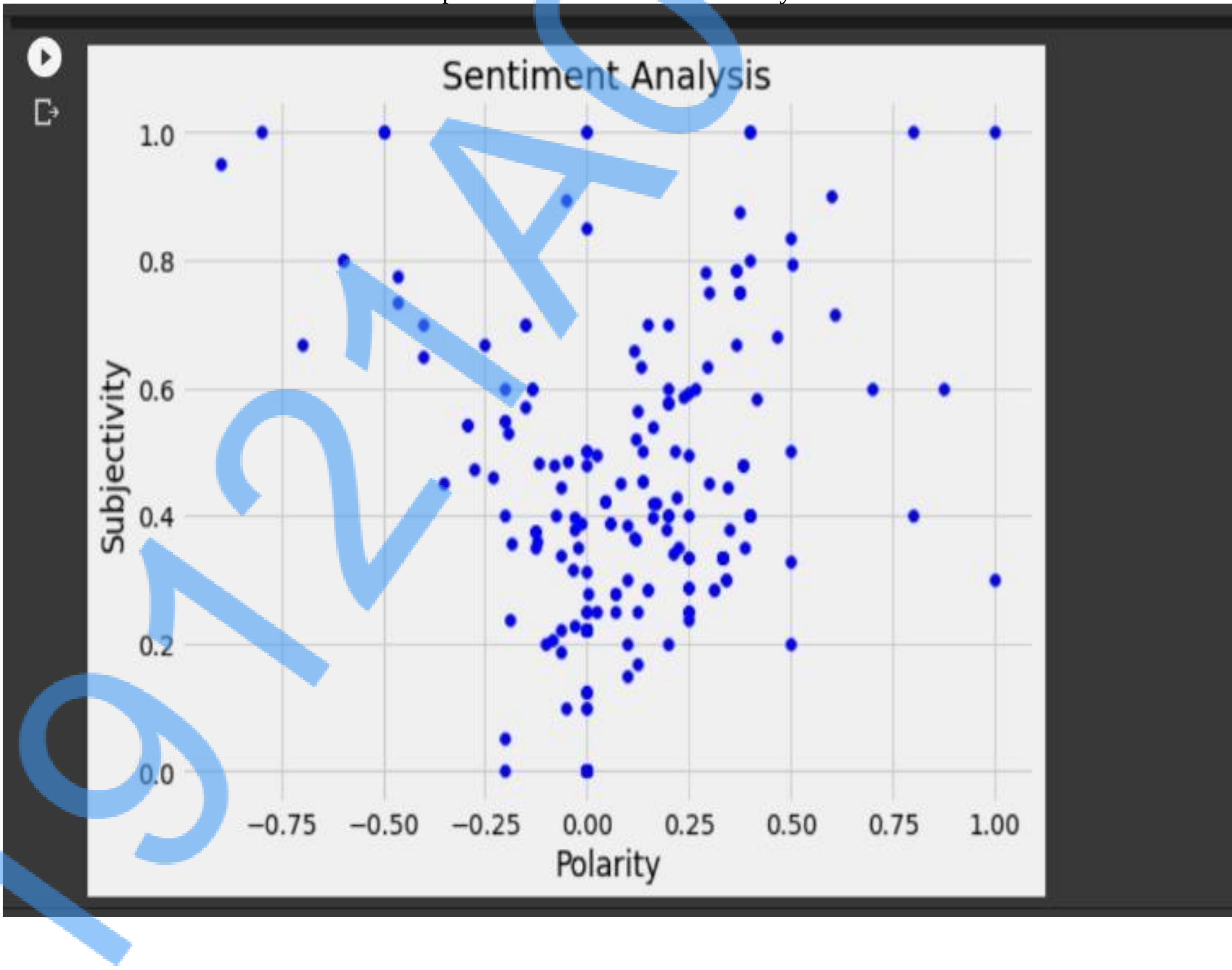
```
                                         Tweets
0    RT @The_MaquinaEN: NEW ERA \nIt truly feels th...
1    RT @MrPatNguyen: Dusk till dawn.\n\n#workfromh...
```

5.3.1 OUTPUT

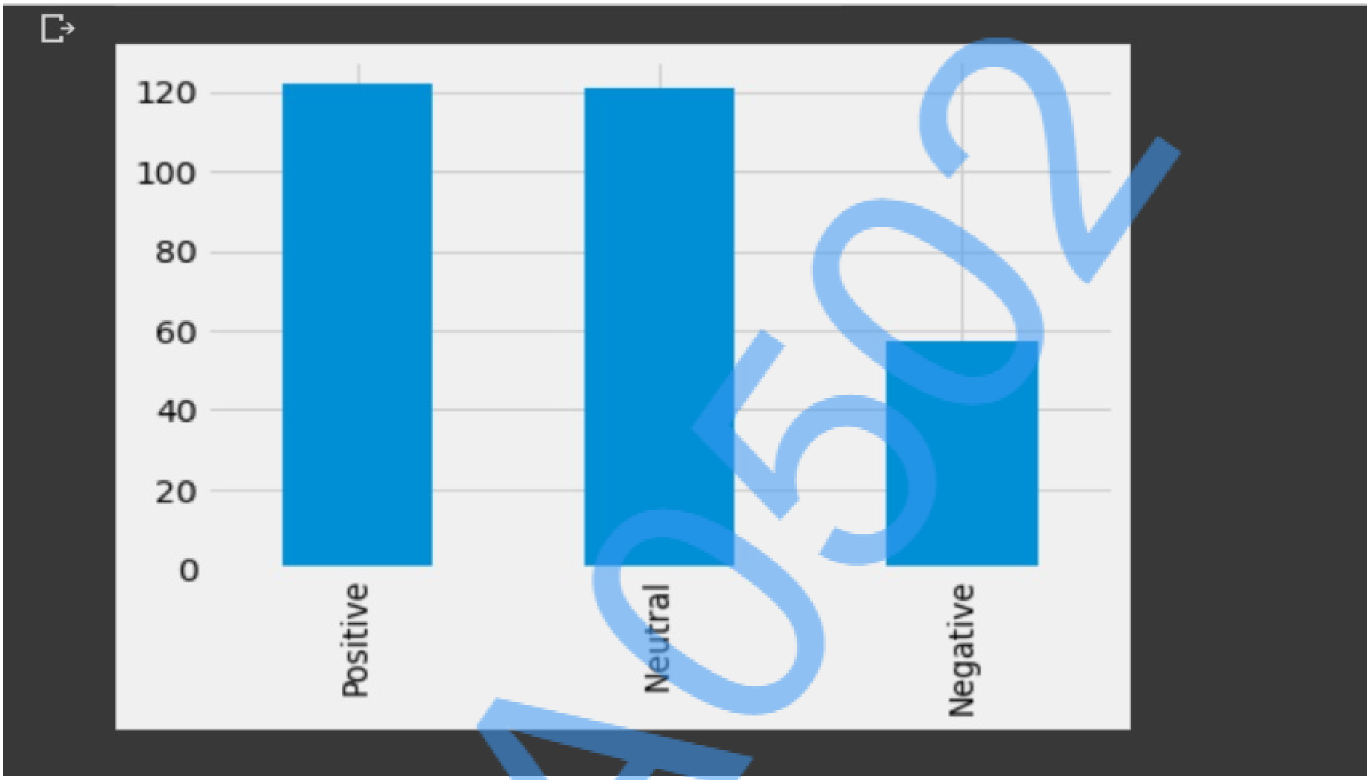
5.3.1.1 Word cloud based on frequency of words

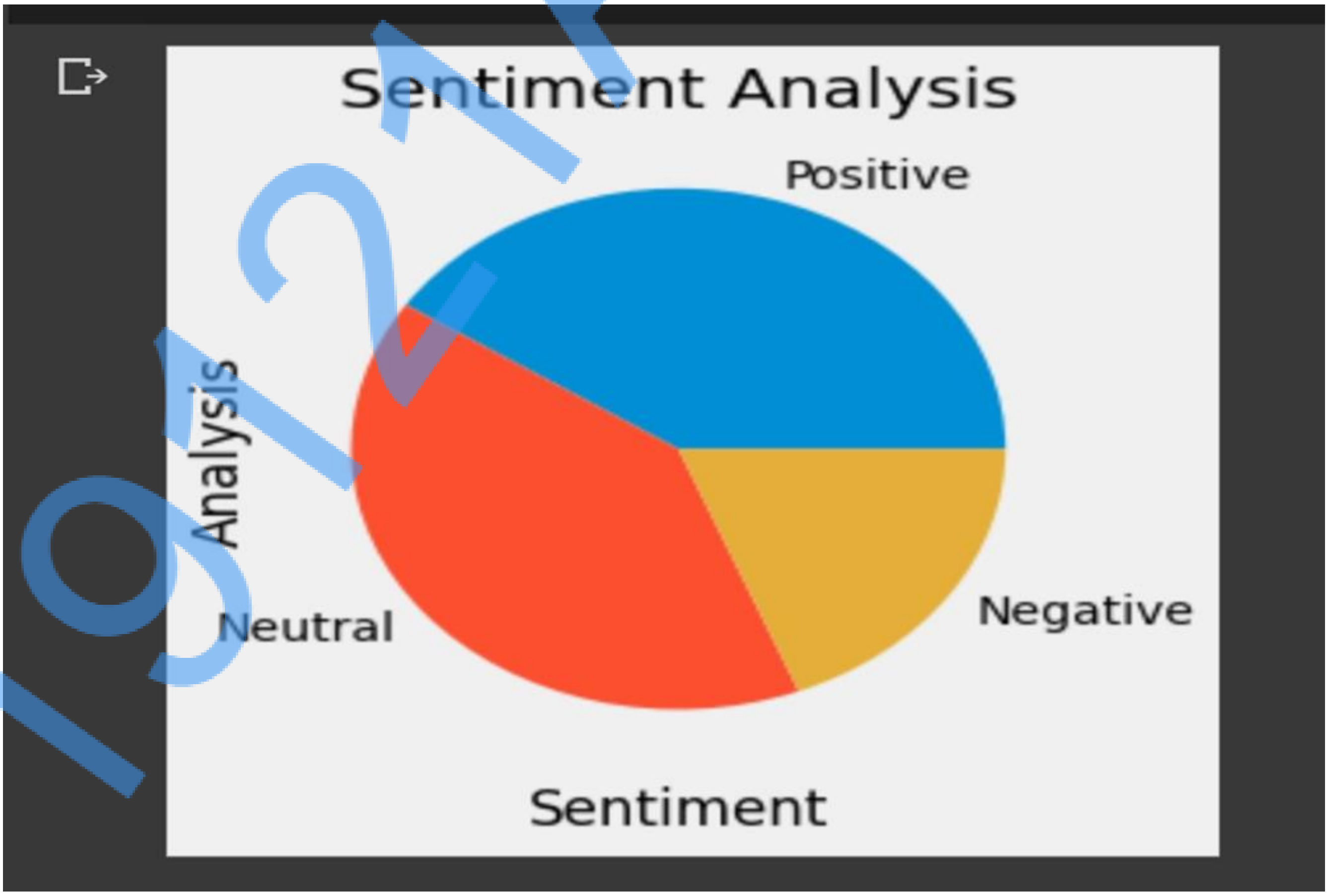


5.3.1.2 Representation of sentimental analysis



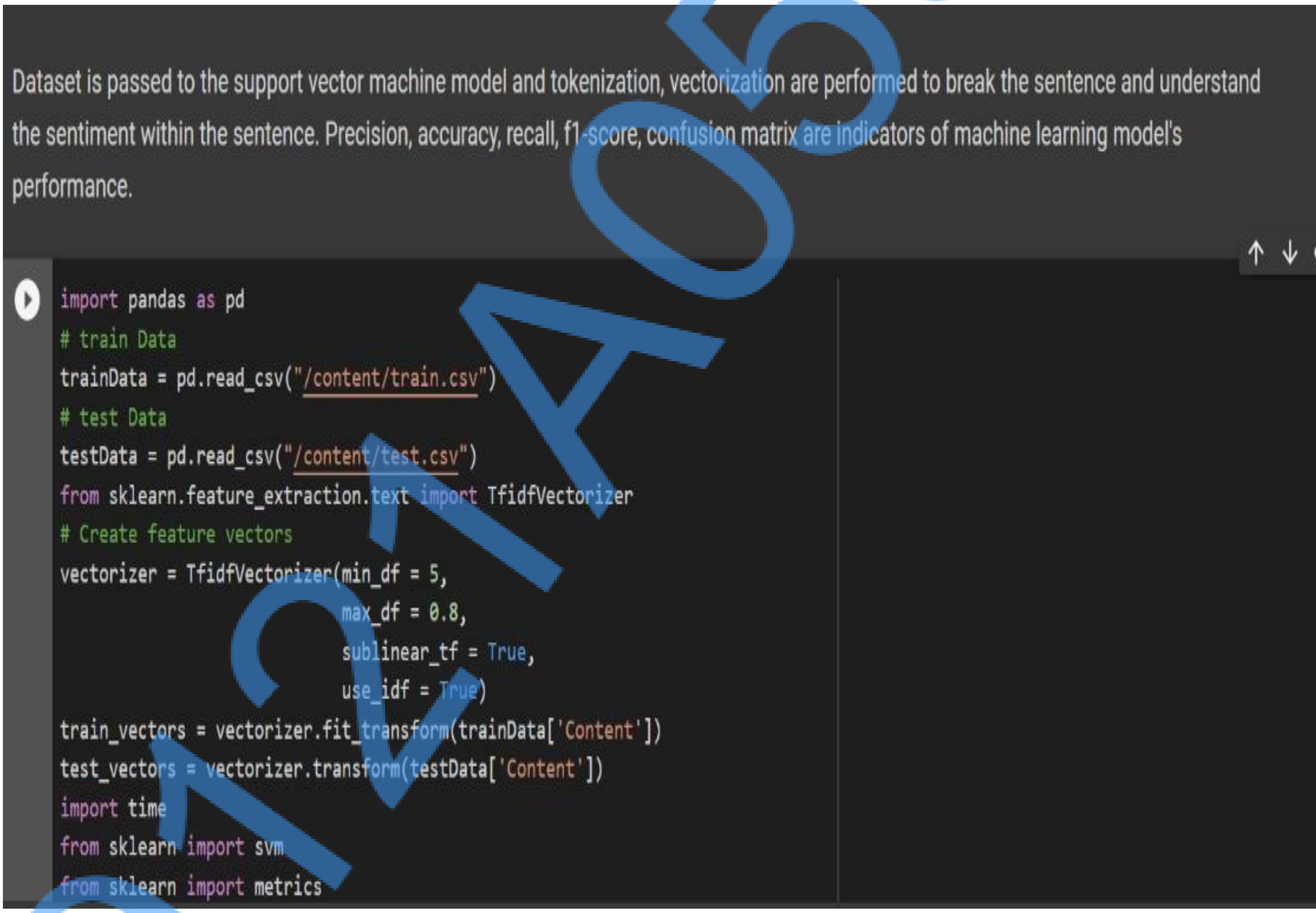
5.3.1.3 Visualization in bar graph and Pie chart





5.4 EXPERIMENT RESULTS AND ANALYSIS

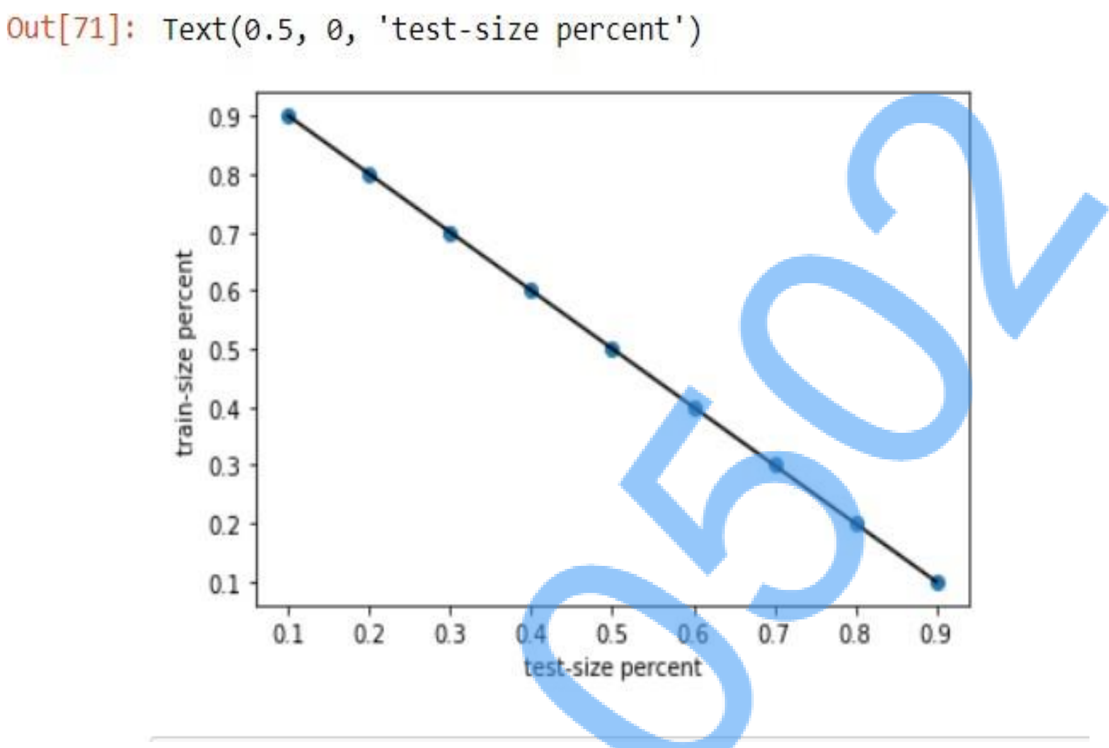
ACCURACY COMPARISON:



Dataset is passed to the support vector machine model and tokenization, vectorization are performed to break the sentence and understand the sentiment within the sentence. Precision, accuracy, recall, f1-score, confusion matrix are indicators of machine learning model's performance.

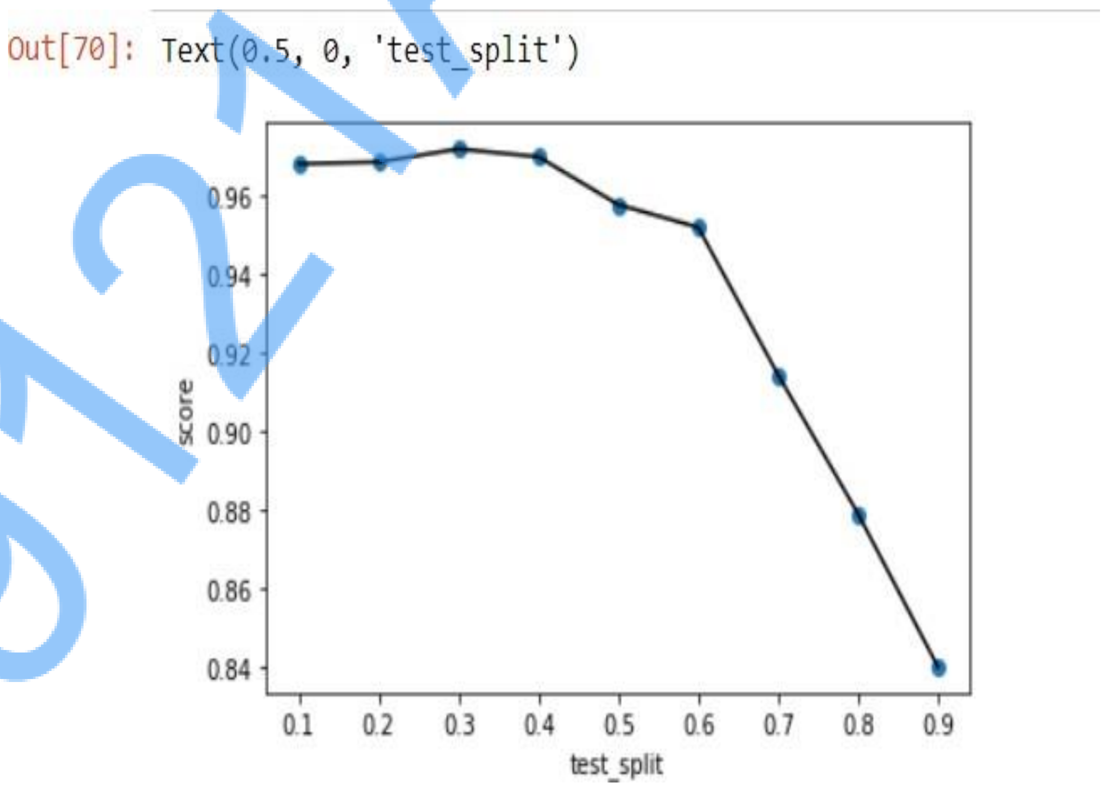
```
import pandas as pd
# train Data
trainData = pd.read_csv("/content/train.csv")
# test Data
testData = pd.read_csv("/content/test.csv")
from sklearn.feature_extraction.text import TfidfVectorizer
# Create feature vectors
vectorizer = TfidfVectorizer(min_df = 5,
                             max_df = 0.8,
                             sublinear_tf = True,
                             use_idf = True)
train_vectors = vectorizer.fit_transform(trainData['Content'])
test_vectors = vectorizer.transform(testData['Content'])
import time
from sklearn import svm
from sklearn import metrics
```

Fig 5.4.1 Passing dataset to model



```
Out[71]: Text(0.5, 0, 'test-size percent')
```

Fig 5.4.2 Train data and Test data(Their values affect final scores Fig 5.4.3)

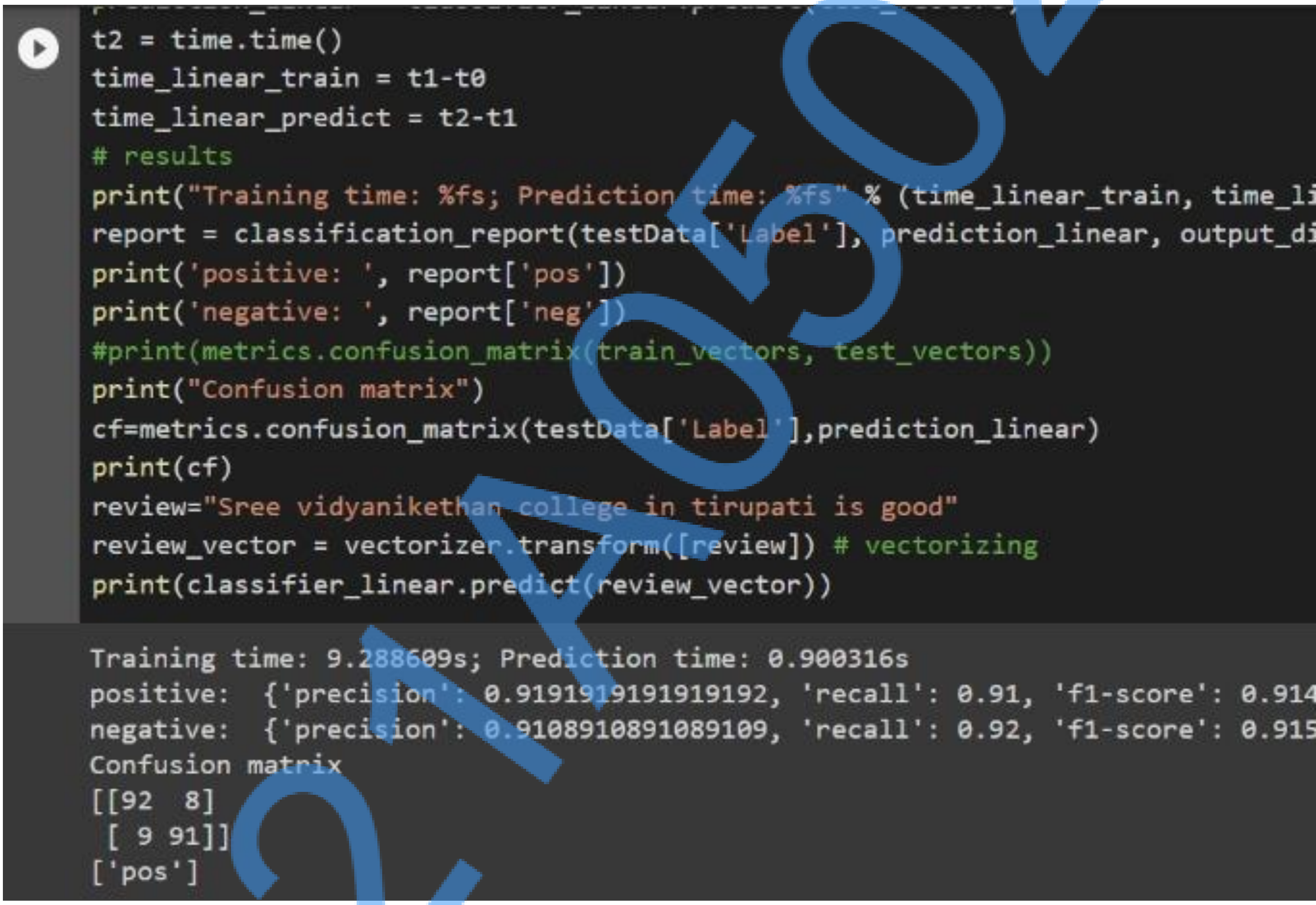


```
Out[70]: Text(0.5, 0, 'test_split')
```

Fig 5.4.3 Scores based on test_split

Thus it can be inferred from these figures that test-split (which means selecting some percent as test data and the rest as train data) affects the final score of our classifier. So choosing an optimal value is a necessity in our case test-split=0.3 ended up with the maximum score.

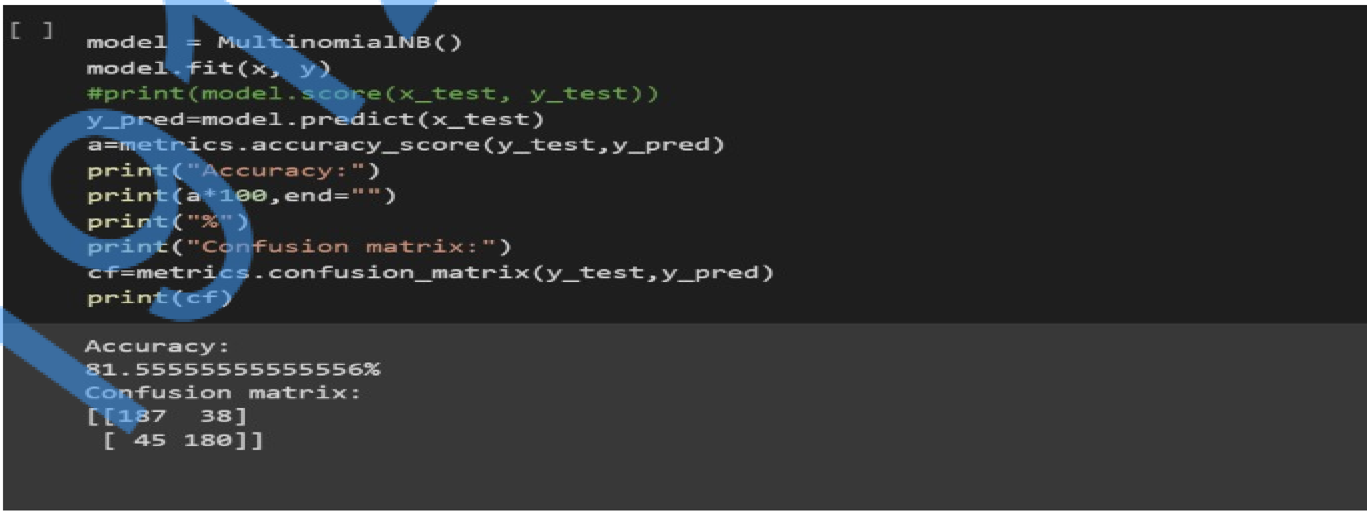
5.4.4 Accuracy of SVM model



```
t2 = time.time()
time_linear_train = t1-t0
time_linear_predict = t2-t1
# results
print("Training time: %fs; Prediction time: %fs" % (time_linear_train, time_li
report = classification_report(testData['Label'], prediction_linear, output_di
print('positive: ', report['pos'])
print('negative: ', report['neg'])
#print(metrics.confusion_matrix(train_vectors, test_vectors))
print("Confusion matrix")
cf=metrics.confusion_matrix(testData['Label'],prediction_linear)
print(cf)
review="Sree vidyanikethan college in tirupati is good"
review_vector = vectorizer.transform([review]) # vectorizing
print(classifier_linear.predict(review_vector))
```

```
Training time: 9.288609s; Prediction time: 0.900316s
positive:  {'precision': 0.9191919191919192, 'recall': 0.91, 'f1-score': 0.914
negative:  {'precision': 0.9108910891089109, 'recall': 0.92, 'f1-score': 0.915
Confusion matrix
[[92  8]
 [ 9 91]]
['pos']
```

5.4.5 Accuracy of Naïve Bayes model



```
[ ]
    model = MultinomialNB()
    model.fit(x, y)
    #print(model.score(x_test, y_test))
    y_pred=model.predict(x_test)
    a=metrics.accuracy_score(y_test,y_pred)
    print("Accuracy:")
    print(a*100,end="")
    print("%")
    print("Confusion matrix:")
    cf=metrics.confusion_matrix(y_test,y_pred)
    print(cf)
```

```
Accuracy:
81.55555555555556%
Confusion matrix:
[[187  38]
 [ 45 180]]
```

6. CONCLUSION AND FUTURE WORK

6.1 CONCLUSION

We furnished results for Sentiment analysis on twitter data . On applying Logistic regression, Bernouille Naive Bayes, Multinomial Naive Bayes, Support vector machine for sentiment analysis Support vector machine stands out with 91.4% accuracy at test_split=0.3. Users topic of interest for sentiment analysis has been considered ,So that they may get to know the statistics of sentiment behind the topic of their own interest. We firmly conclude that implementing sentiment analysis using these algorithms will help in deeper understanding of textual data which can essentially serve a potential platform for businesses .

6.2 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.

7. REFERENCES

[1]. Shiv Dhar, Suyog Pednekar, Kishan Borad- Methods for Sentiment Analysis Computer Engineering, VIVA Institute of Technology, University of Mumbai, India.

[2].Ravinder Ahujaa, Aakarsha Chuga, Shruti Kohlia,Shaurya Guptaa,Pratyush Ahujaa- The Impact of Features Extraction on the Sentiment Analysis Noida, India.

[3].Richa Mathur, Devesh Bandil, Vibhakar Pathak-Analyzing Sentiment of Twitter Data using Machine Learning Algorithm

[4].Machine Learning Tom M. Mitchell

[5].An Idiot's guide to Support vector machines (SVMs) R. Berwick, Village Idiot

[6].https://www.svm-tutorial.com/2014/10/svm-linear-kernel-good-text-classification

[7].https://towardsdatascience.com/naive-bayes-classifier-81d512f50a7c

[8].https://www.ijitee.org/wp-content/uploads/papers/v8i8/H6330068819.pdf