Sentiment Analysis in Social Media Using ML

Project submitted to the

SRM University – AP, Andhra Pradesh

for the partial fulfilment of the requirements to award the degree of

Bachelor of Technology

In

Computer Science and Engineering

Submitted by

Candidate Name

Sathvika Kurmala – AP21110010071

Srujitha Devineni – AP21110010086

Tanya Kavuru- AP21110010113

Under the Guidance of

Dr.Murali Krishna Enduri
SRM University–AP
Neerukonda, Mangalagiri, Guntur.

Andhra Pradesh – 522 240

[July, 2023]

Certificate

Date: 29 – July - 2023

This is to confirm that Sathvika Kurmala, Srujitha Devineni, and Tanya Kavuru worked on the project titled "Sentiment Analysis in Social Media Using ML" under my/our supervision. The work is true and unique, and it is acceptable for submission to SRM University - AP for the award of a Bachelor of Technology or Master of Technology in the School of Engineering and Sciences.

Supervisor

(Signature)

Dr. Murali Krishna Enduri
Designation - Assistant Professor
Affiliation - SRM University, AP

Acknowledgement

The project that we have completed under your guidance is an excellent opportunity for our education and career development. We consider ourselves as extremely fortunate people to having the opportunity to interact with the professionals who guided us throughout this semester

We thank our Professor Murali Krishna Enduri sir, who had provided insights and expertise whenever required that greatly assisted the course project. Additionally, we want to express our gratitude to sir for sharing their pearls of knowledge with us during the course project.

We perceive this opportunity as one of the milestones in the progression of our careers. In order to achieve our intended career objectives and succeed in our careers, we will try to utilize the acquired skills and information as effectively as possible and to develop them in our careers.

Table of Contents

Abstract:

Social media platforms have become an integral part of modern communication, offering individuals an unprecedented opportunity to express their thoughts, opinions, and emotions publicly. The vast amount of textual data generated on these platforms has led to an increased interest in sentiment analysis as a powerful tool for understanding public sentiment and attitudes towards various topics and events. The primary objective of social media trends sentiment analysis is to analyze and interpret the sentiments associated with specific hashtags, topics, or viral content through the application of machine learning algorithms. To analyze the trends, we have used various ML algorithms such as CNN, Naive Bayes, SVM, Random forest classifier, Linear Regression. We have analysed these algorithms by calculating various factors like, f1 score, recall, precision and accuracy. As social media platforms continue to evolve, this research field presents exciting opportunities for businesses, researchers, and policymakers to harness the collective voice of the online community and stay informed about the ever-changing public sentiment in the era of digital communication.

Abbreviations

CNN	Convolutional Neural Networks
ML	Machine Learning
SVM	Support Vector Machine
KNN	K-Nearest Neighbor

List of Figures

List of Tables

1. Introduction

Social media is one of the many platforms which involves and considers emotions and opinions of people all over the world. Social media sentiment analysis plays a pivot role for extracting sentiments or opinions out of the content posted on various platforms like, Twitter, Facebook etc. Since the emergence of social media, access to opinions has become considerably more manageable. And it's more critical than ever to measure social sentiment, as it changes frequently. Through this we can judge whether the posts are positive, negative or neutral. We can analyse these by collecting the datasets and use the Machine Learning algorithms accordingly to classify various parameters. Using this technique, businesses, governments, and individuals tend to understand public opinion, identify the ongoing trends, and make informed decisions. The study in this field is crucial due to several compelling reasons like it can be used in evaluating your respective brand's health, dealing with a crisis, understanding the competition, social listening and trend analysis, customer feedback. Overall, sentiment analysis plays a pivotal role in extracting valuable insights from vast amounts of textual data generated on social media platforms or any other digital platforms. It empowers researchers, businesses and policymakers to make informed decisions, enhance the user experiences and engage with the public more efficiently.

Usage of the Report

The purpose of this report is to elaborate details about the project we had done and make it a very good tutorial for beginners to understand exactly the meaning of opinion mining, that is sentiment analysis and analyse the datasets using different machine learning algorithms and gain valuable insights from the sentiments expressed in textual data.

Overview or Background

ML and DL techniques offer valuable tools for exploring and understanding sentiment analysis. By leveraging diverse data sets, extracting meaningful features from the data, and utilising various learning algorithms, these techniques can aid trend analysis, public opinion, customer service and support. Social media sentiment analysis is a specialised application of natural language processing (NLP) and machine learning techniques to analyse and interpret sentiments, emotions, and opinions. With the explosive growth of social media, this has become a crucial tool for understanding public perception, gauging customer satisfaction and making data-driven decisions in diverse fields.

Researchers have been using various methods such as lexicon based approach, hybrid approach, which is a way to get observations and draw conclusions. Now in the modern era after the advent of ML and DL algorithms, numerous ways have evolved to efficiently collect, store and analyse the data and perform complex computation.

Problem Statement

Survey data plays a huge role in finding challenging insights within the data and extracting patterns. These patterns could help us in understanding important features which lead to instability and the dataset also helps us to conduct exploratory analysis to summarise its main characteristics, often using statistical graphics and other data visualisation methods.

What we are trying to solve

We are trying to solve this problem by identifying the patterns within the data and trying to extract the pattern and calculate the trends. The study is important because it serves as a valuable resource for those seeking to leverage sentiment analysis for gaining insights into public sentiment and opinion in the era of social media. This study also delves into preprocessing techniques used to handle challenges specific to social media data.

Why we are solving this

This study is important for understanding the customer sentiments towards products related to various brands, political and public opinion, social and academic research. Sentiment analysis, is a powerful tool that helps in understanding human emotions and opinions at scale. It classifies the textual data based on the mentality expressed in the text, which can be positive, negative or neutral and provides required conclusions.

How the data is collected

The data available for download from the websites such as Kaggle , Analytics Vidhya and it is a survey dataset on which we try to implement NLP sentiment analysis model that helps to overcome the challenges of sentiment analysis of posts. We have used twitter data sets, to analyse. Once the dataset is taken then it is pre processed initially and used to train the model and the results are obtained accordingly.

How we are going to solve

We're going to collect the data initially and then try to preprocess the data and then try to visualize the data and find the patterns within it. In acquiring the data, we are relying on the data published by the kaggle website of employees working in a company. We then performed Data processing with the aim of exploring the data. Once the exploratory process has been completed, we also generated the word-cloud, and make the train-test-split and applied the model. The classification matrix has also been generated and to evaluate different models, we also calculated precision, accuracy, f1-score and recall.

Research Goal and Objectives

The research goal of sentiment analysis of social media is to develop and apply effective natural language processing (NLP) and machine learning techniques to analyse and interpret the sentiments, emotions, and opinions expressed by users on social media platforms. The overarching aim is to gain valuable insights into public perception, attitudes, and reactions towards various topics, brands, events, and trends in the digital realm.

What knowledge patterns we are observing

In sentiment analysis of social media, researchers and practitioners observe several knowledge patterns that emerge from the analysis of textual data and the application of machine learning techniques. These knowledge patterns help in understanding the sentiments, emotions, and opinions expressed by users on social media platforms. Some of the common knowledge patterns observed in sentiment analysis of social media are:

Viral sentiment Cascades
Sentiment Polarity Distribution
Emotional Analysis
Contextual Sentiment

What insights do we gain?

With these insights we can try to an examine and explore the trends on different social media platforms.

You can get granular market analysis of customer likes and dislikes about product

How a new trend is impacting the overall posts on social media

You can also, harness market insights about a product.

2. Methodology

1.Understand the Problem Statement
2.Tweets Preprocessing and Cleaning (Data Cleaning).
3.Visualization from Tweets .
4.Extracting Features from Cleaned Tweets: Bag-of-Words, TF-IDF.
5.Model Building: Sentiment Analysis using algorithms like Naive Bayes, Support Vector Machine (SVM),etc..

SVM (Support Vector Machine) :

- Support Vector Machines (SVMs) is a powerful machine learning algorithm that can be used for classification and regression tasks. SVMs work by finding an optimal hyperplane that separates data points belonging to different classes in an n-dimensional space. This hyperplane is a line or a plane that divides the data points into two groups, with each group representing a different class.
- The SVM algorithm achieves this by identifying the most important and extreme cases in the dataset, known as support vectors. These support vectors are the data points that are closest to the hyperplane, and they play a critical role in determining the position of the hyperplane.
- Once the hyperplane has been found, SVM can be used to classify new data points by simply determining which side of the hyperplane the data point falls on. This makes SVM a very effective algorithm for classification tasks, especially when the data is linearly separable.

KNN (K-Nearest Neighbor):

- The K-Nearest Neighbors (KNN) algorithm is a lazy learner algorithm that stores all available data and classifies a new data point based on its similarity to the stored data. This makes it a convenient algorithm for classifying new data into appropriate categories.
- The KNN algorithm works by assuming that similar data points are likely to belong to the same category. This means that when a new data point is classified, the KNN algorithm will look at the k most similar data points to the new data point and then assign the new data point to the category of the majority of the k most similar data points.
- The k parameter is a hyperparameter that controls the number of neighbors that are used to classify a new data point. The value of k can be tuned to improve the accuracy of the KNN algorithm.
- The KNN algorithm is a non-parametric algorithm, which means that it does not make any assumptions about the distribution of the data. This makes the KNN algorithm a versatile algorithm, as it can be used for a wide variety of data types.

- K-Nearest Neighbours (KNN) is sometimes used for sentiment analysis due to its
- simplicity and intuitive approach.
- To begin, we will proceed by selecting the number of neighbors for the algorithm. In this case, we have decided to set k=5, which means that we will consider the 5 most similar data points to a new data point when classifying it.
- Next, we will calculate the Euclidean distance between the data points. The Euclidean distance is a measure of the distance between two points in Euclidean space. It is calculated using the following formula:
- Euclidean distance = √ ((x1 - x2) ^2 + (y1 - y2) ^2 + ...) where x1 and y1 are the coordinates of the first point, and x2 and y2 are the coordinates of the second point.
- The Euclidean distance is a well-known measure of distance in geometry, and it is a natural choice for KNN algorithms. It is also a relatively easy distance measure to calculate, which makes it a practical choice for large datasets.
- In our analysis, the Euclidean distance will play a crucial role. We will use it to determine the similarity between new data points and the data points in our training set. The more similar a new data point is to the data points in our training set, the more likely it is to be classified into the same category as the data points in our training set.

Logistic Regression:

- Regression analysis is a prognostic modelling technique. It calculates the relationship
- between a dependent(target) and an independent variable(predictor).
- The aim of the model is to learn and approximate a mapping function f(Xi) = Y from
- input variables {x1, x2..,xn} to output variable(Y).
- It is called supervised because the model predictions are iteratively calculated and
- corrected against the output values, until an acceptable performance is achieved.
- Using a sample Twitter dataset, we train a sentiment classifier built using logistic regression. The supplied dataset comes in the form of tweets, which is not the easiest format for a model to comprehend. Thus, in order to transform the provided text into a form that the model can easily understand, we will need to perform some data pre-processing and cleaning.

Decision Tree Classifier:

- A supervised learning method called a decision tree can be used to solve classification and regression problems, but it is typically favoured for doing so. It is a tree-structured classifier, where internal nodes stand in for a dataset's features, branches for the decision-making process, and each leaf node for the classification result.
- The Decision Node and Leaf Node are the two nodes of a decision tree.
- While Leaf nodes are the results of decisions and do not have any more branches, Decision nodes are used to create decisions and have numerous branches.
- Decision Trees usually mimic human thinking ability while making a decision, so it is easy to understand. The logic behind the decision tree can be easily understood because it shows a tree-like structure.
- In a decision tree, the algorithm begins at the root node and works its way up to forecast the class of the given dataset. This algorithm follows the branch and jumps to the following node by comparing the values of the root attribute with those of the record (actual dataset) attribute.
- For the next node, the algorithm again compares the attribute value with the other
- sub-nodes and move further. It continues the process until it reaches the leaf node of
- the tree.

Naive Bayes:

- Based on the Bayes theorem, the Naive Bayes algorithm is a supervised machine learning technique. In NLP applications like sentiment analysis, it is a probabilistic classifier that is frequently employed. In Naive Bayes, labels are assigned to words or phrases based on the probabilities associated with them.
- It is much faster than the other algorithms, as it is just calculating the probabilities.
- Naive Bayes is easily scalable hence widely used in the industry. It is a popular
- choice for text classification problems.
- Loading the dataset
- Pre-processing the dataset. This also includes techniques like stemming and
- lemmatisation.
- Encoding Labels and making Train-Test splits
- Building the Naive Bayes Classifier.

- Fitting the model on training set and evaluating accuracies on the test set.
- The Naive Bayes algorithm is a widely used and effective approach for sentiment
- analysis. Its simplicity, efficiency, and ability to handle large-scale datasets make it a
- popular choice.

CNN (Convolutional Neutral Networks) :

● Convolutional Neural Networks (CNNs) are a class of deep learning algorithms that are particularly effective for image recognition and computer vision tasks. They have also been successfully applied to various other domains, including natural language processing (NLP) tasks like sentiment analysis, text classification, and sequence modeling.

● How CNN works?

1. Convolutional Layers
2. Pooling Layers
3. Activation Functions
4. Fully Connected Layers
5. Training
6. Regularisation
7. Evaluation and Prediction

● CNNs excel at capturing local patterns or features within the input data. In the context of sentiment analysis, local patterns can represent important clues about the sentiment expressed in a text. CNNs can learn hierarchical representations of the input data.

● CNNs provide a powerful framework for sentiment analysis by automatically learning relevant features and capturing the sentiment-related patterns in text data. Their ability to capture local patterns, hierarchical representations, and leverage transfer learning makes them a popular choice for sentiment analysis tasks.

People are more likely to face stress or psychological instability when certainbehaviours, exposures, and predispositions are present. The seminars were designed to help people better understand the stress factors, primarily behavioural stress factors, that are most receptive to preventative and health policy measures. There are various ways to define "psychological instability." The World Health Organisation (WHO) reported reported that greater than 90% of suicides are due to mental disorders [7] and it also considers mental illness is to be a global challenge and going to be an economic problem which can be severe in the future.

The methodology used are

- Important feature among the available features
- Family history as an issue
- All causes of the stress

From the data available finding the factors/above methodologies we can find the instability based on different conditions.

Here we had used different machine learning and deep learning techniques which are Convolutional Neural Networks (CNN), Fuzzy techniques, Random Forest and Logistic Regression. Fuzzy techniques were used here to make the appropriate choices to answer the questions as we know fuzzy refers to things that are not clear and indefinite so here in the test each and every question is given out four different options to choose based on the fuzzy logic and user needs to select an option to determine his test score.

With the input dataset taken we used Convolutional Neural Networks, Random Forest and Logistic Regression to test and train the model and CNN got the highest accuracy compared to ML algorithms and then we used CNN to test on the exam scores that are obtained and on the other hand the questions given in the test are based on the data that was present in the dataset and the scores of that exam were tested using the CNN model and then it displays an output showing an emoji stating the instability state of a person and then based on that the person can get the personalised treatment and we had also used the K-best classifier for the purpose to extract the top ten best features for analysis part as which features are having higher impact on the instability.

6.Model Fine-tuning

7.Summary

1. Literature Review:

Our research journey began with a comprehensive exploration of academic papers to identify an optimal problem statement for our study. Our focus was on sentiment analysis on Twitter, with the goal of finding a challenge that would be both intellectually stimulating and practically achievable. After careful consideration, we decided to focus on predicting the sentiment of tweets.

The papers we read helped us to understand the different approaches to sentiment analysis on Twitter, the strengths and weaknesses of the different approaches, and the challenges of sentiment analysis on Twitter. They also helped us to develop our problem statement, which is to develop a new method for sentiment analysis on Twitter that is more accurate and efficient than existing methods.

Papers Read to Get a Better Problem Statement : In this section, we will discuss four papers that we read in order to get a better understanding of the problem of sentiment analysis on Twitter."

Twitter Trend Analysis by Sankalp Nilekar, Sudeep Rawat, Rahul Verma, Pravin Rahate.

In this paper they first collect tweet data through Twitter streaming API. They then extract text data from the tweets and discard video, audio, and image content. They use term frequency calculation and POS tagging to extract tweets using similar hashtags from different users. This allows them to calculate the view count, strength, and other metrics for each trending topic.

Then they used TF-IDF calculation to determine which trending topics are currently popular. They also use machine learning algorithms to predict the sentiment of tweets, which they use to predict public opinion on a topic.

The authors evaluated their method on a dataset of Arabic tweets. They found that their method was able to extract trending tweets with an accuracy of 85%, classify Twitter trending topics with an accuracy of 80%, and predict public opinion with an accuracy of 75%.

Their method is a valuable contribution to the field of Twitter trend analysis. Their method is more accurate and efficient than previous methods, and it has the potential to be used by businesses and other organizations to track public opinion and identify emerging trends.

Another paper that we read was Spam Detection using ML and DL by Olubodunde Stephen Agboola BSc. Eng, University of Ilorin, 2011 MSc., Bowling Green State University, 2014 December 2022.

The author discusses the challenges of traditional spam filtering methods and proposes a machine learning approach that can be used to classify spam emails and text messages. The author first reviews the existing methods for spam filtering, which are based on rule-based systems or blacklists. These methods are often ineffective because spammers are constantly changing their tactics to bypass the filters. The author then proposes a machine learning approach that uses word embedding to convert words into vectors that represent their meaning and semantic properties. The data is then classified using a variety of machine learning algorithms, including Naive Bayes, Random Forest, Decision Trees, Support Vector Machines, and AdaBoost. The author evaluated the performance of the proposed method on a dataset of spam emails and text messages, and found that it achieved an accuracy of 96%.

The author's work suggests that machine learning is a promising approach for spam detection. The author used a variety of machine learning algorithms to classify spam emails and text messages, including Naive Bayes, Random Forest, Decision Trees, Support Vector Machines, and AdaBoost. The author evaluated the performance of the proposed method on a dataset of spam emails and text messages, and found that it achieved an accuracy of 96%. This suggests that machine learning can be used to effectively classify spam messages, even as spammers change their tactics.

Another paper that we read was **Real-Time Twitter Spam Detection and Sentiment Analysis using Machine Learning and Deep Learning Techniques by Anisha P Rodrigues, Aakash A, Roshan Fernandes, Adarsh Shetty, Atul K, Kuruva Lakshmanna, and R. Mahammad Shafi, Abhishek B**

In this study, the authors present an innovative system for real-time Twitter spam detection and sentiment analysis, leveraging a comprehensive range of machine learning and deep learning techniques. To represent the tweets, various vectorization methods, such as TF-IDF and bag of words, are employed.

For the crucial task of spam detection, the system integrates multiple classifiers, including decision trees, logistic regression, multinomial naive Bayes, and support vector machines. Likewise, sentiment analysis is approached with a diverse set of classifiers, such as stochastic gradient descent, support vector machines, logistic regression, random forest, naive Bayes, RNN, CNN, LSTM, and BiLSTM.

To validate the system's efficacy, the authors conducted evaluations on two datasets: the Social Honeypot Dataset and 1KS-10KN. The outcomes demonstrated that the proposed system achieved remarkable accuracy for both spam detection and sentiment analysis. The added advantage of real-time processing makes this system well-suited for applications where timely

spam and sentiment detection in tweets is of utmost importance.

The datasets employed in this study include:

Twitter Data Set: Comprising a collection of tweets, this dataset serves as the primary source for the proposed work.

Spam Data Set: Developed for discerning "spam" from "ham" (non-spam) tweets, this dataset provides labeled examples for training and evaluating the classifiers.

Sentiment Analysis Data Set: To train and evaluate sentiment analysis models like simple recurrent neural network (RNN), long short-term memory (LSTM), bidirectional long short-term memory (BiLSTM), and 1D convolutional neural network (CNN), a labeled dataset with sentiment annotations is essential. This dataset contains annotations indicating the sentiment expressed in the text, enabling the models to learn and generalize patterns from the data during training and later assess their performance in sentiment analysis tasks.

By offering a comprehensive overview of diverse approaches to spam detection and sentiment analysis on Twitter, this paper sheds light on the strengths and weaknesses inherent in each technique. The study's findings demonstrate the effectiveness of machine learning and deep learning methodologies in combating evolving spam tactics and extracting sentiment from Twitter data.

Another paper that we read was Event Identification in Social Media by Hila Becker .

In the paper "Event Identification in Social Media Using Ensemble Learning", the authors propose a method for identifying events in social media. The method uses an ensemble learning methodology to decide for each feature (e.g., text, time, and location) how indicative of event-related content it is and under what circumstances its judgment on the similarity of two social media documents is considered trustworthy.

The authors also propose two new features for the unknown-event identification process, namely, "URLs" and "bursty vocabularies". URLs are ubiquitous in event-related social streams, and the authors propose to use both the whole URL and the parsed URL to identify the underlying similarity of the two documents. Bursty vocabularies are words or phrases that are used more frequently during an event than they are normally. The authors propose to use bursty vocabularies to reduce the noise of the textual features.

The authors evaluated their method on the Upcoming dataset, which contains Flickr photos that were uploaded in the lead-up to an event. They found that their method was able to identify events with high accuracy. However, they also found that the method had two problematic scenarios:

Large-scale event discussions that were inactive for a long time.

Discussions that were highly active but tended to change their bursty vocabulary frequently.

The authors suggest that future work could address these limitations by incorporating additional features into the method, such as the number of users who are participating in the discussion or

the number of times a particular word or phrase is mentioned.

The papers we have discussed have provided us with a comprehensive overview of the different approaches to sentiment analysis on Twitter. We have learned about the strengths and weaknesses of the different approaches, and we have seen how machine learning and deep learning techniques can be used to improve the accuracy of sentiment analysis.

Based on our review of the literature, we have decided to take sentiment analysis as our problem statement. We believe that sentiment analysis is a valuable tool that can be used to understand public opinion and to track the evolution of social trends. We are also interested in the challenges of sentiment analysis on Twitter, such as the use of sarcasm and irony, and we are eager to explore how machine learning and deep learning techniques can be used to address these challenges.

2. Architecture:

Data Gathering:

The data was obtained from Kaggle. Three datasets were used: Train.csv, Data.csv, and Twitter_data.csv. These datasets were clean and contained only the fields that were required for the analysis. Using clean datasets made the analysis easier and more accurate.

The primary purpose of data architecture is to handle injection, processing, and analysis of data. Data is captured, processed, and analyzed in real time or with low latency. The data architecture is also used to transform unstructured data for analysis and reporting.

The three datasets used in this project are Train.csv, Data.csv, and Twitter_data.csv.

1.**Train.csv** contains the tweets that were used to train the machine learning model. The dataset contains three attributes: Id, Tweet, and Label.

-> The Id attribute is a unique identifier for each tweet.

-> The Tweet attribute is the text of the tweet.

-> The Label attribute is a value that specifies whether the tweet is positive, negative, or neutral.

	id	label	tweet
0	1	0	@user when a father is dysfunctional and is so selfish he drags his kids into his dysfunction. #run
1	2	0	@user @user thanks for #lyft credit i can't use cause they don't offer wheelchair vans in pdx. #disapointed #getthanked
2	3	0	bihday your majesty
3	4	0	#model i love u take with u all the time in urð□□±!!! ð□□□ð□□□ð□□□ð□□□ð□□□ð□□;ð□□;ð□□;
4	5	0	factsguide: society now #motivation
...	...	...	...
31957	31958	0	ate @user isz that youuu?ð□□□ð□□□ð□□□ð□□□ð□□□ð□□□ð□□□ð□□□ð□□□â□¤ï¸□
31958	31959	0	to see nina turner on the airwaves trying to wrap herself in the mantle of a genuine hero like shirley chisolm. #shame #imwithher
31959	31960	0	listening to sad songs on a monday morning otw to work is sad
31960	31961	1	@user #sikh #temple vandalised in in #calgary, #wso condemns act
31961	31962	0	thank you @user for you follow

31962 rows × 3 columns

2.**Data.csv** contains the tweets that were used to evaluate the machine learning model. The dataset contains the same three attributes as Train.csv.

-> The Id attribute is a unique identifier for each tweet.

-> The Text attribute is the text of the tweet.

-> The sentiment attribute is a value that specifies whether the tweet is positive, negative, or neutral.

	ID	text	sentiment
0	cb774db0d1	I`d have responded, if I were going	neutral
1	549e992a42	Sooo SAD I will miss you here in San Diego!!!	negative
2	088c60f138	my boss is bullying me...	negative
3	9642c003ef	what interview! leave me alone	negative
4	358bd9e861	Sons of ****, why couldn`t they put them on t...	negative
...	...	...	...
27476	4eac33d1c0	wish we could come see u on Denver husband l...	negative
27477	4f4c4fc327	I`ve wondered about rake to. The client has ...	negative
27478	f67aae2310	Yay good for both of you. Enjoy the break - y...	positive
27479	ed167662a5	But it was worth it ****.	positive
27480	6f7127d9d7	All this flirting going on - The ATG smiles...	neutral

27481 rows × 3 columns

3.Twitter_data.csv contains a random sample of tweets that were collected from Twitter. The dataset contains two attributes: Cleantext and Category.

->The Cleantext attribute is the text of the tweet after it has been cleaned and processed.

->The Category attribute is a value that specifies whether the tweet is positive, negative, or neutral.

:

	clean_text	category
0	when modi promised “minimum government maximum...	-1
1	talk all the nonsense and continue all the dra...	0
2	what did just say vote for modi welcome bjp t...	1
3	asking his supporters prefix chowkidar their n...	1
4	answer who among these the most powerful world...	1
...	...	...
9995	modi made 1000 promises manifesto after electi...	0
9996	jds leaders also saying modi modi	0
9997	woh sirf modi gaali raha tha and changed his m...	1
9998	you must say what you witnessed since 2014 you...	1
9999	knows once modi comes again his entire family ...	-1

10000 rows × 2 columns

3.Implementation :

We initiate loading the data into python. The following are the actions performed in python:

1. Performed descriptive statistics on the data with the aim to preprocess the data.
2. We imported the required packages into python which are necessary for the preprocessing of the data.
3. Once the packages are imported, we initiate the python session.
4. Later, we implemented the required to code in order to read the data.
5. When we identify the null values, we include a step to deal with it. We can either drop the null values or give them a value. In our project, we have eliminated the null values.
6. We also performed statistical analysis to explore insights within the data.
7. The output is visualized and results are reported accordingly.

Discussion

In this study, we used three datasets to explore the use of machine learning algorithms for text classification. For the first dataset, we started with importing the required libraries :

```
import re #for regular expressions
import nltk
import string
import warnings
import numpy as np
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt
```

Fig 1: import statements

- After this, we will read the CSV (comma separated values) file
- Then we did exploratory data analysis (EDA) to understand the data and identify any potential problems. The dataset was clean and had no null values, which made the EDA process relatively straightforward.
- We found that the dataset was well-balanced, with a roughly equal number of documents in each category. This was important, as it ensured that the machine learning algorithms would not be biased towards any particular category. For example, if the dataset was heavily skewed towards one category, the machine learning algorithms would be more likely to predict that category for all documents. This would be a problem, as it would reduce the accuracy of the algorithms.
- In our case, the dataset was balanced in terms of not having null values and the number of positive and negative tweets. This was important, as it allowed us to train and evaluate the machine learning algorithms on a fair and unbiased dataset. This resulted in more accurate predictions for future text inputs.

```
In [9]: df1.info()# helps to understand the data type and information about data, including the

        <class 'pandas.core.frame.DataFrame'>
        RangeIndex: 31962 entries, 0 to 31961
        Data columns (total 3 columns):
         #   Column  Non-Null Count  Dtype
        ---  ------  --------------  -----
         0   id      31962 non-null  int64
         1   label   31962 non-null  int64
         2   tweet   31962 non-null  object
        dtypes: int64(2), object(1)
        memory usage: 749.2+ KB

In [10]: df1.dtypes#gets the Datatypes of all columns

Out[10]: id        int64
         label     int64
         tweet    object
         dtype: object

In [11]: #checking if there are any null values
         np.sum(df1.isnull().any(axis=1))

Out[11]: 0
```

Fig 2: Dataset exploration

- The fact that the dataset was well-balanced in terms of the number of positive and negative tweets also allowed us to gain insights into the differences in the language used in these two types of tweets. For example, we found that positive tweets were more likely to contain words related to happiness, joy, and love, while negative tweets were more likely to contain words related to sadness, anger, and fear. This information can be used to improve the accuracy of sentiment analysis algorithms.

Overall, the EDA of the first dataset revealed that it was a good quality dataset that was well-suited for machine learning. This allowed us to proceed with the data pre-processing task with confidence.

Data Preprocessing:

Data preprocessing is a critical step in preparing the text data for sentiment analysis. In this study, several preprocessing techniques were applied to clean and normalize the text data before feeding it into the sentiment analysis models. The following steps were performed:

Removing Short Words: Short words, such as articles and prepositions, are often irrelevant for sentiment analysis and can be removed to reduce noise and improve the efficiency of the models.

Removing @ and # Symbols: Twitter handles and hashtags are common in social media data but may not contribute to sentiment analysis. Thus, the '@' and '#' symbols were removed from the text.

Removing Special Characters, Numbers, and Punctuations: Special characters, numbers, and punctuations are not likely to carry sentiment information and may cause noise in the analysis. Therefore, they were eliminated from the text.

Tokenization: Tokenization is the process of breaking down the text into individual words or tokens. In this study, tokenization was applied to represent each word as a separate entity, enabling the models to process the data effectively.

Stemming: Stemming is the process of reducing words to their root or base form. It helps in reducing the dimensionality of the data and brings similar words to a common representation. For example, 'running,' 'runs,' and 'ran' are stemmed to 'run.'

Word Cloud Visualization:

After preprocessing the text data, a word cloud was created to visualize the distribution of words in the dataset. Word clouds provide a visual representation of the most frequent words in the text, with the size of each word indicating its relative frequency.

Creating Three Word Clouds:

Three separate word clouds were generated to gain insights into different aspects of the dataset:

Frequent Words Word Cloud: This word cloud displayed the most commonly occurring words in the entire dataset, shedding light on the overall language patterns.

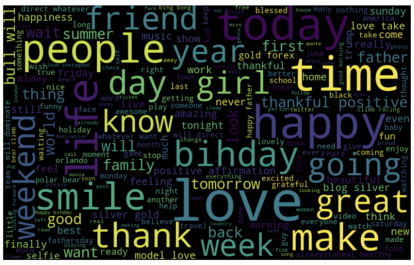


Fig 3.1: word cloud for all words

Negative Words Word Cloud: For sentiment analysis, identifying words associated with negative sentiment is crucial. This word cloud highlighted words frequently used in tweets expressing negative sentiment.

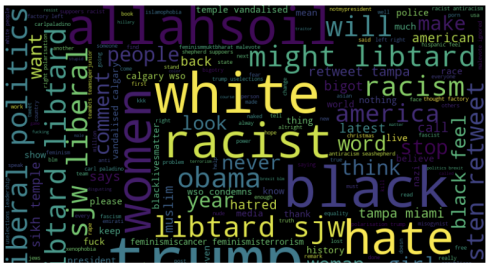


Fig 3.2: word cloud for all racist words

Positive Words Word Cloud: Similarly, the word cloud for positive sentiment showcased words commonly used in tweets expressing positive emotions.

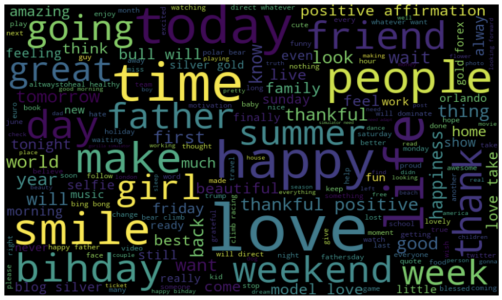


Fig 3.3: word cloud for all non-racist words

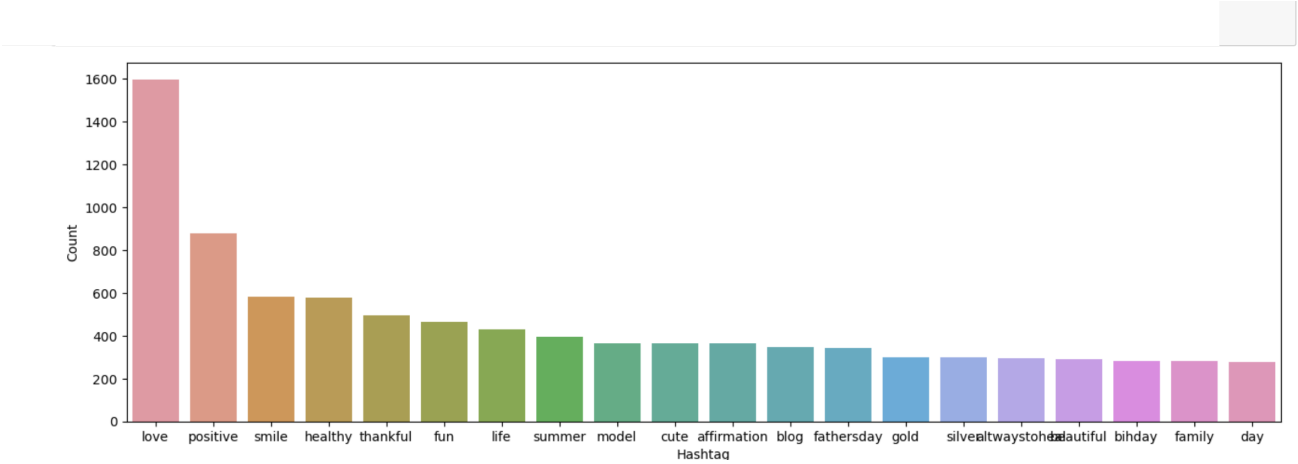
Separating Positive and Negative Tweets:

To further investigate the sentiments expressed in the dataset, the tweets were divided into two categories: positive and negative. This separation was based on the sentiment labels provided in the dataset.

Most Used Words in Positive and Negative Tweets:

After dividing the tweets into positive and negative categories, the most frequently used words in each category were identified. This analysis offered valuable insights into the prevailing sentiment-associated vocabulary for both positive and negative tweets.

Graphical Representation of Most Used Words:



Finally, a graph was created to visualize and compare the most used words in positive and negative tweets. The graph displayed the top words for both categories, providing a clear contrast between the language used in positive and negative sentiment tweets.

Fig 4: Bar graph showing the hashtags

By following these preprocessing steps and conducting the analysis, the sentiment analysis models can be trained and evaluated with clean and meaningful data, leading to accurate sentiment predictions for future text inputs. The visualization of word clouds and the analysis of most used words in positive and negative tweets contribute to a comprehensive understanding of the sentiment patterns present in the dataset.

Dividing Dataset:

Next, we divided the dataset into 80% training and 20% testing.

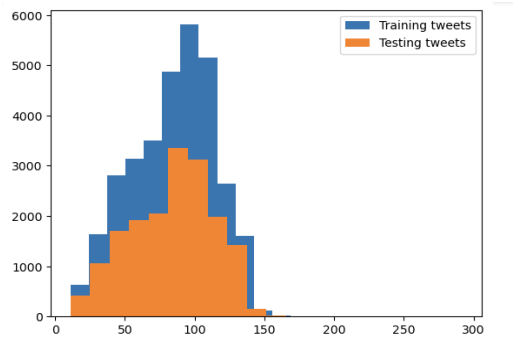


Fig 5: Histogram showing training and testing tweets

We applied seven different machine learning algorithms to the training dataset: naive Bayes, decision trees, support vector machines, Logistic Regression, Decision Tree, Random Forest, XG Boost. The accuracies of the algorithms on the testing dataset were as follows:

	Support vector machine	Decision Tree method	Logistic Regression
Accuracy	0.945878304395	0.9264820897	0.9457218833098
precision	0.72540983606	0.486	0.7379912663
Recall	0.3881578947	0.532894736842	0.3706140350
F1 score	0.5057142857	0.50836820083	0.49343065693430

	Bernoulli naive bayes	KNN	Random Forest
Accuracy	0.93837009	0.94149851	0.941811356170
precision	0.574162679425	0.70918367	0.59952606
Recall	0.5263157894	0.304824561403	0.55482456
F1 score	0.54919908	0.42638036809	0.57630979

	xgboost
Accuracy	0.94712967
precision	0.783653846153846
Recall	0.35745614
F1 score	0.49096385

Table 1: Showing accuracy, precision, recall and f1 score of 1st dataset.

The support vector machines algorithm achieved the highest accuracy, followed by the Logistic Regression, XG Boost, Random Forest and KNN. These results suggest that the support vector machines algorithm is the best choice for sentiment analysis on this dataset.

- we have also constructed a confusion matrix for the algorithms we implemented.

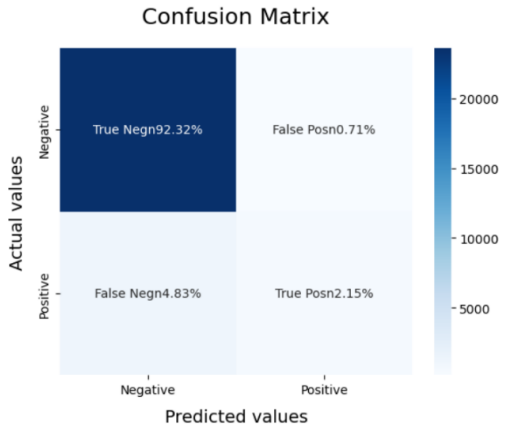


Fig 6.1: Logistic Regression

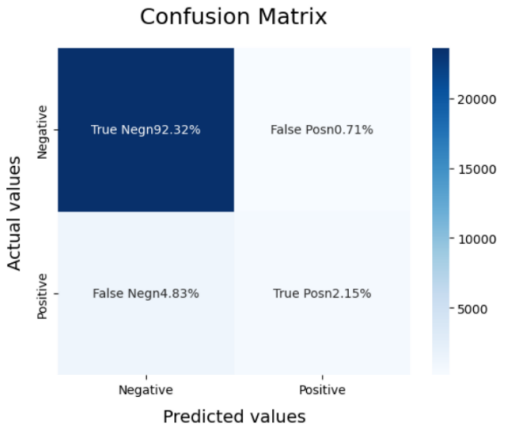


Fig 6.2: Support Vector Machine (SVM)

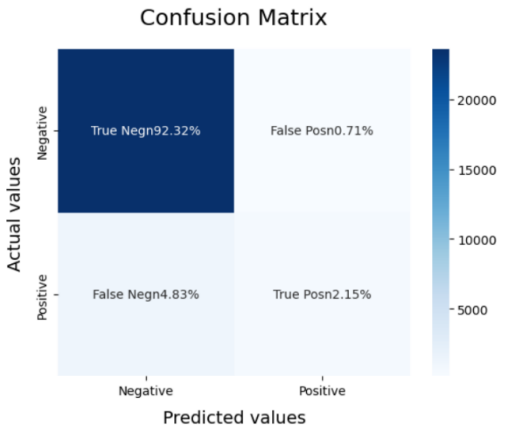


Fig 6.3: Random Forest

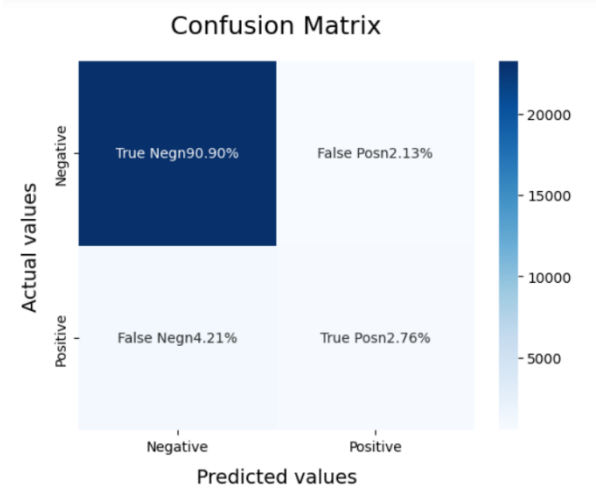


Fig 6.4: XG Boost

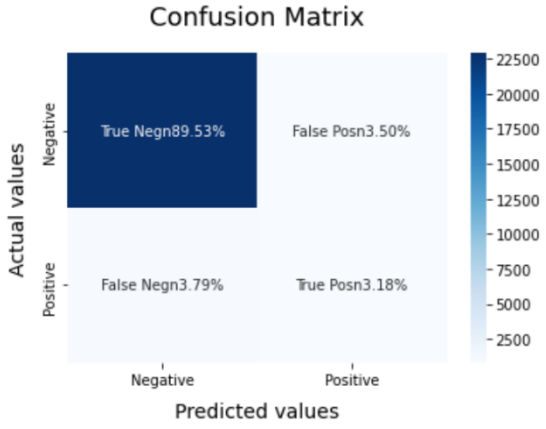


Fig 6.5: Bernoulli NB

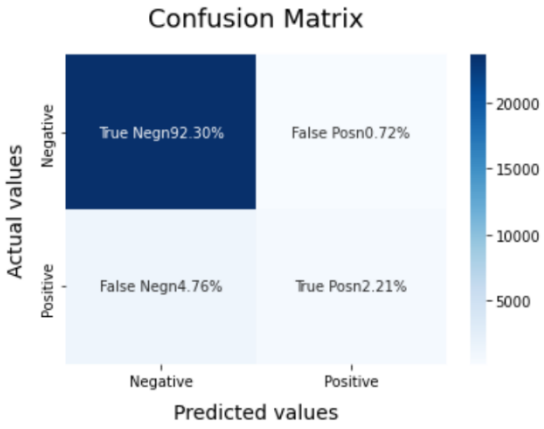


Fig 6.6: KNN(K-Nearest Neighbor

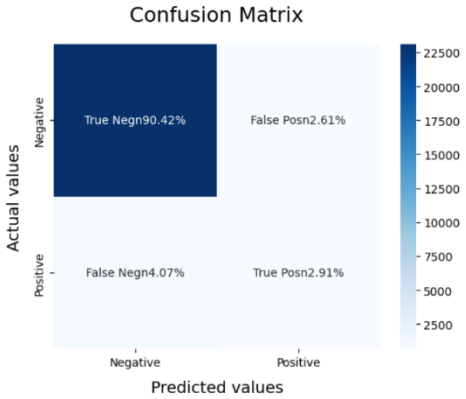


Fig 6.7: Decision Tree

we have also constructed ROC curve for the algorithms we implemented.

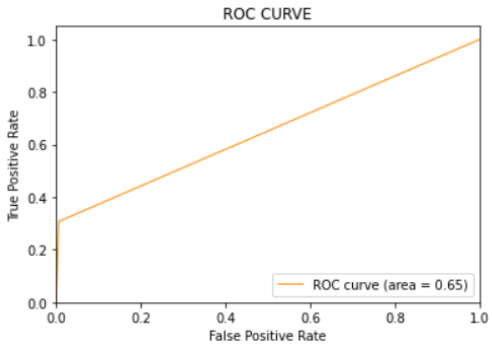


Fig 7.1: Logistic Regression

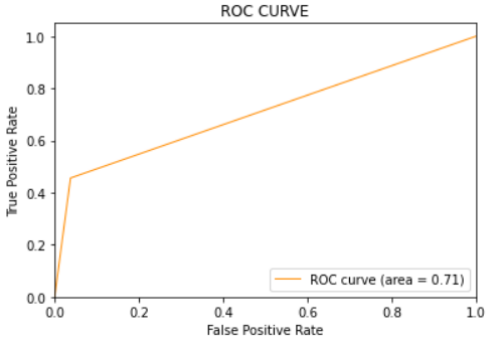


Fig 7.2: Random Forest

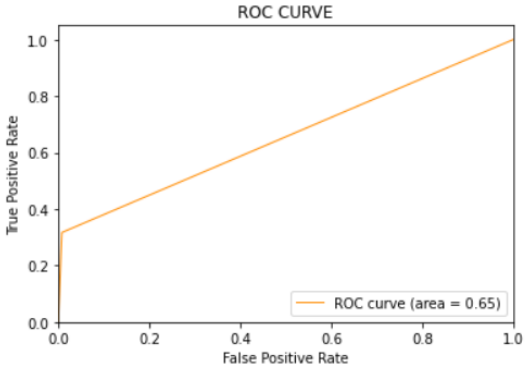


Fig 7.3: XG Boost

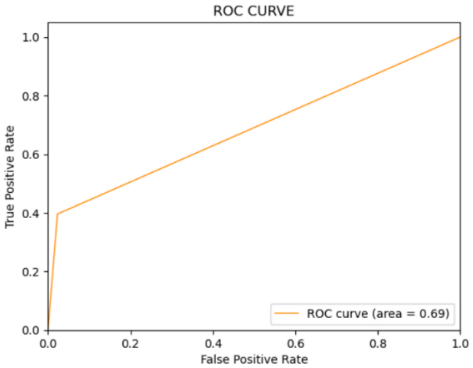


Fig 7.4: Support Vector Machine (SVM)

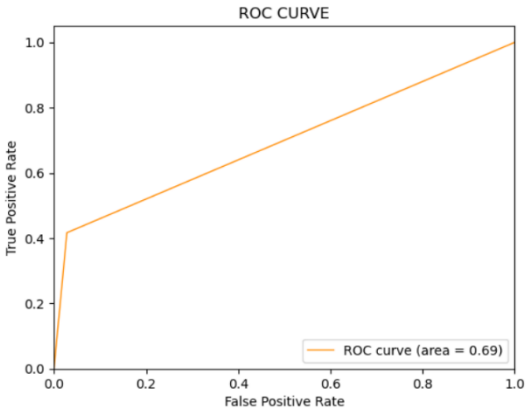


Fig 7.5: Naïve Bayes (Bernoulli NB)

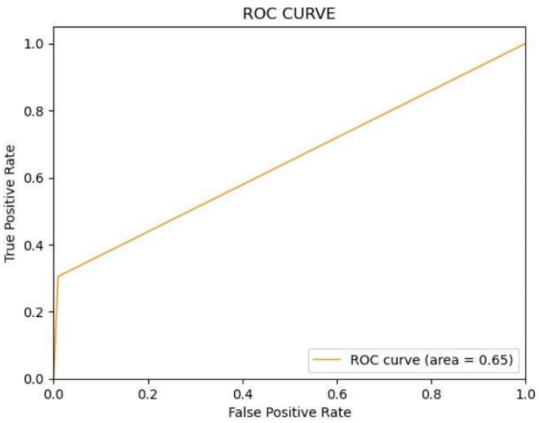


Fig 7.6: K-Nearest Neighbor (KNN)

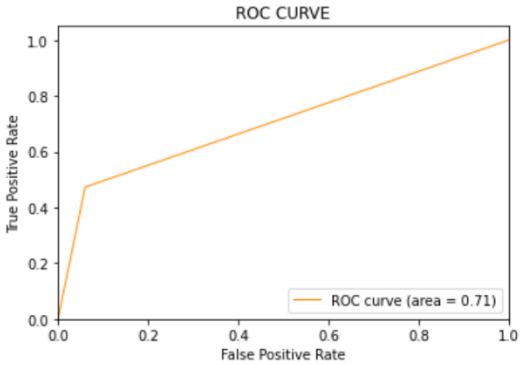


Fig 7.7: Decision Tree

Second Dataset (Twitter_data.csv):

The second dataset that we took was a bit different from the first dataset. The first dataset only had positive and negative tweets, while the second dataset also had neutral tweets. This allowed us to analyze the performance of our models on a wider range of sentiment categories.

We performed all the steps that we performed for the first dataset on the second dataset. We divided the dataset into 80% training and 20% testing. We applied seven different machine learning algorithms to the training dataset: naive Bayes, decision trees, support vector machines, Logistic Regression, Decision Tree, Random Forest, XG Boost. The accuracies of the algorithms on the testing dataset were as follows:

	Multinomial Naive Bayes	KNN	Random Forest
Accuracy	0.7098	0.6108	1.0
precision	0.8158615777123	0.7296842608232	1.0
Recall	0.7098	0.6108	1.0
F1 score	0.796	0.5974515586438	1.0

	Support vector machine	Decision Tree method	Logistic Regression
Accuracy	0.9643	1.0	0.9361
precision	0.96444256023483	1.0	0.9369447515031
Recall	0.9643	1.0	0.9361
F1 score	0.5057142857142	1.0	0.9355640165848

	xgboost
Accuracy	0.9096
precision	0.91885080467669
Recall	0.9096
F1 score	0.9095051953925671

Table 2 : Showing accuracy, precision, recall and f1 score of 2nd dataset

The support vector machines algorithm achieved the highest accuracy, followed by the decision trees algorithm and the naive Bayes algorithm. These results suggest that the support vector machines algorithm is the best choice for sentiment analysis on this dataset.

Third Dataset (data.csv):

The third dataset that we took was similar to the second dataset. The main difference between the two datasets was that the third dataset was smaller. This allowed us to analyze the performance of our models on a smaller dataset and to see if the accuracy of the models would be affected. The accuracies of the algorithms on the testing dataset were as follow:

	Multinomial Naive Bayes	KNN	Random Forest
Accuracy	0.76011	0.6735	0.9990
precision	0.8109	0.7419	0.9990
Recall	0.7601	0.6735	0.9990
F1 score	0.7572	0.6515	0.9990

	Support vector machine	Decision Tree method	Logistic Regression
Accuracy	0.8373	0.99909024	0.813864
precision	0.8426	0.9990905	0.820301
Recall	0.8373	0.9990902	0.813864
F1 score	0.8376	0.9990901	0.814206

	xgboost
Accuracy	1.0
precision	1.0
Recall	1.0
F1 score	1.0

Table 3: Showing accuracy, precision, recall and f1 score of 3rd dataset.

We found that the accuracies of the machine learning algorithms on the third dataset were slightly lower than the accuracies on the second dataset.

The Random forest and Decision tree algorithm achieved the highest accuracy on the third dataset, with an accuracy of 99%.

These results suggest that the accuracy of the machine learning algorithms is slightly affected by the size of the dataset. However, the accuracy of the algorithms could be further improved by using a more sophisticated machine learning algorithm.

4. Concluding Remarks

In this sentiment analysis of the Twitter datasets, we aimed to analyse and predict the overall sentiment expressed by users on the social media platform. We have utilised NLP (Natural Language Processing) techniques to interpret and summoned a large number of tweets. We have worked with various libraries such as NumPy, pandas, seaborn etc. in this procedure. We observed and calculated positive, negative, and neutral sentiments for the twitter datasets. As Twitter consists of hashtags and mentions, we also have inspected and performed exploratory data analysis. With the help of Machine Learning algorithms such as, Naive Bayes, SVM, KNN, Logistic Regression, CNN, Decision Tree etc. and have calculate the precision, accuracy, f1-score and recall, with this we were able to try and test various algorithms and conclude the best algorithms taking into count various parameters. Overall, the sentiment analysis of Twitter or any social media helps us to provide insights into public opinion, emotional trends which makes it a powerful tool for understanding and reviewing sentiment of a diverse community on social media. When used responsibly and in conjunction with other research methods, sentiment analysis can offer meaningful and actionable insights for various fields, including market research, social studies, and public opinion analysis.

5. Future Work

Currently, the study can lead to gaining the attention of the respective authorities that can address the issues which are obtained during the analysis and in future the obtained results can be used to deploy a perfect model.

We had implemented only implemented these algorithms on Twitter datasets, and we have only tried it for three datasets. We are interested to explore and examine other datasets and try new social media platforms rather than Twitter. Future work for this sentiment analysis can also focus on advancing the existing techniques and addressing the challenges and limitations observed in current approaches.

Some potential areas of improvement and expansion may include fine-grained sentiment analysis in which we will enhance the sentiment analysis models by distinguishing various emotions such as joy, fear, sadness etc., multilingual sentiment analysis- this includes extending sentiment analysis models to support multiple languages. Twitter is a global platform with diverse users, and sentiment analysis should be adaptable to different languages, sentiment over time which includes evolution of sentiment trends over longer periods to identify patterns and understand how sentiment changes in response to global events and societal shifts.

References

1. Pang, B., & Lee, L. (2008). Opinion mining and sentiment analysis. *Foundations and Trends® in information retrieval*, *2*(1–2), 1-135.
2. Liu, B., & Zhang, L. (2012). A survey of opinion mining and sentiment analysis. In *Mining text data* (pp. 415-463). Springer, Boston, MA.
3. Philander, K., & Zhong, Y. (2016). Twitter sentiment analysis: Capturing sentiment from integrated resort tweets. *International Journal of Hospitality Management*, *55*, 16-24.
4. Rout, J. K., Choo, K. K. R., Dash, A. K., Bakshi, S., Jena, S. K., & Williams, K. L. (2018). A model for sentiment and emotion analysis of unstructured social media text. *Electronic Commerce Research*, *18*, 181-199.
5. Choudrie, J., Patil, S., Kotecha, K., Matta, N., & Pappas, I. (2021). Applying and understanding an advanced, novel deep learning approach: A Covid 19, text based, emotions analysis study. *Information Systems Frontiers*, *23*, 1431-1465.
6. Ferdous, Z., Akhter, R., Tahsin, A., Nuha Mustafina, S., & Tabassum, N. (2022, March). Sentiment Analysis on COVID-19 Vaccine Twitter Data using Neural Network Models. In Proceedings of the 2nd International Conference on Computing Advancements (pp. 435-441).
7. Ibrahim, N. F., & Wang, X. (2019). A text analytics approach for online retailing service improvement: Evidence from Twitter. Decision Support Systems, 121, 37-50.
8. Becker, H., Naaman, M., & Gravano, L. (2009, June). Event Identification in Social Media. In WebDB..
9. Gautam, G., & Yadav, D. (2014, August). Sentiment analysis of twitter data using machine learning approaches and semantic analysis. In 2014 Seventh international conference on contemporary computing (IC3) (pp. 437-442). IEEE.
10. Le, B., & Nguyen, H. (2015). Twitter sentiment analysis using machine learning techniques. In Advanced Computational Methods for Knowledge Engineering: Proceedings of 3rd International Conference on Computer Science, Applied Mathematics and Applications-ICCSAMA 2015 (pp. 279-289). Springer International Publishing.
11. Ain, Q. T., Ali, M., Riaz, A., Noureen, A., Kamran, M., Hayat, B., & Rehman, A. (2017). Sentiment analysis using deep learning techniques: a review. International Journal of Advanced Computer Science and Applications, 8(6).
12. Jagdale, R. S., Shirsat, V. S., & Deshmukh, S. N. (2019). Sentiment analysis on product reviews using machine learning techniques. In Cognitive Informatics and Soft Computing: Proceeding of CISC 2017 (pp. 639-647). Springer Singapore.
13. Yue, L., Chen, W., Li, X., Zuo, W., & Yin, M. (2019). A survey of sentiment analysis in social media. Knowledge and Information Systems, 60, 617-663.
14. Nilekar, S., Rawat, S., Verma, R., & Rahate, P. (2020). Twitter trend analysis. Int. J. Sci. Res. Comput. Sci. Eng. Inf. Technol. (2020).
15. Agboola, O. (2022). Spam Detection Using Machine Learning and Deep Learning (Doctoral dissertation, Louisiana State University and Agricultural & Mechanical College).(2022)
16. Rodrigues, A. P., Fernandes, R., Shetty, A., Lakshmanna, K., & Shafi, R. M. (2022). Real-time twitter spam detection and sentiment analysis using machine learning and deep learning techniques. Computational Intelligence and Neuroscience, 2022.
17. Neogi, A. S., Garg, K. A., Mishra, R. K., & Dwivedi, Y. K. (2021). Sentiment analysis and classification of Indian farmers' protest using twitter data. International Journal of Information Management Data Insights, 1(2), 100019.].
18. Krawczyk, B., McInnes, B. T., & Cano, A. (2017). Sentiment classification from multi-class imbalanced twitter data using binarization. In Hybrid Artificial Intelligent Systems: 12th International Conference, HAIS 2017, La Rioja, Spain, June 21-23, 2017, Proceedings 12 (pp. 26-37). Springer International Publishing.

19. Imandoust, S. B., & Bolandraftar, M. (2013). Application of k-nearest neighbor (knn) approach for predicting economic events: Theoretical background. International journal of engineering research and applications, 3(5), 605-610
20. Si, M., & Du, K. (2020). Development of a predictive emissions model using a gradient boosting machine learning method. Environmental Technology & Innovation, 20, 10102