ANALYSIS OF TWITTER SENTIMENTS USING MACHINE LEARNING ALGORITHMS

Submitted in partial fulfillment of the requirements for the award of
Bachelor of Engineering degree in Computer Science and Engineering

By

AMIRTHA VARSINI K S (Reg No - 39110053)
VIDHYA S (Reg No - 39111090)

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

SCHOOL OF COMPUTING

SATHYABAMA
INSTITUTE OF SCIENCE AND TECHNOLOGY
(DEEMED TO BE UNIVERSITY)
Accredited with Grade "A" by NAAC | 12B Status by UGC | Approved by AICTE
JEPPIAAR NAGAR, RAJIV GANDHI SALAI,
CHENNAI - 600119

APRIL- 2023

SATHYABAMA
INSTITUTE OF SCIENCE AND TECHNOLOGY
(DEEMED TO BE UNIVERSITY)
Accredited with Grade "A" by NAAC | 12B Status by UGC | Approved by AICTE

www.sathyabama.ac.in

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

BONAFIDE CERTIFICATE

This is to certify that this Project Report is the bonafide work of **Amirtha Varsini K S (39110053) and Vidhya S(39111090)** who done the project work as a team who carried out the Project phase-2 entitled **"ANALYSIS OF TWITTER SENTIMENTS USING MACHINE LEARNING ALGORITHMS"** under my supervision from January 2023 to April 2023.

Internal Guide
Dr.P. ASHA , M.E., Ph.D.

Head of the Department
Dr. L. LAKSHMANAN, M.E., Ph.D.

Submitted for Viva voce Examination held on 20.04.2023

SATHYABAMA INSTITUTE OF SCIENCE AND TECHNOLOGY
DEPARTMENT OF CSE

Internal Examiner **External Examiner**

DECLARATION

I, **AMIRTHA VARSINI K S (Reg No-39110053),** hereby declare that the Project Report entitled "**ANALYSIS OF TWITTER SENTIMENTS USING MACHINE LEARNING ALGORITHMS"** done by me under the guidance of **Dr.P.ASHA M.E., Ph.D.** is submitted in partial fulfillment of the requirements for the award of Bachelor of Engineering degree in **Computer Science and Engineering**.

amirtha varsini k s

DATE: 20.4.23

VIDHYA S

PLACE: Chennai **SIGNATURE OF THE CANDIDATE**

ACKNOWLEDGEMENT

I am pleased to acknowledge my sincere thanks to **Board of Management** of **SATHYABAMA** for their kind encouragement in doing this project and for completing it successfully. I am grateful to them.

I convey my thanks to **Dr. T. Sasikala M.E., Ph.D**, Dean, School of Computing, **Dr. L. Lakshmanan M.E., Ph.D.,** Head of the Department of Computer Science and Engineering for providing me necessary support and details at the right time during the progressive reviews.

I would like to express my sincere and deep sense of gratitude to my Project Guide **Dr.P. ASHA M.E.,Ph.D,** for her valuable guidance, suggestions and constant encouragement paved way for the successful completion of my phase-2 project work.

I wish to express my thanks to all Teaching and Non-teaching staff members of the **Department of Computer Science and Engineering** who were helpful in many ways for the completion of the project.

ABSTRACT

With the advancement of web technology and its growth, there is a huge volume of data present in the web for internet users and a lot of data is generated too. Internet has become a platform for online learning, exchanging ideas and sharing opinions. Social networking sites like Twitter, Facebook, Google+ are rapidly gaining popularity as they allow people to share and express their views about topics, have discussion with different communities, or post messages across the world. There has been lot of work in the field of sentiment analysis of twitter data. This survey focuses mainly on sentiment analysis of twitter data which is helpful to analyze the information in the tweets where opinions are highly unstructured, heterogeneous and are either positive or negative, or neutral in some cases. Social media have received more attention nowadays. Public and private opinion about a wide variety of subjects are expressed and spread continually via numerous social media. Twitter is one of the social media that is gaining popularity. So this study provides the ratio of negative positive and neutral comments from twitter . Twitter sentiment analysis is helpful for classifying the tweets depending upon the polarity which is positive negative and neutral. Analysing the twitter data for this analysis is tough because it has a large amount of data as the twitter contains data on various topics. For our prediction, supervised classification algorithms have been used. Using distinctive supervised machine learning algorithms such as K-nearest neighbor, support vector machine, naïve bayes, logistic regression , decision tree, and random forest algorithms were used to analyze the sentiment of tweets. These machine learning algorithms has been compared, and the most effective method is chosen to predict the outcome.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF ABBREVIATION

S.No	ABBREVIATIONS	EXPANSION
1	API	Application programming Interface
2	DT	Decision Tree
3	KNN	K-Nearest Neighbour
4	LR	Logistic Regression
5	ML	Machine Learning
6	NB	Naïve Bayes
7	RF	Random Forest
8	SA	Sentiment Analysis
9	SVM	Support Vector Machine

CHAPTER 1

INTRODUCTION

Millions of individuals use social media, claims network sites to reveal their thoughts, feelings, and emotions regarding the people's daily existence. However, anything can be written about, including social interactions or product reviews. By means of Online communities offer a dialogue forum where Consumers inform and sway other people. Furthermore, social media offers a chance for business that gives a social networking platform for connecting with its customers to promote or communicate with consumers directly in order to the viewpoint of the customer regarding goods and services.

Contrarily, consumers have complete authority over purchasing decisions. what viewers want to see and how viewers react With this, the business's success and failure are made public resulting through word of mouth The social network, however may alter cc ix er behavior and decision-making, As an illustration, notes that 87% of internet users are Customers' choices and purchases are impacted by them review.

In order for the organization to catch up more quickly with what Customers believe that organizing would be more advantageous should be quick to respond and develop a winning plan competitors of theirs People are using social media sites like Twitter, which produce large volumes of opinion writings in the form of tweets and are available for sentiment analysis, while the World Wide Web expands quickly.

From a human perspective, this corresponds to a vast volume of information, making it challenging to quickly extract sentences, read them, analyze tweet by tweet, summarize them, and organize them into an understandable style.

Additionally, social media gives businesses a chance by offering them a platform to engage with their customers for advertising. People heavily rely on user-generated content from the internet when making decisions. For instance, if someone wants to purchase a good or use a service, they will research it online and discuss it on social media before making a choice.

The volume of user-generated content is too great for a typical user to process. Since this needs to be automated, many different sentiment analysis approaches are in use Before a user purchases a product, sentiment analysis (SA) lets them know if the product's information is good or not. Marketers and businesses utilize these analytical results to learn about their products or services in such a way that it can be offered as per the user’s requirements.

The fundamental goals of textual information retrieval strategies are to process, search for, or examine the factual material that is already there.

Even if facts have an objective component, some other literary contents exhibit subjective traits. Sentiment Analysis's fundamental components opinions, sentiments, assessments, attitudes, and emotions are primarily represented by these contents (SA). In large part because of the enormous expansion in the amount of information available online from sources like blogs and social networks, it presents many challenging chances to design new applications.

For instance, by using SA and taking into account factors such as positive or negative attitudes about the goods, recommendations of items proposed by a recommendation system can be predicted. This technique could be used for different purposes such as politicians could use it for analyzing what kind of sentiments people from different areas are carrying towards him/her and hence could invest more in those areas.

An example of this is recent Trump elections, where he hired a group of analysts for this specific purpose. Sentiment analysis could also be applied in the field of business marketing. With the help of this technology different business organizations capture the feelings of people regarding their products and of that of their competitors. Organizations employ their strategies with accordance to this knowledge only. Leaving market research aside , analysis of sentiments could play a vital part in Service industries As it could analyze a full fledged customer experience and could reveal customer feeling, which could prove to be very beneficial.

1.1 ***OBJECTIVE***

The main objective of this project is to analyze the sentiment of the tweets and convert it into a web application by using flask framework.

1.2 ***OUTLINE OF THE PROJECT***

First, we must convert our Twitter account to a developer account. Then, we must request an API and complete an application for a Twitter developer account. Now, we must choose the relevant use-case and offer a project description. The next step is to link an existing App or build a new App. You will receive your Bearer Token and API Keys, which you can use to access to the Twitter API v2's new endpoints. Then after we get the acceptance letter from twitter ,we can proceed by converting the tweets into a dataset by using python language in Visual studio code, the raw data will next undergo preprocessing on the Google Colab platform. After that, several Machine Learning techniques will be used to compare and analyse which model performs best. And the output is predicted, the ratio of the positive negative and neutral tweets is found. Then the web application is build using flask which is a python framework.

1.3 ***APPLICATION OF TWITTER SENTIMENT ANALYSIS***

Some popular applications of sentiment analysis include,

- ✓ Social media monitoring
- ✓ Customer support management
- ✓ Analysing customer feedback
- ✓ Provide an overview of an brands opinion
- ✓ Understand your competitors strategy
- ✓ Market research
- ✓ Improve crisis management
- ✓ Evaluate social media post
- ✓ Brand strength
- ✓ Product Popularity

✓ Movie reviews

CHAPTER 2

AIM AND SCOPE OF THE PROJECT

2.1 AIM

The main aim of our project is collecting the tweets directly by using the twitter Application Programming Interface (API) from the twitter and analyse the sentiment of the tweets by using the polarity value and based on the polarity value we build the machine learning model and compare various Machine Learning classification algorithms and find out wich gives the highest accuracy and to deploy the ML model into web application which helps the users to identify the positive and negative tweets.

2.2 SCOPE

The scope of our project is to help online users to find the positive and negative informations of a certain topic. Twitter Sentiment Analysis is the process of determining the emotional tone behind a series of words specifically on Twitter. A sentiment analysis tool is an automated technique that extracts meaningful customer information related to their attitudes, emotions and opinions. Twitter sentiment analysis allows users to keep track of whats being said about a product or service on social media and can help detect angry customers or negative mentions before they escalate. This primarily focuses on sentiment analysis of twitter data, which is useful for analysing information in tweets. Twitter gives businesses a quick and efficient approach to examine user opinions on issues that are crucial to their performance in the marketplace.

CHAPTER 3

LITERATURE SURVEY

3.1 SURVEY

These days analysis of feelings from twitter is on constant appraisal within the research community as its applications have a huge influence over the working of different industries today. The main challenge faced by this type of analysis is the variation of speech and complex structure of data when extracted.

Aliza Sarlan, Shuib and Chayanit conducted experiments on twitter data in which they simply extracted the tweets in Jason format and used python lexicon dictionary to assign polarity to the tweets. On the other hand Mandava Geeta, Bhargavav and Duvvada turned it up a notch and used learning methods for the same purpose and achieved a better accuracy of result. For this they collected data regarding cryptocurrency and applied algorithms like naïve bayes and SVM on it. These experiments further concluded that naïve bayes classifier has more accuracy then SVM.

Another research was conducted by Agarwal, Xie,Vovshaa, I., Rambow, O., and also Passonneau in which a unigram model was used as a baseline and was compared with other models such as one, model based on features and another model based on kernel tree . The experiments revealed that feature based model out performed the unigram model with a negligible margin where as both unigram as well as feature based models were outperformed by kernel tree based model with a significant margin.

Akshi Kumar and Teeja Mary Sebastian proceeded with an approach which was a combination of corpus based as well as lexicon based approach . This combination is very rarely found in the work that has being done in this field as machine learning techniques are taking over. In their experiments they have used adjectives and verbs as their features and have used corpus based techniques for finding the semantic orientation of various adjectives present in the tweets and as for the verbs they have used lexicon dictionary. A linear equation is used to convey the total sentiment polarity of tweets.

K. Arun et al gathered data on different aspects of demonetization from twitter. They used R language as a tool for analyzing these tweets. Not only were the tweets analyzed but the result was visualized using different projections such as word cloud and other different plots. These plots showed that the number of people accepting demonetization is more then the number of people rejecting it.

Vaibhavi N. Patodkar, Imran R. Shaikh aimed to predict the emotions behind the audience watching a random tv show as positive or negative. For this purpose they extracted comments regarding some random tv shows and used these as data set for training and testing the model. The model they choose was naïve bayes classifier for which a result was displayed using a pie chart. This pie chart concluded that the polarity of tweets with respect to negative is more then that of positive.

As stated in the section above sentiment analysis could be used for politics . Tumasjan et al came across the field and its benefits in election and used it for predicting the results in 2009 for German federal elections. They extracted approximately 100,000 tweets for this purpose regarding many political parties of that time and area. Then analyzed the tweets in order to gain sentiments for them .For this they used a software popularly known as(Linguistic Inquiry and Word Count) LIWC2007. This software uses textual analysis as a base to derive sentiments. The results obtained by this analysis were very much similar to the actual results of the elections.

Another interesting research was carried out by Dr Rajiv along with some of his mates .They have applied the technique of sentiment analysis in a brand new way , where they have used this technique to better situations in crises situations. They collected the data of 2014 about a deluge which occurred in Kashmir at that time. Data set collected by them consisted of almost 8490 tweets on which naïve bayes classification technique was implemented. Their research showed that applying analysis of feeling in these situations of crises could help the government in saving lives.

3.2 INFERENCES FROM LITREATURE SURVEY

Sentiment Analysis(SA) approach is classified into two types lexicon based and Machine Learning based approaches. Vishal A Kharde and S.S SonaWane has used both Lexicon and Machine Learning based approaches but they got better accuracy from Machine Learning approach.

Aliza Sarlan, Shuib and Chayanit conducted experiments on twitter data in which they simply extracted the tweets in Jason format and used python lexicon dictionary to assign polarity to the tweets.

According to our research, Support Vector Machine(SVM), K-Nearest Neighbour (KNN) and Naïve Bayes algorithms were used.

Classification Algorithms were used inorder to classify the tweets into positive , negative and neutral tweets.

Authors Abdullah alseedi and mohammad Zubair khan used Support Vector Machine(SVM) Algorithm and Author Faizan used K-Nearest Neighbour (KNN) algorithm. By comparing both the research papers we conclude that Support Vector Machine (SVM) has got the better accuracy than K-Nearest Neighbour (KNN) algorithm.

CHAPTER 4

PROPOSED METHODOLOGY

4.1 SOFTWARE REQUIREMENTS

- Python language is used
- Platfrom – Windows 11
- Visual studio code – For converting the twitter data into an dataset using the python code and building a web application
- Google colab – Were used used for building ML models
- Flask Framework – To deploy ML models into web application

4.2 HARDWARE REQUIREMENTS

The hardware requirements are needed to serve as the basis for implementation of the system and hence should be an absolute and coherent specification of the entire system. The software engineers use the hardware requirements as the starting point for the system design. It indicates what the system performs and what the system should execute.

PROCESSOR – Core i5-1035G1

RAM – 8 gb

4.3 PYTHON

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.

4.4 APPLICATIONS OF PYTHON

The major applications of Python Include

- ✓ Web development
- ✓ Video game development
- ✓ Data science
- ✓ Software development
- ✓ Machine Learning
- ✓ Web application
- ✓ Artificial Intelligence
- ✓ Digital Image Processing
- ✓ Embedded system
- ✓ Automation
- ✓ E-commerce
- ✓ TensorFlow
- ✓ Keras
- ✓ Data and Information Visualization
- ✓ Tkinter

4.5 FEATURES OF PYTHON

Python provides many useful features which make it popular and valuable from the other programming languages. It supports object-oriented programming, procedural programming approaches and provides dynamic memory allocation. We have listed below a few essential features.

1) Easy to Learn and Use

Python is easy to learn as compared to other programming languages. Its syntax is straightforward and much the same as the English language. There is no use of the semicolon or curly-bracket, the indentation defines the code block. It is the recommended programming language for beginners.

2) Expressive Language

Python can perform complex tasks using a few lines of code. A simple example, the hello world program you simply type **print("Hello World")**. It will take only one line to execute, while Java or C takes multiple lines.

3) Interpreted Language

Python is an interpreted language; it means the Python program is executed one line at a time. The advantage of being interpreted language, it makes debugging easy and portable.

4) Cross-platform Language

Python can run equally on different platforms such as Windows, Linux, UNIX, and Macintosh, etc. So, we can say that Python is a portable language. It enables programmers to develop the software for several competing platforms by writing a program only once.

5) Free and Open Source

It has a large community across the world that is dedicatedly working towards make new python modules and functions. Anyone can contribute to the Python community. The open-source means, "Anyone can download its source code without paying any penny."

6) Object-Oriented Language

Python supports object-oriented language and concepts of classes and objects come into existence. It supports inheritance, polymorphism, and encapsulation, etc. The object-oriented procedure helps to programmer to write reusable code and develop applications in less code.

7) Extensible

It implies that other languages such as C/C++ can be used to compile the code and thus it can be used further in our Python code. It converts the program into byte code, and any platform can use that byte code.

8) Large Standard Library

It provides a vast range of libraries for the various fields such as machine learning, web developer, and also for the scripting. There are various machine learning libraries, such as Tensor flow, Pandas, Numpy, Keras, and Pytorch, etc. Django, flask, pyramids are the popular framework for Python web development.

9) GUI Programming Support

Graphical User Interface is used for the developing Desktop application. PyQT5, Tkinter, Kivy are the libraries which are used for developing the web application.

10) Integrated

It can be easily integrated with languages like C, C++, and JAVA, etc. Python runs code line by line like C,C++ Java. It makes easy to debug the code.

11) Embeddable

The code of the other programming language can use in the Python source code. We can use Python source code in another programming language as well. It can embed other language into our code.

12) Dynamic Memory Allocation

In Python, we don't need to specify the data-type of the variable. When we assign some value to the variable, it automatically allocates the memory to the variable at run time. Suppose we are assigned integer value 15 to **x,** then we don't need to write **int x = 15.** Just write x = 15.

4.6 ARCHITECTURE OVERALL DESIGNS OF PROPOSED SYSTEM

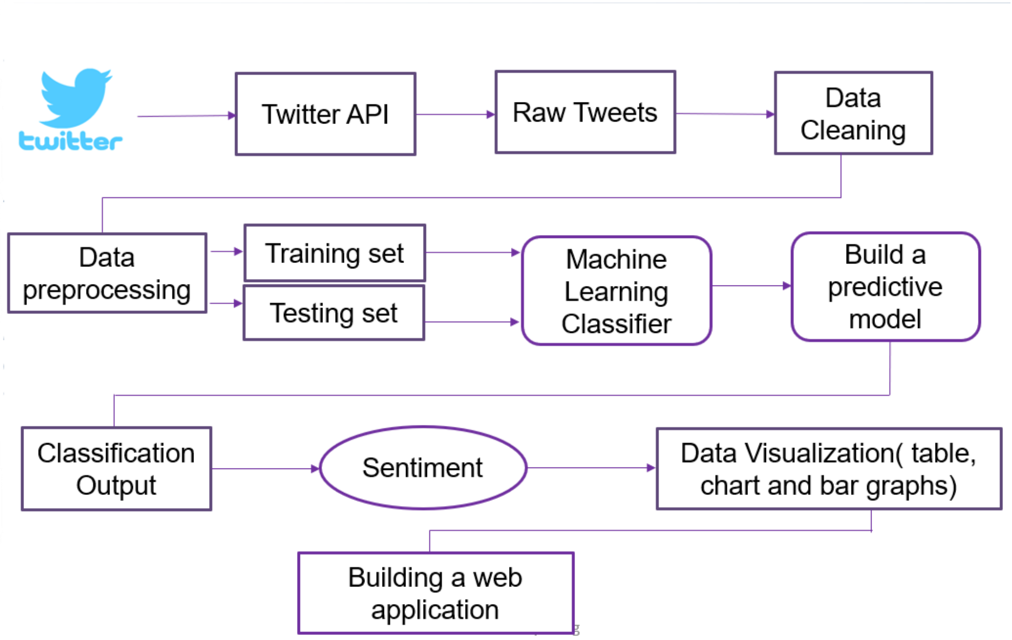


Fig 4.1.System Architecture

The block diagram of the proposed system has been shown in the above figures. At first the API is requested and collected from the twitter, then it is converted into a dataset, then the raw data is cleaned using data cleaning and data preprocessing to a machine understandable format . Then the dataset is split into training and testing data. The data which is split were given to machine learning algorithms and it has build a predictive model and shows the classification output. The output consists of positive negative or neutral tweets. Then the output is visualized , and web application is build.

4.7 ALGORITHM STEPS

STEP 1 **:** Convert the twitter account to developer account.

STEP 2 **:** Request API from Twitter developer account and get API keys and the Bearer Token.

STEP 3 **:** Using Visual studio convert the data collected from the twitter using API key into dataset.

STEP 4 **:** Build various Machine Learning model and compare which gives the highest accuracy.

STEP 5 **:** Deploy Machine Learning model into web application using flask framework.

After deploying ML model into web application , get the input from the user and it has to analyze that the particular tweet is positive,negative or neutral . The analysis is based on the polarity value , if the polarity value is equal to 0 it is a neutral tweet , if it is less than 0 , it is negative and if it is greater than 0 it is positive.

CHAPTER 5
PROJECT DESIGN AND IMPLEMENTATION

5.1 EXISTING SYSTEM

When examining all the previous approaches utilised in earlier systems, the most prevalent issue is that the majority of these systems simply utilized a pre-defined dataset from Twitter to execute this problem statement. Additionally, they only used dynamic numbers and carried out sentiment analysis, thus they can only determine whether a tweet is positive or negative. They have established a method for analysing text from Twitter, particularly for the recently developed field of sentiment analysis. Some of the systems are also not very accurate. They didn't compare any algorithms for machine learning.

5.2 EXISTING MODEL DISADVANTAGES

- Web applications were not created due to the limitations of Dijango, which can only work on linux server or lamp. Thus it cannot be realized – Aliza Sarlan.
- In this case the author was not able to get better accuracy because of using K-Nearest Neighbour algorithm – Faizan.
- Both Lexicon and Machine Learning(ML) based approaches were used,

Lexicon aproach got the lower accuracy – S.S SonaWan , Vishal A Kharde.

5.3 PROPOSED WORK

Looking at the disadvantages of all the above methodologies used in previous systems, the most common point that pops up is most of these systems implemented this problem statement using only a pre-defined dataset from twitter. Also they had only dynamic values and only the sentiment analysis is carried, that is they say only the tweet is good or bad. They have demonstrated a system for the analysis of textual twitter data in particular for the emerging field of sentiment analysis. They have not created any web applications. Also some of the systems does not have good accuracy. They have not compared any Machine learning algorithms. Our system aims to overcome all these issues with the previously existing systems. Our basic idea is to collect API from twitter directly and we will not use any pre-defined datasets. The older systems will not provide any ratio but our system will provide the ratio between good and bad tweets and it will be static. And we will be creating an web application. A variety of Machine Learning algorithms like K-nearest neighbor, support vector machine, Naïve bayes, logistic regression , decision tree, and random forest algorithms will be used and compared to get better performance.

5.4 PROJECT MANAGEMENT PLAN

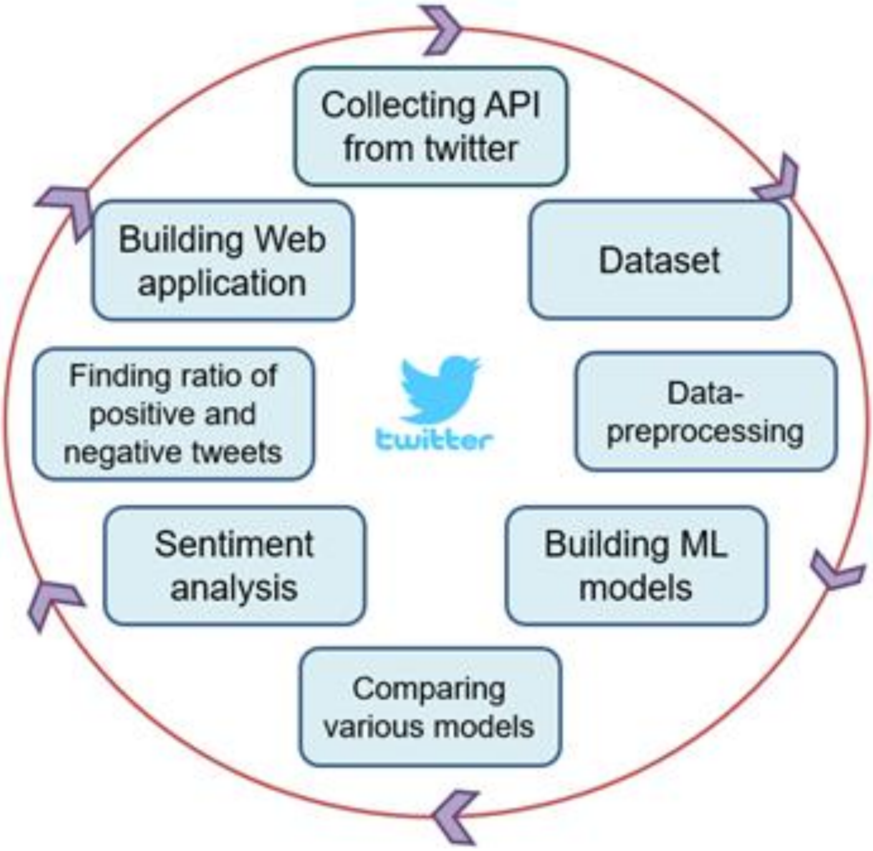


Fig 5.1. Project management plan

5.5 METHODOLOGY

5.5.1 <u>Collecting API</u> **–** A Project and App, which can be created in the Twitter developer account. A bearer token for the App's enrollment. You can find the bearer token with the keys and tokens of the App provided in the developer portal. Then after getting data from twitter it is made into a dataset through Visual Studio Code

5.5.2 ***<u>Data Analyzing</u>*** **-** In Machine Learning, Data Analysis is the process of inspecting, cleansing, transforming, and modeling data with the goal of discovering useful information by informing conclusions and supporting decision making.

5.5.3 ***Data-<u>Preprocessing</u>*** **-** Data preprocessing in Machine Learning refers to the technique of preparing (cleaning and organizing) the raw data to make it suitable for a building and training Machine Learning models.

5.5.4 ***Comparing Various ML models*** **-** Each model or any machine learning algorithm has several features that process the data in different ways. Often the data that is fed to these algorithms is also different depending on previous experiment stages. But, since machine learning teams and developers usually record their experiments, there's sample data available for comparison. The challenge is to understand which parameters, data, and metadata must be considered to arrive at the final choice. It's the classic paradox of having an overwhelming amount of details with no clarity. In our model various ML algorithms like K-nearest neighbor, support vector machine, naïve bayes, logistic regression , decision tree, and random forest algorithms.

5.5.5 ***Sentiment Analysis*** **–** From the dataset each tweet is checked using polarity check and the positive negative and neutral tweets are detected. The ratio of positive and negative tweet is calculated.

5.5.6 ***Web Application*** **–** A web application is made by using flask framework in visual studio code out of the best performed model.

CHAPTER 6

RESULTS AND DISCUSSION

6.1 RESULTS

Data visualization of Positive words that were used repeatedly in the tweets collected.

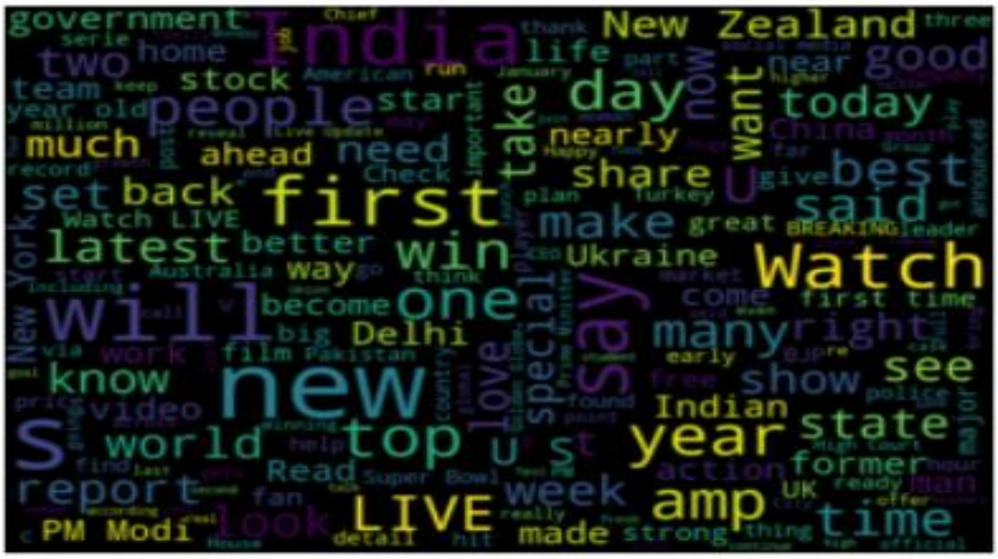


Fig 6.1. Positive words in tweets

Data visualization of negative words that were used repeatedly in the tweets collected.

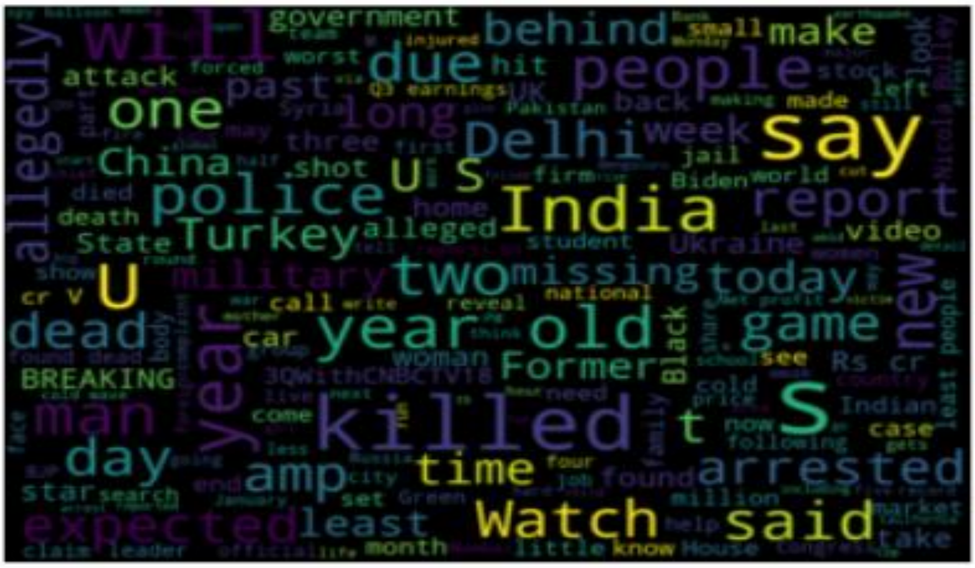


Fig 6.2. Negative words in tweets

Data visualization of neutral words that were used repeatedly in the tweets collected.

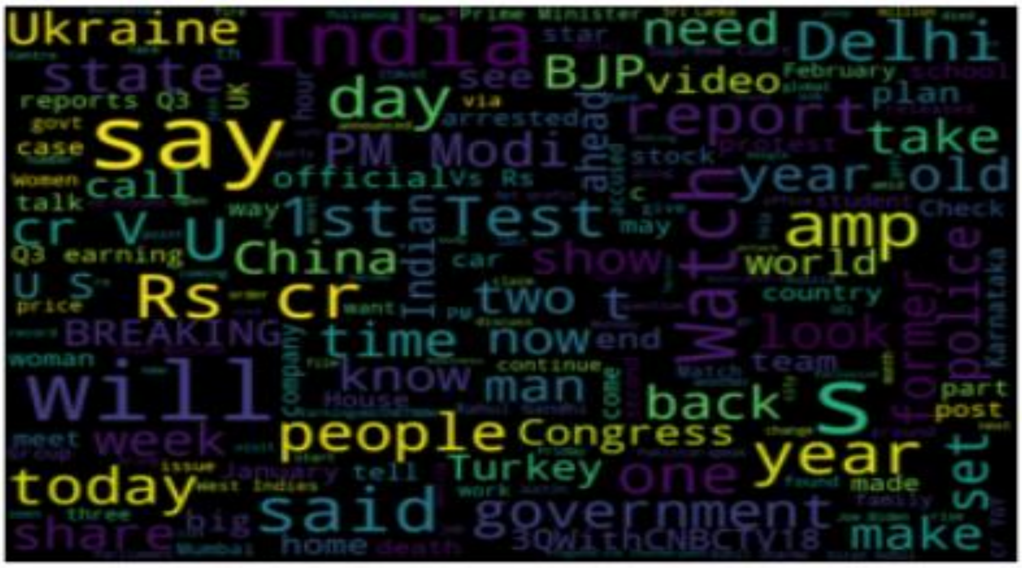


Fig 6.3. Neutral words in tweets

Heat map Analysis is a process of reviewing and analysing user behaviour.

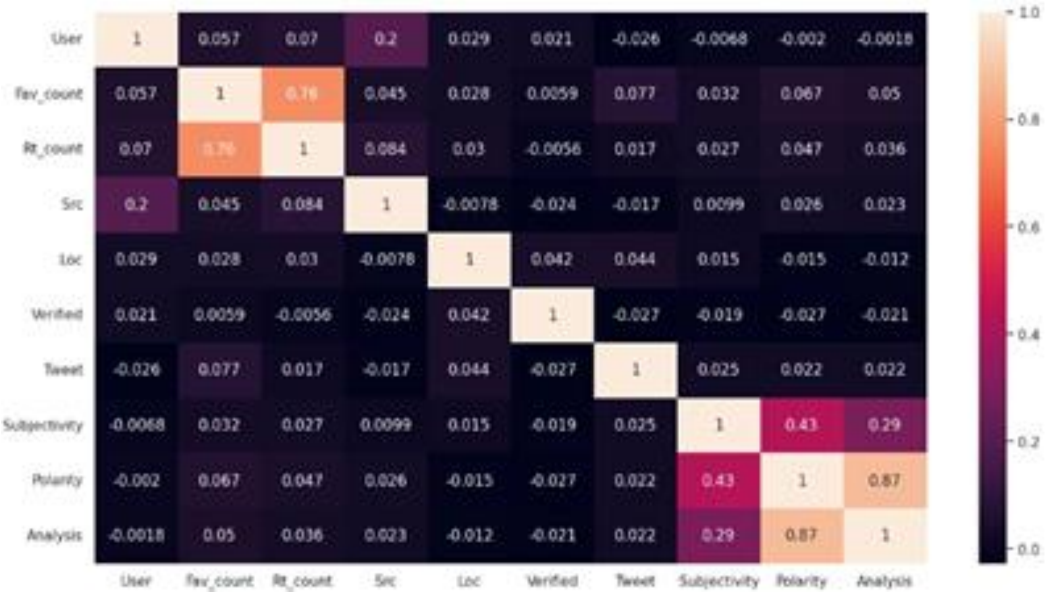


Fig 6.4. Heat map

The sentiment analysis is done, and a bar graph for tweets that were positive, negative, and neutral was displayed.

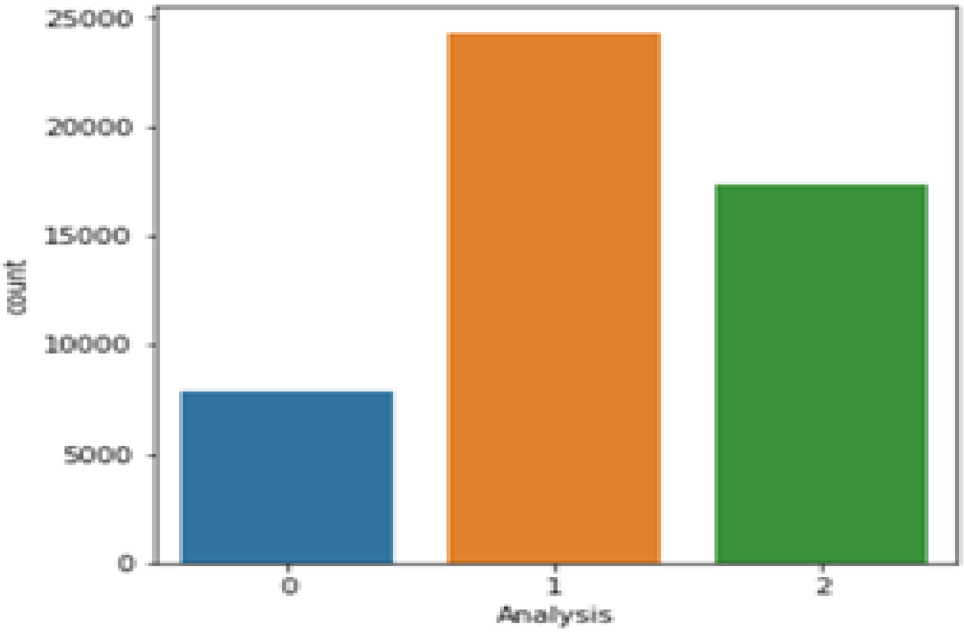


Fig 6.5. Sentiment Analysis

In this proposed work initially we analyze the polarity of the tweets based on the polarity value we segregate the tweets into positive, negative and neutral. And then the ratio of the tweets were found.

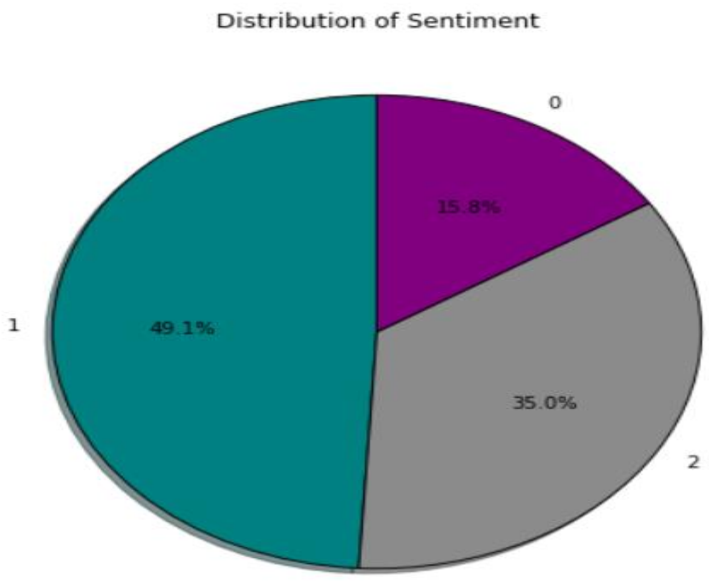


Fig 6.6. Ratio of tweets

By using six various ML algorithms and we have compared all the model to find the best performing model, and we have found that Random forest gives the better performance.

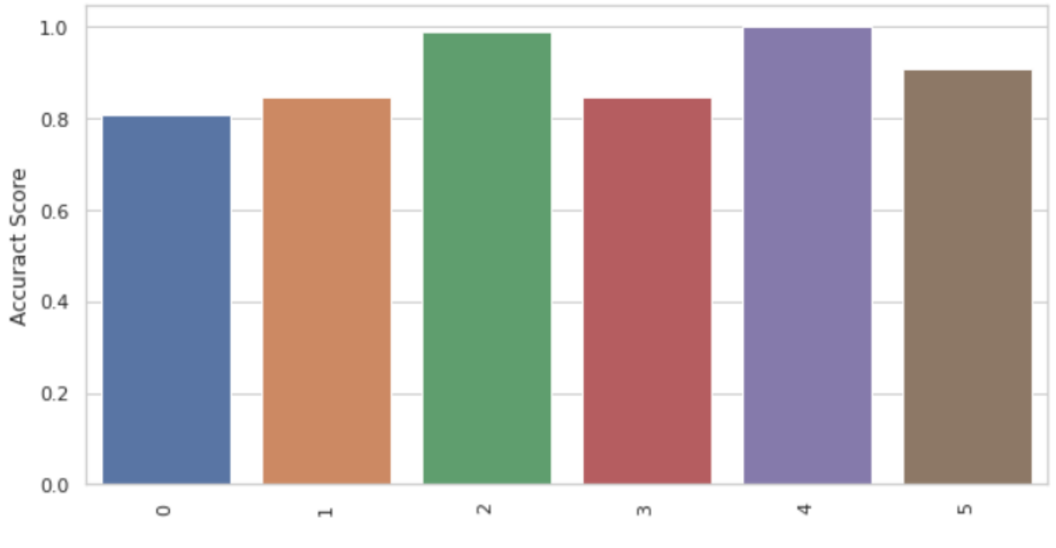


Fig 6.7. Comparision of ML models

6.2 PERFORMANCE

In this proposed work we have build an web application using flask frame work in visual studio code.

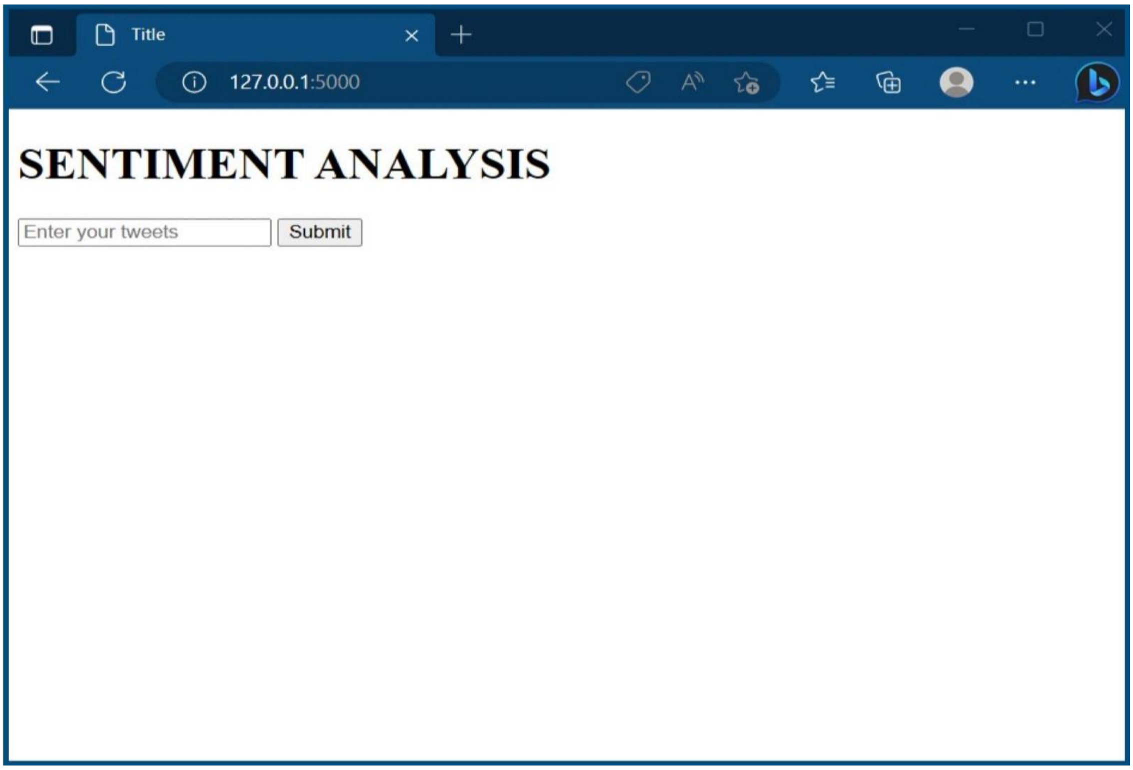


Fig 6.8. Main page

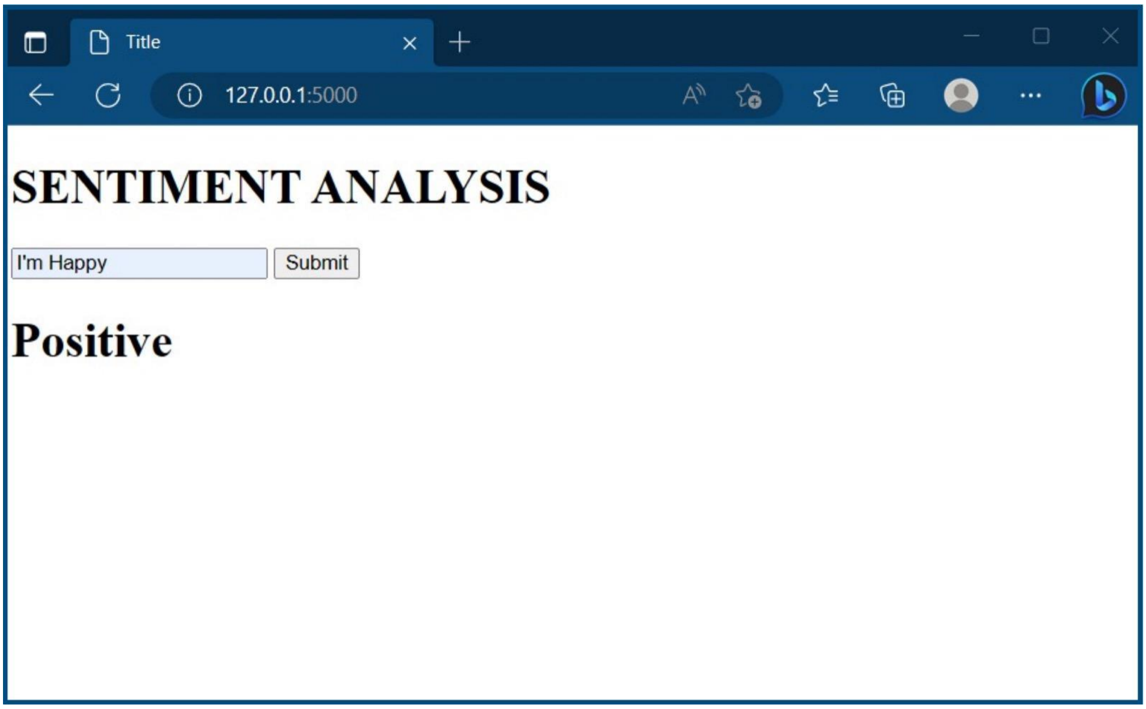
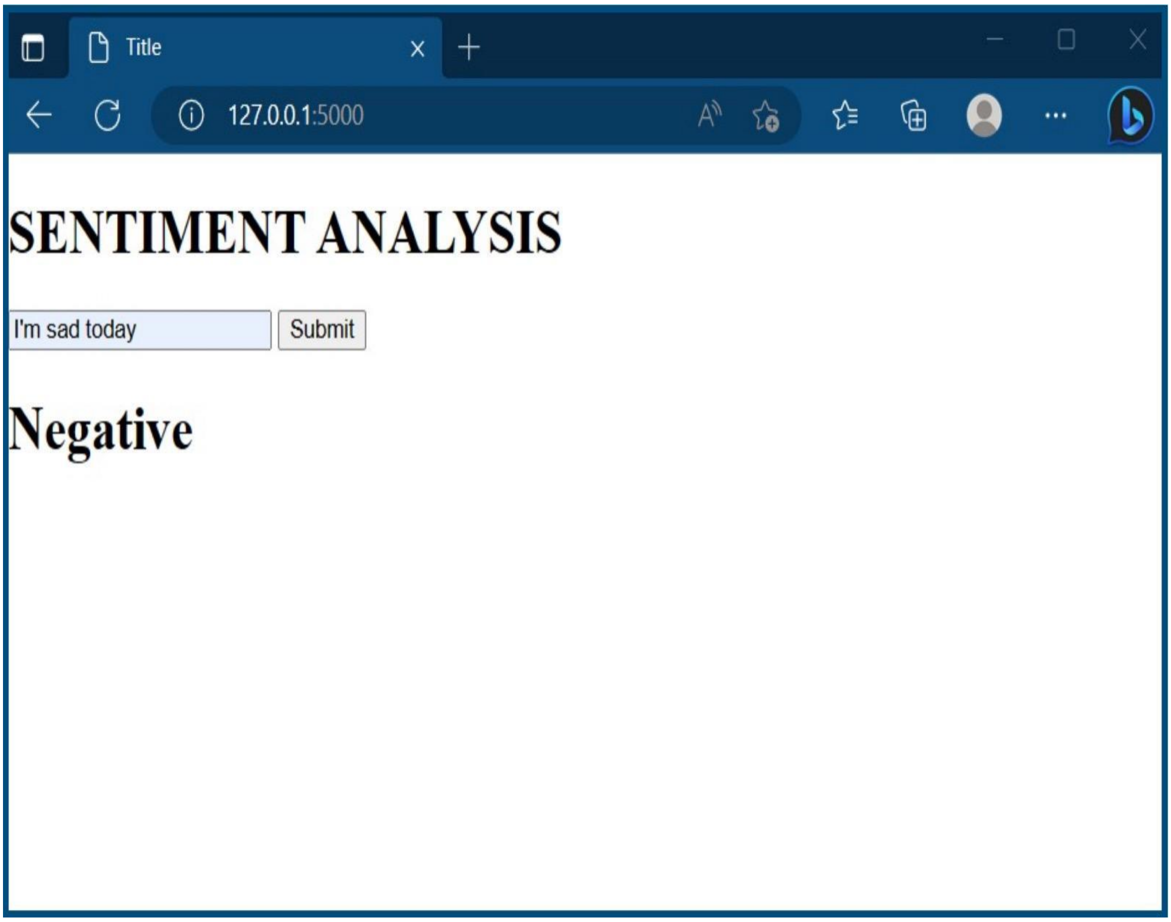


Fig 6.9. Positive Prediction



SENTIMENT ANALYSIS

I'm sad today

Negative

Fig 6.10.Negative Prediction

CHAPTER 7

CONCLUSION

This study displays the no of positive negative and neutral tweets, Nowadays more people are into social media. So if people wanted to buy any product they refer to social media first , and this project makes it easy for them to identify the positive negative and neutral suggestions. This also helps many business, it allows owners to know the reviews about their product and find the angry customers and help them produce better products. This analysis is also used in elections to check who is going to win. The sentimental analysis is a technique used to understand the emotional tone of text . It can be used to identify positive, negative and neutral sentiments in a piece of writing. This information can be useful for business owners who want to understand how their customers feel about their company. Tracking real time customer feedback and sentiment about an organization brand product and service. Providing feedback on ways to improve product service and customer experience.

REFERENCES

1) Bishwo Prakash Pokharel M.Phil ,Twitter Sentiment analysis during COVID-19 Outbreak in Nepal in 2020/6/11.

2)Hassan nazeer Chaudhry, Yasir javed, Farzana kulsoom, Zahid mehmood, zafar Iqbal khan, umar Shoaib ”Twitter sentiment analysis”,2021.

3) Jose Ramon Saura a, Domingo Ribeiro-Soriano b, Pablo Zegarra Saldana, “Exploring the challenges of remote work on Twitter users’ sentiments: From digital technology development to a post-pandemic era” https://doi.org/10.1016/j.jbusres.2021.12.052.

4) Kamaran H. Manguri, Rebaz N. Ramadhan, Pshko R. Mohammed Amin “Twitter Sentiment Analysis on Worldwide COVID-19 Outbreaks” Published 19 May 2020.

5) Authors Maryum Bibi, Wajid Aziz, Majid Almaraashi, Imtiaz Hussain Khan, Malik Sajjad Ahmed Nadeem, Nazneen Habib “A cooperative binary-clustering framework based on majority voting for Twitter sentiment analysis” Published on 2020/3/27 Journal IEEE Access.

6) Pallavi Tiwari, Deepak Upadhyay, Bhaskar pant, Noor Mohd “Twitter Sentiment Analysis Using Machine Learning and Deep Learning” October 2020.

7) Park Ji-woong professor of Hanyang University study on”Twitter sentiment analysis “ published on 20228) Pranati Rakshit, Sumit gupta, Tarpan das Twiter sentiment analysis using NLP” published on 2022.

8)) Pranati Rakshit, Sumit gupta, Tarpan das Twiter sentiment analysis using NLP” published on 2022.

9)Shruti Kulkarni Anisaaara Nadaph, Dr. Geetika Narang In recognition of the publication of the paper entitled “Survey on Twitter Sentiment Analysis using Supervised Machine Learning Algorithms” Published in Volume 7 Issue 5, May-2022.

APPENDIX

A.SOURCE CODE

```
 [twitter]

api_key = 2PZm8X6c0L5tFaqoikRwe4e5Y
```

```
api_key_secret =
EL5qMCQBcH6SnJsLv9UJQIHIyV7XuDoDDoYVLZIhfOMDzsv8Db
access_token = 1403974271299252225-OZqrzOxwt8SyIdZQc8FgekeVgCXA5K
access_token_secret =
qRmYrfpr4h8RehtGo0E56VrdIJ6GqUCWRfXH5k4OTfO4Q
```

```
import tweepy
import configparser
import pandas as pd

# read configs
config = configparser.ConfigParser()
config.read('config.ini')

api_key = config['twitter']['api_key']
api_key_secret = config['twitter']['api_key_secret']

access_token = config['twitter']['access_token']
access_token_secret = config['twitter']['access_token_secret']

# authentication
auth = tweepy.OAuthHandler(api_key, api_key_secret)
auth.set_access_token(access_token, access_token_secret)

api = tweepy.API(auth)

public_tweets = api.home_timeline(count=200)

# create dataframe
columns = ['Time', 'User', 'Fav_count', 'Rt_count', 'Src', 'Loc', 'Created', 'Verified',
'Tweet']
data = []
for tweet in public_tweets:
    data.append([tweet.created_at, tweet.user.screen_name, tweet.favorite_count,
tweet.retweet_count, tweet.source, tweet.user.location,
    tweet.user.created_at, tweet.user.verified, tweet.text])

df = pd.DataFrame(data, columns=columns)

df.to_csv('tweets272.csv')

import tweepy
from textblob import TextBlob
from wordcloud import WordCloud
import pandas as pd
import numpy as np
import seaborn as sns
import re
import sklearn
import matplotlib.pyplot as plt
from google.colab import files
upload = files.upload()

dataframe = pd.read_excel('Tweets.xlsx')
dataframe.head()
dataframe.tail()
dataframe.shape
df = dataframe.drop_duplicates()
print(df)
df.shape
print("Rowa : ",df.shape[0])
print("Columns : ",df.shape[1])
df.describe()
df.isnull().sum()
df["Loc"].fillna("missing", inplace=True)
df.isnull().sum()
df.dtypes
df.info()
df['Fav_count'].value_counts()
df.columns

def cleanTxt(text):
  text = re.sub(r'@[A-Za-z0-9]+', '', text)
  text = re.sub(r'#', '', text)
  text = re.sub(r'RT[\s]+', '', text)
  text = re.sub(r'https?:\/\/\S+', '', text)
  return text
df['Tweet']= df['Tweet'].apply(cleanTxt)
df
def getSubjectivity(text):
  return TextBlob(text).sentiment.subjectivity

def getPolarity(text):
  return TextBlob(text).sentiment.polarity

df['Subjectivity'] = df['Tweet'].apply(getSubjectivity)
df['Polarity'] = df['Tweet'].apply(getPolarity)
df
df.shape

plt.figure(figsize=(8,6))
for i in range(0, df.shape[0]):
  plt.scatter(df['Polarity'][i], df['Subjectivity'][i], color='Purple' )

plt.title('Sentiment Analysis')
plt.xlabel('Polarity')
plt.ylabel('Subjectivity')
plt.show()
```

```
allWords = ' '.join( [twts for twts in df['Tweet']])
wordCloud = WordCloud(width = 1000, height=700, random_state =21, max_font_size
= 119).generate(allWords)
plt.imshow(wordCloud, interpolation = "bilinear")
plt.axis('off')
plt.show()

def getAnalysis(score):
  if score < 0:
    return 'Negative'
  elif score == 0:
    return 'Neutral'
  else:
    return 'Positive'

df['Analysis'] = df['Polarity'].apply(getAnalysis)
df

allWords = ' '.join( [twts for twts in df['Tweet'][df['Analysis']=='Positive']])
wordCloud = WordCloud(width = 1000, height=700, random_state =21, max_font_size
= 119).generate(allWords)
plt.imshow(wordCloud, interpolation = "bilinear")
plt.axis('off')
plt.show()

allWords = ' '.join( [twts for twts in df['Tweet'][df['Analysis']=='Negative']])
wordCloud = WordCloud(width = 1000, height=700, random_state =21, max_font_size
= 119).generate(allWords)
plt.imshow(wordCloud, interpolation = "bilinear")
plt.axis('off')
plt.show()

allWords = ' '.join( [twts for twts in df['Tweet'][df['Analysis']=='Neutral']])
```

```
wordCloud = WordCloud(width = 1000, height=700, random_state =21, max_font_size = 119).generate(allWords)
plt.imshow(wordCloud, interpolation = "bilinear")
plt.axis('off')
plt.show()

ptweets = df[df.Analysis == 'Positive']
ptweets = ptweets['Tweet']
ptweets

ptweets = df[df.Analysis == 'Negative']
ptweets = ptweets['Tweet']
ptweets

ptweets = df[df.Analysis == 'Neutral']
ptweets = ptweets['Tweet']
ptweets

ptweets = df[df.Analysis == 'Positive']
ptweets = ptweets['Tweet']

round(   (ptweets.shape[0] / df.shape[0]) *100 , 1)

ptweets = df[df.Analysis == 'Negative']
ptweets = ptweets['Tweet']

round(   (ptweets.shape[0] / df.shape[0]) *100 , 1)

ptweets = df[df.Analysis == 'Neutral']
ptweets = ptweets['Tweet']

round(   (ptweets.shape[0] / df.shape[0]) *100 , 1)
```

```
from sklearn.preprocessing import LabelEncoder
le_User = LabelEncoder()
le_Fav_count = LabelEncoder()
le_Rt_count = LabelEncoder()
le_Src = LabelEncoder()
le_Loc = LabelEncoder()
le_Verified = LabelEncoder()
le_Tweet = LabelEncoder()
le_Subjectivity = LabelEncoder()
le_Polarity = LabelEncoder()
le_Analysis = LabelEncoder()

df['User'] = le_User.fit_transform(df['User'])
df['Fav_count'] = le_Fav_count.fit_transform(df['Fav_count'])
df['Rt_count'] = le_Rt_count.fit_transform(df['Rt_count'])
df['Src'] = le_Src.fit_transform(df['Src'])
df['Loc'] = le_Loc.fit_transform(df['Loc'])
df['Verified'] = le_Verified.fit_transform(df['Verified'])
df['Tweet'] = le_Tweet.fit_transform(df['Tweet'])
df['Subjectivity'] = le_Subjectivity.fit_transform(df['Subjectivity'])
df['Polarity'] = le_Polarity.fit_transform(df['Polarity'])
df['Analysis'] = le_Analysis.fit_transform(df['Analysis'])

df.head()
X = df.loc[:,['User', 'Fav_count', 'Rt_count', 'Src', 'Loc',
       'Verified', 'Tweet', 'Subjectivity', 'Polarity']]
y = df.loc[:,'Analysis']

fig = plt.figure(figsize=(5,5))
sns.countplot(x = 'Analysis', data = df)

fig = plt.figure(figsize=(7,7))
colors = ("Teal", "#8b8b8b", "Purple")
```

```
wp = {'linewidth':1, 'edgecolor':"black"}
tags = df['Analysis'].value_counts()
explode = (0,0,0)
tags.plot(kind='pie', autopct='%1.1f%%', shadow=True, colors= colors,
        startangle=90, wedgeprops = wp, explode = explode, label='')
plt.title('Distribution of Sentiment')

df.columns
sns.set(rc = {'figure.figsize':(15,8)})
sns.heatmap(df.corr(),annot=True)

df.hist(figsize=(10,10),bins=10)

X.head()
y.head()
from sklearn.model_selection import train_test_split
trainingSet, testSet = train_test_split(df, test_size=0.2)

train_df = trainingSet
test_df = testSet

print(X_train.shape)
print(y_train.shape)
print(X_test.shape)
print(y_test.shape)

X_train.head()

X_test.head()
y_train.head()
y_test.head()
y_train.value_counts()
y_test.value_counts()
```

```
from sklearn.linear_model import LogisticRegression
logreg = LogisticRegression()
logreg.fit(X_train,y_train)
y_pred = logreg.predict(X_test)

from sklearn import metrics
cnf_matrix = metrics.confusion_matrix(y_test, y_pred)
cnf_matrix

accuracy_score_lr = accuracy_score(y_test, y_pred)
print("accuracy_score_lr:", accuracy_score_lr)

from sklearn.model_selection import train_test_split

X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)

from sklearn.tree import DecisionTreeClassifier

# Instantiate a DecisionTreeClassifier 'dt' with a maximum depth of 6
dt = DecisionTreeClassifier(max_depth=1, random_state=42)

# Fit dt to the training set
dt.fit(X_train, y_train)

# Predict test set labels
y_pred = dt.predict(X_test)
print(y_pred[0:5])

y_pred
y_test

from sklearn.metrics import accuracy_score
```

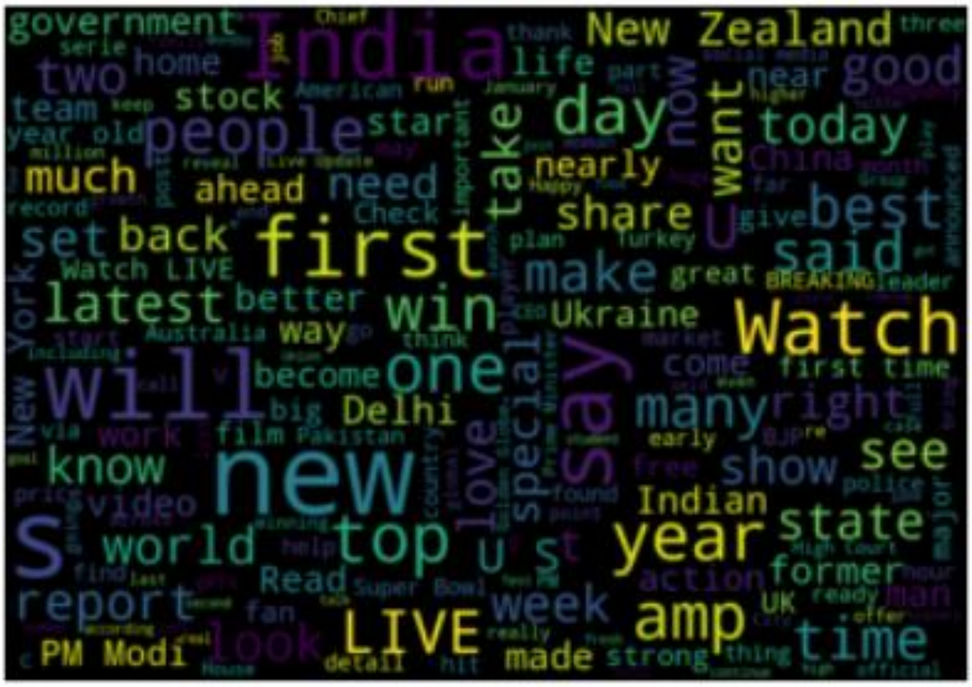
```
# Predict test set labels
y_pred = dt.predict(X_test)

# Compute test set accuracy
acc = accuracy_score(y_test, y_pred)
print("Accuracy_score_dt: {:.2f}".format(acc))

from sklearn.neighbors import KneighborsClassifier
X_train, X_test, y_train, y_test = train_test_split(X, y, random_state=1)

knn = KNeighborsClassifier(n_neighbors=222, metric='euclidean')
knn.fit(X_train, y_train)
Knn_pred = knn.predict(X_test)

from sklearn.metrics import confusion_matrix
from sklearn.metrics import accuracy_score

confusion_matrix(y_test, Knn_pred)
accuracy_score_knn = accuracy_score(y_test, Knn_pred)
print("accuracy_score_knn:", accuracy_score_knn)

from sklearn.naive_bayes import GaussianNB
NB = GaussianNB()
NB.fit(X_train, y_train)

NB_pred = NB.predict(X_test)
NB_pred
from sklearn.metrics import confusion_matrix
cm = confusion_matrix(y_test, NB_pred)
cm
accuracy_score_nb = accuracy_score(y_test, NB_pred)
print("accuracy_score_nb:", accuracy_score_nb)
```

```
from sklearn.svm import SVC # "Support vector classifier"
SVC = SVC(kernel='rbf', random_state=0)
SVC.fit(X_train, y_train)

SVC_pred= SVC.predict(X_test)
from sklearn.metrics import confusion_matrix
cm= confusion_matrix(y_test, SVC_pred)
cm
accuracy_score_svm = accuracy_score(y_test, SVC_pred)
print("accuracy_score_svm:", accuracy_score_svm)

from sklearn.ensemble import RandomForestClassifier
RF= RandomForestClassifier(n_estimators= 200, criterion="entropy")
RF.fit(X_train, y_train)

RF_pred= RF.predict(X_test)
from sklearn.metrics import confusion_matrix
cm= confusion_matrix(y_test, RF_pred)

cm

accuracy_score_rf = accuracy_score(y_test, RF_pred)
print("accuracy_score_rf:", accuracy_score_rf)

models = pd.DataFrame({
   "Model": ["Logistic Regression",
             "Decision Tree",
             "Naive Bayes",
             "KNN",
             "Support Vector machine",
             "Random forest"],
   "Accuract Score" :[accuracy_score_lr, acc,
accuracy_score_nb,accuracy_score_knn,accuracy_score_svm,accuracy_score_rf]
})

Models
plt.figure(figsize=(10, 5))
sns.set_theme(style="whitegrid")
ax = sns.barplot(x=models.index, y="Accuract Score", data=models)
plt.xticks(rotation=90)

import pickle
filename= 'RF_pred'

pickle.dump(RF,open(filename,'wb'))

model=pickle.load(open(filename,'rb'))

RF = pickle.load(open(filename,'rb'))
RF.predict(X_test)



from flask import Flask, render_template, request
import pickle
import numpy as np
import pandas as pd

app = Flask(__name__)

print("hello world")

file = open('models/models/RF_pred.pkl','rb')
model = pickle.load(file)
```

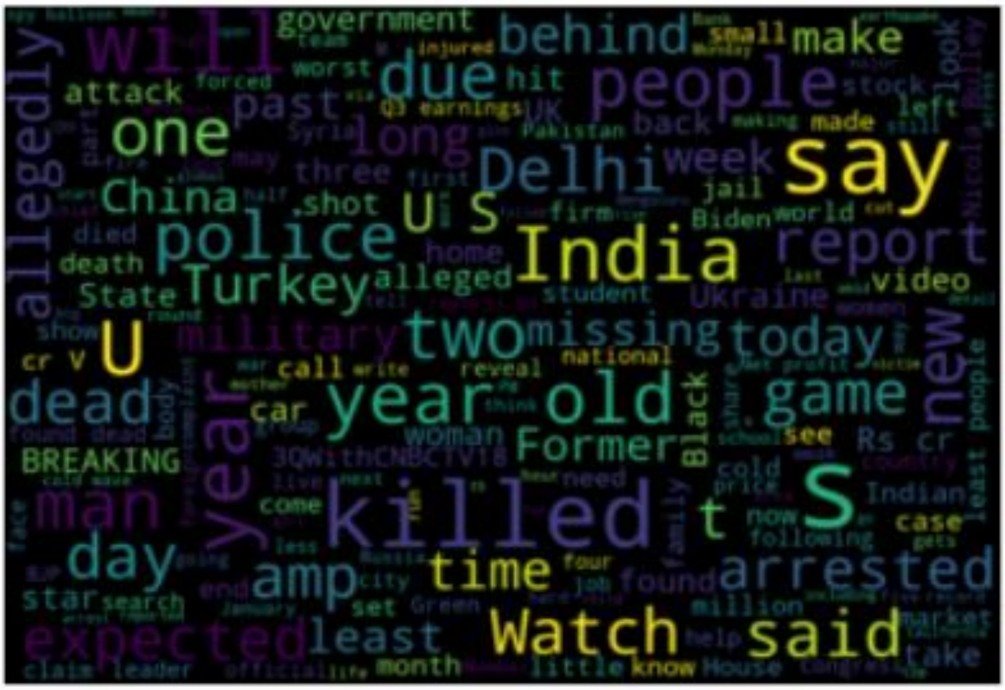
```
@app.route('/')
def home():
    return render_template('index.html', **locals())

@app.route('/', methods=['POST', 'GET'])

def index():
    Polarity = str(request.form.get('inp'))
    print(Polarity)

    return render_template('after.html',Polarity=Polarity)



if __name__=='__main__':
    app.run(debug=True)

<html>
    <body bgcolor="Teal">

        <center>
            <br></br><br></br><br></br>

        <h1>TWITTER SENTIMENT ANALYSIS</h1><br></br>


        <form methods="POST", action="{{url_for('index')}}">
            <b><input type="text"name="inp" placeholder="Type the
tweet"required="required"/><br></br><br></br>
                <input type="int"name="inp" placeholder="Type the
polarity"required="required"/>

            <button type="submit" class="btn btn-primary btn-block btn-large"><h3>Predict
```

```
the Sentiment Analysis</h3></button><br></br>
            </b>
        </form>

        </center>
    </body>
</html>



<html>
    <body bgcolor=olive>
        <center>
        <br></br><br></br><br></br>
        <h1> PREDICTION </h1>
        {% block content %}
        {%if Polarity == 0%}

        <h1>Negative □</h1>



        {%elif Polarity == 1%}
        <h1>Neutral □</h1>



        {%else%}
        <h1>Positive □</h1>

        {%endif%}
        {% endblock %}

        <br><br>
        <a href='/'>go back to home page</a>
        </center>
    </body>
</html>
```

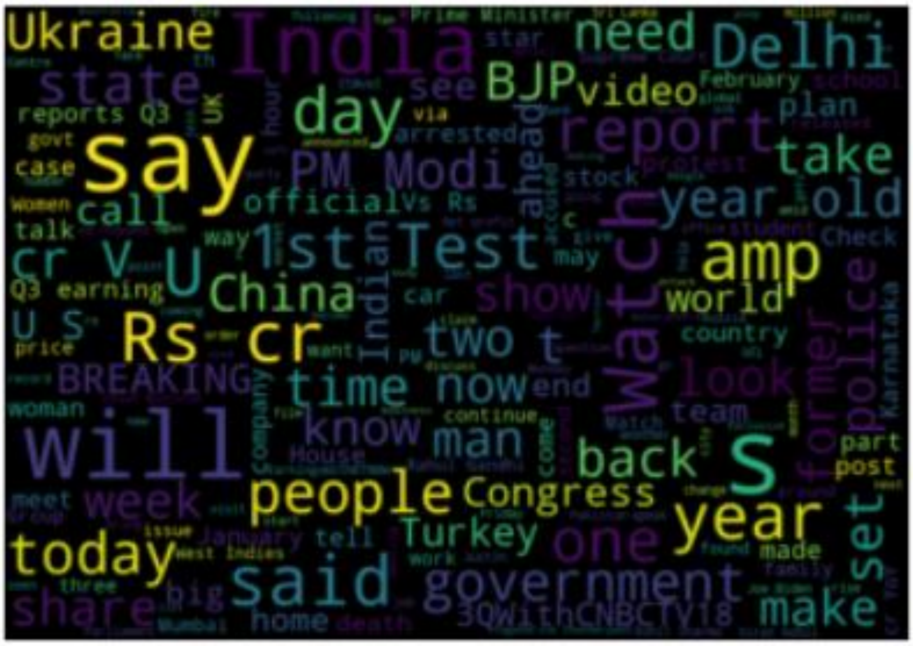
B.SCREENSHOTS



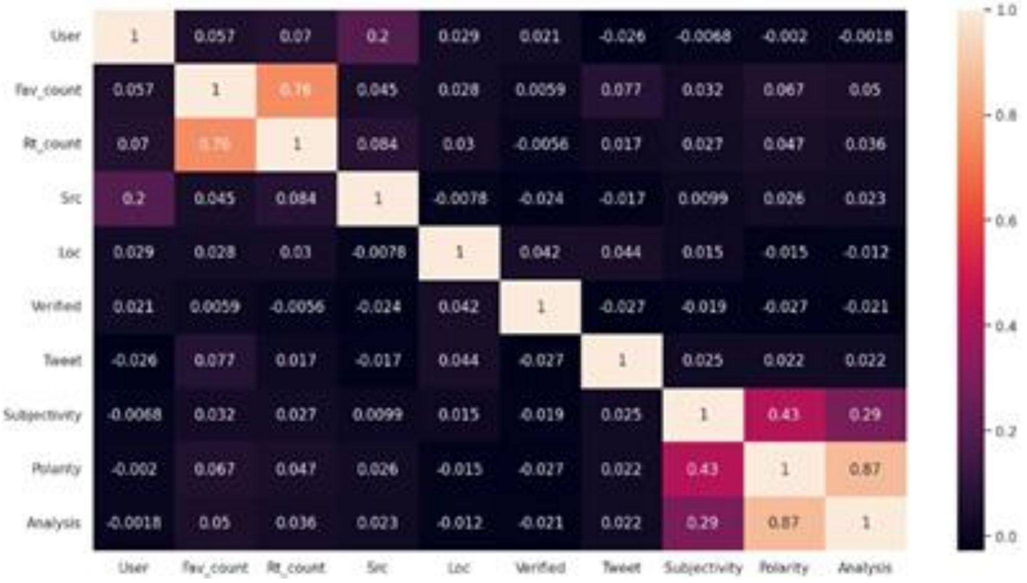
Data visualization of Positive words that were used repeatedly in the tweets collected



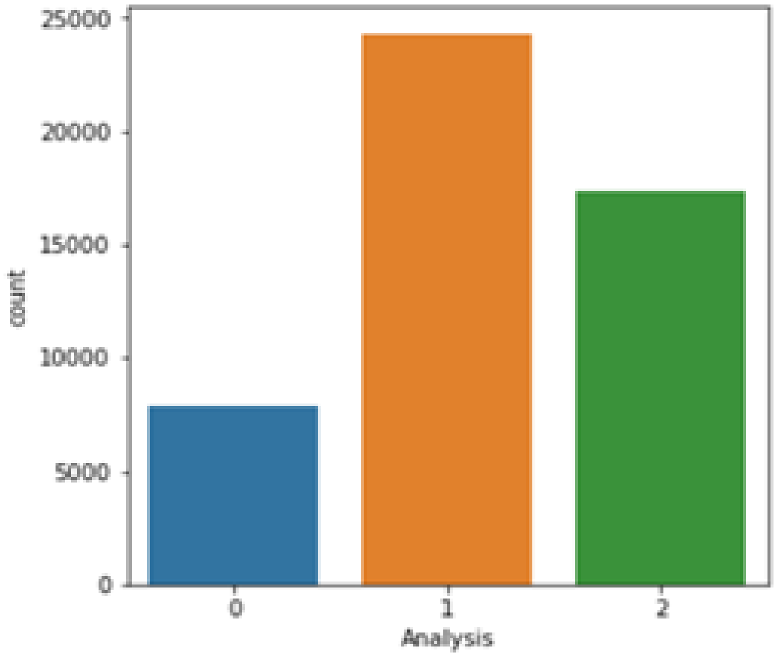
Data visualization of negative words that were used repeatedly in the tweets collected



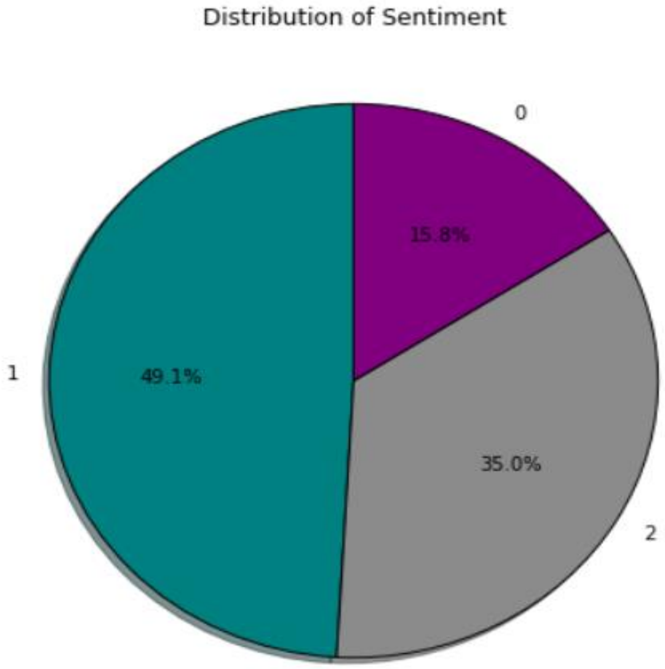
Data visualization of neutral words that were used repeatedly in the tweets collected



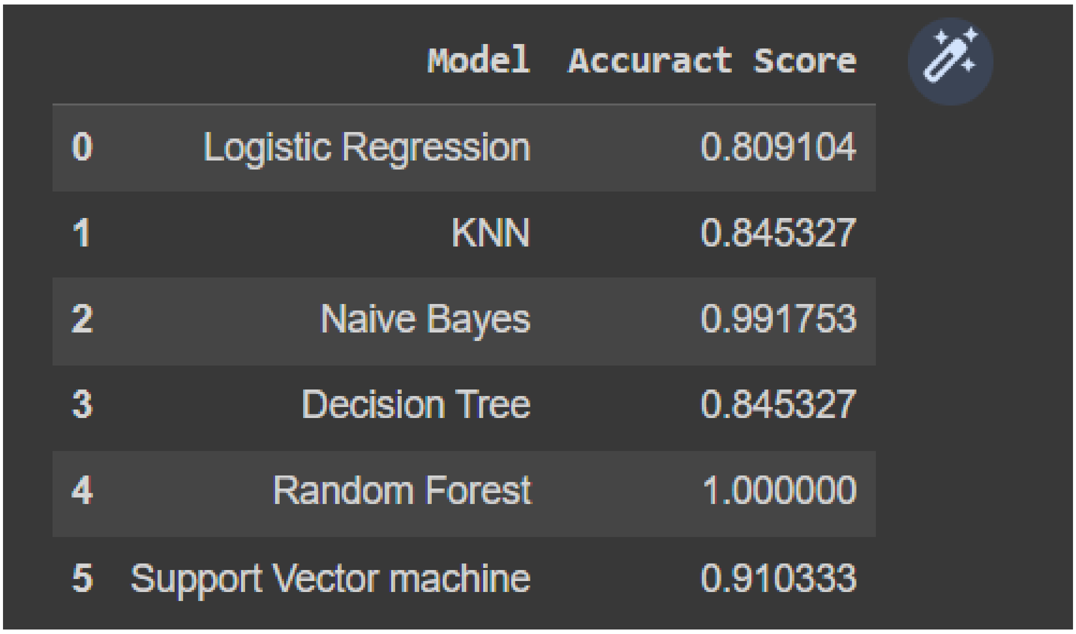
Heat map visualization



The sentiment analysis is done, and a bar graph for tweets that were positive, negative, and neutral was displayed.

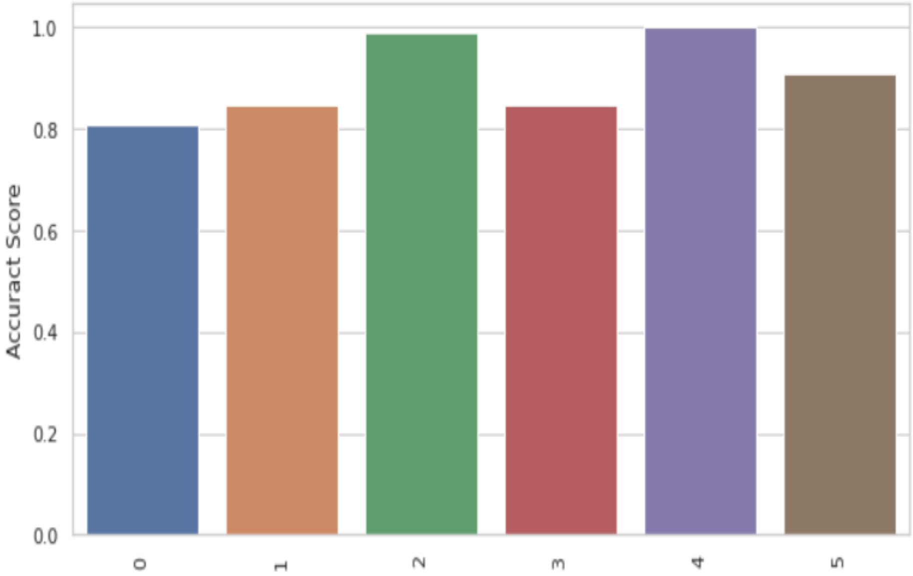


The ratio of the tweets were displayed in a pie chart which represents positive negative and neutral tweets.



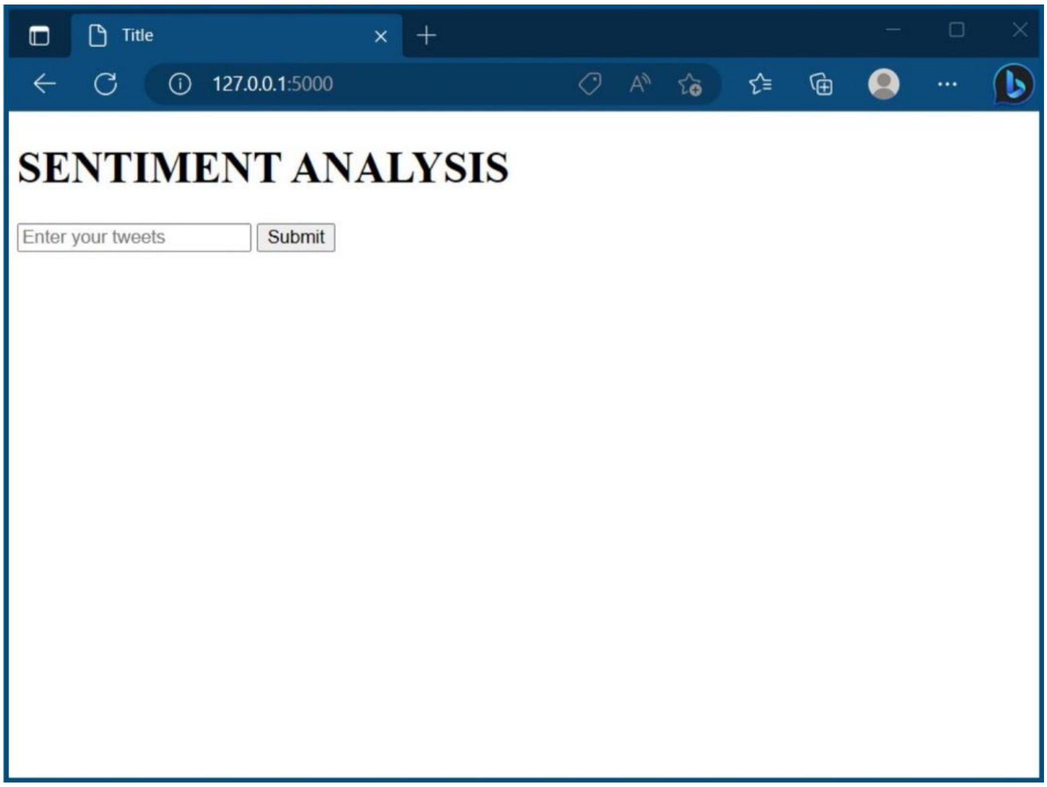
	Model	Accuract Score
0	Logistic Regression	0.809104
1	KNN	0.845327
2	Naive Bayes	0.991753
3	Decision Tree	0.845327
4	Random Forest	1.000000
5	Support Vector machine	0.910333

We have build and compared the top 6 machine learning algorithms.

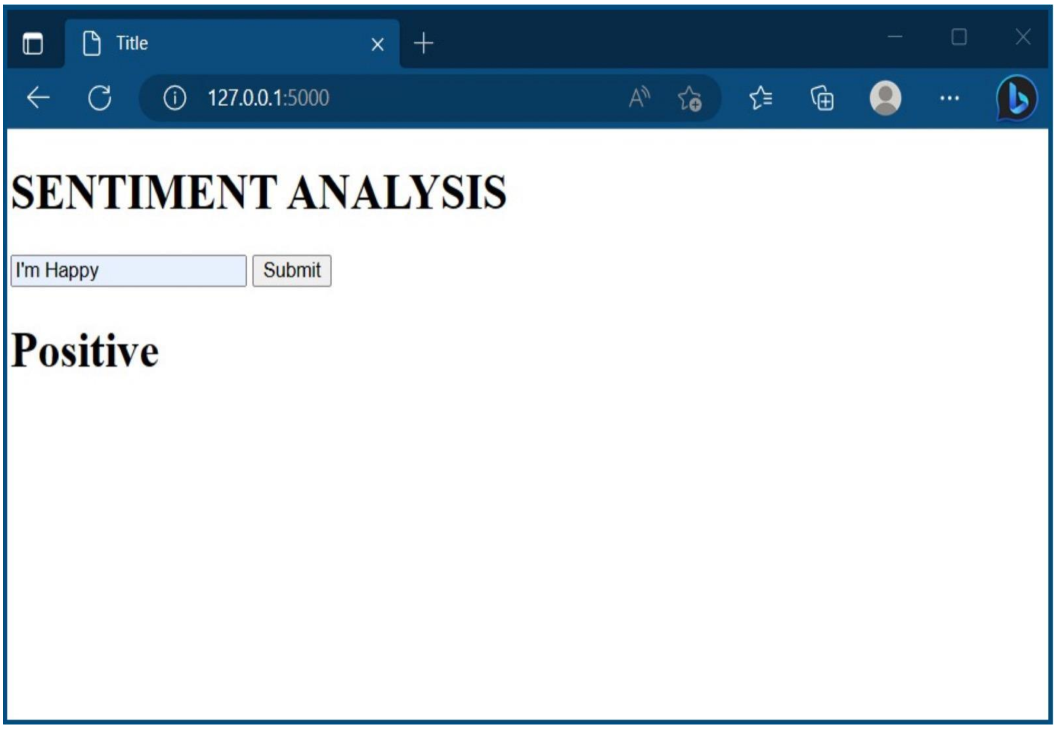


We got to the conclusion that Random forest performs better with an accuracy of 100% and logistic regression performs poorly.

Web outputs:



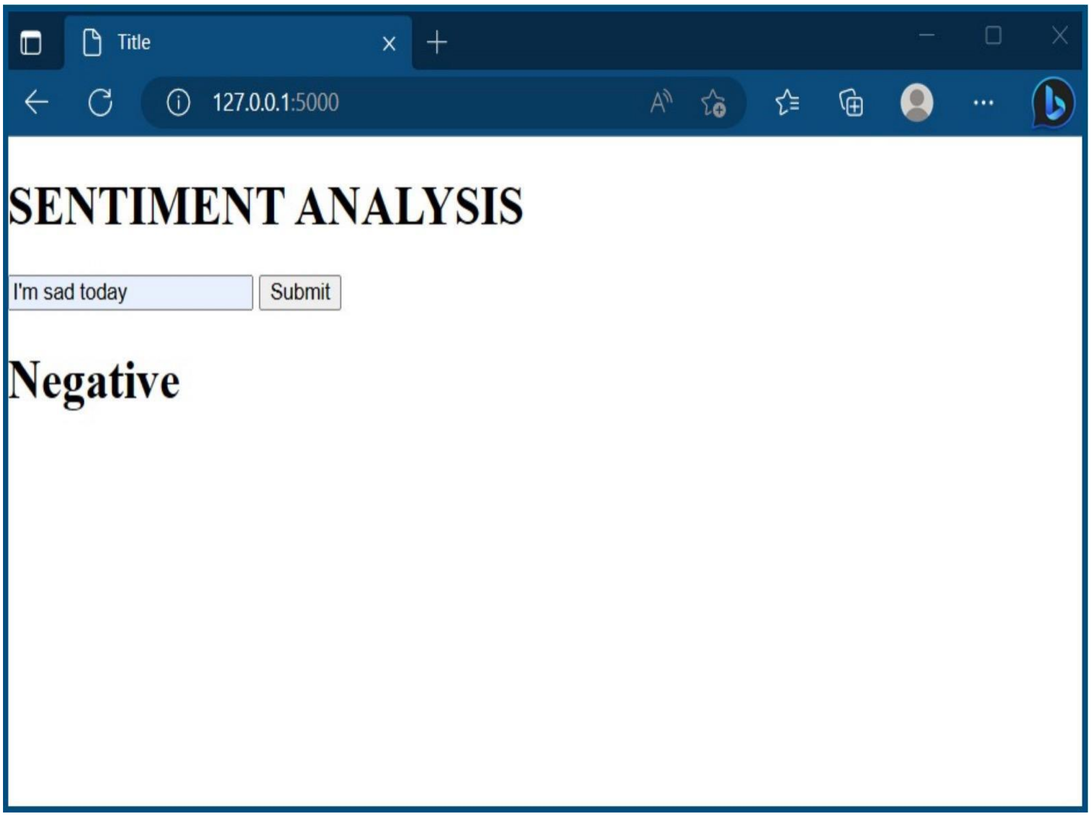
SENTIMENT ANALYSIS



SENTIMENT ANALYSIS

I'm Happy

Positive



SENTIMENT ANALYSIS

I'm sad today

Negative

C. PUBLICATION

ANALYSIS OF TWITTER SENTIMENTS USING MACHINE LEARNING ALGORITHMS

-Amirtha Varsini K S, Vidhya S
-Dr.P.Asha
Computer Science and Engineering department
Sathyabama Institute of Science and Technology

ABSTRACT: There is a plenty of data exist on the web for internet users at a time of the development of various web technologies, the internet is now grew into a place for online education, sharing ideas of each people and also for sharing ones opinion, the internet sites for instance google+, facebook, Twitter are growing like a wind because they allow users to express their opinion freely on various topics, it also acknowledges users to get into an conversation with other people or groups, and also share their messages across the world. The study of sentiment analysis in twitter data has driven a lot of people. It is helpful for classifying the tweets depending upon the polarity which is positive negative and neutral. Analysing the twitter data for this analysis is tough because it has a large amount of data as the twitter contains data on various topics. For our prediction, supervised classification algorithms have been used. Using distinctive supervised machine learning algorithms such as K-nearest neighbor, support vector machine, naïve bayes, logistic regression , decision tree, and random forest algorithms were used to analyze the sentiment of tweets. These machine learning algorithms has been compared, and the most effective method is chosen to predict the outcome.

***Keywords: Twitter, Machine Learning, SVM, Twitter API, KNN, Sentiment Analysis, Naïve bayes, Tweets, Classification*.**

1. INTRODUCTION

Social media and network sites are used by countless of people to express their thoughts, sentiments, and emotions about people's day-to-day lives. Any topic, including social interactions and product reviews, can be written about. Online communities provide a venue for discussion where users can enlighten and influence others.

The tweets from internet is a vast amount of data, so to an human it might be an very difficult task to quickly analyze the data, to read , summarize and organize them by checking each tweet and making them into an understandable format. Nowadays most people rely on internet for fast growth in business. Many people tend to refer social media to know about a product and to know the reviews given by other customers on a product, there is a immeasurable amount of comments on internet by a lots of users. So this analysis can help the owners to know the positive negative comments on their product and it might help them to improve their product which will give them a huge profit. Sentiment analysis plays a major role in producing the information for the users.

For instance, SA can be used to forecast recommendations for products put forward by recommendation systems while accounting for variables like positive or negative sentiments about the goods. This method could be applied for a variety of things, such as helping politicians understand the attitudes that people in various regions have toward them so they can increase their investment in certain regions. Sentiment analysis could also be applied in the field of business marketing. Analysis of sentiments could be extremely important in the service sector because it can expose consumer feelings and examine the whole scope of the customer experience.

2. LITERATURE REVIEW

Hassan nazeer Chaudhry [1] et.al used Twitter sentiment to gauge public opinion before, during, and after elections and was then contrasted with the results of the vote. The Naive Bayes Classifier was used to gather public opinion after we constructed a dataset using the tweets API, pre-processed the data, and used Term Frequency-Inverse Document Frequency to extract the relevant features. The accuracy and precision of our sentiment classifier are 94.58% and a precision of 93.10%.

Bishwo Prakash Pokharel [2] et.al He analysed the sentiment of tweets posted on the Twitter social media platform, The tweets were gathered, prepared, and used with Google Colab for text mining and sentiment analysis.Following sentiment analysis using the two designated hashtags such as #COVID-19 and #coronavirus, a graphical representation of the data has been provided. The information is gathered from people who, between May 21 and May 31, 2020, shared Nepal as their location.

Kamaran H.Manguri [3] et.al In this work, Twitter data was obtained from the social media site using the Tweepy library. Sentiment analysis was then carried out using the Python TextBlob module. The data has been supplied with a graphical form following the measuring sentiment analysis. The information gathered on Twitter is based on the use of two specific hashtags, "COVID-19, coronavirus.". A visual depiction of the results and additional explanation are given at the conclusion.

Jose Ramon saura [4] et.al This study, they have conducted a sentiment analysis created with Textblob, a machine learning tool, using Computer-Aided Text Analysis and Natural Language Processing. Then, they used the Latent Dirichlet allocation model, a mathematical approach for topic modelling.

Park Ji-woong [5].This study seeks to forecast the election outcome using data from both tweets and Twitter user descriptions. This study concluded that utilising Twitter user descriptions as data for forecasts was unsatisfactory and that tweets are very successful at forecasting the outcome of the presidential election.

Shruti Kulkarni [6] et.al This analysis's objective is to show how to combine topical information from Twitter discussions with sentiment dispersion patterns to improve sentiment analysis on Twitter data .SentiDiff, a consistent approach for predicting sentiment polarity expressed in Twitter tweets, is then shown, taking into account interactions between subject information in Twitter messages and the sentiment dispersion patterns. The data indicate that this is the first study to exploit sentiment diffusion patterns to advance Twitter sentiment/emotion analysis.

Deepak Upadhyay [7] et.al However, the focus of this article will be on sentiment analysis of Twitter data. To gain a deeper grasp of public mood, They used text mining. Python is used in this study to gather, collect, and analyse tweets. Three distinct machine learning algorithms and one deep learning algorithm are then utilised for sentiment analysis, and they were compared to see which method or model performs with the highest degree of accuracy.

Pranati Rakshit [8] et.al The goal of this paper is to analyse the polarity of a tweet utilising a large amount of data obtained by natural language processing (NLP). They applied machine learning methods to identify the tweet's polarity. To determine the polarity of tweets, they used the Multinomial Naive-Bayes, Complement Naive-Bayes, and Logistic Regression classifier. The 1.6 million-entry dataset produced the greatest results when we used logistic regression. 78.05% is the highest accuracy.

Maryum bibi [9] et.al This research proposes a novel unsupervised learning approach for sentiment analysis on Twitter that is based on concept-based and hierarchical clustering. The common hierarchical clustering algorithms single linkage, complete linkage, and average linkage are serially combined. They have also tried with well-known classifiers. The performance of understudied approaches is measured by accuracy (the percentage of accurate predictions). The performance of knowledge discovery approaches is demonstrated empirically to be on par with supervised learning techniques.

3. REQUIREMENTS

The API and its key were obtained from Twitter and Visual Studio code was used to turn the unprocessed data into a dataset. The programming language Python has been used. Google Colab was the platform used for the coding.

4. EXISTING SYSTEM

VADER approach techniques were utilised in earlier works, but they didn't produce good accuracy. No authors have contrasted any machine learning algorithms up to this point. The naive Bayes algorithm is found to be the best algorithm. Some authors' who used K-nearest neighbour methods have got poorer accuracy.

5. PROPOSED SYSTEM

When examining the drawbacks of all the previous approaches utilised in earlier systems, the most prevalent issue is that the majority of these systems simply utilized a pre-defined dataset from Twitter to execute this problem statement. Additionally, they only used dynamic numbers and carried out sentiment analysis, thus they can only determine whether a tweet is positive or negative. They have established a method for analysing text from Twitter, particularly for the recently developed field of sentiment analysis. Some of the systems are also not very accurate. They didn't compare any algorithms of machine learning. All these problems in the pre-existing systems were addressed by our system. We did not use any pre-defined datasets, our main concept is to collect data using Twitter API directly. The more traditional solutions will not provide any ratios, but our system has provided a constant ratio between positive and negative tweets. We have also developed a web application.

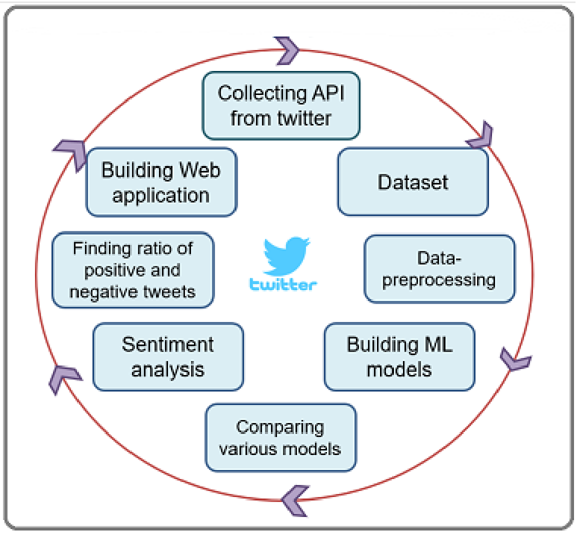


Fig.1. Project plan

The above figure (fig.1) shows the basic concept of this project. To improve performance, a range of machine learning algorithms, including Naive Bayes, Support Vector Machine, K-Nearest Neighbor, Logistic Regression, Random forest, Decision tree has been utilised and compared.

6. METHODOLOGY

Twitter account is converted to a developer account, after which Twitter's API and API key are requested. Once the API key has been obtained, it is utilised in visual studio code to gather tweets from Twitter. The tweets are then transformed into a dataset of 50950 rows and 7 columns. Then data analysing step is done. In Machine Learning, Data Analysis is the process of analyzing the data to find if there are any null values, duplicates and to know more details about the dataset.

Then data preprocessing is done. Data preprocessing in Machine Learning refers to the technique of preparing (cleaning and organizing) the raw data to make it suitable for a building and training Machine Learning models. Due to the fact that categorical data cannot be comprehended by machines, label encoding is performed during data preprocessing to transform categorical data into numerical data. The below figure(fig.2) explains each and every steps involved in this project

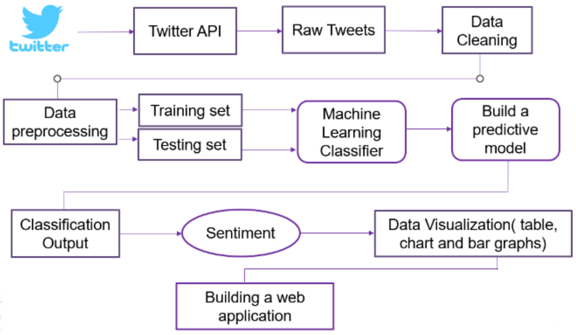


Fig.2. System architecture

Then, sentiment analysis is carried out From the dataset each tweet is checked using polarity check and the positive negative and neutral tweets were detected, and approximately 24000 neutral tweets, 17000 positive and 8000 negative tweets are obtained.

Finally, the top six machine learning algorithms is build and compared to get the best performing algorithm.

7. RESULT

We have obtained the Twitter API and converted the raw data into a dataset that includes tweets, time, user name, location, generated time, etc. using Python in Visual Studio Code.

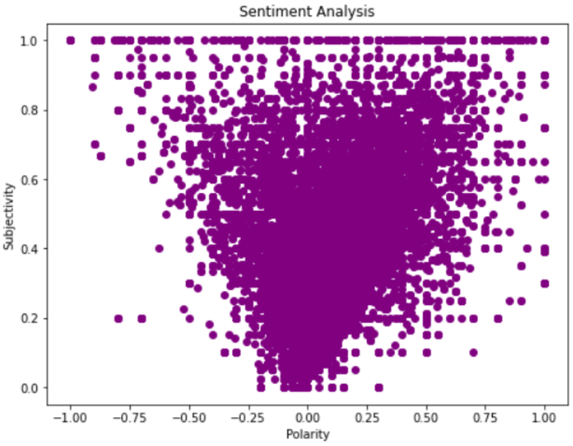


Fig.3.Subjctivity and polaity

The above figure (fig.3) shows the subjectivity and polarity of the tweets gathered

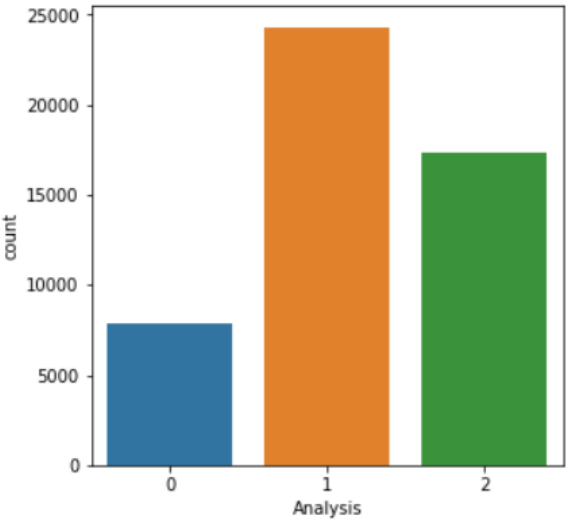


Fig.4.Sentiment analysis

In figure (fig.4)positive negative neutral tweets are represented in a bar diagram

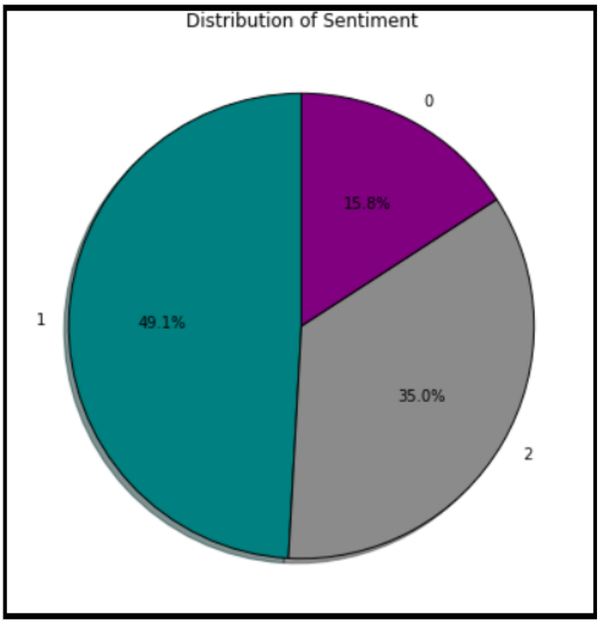
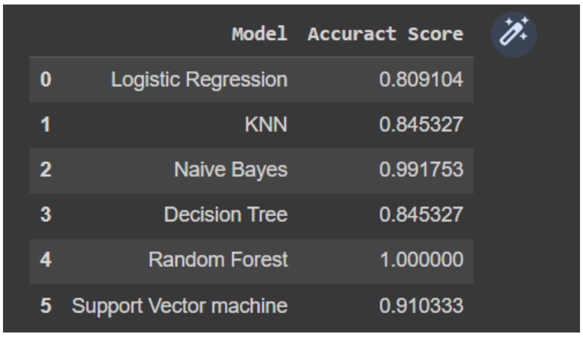


Fig.5.Ratio of tweets

In the above figure (fig.5) The ratio of the tweets were displayed in a pie chart which represents positive negative and neutral tweets.



	Model	Accuract Score
0	Logistic Regression	0.809104
1	KNN	0.845327
2	Naive Bayes	0.991753
3	Decision Tree	0.845327
4	Random Forest	1.000000
5	Support Vector machine	0.910333

Fig.6.Accuracy of the models

Finally, after building and comparing the top six machine learning algorithms.The figure (fig.6) shows their accuracy

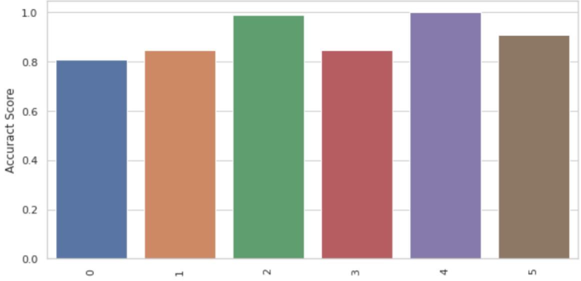


Fig.7.Comparision of the models

The figure (fig.7) shows the comparison of the models in a bar graph. We got to the conclusion that Random forest performs better with an accuracy of 100% and logistic regression performs poorly.

8. CONCLUSION

This study displays the no of positive negative and neutral tweets, Nowadays more people are into social media. So if people wanted to buy any product they refer to social media first , and this project makes it easy for them to identify the positive negative and neutral suggestions. This also helps many business, it allows owners to know the reviews about their product and find the angry customers and help them produce better products. This analysis is also used in elections to check who is going to win.

9. REFERENCE

1)Hassan nazeer Chaudhry, Yasir javed, Farzana kulsoom, Zahid mehmood, zafar Iqbal khan, umar Shoaib "Twitter sentiment analysis",2021

2)Twitter Sentiment analysis during COVID-19 Outbreak in NepalBishwo Prakash Pokharel M.Phil in 2020/6/11

3) "Twitter Sentiment Analysis on Worldwide COVID-19 Outbreaks" Kamaran

H. Manguri, Rebaz N. Ramadhan, Pshko R. Mohammed Amin Published 19 May 2020

4) “Exploring the challenges of remote work on Twitter users’ sentiments: From digital technology development to a post-pandemic era”Jose Ramon Saura a, Domingo Ribeiro-Soriano b, Pablo Zegarra Saldanahttps://doi.org/10.1016/j.jbusres.2021.12.052

5) Park Ji-woong professor of Hanyang University study on”Twitter sentiment analysis “ published on 2022

6)Shruti Kulkarni Anisaaara Nadaph, Dr. Geetika Narang In recognition of the publication of the paper entitled “Survey on Twitter Sentiment Analysis using Supervised Machine Learning Algorithms” Published in Volume 7 Issue 5, May-2022

7) “Twitter Sentiment Analysis Using Machine Learning and Deep Learning” Pallavi Tiwari, Deepak Upadhyay, Bhaskar pant, Noor Mohd October 2020

8) Pranati Rakshit, Sumit gupta, Tarpan das Twiter sentiment analysis using NLP” published on 2022

9) “A cooperative binary-clustering framework based on majority voting for Twitter sentiment analysis” Authors Maryum Bibi, Wajid Aziz, Majid Almaraashi, Imtiaz Hussain Khan, Malik Sajjad Ahmed Nadeem, Nazneen Habib Published on 2020/3/27Journal IEEE Access

RE-2022-87016-plag-report

by Research Experts - Turnitin Report

Submission date: 08-Apr-2023 12:59PM (UTC-0400)
Submission ID: 2059073403
File name: RE-2022-87016.docx (326.68K)
Word count: 2100
Character count: 11692

ANALYSIS OF TWITTER SENTIMENTS USING MACHINE LEARNING ALGORITHMS

-Amirtha Varsini K S, Vidhya S

-Dr.P.Asha

Computer Science and Engineering department

Sathyabama Institute of Science and Technology

ABSTRACT: There is a plenty of data exist on the web for internet users at a time of the development of various web technologies, the internet is now grew into a place for online education, sharing ideas of each people and also for sharing ones opinion, the internet sites for instance google+, facebook, Twitter are growing like a wind because they allow users to express their opinion freely on various topics, it also acknowledges users to get into an conversation with other people or groups, and also share their messages across the world. The study of sentiment analysis in twitter data has driven a lot of people. It is helpful for classifying the tweets depending upon the polarity which is positive negative and neutral. Analysing the twitter data for this analysis is tough because it has a large amount of data as the twitter contains data on various topics. For our prediction, supervised classification algorithms have been used. Using distinctive supervised machine learning algorithms such as K-nearest neighbor, support vector machine, naïve bayes, logistic regression , decision tree, and random forest algorithms were used to analyze the sentiment of tweets. These machine learning algorithms has been compared, and the most effective method is chosen to predict the outcome.

Keywords: Twitter, Machine Learning, SVM, Twitter API, KNN, Sentiment Analysis, Naïve bayes, Tweets, Classification.

1. INTRODUCTION

Social media and network sites are used by countless of people to express their thoughts, sentiments, and emotions about people's day-to-day lives. Any topic, including social interactions and product reviews, can be written about. Online communities provide a venue for discussion where users can enlighten and influence others.

The tweets from internet is a vast amount of data, so to an human it might be an very difficult task to quickly analyze the data, to read ,summarize and organize them by checking each tweet and making them into an understandable format. Nowadays most people rely on internet for fast growth in business. Many people tend to refer social media to know about a product and to know the reviews given by other customers on a product, there is a immeasurable amount of comments on internet by a lots of users. So this analysis can help the owners to know the positive negative comments on their product and it might help them to improve their product which will give them a huge profit. Sentiment analysis plays a major role in producing the information for the users.

For instance, SA can be used to forecast recommendations for products put forward by recommendation systems while accounting for variables like positive or negative sentiments about the goods. This method could be applied for a variety of things, such as helping politicians understand the attitudes that people in various regions have toward them so they can increase their investment in certain regions. Sentiment analysis could also be applied in the field of business marketing. Analysis of sentiments could be extremely important in the service sector because it can expose consumer feelings and examine the whole scope of the customer experience.

2. LITERATURE REVIEW

Hassan nazeer Chaudhry [1] et.al used Twitter sentiment to gauge public opinion before, during, and after elections and was then contrasted with the results of the vote. The Naive Bayes Classifier was used to gather public opinion after we constructed a dataset using the tweets API, pre-processed the data, and used Term Frequency-Inverse Document Frequency to extract the relevant features. The accuracy and precision of our sentiment classifier are 94.58% and a precision of 93.10%.

Bishwo Prakash Pokharel [2] et.al He analysed the sentiment of tweets posted on the Twitter social media platform, The tweets were gathered, prepared, and used with Google Colab for text mining and sentiment analysis.Following sentiment analysis using the two designated hashtags such as #COVID-19 and #coronavirus, a graphical representation of the data has been provided. The information is gathered from people who, between May 21 and May 31, 2020, shared Nepal as their location.

Kamaran H.Manguri [3] et.al In this work, Twitter data was obtained from the social media site using the Tweepy library. Sentiment analysis was then carried out using the Python TextBlob module. The data has been supplied with a graphical form following the measuring sentiment analysis. The information gathered on Twitter is based on the use of two specific hashtags, "COVID-19, coronavirus.". A visual depiction of the results and additional explanation are given at the conclusion.

Jose Ramon saura [4] et.al This study, they have conducted a sentiment analysis created with Textblob, a machine learning tool, using Computer-Aided Text Analysis and Natural Language Processing. Then, they used the Latent Dirichlet allocation model, a mathematical approach for topic modelling.

Park Ji-woong [5].This study seeks to forecast the election outcome using data from both tweets and Twitter user descriptions. This study concluded that utilising Twitter user descriptions as data for forecasts was unsatisfactory and that tweets are very successful at forecasting the outcome of the presidential election.

Shruti Kulkarni [6] et.al This analysis's objective is to show how to combine topical information from Twitter discussions with sentiment dispersion patterns to improve sentiment analysis on Twitter data .SentiDiff, a consistent approach for predicting sentiment polarity expressed in Twitter tweets, is then shown, taking into account interactions between subject information in Twitter messages and the sentiment dispersion patterns. The data indicate that this is the first study to exploit sentiment diffusion patterns to advance Twitter sentiment/emotion analysis.

Deepak Upadhyay [7] et.al However, the focus of this article will be on sentiment analysis of Twitter data. To gain a deeper grasp of public mood, They used text mining. Python is used in this study to gather, collect, and analyse tweets. Three distinct machine learning algorithms and one deep learning algorithm are then utilised for sentiment analysis, and they were compared to see which method or model performs with the highest degree of accuracy.

Pranati Rakshit [8] et.al The goal of this paper is to analyse the polarity of a tweet utilising a large amount of data obtained by natural language processing (NLP). They applied machine learning methods to identify the tweet's polarity. To determine the polarity of tweets, they used the Multinomial Naive-Bayes, Complement Naive-Bayes, and Logistic Regression classifier. The 1.6 million-entry dataset produced the greatest results when we used logistic regression. 78.05% is the highest accuracy.

Maryum bibi [9] et.al This research proposes a novel unsupervised learning approach for sentiment analysis on Twitter that is based on concept-based and hierarchical clustering. The common hierarchical clustering algorithms single linkage, complete linkage, and average linkage are serially combined. They have also tried with well-known classifiers. The performance of understudied approaches is measured by accuracy (the percentage of accurate predictions). The performance of knowledge discovery approaches is demonstrated empirically to be on par with supervised learning techniques.

3. REQUIREMENTS

The API and its key were obtained from Twitter and Visual Studio code was used to turn the unprocessed data into a dataset. The programming language Python has been used. Google Colab was the platform used for the coding.

4. EXISTING SYSTEM

VADER approach techniques were utilised in earlier works, but they didn't produce good accuracy. No authors have contrasted any machine learning algorithms up to this point. The naive Bayes algorithm is found to be the best algorithm. Some authors' who used K-nearest neighbour methods have got poorer accuracy.

5. PROPOSED SYSTEM

When examining the drawbacks of all the previous approaches utilised in earlier systems, the most prevalent issue is that the majority of these systems simply utilized a pre-defined dataset from Twitter to execute this problem statement. Additionally, they only used dynamic numbers and carried out sentiment analysis, thus they can only determine whether a tweet is positive or negative. They have established a method for analysing text from Twitter, particularly for the recently developed field of sentiment analysis. Some of the systems are also not very accurate. They didn't compare any algorithms of machine learning. All these problems in the pre-existing systems were addressed by our system. We did not use any pre-defined datasets, our main concept is to collect data using Twitter API directly. The more traditional solutions will not provide any ratios, but our system has provided a constant ratio between positive and negative tweets. We have also developed a web application.

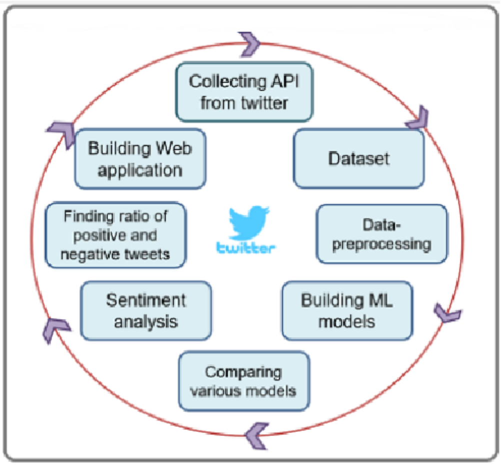


Fig.1. Project plan

The above figure (fig.1) shows the basic concept of this project. To improve performance, a range of machine learning algorithms, including Naive Bayes, Support Vector Machine, K-Nearest Neighbor, Logistic Regression, Random forest, Decision tree has been utilised and compared.

6. METHODOLOGY

Twitter account is converted to a developer account, after which Twitter's API and API key are requested. Once the API key has been obtained, it is utilised in visual studio code to gather tweets from Twitter. The tweets are then transformed into a dataset of 50950 rows and 7 columns. Then data analysing step is done. In Machine Learning, Data Analysis is the process of analyzing the data to find if there are any null values, duplicates and to know more details about the dataset.

Then data preprocessing is done. Data preprocessing in Machine Learning refers to the technique of preparing (cleaning [redacted]ng is [redacted]rformed during [redacted] preprocessing to transform categorical data into numerical data. The below figure(fig.2) explains each and every steps involved in this project

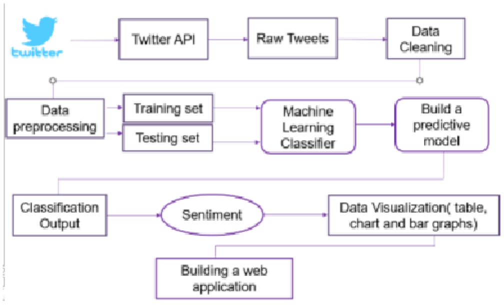


Fig.2. System architecture

Then, sentiment analysis is carried out From the dataset each tweet is checked using polarity check and the positive negative and neutral tweets were detected, and approximately 24000 neutral tweets, 17000 positive and 8000 negative tweets are obtained.

Finally, the top six machine learning algorithms is build and compared to get the best performing algorithm.

7. RESULT

We have obtained the Twitter API and converted the raw data into a dataset that includes tweets, time, user name, location, generated time, etc. using Python in Visual Studio Code.

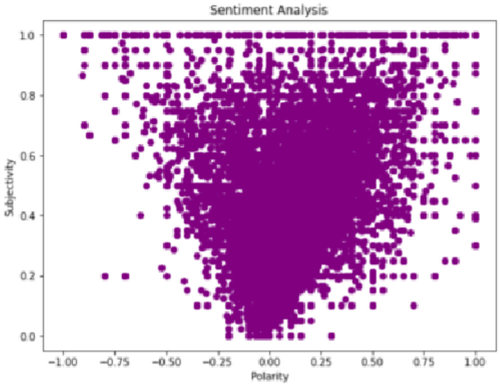


Fig.3 Subjctivity and polaity

The above figure (fig.3) shows the subjectivity and polarity of the tweets gathered

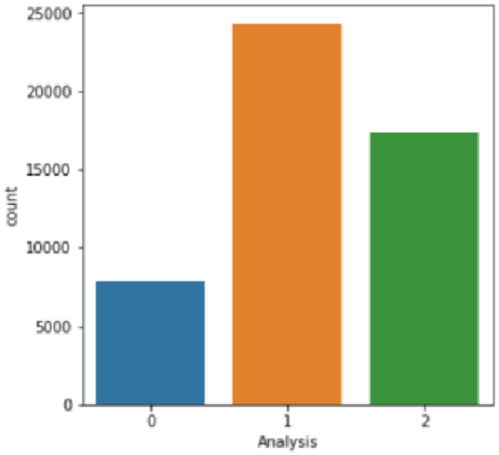


Fig.4 Sentiment analysis

In figure (fig.4)positive negative neutral tweets are represented in a bar diagram

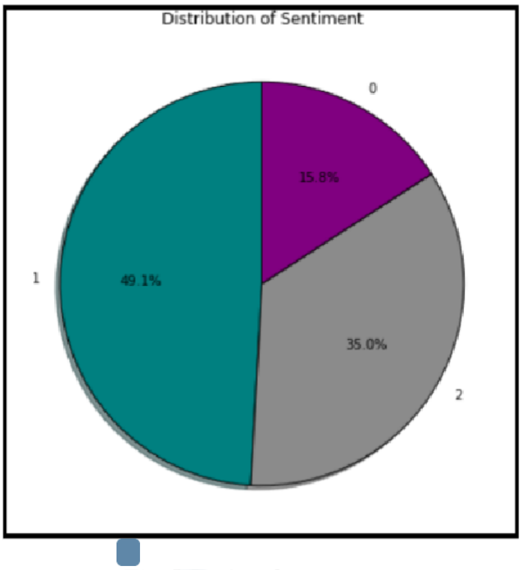
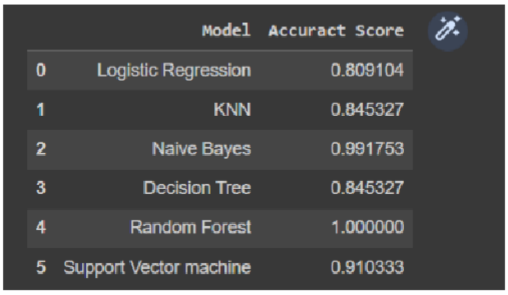


Fig.5 Ratio of tweets

In the above figure (fig.5) The ratio of the tweets were displayed in a pie chart which represents positive negative and neutral tweets.



	Model	Accuract Score
0	Logistic Regression	0.809104
1	KNN	0.845327
2	Naive Bayes	0.991753
3	Decision Tree	0.845327
4	Random Forest	1.000000
5	Support Vector machine	0.910333

Fig.6 Accuracy of the models

Finally, after building and comparing the top six machine learning algorithms.The figure (fig.6) shows their accuracy

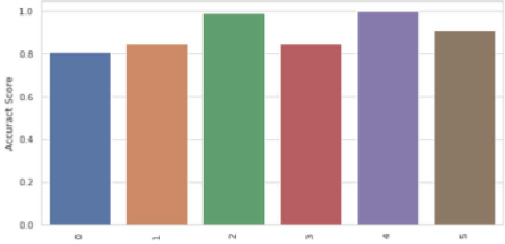


Fig.7.Comparision of the models

The figure (fig.7) shows the comparison of the models in a bar graph. We got to the conclusion that Random forest performs better with an accuracy of 100% and logistic regression performs poorly.

8. CONCLUSION

This study displays the no of positive negative and neutral tweets, Nowadays more people are into social media. So if people wanted to buy any product they refer to social media first , and this project makes it easy for them to identify the positive negative and neutral suggestions. This also helps many business, it allows owners to know the reviews about their product and find the angry customers and help them produce better products. This analysis is also used in elections to check who is going to win.

9. REFERENCE

1)Hassan nazeer Chaudhry, Yasir javed, Farzana kulsoom, Zahid mehmood, zafar Iqbal khan, umar Shoaib "Twitter sentiment analysis",2021

2)Twitter Sentiment analysis during COVID-19 Outbreak in NepalBishwo Prakash Pokharel M.Phil in 2020/6/11

3) "Twitter Sentiment Analysis on Worldwide COVID-19 Outbreaks" Kamaran H. Manguri, Rebaz N. Ramadhan, Pshko R. Mohammed Amin Published 19 May 2020

4) "Exploring the challenges of remote work on Twitter users' sentiments: From digital technology development to a post-pandemic era"Jose Ramon Saura a, Domingo Ribeiro-Soriano b, Pablo Zegarra Saldanahttps://doi.org/10.1016/j.jbusres.2021.12.052

5) Park Ji-woong professor of Hanyang University study on"Twitter sentiment analysis " published on 2022

6)Shruti Kulkarni Anisaaara Nadaph, Dr. Geetika Narang In recognition of the publication of the paper entitled "Survey on Twitter Sentiment Analysis using Supervised Machine Learning Algorithms" Published in Volume 7 Issue 5, May-2022

7) "Twitter Sentiment Analysis Using Machine Learning and Deep Learning" Pallavi Tiwari, Deepak Upadhyay, Bhaskar pant, Noor Mohd October 2020

8) Pranati Rakshit, Sumit gupta, Tarpan das Twiter sentiment analysis using NLP” published on 2022

9) “A cooperative binary-clustering framework based on majority voting for Twitter sentiment analysis” Authors Maryum Bibi, Wajid Aziz, Majid Almaraashi, Imtiaz Hussain Khan, Malik Sajjad Ahmed Nadeem, Nazneen Habib Published on 2020/3/27Journal IEEE Access

RE-2022-87016-plag-report

ORIGINALITY REPORT

1 %	8%	10%	5%	3
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS	

PRIMARY SOURCES

1 Md Babul Islam, Swarna Hasibunnahar, Piyush Kumar Shukla, Prashant Kumar Shukla, Vaibhav Jain. "Pandemic Outbreak Time: Evaluation of Public Tweet Opinion by Machine Learning", 2022 IEEE International Conference on Current Development in Engineering and Technology (CCET), 2022
Publication

www.researchgate.net
Internet Source

3

%

2

www.mdpi.com
Internet Source

%

2

www.codersarts.com
Internet Source

%

1

etd.aau.edu.et
Internet Source

%

1

gerardcambon.net
Internet Source

%

1

7 Submitted to Liverpool John Moores University

Student Paper

8 Marco Avvenuti, Salvatore Bellomo, Stefano Cresci, Leonardo Nizzoli, Maurizio Tesconi. "Towards better social crisis data with HERMES: Hybrid sensing for EmeRgency ManagEment System", Pervasive and Mobile Computing, 2020 1%

Publication

staging.friendsofunfpa.org

Internet Source

9 "Advances in Computing and Data Sciences", Springer Science and Business Media LLC, 2022 1%

Publication

10 <1%

Jose Ramon Saura, Domingo Ribeiro-Soriano, Pablo Zegarra Saldaña. "Exploring the challenges of remote work on Twitter users' sentiments: From digital technology development to a post-pandemic era", Journal of Business Research, 2022

11 <1%

Publication

Pallavi Tiwari, Deepak Upadhyay, Bhaskar Pant, Noor Mohd. "Chapter 57 Twitter Sentiment Analysis Using Machine Learning and Deep Learning", Springer Science and Business Media LLC, 2023

Publication

12 <1%

Exclude quotes On Exclude
bibliography On