

Project Report

**Sentimental analysis using a machine learning and natural language
process to diseases breakout prediction**

Submitted by

Usman Ali

2020-uam-1873

2020-2024

Muhammad Shahbaz

2020-uam-1881

2020-2024

Muhammad Kashif

2020-uam-1882

2020-2024

Supervised by

Ms. Manal Ahmad



**INSTITUTE OF COMPUTING
MNS-UNIVERSITY OF AGRICULTURE, MULTAN PAKISTAN**

FINAL APPROVAL

This is to certify that we have read this report submitted by **Usman Ali(2020-uam1873)**, **Muhammad Shahbaz(2020-uam-1881)**, **Muhammad Kashif(2020-uam-1882)** and it is our judgment that this report is of sufficient standard to warrant its acceptance by MNS-University of Agriculture, Multan for the degree of BS (Computer Science).

Committee:

1. External Examiner

Dr. M. Umar Chaudhry
Assistant Professor
Bahauddin Zakariya University, Multan

2. Supervisor

Ms. Manal Ahmad
Research Officer
Institute of Computing

3. Head of Department

Prof. Dr. Salman Qadri
Director,
Institute of Computing
MNS-University of Agriculture, Multan

DECLARATION

This is to certify that *Usman Ali (2020-uam-1873), Muhammad Shahbaz (2020-uam-1881), Muhammad Kashif (2020-uam-1882)*, Session (2020-2024) have worked on and completed their software project “*Sentimental analysis using a machine learning and natural language process to diseases breakout prediction*” at the, MNS-University of Agriculture, Multan, in partial fulfillment of the requirements for the degree of BS (Computer Science).

Date:_____

Signature: _____

Usman Ali
2020-uam-1873

Signature: _____

Muhammad Shahbaz
2020-uam-1881

Signature: _____

Muhammad Kashif
2020-uam-1882

DEDICATION

My **Loving Parents**, Teachers, family and my friends who always encouraged me to work hard and guided me towards the right destination, and their love and prayers always accompanied me and guided me like a shining star whatever I was in darkness enabled me to reach this stage. Lastly, I am very thankful to our department who always supports me when I need help regarding my studies.

ACKNOWLEDGMENT

We would like to express our deepest obligation to all those who provided us the opportunity to complete this report. A special thanks we give to our final year project Supervisor, **Ms. Manal Ahmad**, whose contribution in stimulating ideas and praise, helped us to organize our project specially in writing this report.

Furthermore, we would also like to acknowledge with much gratitude the crucial role of the staff of **Institute of Computing of MNS University of Agriculture, Multan** who gave the permission to use all required equipment and the necessary materials to complete the task.

PROJECT BRIEF

PROJECT NAME	Sentimental analysis using a machine learning and natural language process to diseases breakout prediction
UNIVERSITY NAME	MNS-UNIVERSITY OF AGRICULTURE, MULTAN
UNDERTAKEN BY	Usman Ali, Muhammad Shahbaz, Muhammad Kashif
SUPERVISED BY	Ms. Manal Ahmad
STARTING DATE	Feb 1, 2024
COMPLETION DATE	June 20,2024
COMPUTER USED	Intel(R) Core (TM) i5-1235U 1.30 GHz, 8GB RAM , 256GB Solid State disk
OPERATING SYSTEM	MS Windows X
SOURCE LANGUAGE(S)	Python, Flutter, Dart,Flask
DBMS Used	Python CSV Excel File.

PLAGIARISM UNDERTAKING

I solemnly declare that the work presented in the report titled “Sentimental analysis using a machine learning and natural language process to diseases breakout prediction” is solely our work with no significant contribution from any other person. Small contributions/help wherever taken has been duly acknowledged and that complete report has been written by us.

I understand the zero-tolerance policy of the HEC and MNS-University of Agriculture, Multan towards plagiarism. Therefore, I as an Author of the above titled report declare that no portion of my report has been plagiarized and any material used as reference is properly referred/recited.

I undertake that if I am found guilty of any formal plagiarism in the above titled report even after award of the degree, the University reserves the rights to withdraw/revoke my degree and that HEC and the University has the right to publish my name on the HEC/University Website on which names of students are placed who submitted plagiarized report.

Student /Author Signature: _____

Name: _____

Student /Author Signature: _____

Name: _____

Student /Author Signature: _____

Name: _____

ABSTRACT

"Predicting sentiment of COVID-19 tweets" is a sophisticated Android application tailored to analyze and predict the sentiment embedded in tweets related to the COVID-19 pandemic. Its primary function revolves around analyzing the sentiment—whether positive, negative, or neutral—expressed in those tweets. Users can effortlessly enter tweets related to COVID-19, prompting the app to perform in-depth sentiment analysis in real-time. Through this process, users get instant insight into the prevailing sentiment on social media platforms regarding the pandemic. In addition, the app offers comprehensive trend analysis that allows users to track changes in sentiment over time, facilitating a deeper understanding of evolving public perception. To improve the user experience, the application presents its sentiment analysis results in visually engaging tables and graphs that make data interpretation more accessible. Customization options allow users to tailor sentiment analysis to specific aspects of the COVID-19 debate, such as vaccination attitudes or government responses. In addition, the app provides actionable insights and recommendations derived from sentiment analysis results, enabling users to understand prevailing sentiments and make informed decisions accordingly. The app encourages user engagement and facilitates participation through opportunities to share feelings, engage in discussions and contribute to the sentiment analysis process. Crucially, the app prioritizes user privacy and data security and follows strict privacy protocols to protect user information. The "Predicting sentiment of COVID-19 tweets" application essentially serves as a valuable tool for users who want to understand and navigate the complex landscape of public sentiment surrounding the COVID-19 pandemic on social media platforms.

Table of Contents

1.1	Introduction	1
1.2	Scope of the Project.....	3
1.3	Aims of the Project.....	4
1.4	Problem Statement.....	5
2.1	LITERATURE REVIEW	7
2.1.1	Sentiment Analysis in HealthCare	8
2.1.2	Sentiment Analysis in HealthCare	9
2.1.3	Natural Language Processing (NLP) for Text Analysis	10
3.1	Introduction to System specification	12
3.1.1	Methodology.....	12
3.1.2	Data Collection Sources	13
3.1.3	Data Preprocessing Techniques.....	13
3.1.4	Feature Extraction Methods.....	14
3.1.1	Machine Learning Models for Prediction	14
3.1.1.1	Random Forest:	15
3.1.1.2	Support Vector Machines (SVM):	15
3.1.1.3	SGD (Stochastic Gradient Descent) Algorithm:.....	15
3.1.1.4	Logistic Regression:.....	15
3.1.1.5	CatBoost (Categorical Boosting):.....	15
3.1.1.6	XGBoost (Extreme Gradient Boost):.....	15
3.1.1.7	Naïve Bayes:.....	16
4.1	Introduction	18
4.4.1	Sentiment Result	18
4.4.2	Percentage of missing values in column	19
4.4.3	Percentage of missing values in column	19
4.4.4	Percentage of missing values in column	20
4.4.5	Word Cloud Graph.....	21
4.4.5	Understanding the impact of Hashtags on tweets sentiment	22
4.4.5.1	Extremely Positive.....	22
4.4.5.2	Positive	22
4.4.5.3	Neutral	23
4.4.5.4	Negative	24

4.4.5.5	Extremely Negative	24
4.4.6	Spitting Our Dataset into Training and Testing Dataset.....	25
4.4.7	Use of Counter Vectorizer for Multi Class Classification	25
4.4.8	Machine Learning Algorithm	25
4.4.8.1	Naive Bayes Classifier for MULTICLASS Classification.....	25
4.4.8.2	Stochastic Gradient Descent-SGD Classifier(MULTICLASS CLASSIFICATION)	27
4.4.8.3	RANDOM FOREST CLASSIFIER (For Multiclass Classification)	29
4.4.8.4	Extreme Gradient Boosting (For Multiclass Classification).....	30
4.4.8.5	Support vector machine (For Multiclass Classification).....	32
4.4.8.6	Logistic Regression (For Multiclass Classification)	34
4.4.8.7	CATBOOST MODEL (For Multiclass Classification)	35
4.4.8.8	All the multiclass models test accuracy in descending order:	37
4.4.9	Machine Learning Algorithm for Binary Classification.....	37
4.4.9.1	NAIVE BAYES CLASSIFIER FOR BINARY CLASSIFICATION.	37
4.4.9.2	RANDOM FOREST CLASSIFIER FOR BINARY CLASSIFICATION ...	39
4.4.9.3	LOGISTIC REGRESSION (BINARY CLASSIFICATION).....	40
4.4.9.4	CATBOOST ALGORITHM FOR BINARY CLASSIFICATION	41
4.4.9.5	XG BOOST(BINARY CLASSIFICATION)	41
4.4.9.6	SUPPORT VECTOR MACHINE (BINARY CLASSIFICATION).....	42
4.4.10	Mobile Application for Sentiment	44
4.4.10.1	Start Screen	44
4.4.10.2	Sign Up Screen	45
4.4.10.3	Login Screen	46
4.4.10.4	Start Screen	47
4.4.10.5	Sentiment Result.....	47
5.1	Discussion and Conclusion.....	50
5.1.1	Discussion	50
5.1.2	Conclusion	51
5.2	Future Work	51
References.....		52

List of Figures

2.1.1 Flow Chart of Sentiment	9
4.1.1 Result of Sentiment.....	18
4.1.2 Graph of Sentiment	18
4.1.3 Graph of percentage of missing Values in Column	19
4.1.4 Missing Values Columns Heat map	20
4.1.5 Piechart of Top 15 Locations	20
4.1.6 Word Cloud Graph of tweet Columns	21
4.1.7 Extremely Positive Graph according to hashtags	22
4.1.8 Extremely Positive Graph according to hashtags	23
4.1.9 Neutral Graph according to hashtags.....	23
4.1.10 Negative Graph according to hashtags	24
4.1.11 Extremely Negative Graph according to hashtags.....	24
4.1.12 Result of Naive Bayes Classifier for MULTICLASS Classification	26
4.1.13 Result of Stochastic Gradient Descent-SGD Classifier	28
4.1.14 Result of RANDOM FOREST CLASSIFIER	29
4.1.15 Result of Extreme Gradient Boosting.....	31
4.1.16 Result of Support vector machine	32
4.1.17 Result of Logistic Regression	34
4.1.18 Result of CATBOOST MODEL	35
4.1.19All the multiclass models test accuracy in descending order:	37
4.1.20 Result of NAIVE BAYES	38
4.1.21 Result of RANDOM FOREST CLASSIFIER	39
4.1.22 Result of LOGISTIC REGRESSION	40
4.1.23Result of CATBOOST ALGORITHM	41
4.1.24 Result of XG BOOST	42
4.1.25 Result of SVM.....	42
4.1.27 Start Screen	44
4.1.28 SignUp Screen.....	45
4.1.29 Login Screen	46
4.1.31 Start Screen	47
4.1.32 Sentiment Result.....	48

List of Tables

Table 1 Spitting Our Dataset into Training and Testing Dataset.....	25
Table 2 Use of Counter Vectorizer for Multi Class Classification	25
Table 3 Extremely Negative of Naive Bayes Classifier for MULTICLASS Classification...26	
Table 4 Extremely Positive of Naive Bayes Classifier for MULTICLASS Classification. ...26	
Table 5 Negative of Naive Bayes Classifier for MULTICLASS Classification.....27	
Table 6 Neutral of Naive Bayes Classifier for MULTICLASS Classification	27
Table 7 Positive of Naive Bayes Classifier for MULTICLASS Classification.	27
Table 8 Extremely Negative of SGD	28
Table 9 Extremely Positive of SGD	28
Table 10 Negative of SGD	28
Table 11 Neutral of SGD Classifier	29
Table 12 Extremely Positive of SGD	29
Table 13 Extremely Negative of RANDOM FOREST CLASSIFIER.....	29
Table 14 Extremely Positive of RANDOM FOREST CLASSIFIER	30
Table 15 Negative of RANDOM FOREST CLASSIFIER.....	30
Table 16 Neutral of RANDOM FOREST CLASSIFIER	30
Table 17 Positive of RANDOM FOREST CLASSIFIER	30
Table 18 Class 0 of Extreme Gradient Boosting	31
Table 19 Class 1 of Extreme Gradient Boosting	31
Table 20 Class 2 of Extreme Gradient Boosting	31
Table 21 Class 3 of Extreme Gradient Boosting	31
Table 22 Class 4 of Extreme Gradient Boosting	32
Table 23 extremely negative of Support vector machine.....	32
Table 24 extremely positive of Support vector machine	33
Table 25 Negative of Support vector machine	33
Table 26 Neutral of Support vector machine	33
Table 27 positive of Support vector machine	34
Table 28 Extremely negative of Logistic Regression.....	34
Table 29 Extremely Positive of Logistic Regression	34
Table 30 Negative of Logistic Regression	35
Table 31 Neutral of Logistic Regression.....	35
Table 32 Positive of Logistic Regression	35
Table 33 Extremely Negative of Catboost model	36
Table 34 Extremely Positive of Catboost model.....	36
Table 35 Negative of Catboost model	36
Table 36 Neutral of Catboost model.....	36
Table 37 Positive of Catboost model.....	36
Table 38 class 0 of NAIVE BAYES	38
Table 39 class 1 of NAIVE BAYES	38
Table 40 class 0 of RANDOM FOREST.....	39
Table 41 class 1 of RANDOM FOREST.....	39

Table 42 class 0 of LOGISTIC REGRESSION	40
Table 43 class 1 of LOGISTIC REGRESSION	40
Table 44 class 0 of CATBOOST ALGORITHM	41
Table 45 Class 1 CatBoost Model.....	41
Table 46 class 0 of XG BOOST.....	42
Table 47 class 1 of XG BOOST.....	42
Table 48 class 0 of SVM	43
Table 49 class 1 of SVM.....	43

Chapter 1

INTRODUCTION

1.1 Introduction

The coronavirus known as COVID-19 is a virus that broke out worldwide in 2019 from Wuhan, China, and for a long time was the most widespread disease and the most widely discussed in the world. COVID-19 has affected many sectors of the world from healthcare, economics and education to name a few. Like this time study, more than 119 million individuals were infected with the disease, 67.2 million were infected has recovered and more than 2.63 million deaths have been recorded worldwide.

COVID-19 originally known as Corona Virus Disease of 2019, has been declared as a pandemic by World Health Organization (WHO) on 11th March 2020. Unprecedented pressures have mounted on each country to make compelling requisites for controlling the population by assessing the cases and properly utilizing available resources. The mental and physical health of the global population is found to be directly proportional to this pandemic disease. It is the need of the hour to implement different measures to safeguard the countries by demystifying the pertinent facts and information.

Since the outbreak of the COVID-19 pandemic in February 2019, discussions about COVID-19 have become widespread on Twitter, the renowned micro-blogging platform that allows users around the world to share and engage with millions of short 140-character messages. limit on different topics. In particular, individuals who are in the limelight or are widely recognized often share their status updates, which then gather responses such as retweets, discussions or reactions from their followers. It is worth noting that the influence of a tweet, especially from a prominent figure, can go beyond mere discourse and potentially affect sectors such as cryptocurrencies or the stock market (Ante, 2021). Consequently, analyzing the sentiments expressed on Twitter regarding the issues of COVID-19 on a global scale is of considerable importance. This showed that it is perhaps the most natural focus of infection in the last twenty years in the century. With statistics obtained from the use of social media, especially Facebook and Twitter, the advent of technology has allowed us to understand the impact of this pandemic in many industries. It is well known that information spreads quickly Internet, resulting in a wide range of emotions especially among social network users on microblogs such as Twitter. In the twentieth century, social connection developed to a technical level that allows people to connect with others around the world to drive corporate adoption. Twitter is one of the social programs that is widely used for polling, with 100 million users posting 250 million tweets. Infection with COVID-19 has become a threat not only to public

health, but also to global development. COVID-19 can be a highly contagious disease that targets the respiratory system and lung. According to studies, COVID-19 is related to the coronavirus and shares characteristics with the disease that first appeared in 2003 under the name severe acute respiratory syndrome (SARS). Starting with China, the SARS virus infected 29 countries.

In our application, the primary goal is to perform sentiment analysis on tweets related to COVID-19 to discern the predominant sentiments expressed by patients. Using natural language processing techniques and machine learning algorithms, we try to accurately predict the sentiment contained in these tweets. Understanding the feelings of patients in the midst of the COVID-19 pandemic is invaluable to various stakeholders, including health care providers, policy makers, and the general public. In the field of COVID-19 sentiment analysis on Twitter data, we focus on understanding public perceptions and emotions related to the pandemic. Sentiment analysis plays a key role in assessing the general mood, concerns and opinions expressed by individuals on social media platforms. COVID-19 sentiment analysis involves using machine learning techniques to analyze massive amounts of Twitter data to gain insights into public sentiment towards the pandemic.

1.2 Scope of the Project

The scope of our project includes the development of a robust sentiment analysis system specifically tailored for patient tweets related to COVID-19. This includes collecting various datasets of tweets related to COVID-19, pre-processing text data to ensure quality and relevance, extracting features using NLP techniques and implementing machine learning models to predict sentiment. We aim to cover a broad spectrum of feelings, including but not limited to positive, negative and neutral, to provide a comprehensive understanding of patient perspectives. Our project further includes evaluating and fine-tuning machine learning algorithms to optimize predictive accuracy. We intend to explore different techniques such as TF-IDF vectorization, word embedding and file methods to improve the model performance. In addition, the project may involve developing a user-friendly interface to facilitate easy access and interpretation of sentiment analysis results. While the primary focus is on sentiment analysis, we may also explore additional tasks such as topic modeling, analyzing trends in sentiment over time, and identifying correlations between sentiment and various demographic or geographic factors. However, the scope of the project will be defined and refined based on available resources, time constraints, and the specific needs and goals of the stakeholders.

Overall, the project aims to provide a reliable and insightful sentiment analysis system that can contribute to a better understanding of patient sentiments during the COVID-19 pandemic and ultimately aid in informed decision-making and resource allocation.

1.3 Aims of the Project

This project aims to perform sentiment analysis on global tweets focusing on key topics such as National Immunization Program, Movement Control Order (MCO) and daily cases of COVID-19. Twitter's real-time streaming API will be used to collect the latest tweets from around the world, providing the most up-to-date results on targeted issues. However, only geotagged tweets will be streamed to specifically capture tweets originating from different countries around the world. It is worth noting that there may be cases where tweets without geotags are not included in the analysis. A polarity index will be used to assess the overall sentiment of tweets and classify them as positive, Extremely Positive, negative, Extremely Negative or neutral. At the same time, the subjectivity index will measure the emotional tone of the tweets, which includes feelings such as happiness, sadness, anger and confusion. These indices will be calculated using the Naïve-Bayes algorithm. The results of the sentiment analysis will be visually presented through pie charts, facilitating a clearer understanding of the global sentiment landscape regarding the given topics. In conclusion, this project aims to provide insight into global sentiment around key issues related to the COVID-19 pandemic, enabling a comprehensive understanding of public opinion and emotions in different countries around the world. The primary objective of this project is to develop an efficient, time-efficient and user-friendly system for conducting sentiment analysis on COVID-19. Unlike conventional sentiment analysis methods, this system aims to streamline the process and reduce the workload associated with manual analysis. With widespread adoption of the technology, individuals may soon be able to make payments in stores using facial recognition, eliminating the need for physical contact with credit cards or cash. This innovative approach not only saves time when paying, but also complies with security measures in a post-Covid world, offering contactless and hassle-free payments. Facial recognition technology enables quick identification, with the entire process taking just seconds. This fast verification process is especially beneficial for companies looking for both security and efficiency in an era marked by cyber threats and hacking incidents. In law enforcement, facial recognition technology holds promise for improving the accuracy and fairness of identification procedures. By automating the process of identifying suspects in crowds, it has the potential to reduce bias and reduce unnecessary stops and searches of law-abiding citizens. Additionally, most facial recognition solutions are compatible with existing security software, making integration seamless and cost-effective. This minimizes the need for additional investment in technology implementation.

1.4 Problem Statement

Current methods of analyzing public opinion on COVID-19 around the world are insufficient to provide accurate information to government bodies and the public. These mechanisms do not provide a comprehensive overview of global sentiment on critical issues such as the National Immunization Program, the Movement Control Order (MCO) and daily case counts. The current increase in cases of COVID-19 worldwide is exacerbating this problem and presenting additional challenges in accurately assessing public opinion and concerns. In the early to mid-stages of the pandemic worldwide, studies have shown that lower-income individuals are disproportionately affected by moderate to severe anxiety. However, these statistics do not fully capture the wider impact of COVID-19 on poverty rates globally. It is imperative that governments around the world address the psychological, behavioral and economic consequences of the pandemic on their citizens. Additionally, the spread of fake news poses a significant threat to shaping public sentiment around the world. Spreading misinformation through social media platforms can distort perceptions and lead to harmful outcomes, such as promoting ineffective home remedies and fueling anti-vaccination sentiment. In various countries, platforms like WhatsApp serve as common channels for sharing misinformation, including dangerous practices such as drinking bleach as a supposed cure for COVID-19. Additionally, global vaccination efforts face challenges in gaining public trust and participation. For example, only a fraction of the global population has registered for national vaccination programs, with significant numbers of people missing vaccination appointments in some regions (News Strait Times, 2021). These sentiment-related issues are hindering effective efforts to control the pandemic around the world, requiring governments to consider public opinion and sentiment when designing future strategies to combat COVID-19. In summary, the global landscape of public sentiment around COVID-19 presents multifaceted challenges, including disparities in vulnerability, the spread of misinformation, and vaccination hesitancy. Addressing these challenges requires a concerted effort by governments and stakeholders worldwide to accurately understand and address public concerns, ultimately contributing to more effective pandemic response and mitigation strategies.

Chapter 2

BACKGROUND

2.1 LITERATURE REVIEW

The five main themes associated with the COVID-19 pandemic were: economy and trade, health care, emotional support, psychological stress and social change. These topics brought the greatest worries to the inhabitants [1]. Gao et al. [2] examined outcomes related to individuals' mental well-being, who have been constantly exposed to online media during this pandemic. It was stated that a large number of mental health problems, such as anxiety or bitterness, were well associated with increased social media use during COVID-19. In Konac et al. [3] the use of social media as a source of information has been linked to the spread of conspiracies regarding the pandemic, along with the promotion of several unverified health claims protective practices. The difficult situation where people cannot leave their homes requires investigating what people think about during a pandemic [4]. The overall sentiment conveyed in tweets during the pandemic was more positive, meaning that the public remained hopeful even when faced with a global public health problem. The highest level of positive moods indicates that many people were carefree on the severity of COVID-19 in the early phase of the pandemic. One important one of note is that tweets generated in states with lower infection rates were generally positive, while states directly affected by the pandemic were negative- sentimental phrases recommend that tweets can be an approach express negative feelings about the consequences of the COVID-19 restrictions. Negative erms usually include "know" and "think," words that identify with information and information sharing [1].

Since the outbreak of the COVID-19 pandemic, scientists have been actively investigating its origins, impacts and trends. In particular, the analysis of sentiments expressed on Twitter, which uses various machine learning (ML), deep learning (DL) and natural language processing (NLP) techniques, has become a focal point. However, extracting meaningful insights from the vast amount of noisy data available on social media presents a significant challenge that underscores the necessity of ML and DL tools for this task. Twitter serves as a major platform for gathering news and opinions in real time [5]. For example, one study focused on India-specific sentiment analysis and used 24,000 tweets related to COVID-19 to visualize public sentiment toward the pandemic [6]. Meanwhile, other research has delved into categorizing tweets as positive, negative, or neutral and analyzing sentiment using different feature sets and classifiers. Notably, they achieved a remarkable accuracy of 94.80% with the Bidirectional Encoder Representations from Transformers (BERT) model (1). In addition to sentiment analysis, researchers have explored the psychological implications of the pandemic, highlighting heightened levels of anxiety and feelings

of crisis among individuals due to news of COVID-19 [7]. In addition, there is extensive research on the industrial and economic consequences of the crisis in different sectors and countries [8]. Over the years, sentiment analysis based on social media data, including platforms such as Twitter, Facebook, Reddit, and YouTube, has found numerous applications [9]. However, these analyzes often reveal gaps in the collected data, leading to the exploration of different ML and DL classifiers to handle both short and long text information. While models such as logistic regression and Naive Bayes provide satisfactory results for short texts, they struggle with longer texts [10]. As people increasingly rely on social media for news and expression during the pandemic, there is an increasing influx of opinions, emotions, and feelings shared through online posts [11]. This underscores the importance of continued research in understanding and interpreting the feelings and behaviors reflected on social media platforms in the midst of this global health crisis.

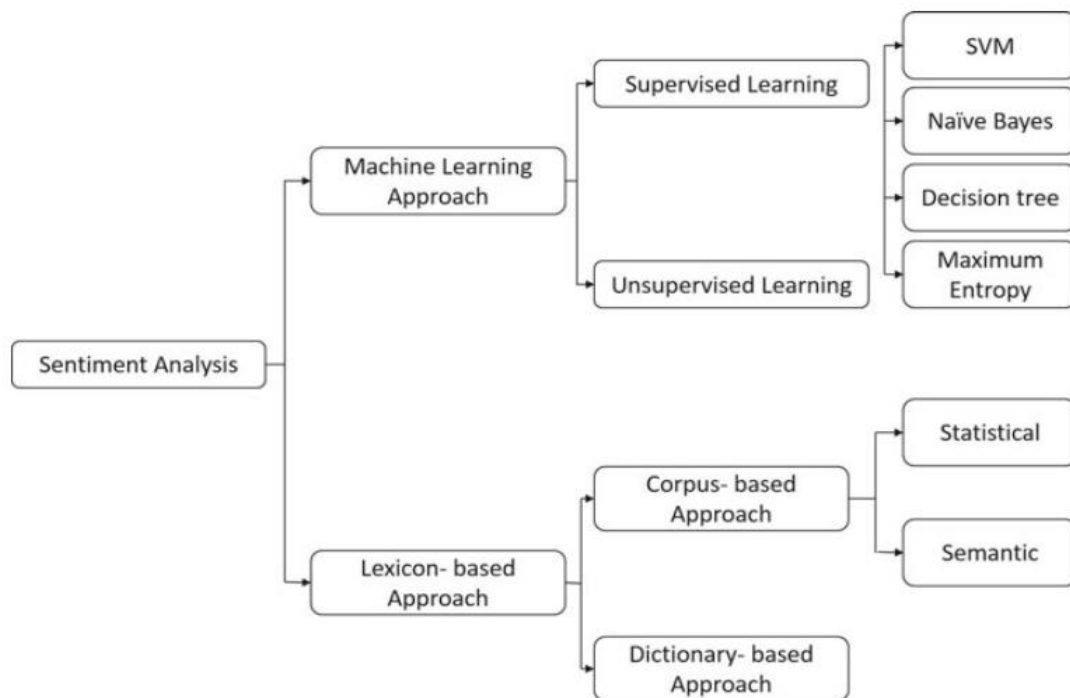
2.1.1 Sentiment Analysis in HealthCare

Understanding patients' experiences through sentiment analysis is crucial for them to make decisions about their health concerns, such as choosing hospitals, clinics and medications. This information is also valuable for hospitals to understand the interests, problems and preferences of patients and enable them to provide better care. Patients share their experiences, emotions and feelings, which are the essence of sentiment analysis. Sentiment analysis involves understanding people's feelings about a subject and its features. The vast amount of medical content available online is free, but manual analysis of such a large amount of data is inefficient. Evaluation often focuses on automatically identifying sentiments as positive or negative. Currently, sentiment analysis is replacing traditional surveys conducted by organizations to gauge public opinion about their products and services, which helps in marketing strategies and product improvement. Automated sentiment analysis is essential due to the availability of vast content online. It requires a deep understanding of natural languages because feelings and emotions play a significant role in decision-making, learning, communication and situational awareness. With the growth of regionally specific content on social media, processing and understanding vernacular content is increasingly important. Sentiment analysis techniques automate this process without human intervention. Previously, researchers used questionnaires and surveys to analyze sentiment, which were costly and time-consuming. However, these methods did not address the real problems of patients. Sentiment analysis delves into the sentiments of patients expressed in many online

documents across multiple platforms. The main goal of sentiment analysis is to categorize medical decisions as positive or negative. However, deeper analysis can reveal subtle aspects of medical problems. This article highlights the importance of sentiment analysis based on a large number of patients' perspectives on their diseases, medications and health problem.

2.1.2 Sentiment Analysis in HealthCare

In this study [12], the authors reviewed aspects of sentiment analysis in the medical field and



0.1 Flow Chart of Sentiment

explored potential use cases. They summarized the state-of-the-art approaches in the health care setting through their research review. Their goal was to understand the linguistic nuances of sentiment expressed in medical texts and to identify open research challenges in sentiment analysis in medicine. To do this, they conducted a quantitative assessment focusing on the language used and sentiment combinations on a dataset containing clinical narratives and medical community discussions collected from six different sources. Sentiment analysis in healthcare is critical to solving healthcare patient problems [13]. Analyzing patients' feelings helps to quickly solve their problems, helping decision-makers to design plans and implement beneficial changes. Sentiment

analysis finds application in various fields, and studies focusing on healthcare analytics highlight the strengths of treatments and services [14,15].

2.1.3 Natural Language Processing (NLP) for Text Analysis

Natural language processing is an area of research within artificial intelligence (AI) that deals with enabling computers to understand natural language (spoken and written) in the same way that a human would [18]. To achieve the above-mentioned goal, knowledge from computational linguistics (rule-based modeling of human language), statistics, machine learning and deep learning [19] is used either separately or combined. The term Exposom was first introduced [20], which defined a field of research that makes systematic measurements of exposures (e.g. occupational, physical environment or socio-economic factors) to which a person is exposed throughout life (from birth to death).) and affects their health outcomes [20]. However, the term Exposom itself has not yet been fully integrated into all areas of exposure research, where the term "exposure research" is often used when referring to the same or similar concepts. At the same time, text mining and NLP techniques are increasingly applied in various fields of exposure-related research. While there are a number of surveys and literature reviews on NLP and its various subtasks, there is no overview of NLP and text mining techniques used in occupational and environmental exposure research. This review fills this gap by providing a description of existing tools based on NLP and text mining techniques that have been used in occupational and environmental exposure research.

Chapter 3

METHDOLOGY

3.1 Introduction to System specification

we delved into the existing literature on sentiment analysis using machine learning and natural language processing (NLP) techniques for disease outbreak prediction. We have explored the various methodologies and approaches used by researchers in this field. Now let's discuss our methodology for performing sentiment analysis to predict disease outbreaks.

After reviewing related works in sentiment analysis and disease prediction, we came across several algorithms that could be applied to our project. These algorithms ranged from traditional machine learning models. After careful consideration and analysis, we have decided to use a specific set of algorithms that we believe would best suit our goals. We used machine learning algorithms such as Support Vector Machines (SVM), Naive Bayes, Logistic Regression, Stochastic Gradient Decent SGD, Catboost Algorithm, XG boost and Random Forests for sentiment analysis. These algorithms have proven effective in sentiment classification in textual data, allowing us to identify positive, negative, or neutral sentiments related to disease outbreaks. In addition, we used natural language processing techniques to pre-process and analyze the text data. This included tokenization, stemming, stop-word removal and feature extraction using methods such as TF-IDF (Term Frequency-Inverse Document Frequency) and word embedding (e.g. Word2Vec). To predict disease outbreaks based on sentiment analysis, we integrated these machine learning and NLP components into a cohesive framework. This framework allowed us to process large volumes of textual data from various sources such as social media, news articles, and healthcare databases.

3.1.1 Methodology

The methodology for performing sentiment analysis using machine learning and natural language processing (NLP) techniques for disease outbreak prediction involves several key steps.

First, data collection is essential, including sources such as Twitter tweet data records. This data is then pre-processed to remove noise, tokenize and standardize textual representations using techniques such as stemming or lemmatization. This step may also include consideration of domain-specific features related to diseases, symptoms, treatment, and public sentiment. The sentiment analysis phase uses different techniques, including rule-based methods, machine learning classifiers (SVM, Naive Bayes, Random Forest). These models are trained on labeled data to classify sentiments into positive, negative, or neutral categories. To predict disease outbreaks, machine learning models are built using sentiment analysis results as features. Classification

algorithms such as SVM, logistic regression, or decision trees are explored to predict disease outbreaks based on sentiment trends identified in the data. Evaluation metrics such as accuracy, precision, recall, and F1-score are defined to assess the performance of both sentiment analysis and disease prediction models. To verify the generalizability of the model, cross-validation and testing on separate data sets is performed. Integration with existing healthcare systems or public health surveillance platforms is essential for practical deployment. Data privacy, security and compliance with healthcare regulations such as HIPAA are priorities for integration. Continuous learning and improvement mechanisms are implemented to adapt the models to evolving moods and disease patterns. Feedback from health professionals, researchers and users is collected to refine the model and increase the accuracy of predictions. Finally, the system is deployed in a production environment and continuous monitoring ensures optimal performance, data quality and model stability over time.

3.1.2 Data Collection Sources

The COVID-19 pandemic has been a major global health crisis, and sentiment analysis using machine learning and natural language processing has played a key role in understanding public perception and predicting trends in outbreaks related to COVID-19. Data collection for the sentiment analysis on COVID-19 involved gathering information from various sources. Social media platforms such as Twitter provided real-time data on public discussions, concerns and opinions about the pandemic. Analysis of hashtags, trends and user comments helped gauge sentiment on topics such as lockdowns, vaccinations, masking and government response.

Machine learning algorithms were then used to analyze this data and develop predictive models for trends in the spread of COVID-19. These models have helped predict infection rates, identify high-risk areas, and evaluate the effectiveness of public health interventions. Sentiment analysis using machine learning and natural language processing has played a vital role in tracking public sentiment, informing decision-making, and mitigating the impact of the COVID-19 pandemic.

3.1.3 Data Preprocessing Techniques

Data preprocessing techniques are essential steps in preparing raw data for machine learning and data analysis tasks. These techniques include a series of operations aimed at cleaning, transforming and organizing data so that it is suitable for further analysis and modeling. One of the primary data

preprocessing techniques is data cleaning, which involves handling missing values, removing duplicates, and correcting inconsistencies in the data. Missing values can be resolved using imputation techniques such as mean, median, or mode imputation, or by removing rows or columns with a high percentage of missing values. Duplicates are identified and eliminated to ensure data accuracy, while inconsistencies such as typos and formatting errors are corrected. Another crucial step is data transformation, which includes normalization and standardization. Normalization scales numeric elements to a standard range, such as 0 to 1, to ensure that all elements contribute equally to the analysis. Standardization transforms the data to have a mean of 0 and a standard deviation of 1, which is particularly useful for feature range-sensitive algorithms such as support vector machines and k-nearest neighbors.

3.1.4 Feature Extraction Methods

Feature extraction methods are techniques used to reduce the dimensionality of data by extracting relevant features that capture essential information for analysis and modeling tasks. These methods play a key role in machine learning and data analysis, especially when dealing with high-dimensional datasets where the number of features exceeds the number of samples.

Common feature extraction methods include principal component analysis (PCA), which transforms correlated variables into linearly uncorrelated components while retaining the most significant variance in the data. Another technique is linear discriminant analysis (LDA), which aims to find linear combinations of features that best separate different classes in supervised learning tasks. Feature extraction is essential to reduce computational complexity, improve model performance, and improve interpretability by focusing on the most informative features relevant to underlying patterns in the data.

3.1.1 Machine Learning Models for Prediction

Machine learning models for prediction include a series of algorithms designed to learn patterns and relationships from data and make predictions or decisions based on those learned patterns. These models are used in a variety of fields, including healthcare, finance, marketing, and more. Common machine learning models for prediction include:

3.1.1.1 Random Forest:

A set of decision trees that combines multiple trees to improve prediction accuracy and reduce overfitting.

3.1.1.2 Support Vector Machines (SVM):

A supervised learning model that finds the optimal hyperplane for class separation in classification tasks and can be extended to regression (SVR) for numerical predictions.

3.1.1.3 SGD (Stochastic Gradient Descent) Algorithm:

An iterative method for optimizing an objective function by updating parameters using loss function gradients, often used in large-scale machine learning problems due to its efficiency. Updates parameters for each training example one by one, which can lead to faster convergence compared to batch gradient.

3.1.1.4 Logistic Regression:

A statistical method for binary classification that models the probability of a binary outcome based on one or more predictor variables using a logistic function. Finds the most appropriate model to describe the relationship between the dependent binary variable and the independent variables.

3.1.1.5 CatBoost (Categorical Boosting):

A gradient boosting algorithm that automatically processes categorical features without the need for extensive preprocessing, designed to provide robust performance with less parameter tuning. Uses ordered boosting and a combination of categorical and numeric functions to improve accuracy and speed.

3.1.1.6 XGBoost (Extreme Gradient Boost):

An efficient and scalable implementation of gradient boosting that includes regularization techniques to avoid overfitting and supports parallel processing for faster computations. Known for its performance and accuracy in structured/tabular data, it provides various hyper parameters to fine-tune models.

3.1.1.7 Naïve Bayes:

A simple probabilistic classifier based on Bayes theorem with strong (naive) assumptions of independence between features. Despite its simplicity, it is highly effective for certain types of problems, especially text classification and spam detection.

Chapter 4

RESULTS

4.1 Introduction

The COVID-19 pandemic has highlighted the need for innovative approaches to predicting and managing disease outbreaks. As the virus spread around the world, a huge amount of textual data was generated on social networks, news platforms and other digital communications, reflecting public opinion and reactions to the unfolding crisis. This study focuses on using sentiment analysis through machine learning (ML) and natural language processing (NLP) techniques to predict the outbreak of COVID-19, specifically using data obtained from Twitter. By analyzing the sentiments expressed in tweets, we aim to uncover patterns that precede spikes in COVID-19 cases. This approach uses computational linguistics and predictive modeling to provide early warning signals, potentially improving public health responses and mitigating the impact of future epidemics. Through this research, we aim to demonstrate how sentiment analysis can be a valuable tool in the anti-pandemic arsenal, offering insights that are timely and actionable.

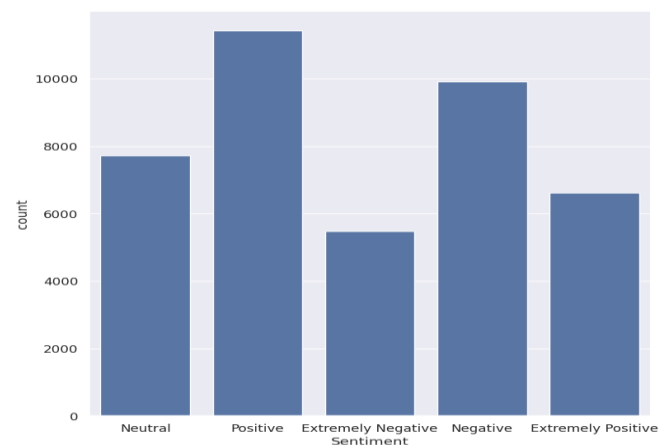
4.4.1 Sentiment Result

Our dataset contains a total of 41,157 tweets, divided into different sentiment labels. Among them, 11,422 tweets are marked as positive, reflecting a generally favorable response. Negative sentiments are represented by 9,917 tweets, indicating a substantial portion of unfavorable opinions. Additionally, there are 7,713 tweets that fall into the neutral category, neither positive nor negative. The dataset also contains more extreme sentiments, with 6,624 tweets classified as extremely positive, indicating strong favorable sentiment, and 5,481 tweets labeled as extremely negative, indicating a strong unfavorable reaction. This diverse distribution of sentiment provides a comprehensive overview of the varying degrees of response expressed in tweets.

```
# There are 5 unique sentiment types in our dataset  
df['Sentiment'].value_counts()
```

```
➡ Sentiment  
Positive      11422  
Negative      9917  
Neutral       7713  
Extremely Positive  6624  
Extremely Negative  5481  
Name: count, dtype: int64
```

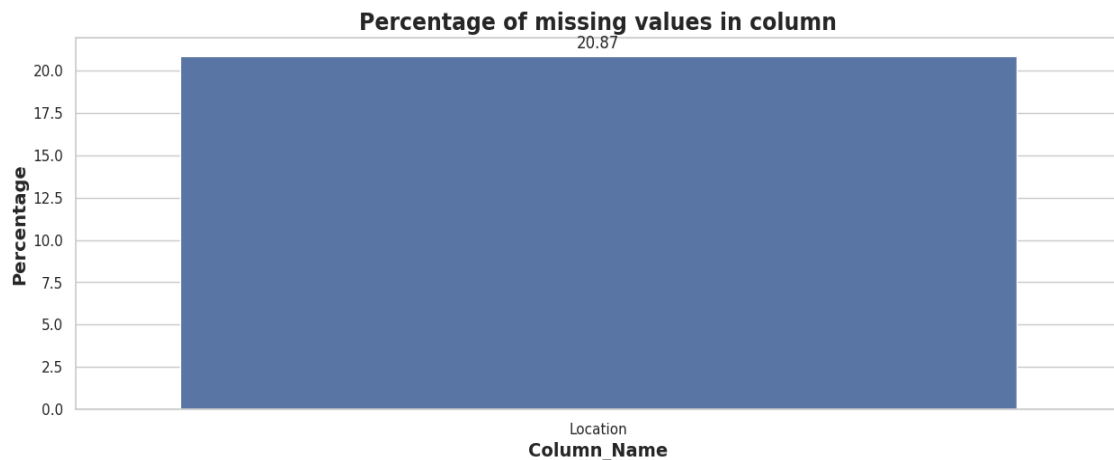
0.2 Result of Sentiment



0.3 Graph of Sentiment

4.4.2 Percentage of missing values in column

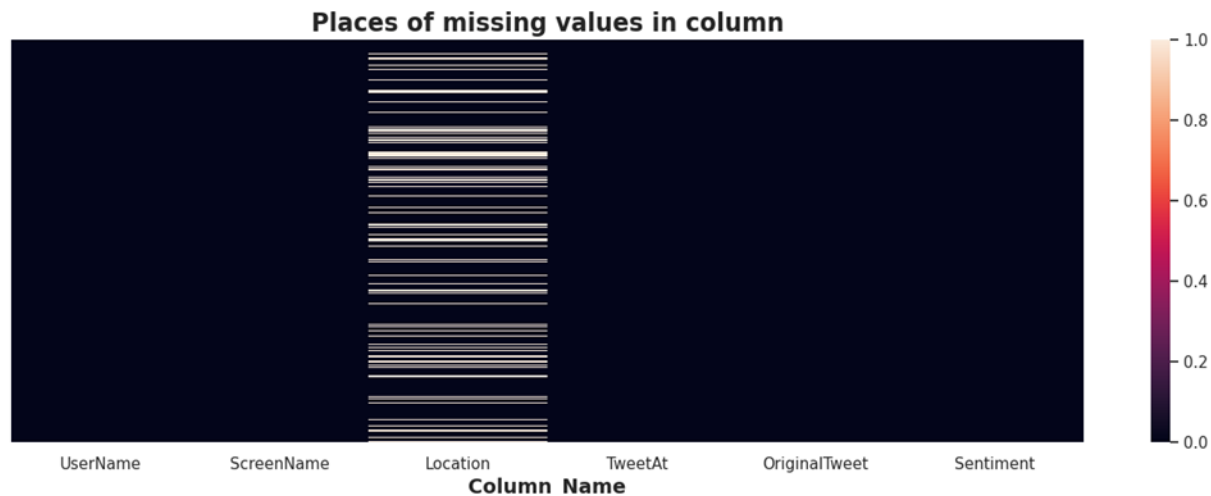
The provided code snippet is a Python script using the matplotlib and seaborn libraries to create a bar plot that visualizes the percentage of missing values in different columns of a dataset. The script generates a bar graph to visualize the percentage of missing values in different columns of the dataset. It starts with setting the character size and drawing aesthetic using the seaborn styling options. It then creates a bar chart from the missing values data frame that contains the columns and their respective percentages of missing data. The chart is marked with exact percentages at the top of each column for clarity. Finally, the axes are labeled and a title is added to provide context before the graph is displayed to the user. This visualization helps in quickly identifying columns with a high proportion of missing data, which is crucial for data cleaning and preprocessing steps.



0.4 Graph of percentage of missing Values in Column

4.4.3 Percentage of missing values in column

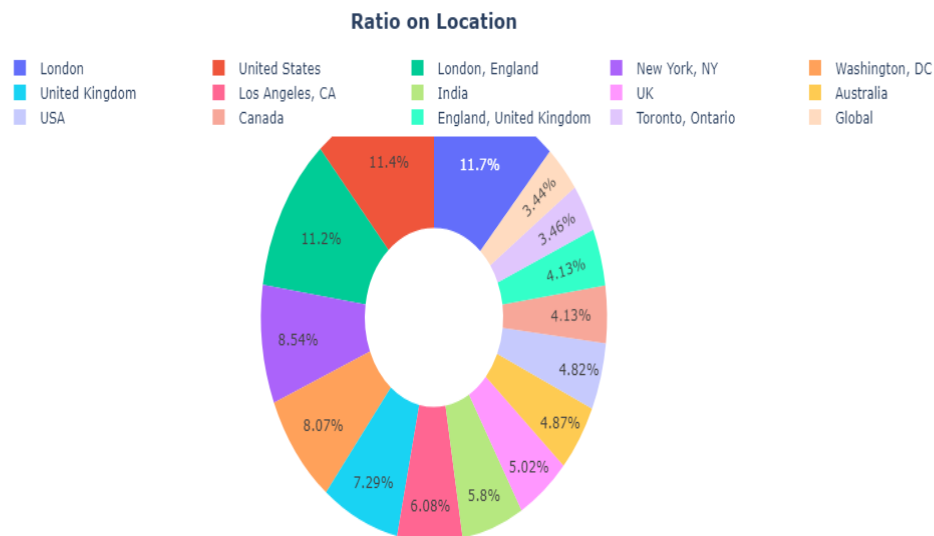
A heat map to visualize the locations of missing values in the data frame. It starts by setting the figure size to 17 by 5 inches for a clear and wide view. A heat map is created using Seaborn, where missing values are represented by one color (usually lighter) and non-missing values by another (usually darker). A color bar is included to indicate what each color represents. The y-axis labels are removed for simplicity, while the x-axis is labeled "Location" to identify the column. A title is added to give context to the visualization. Finally, the plot is displayed to the user. This visualization is useful for quickly identifying patterns or concentrations of missing data across a dataset.



0.5 Missing Values Columns Heat map

4.4.4 Percentage of missing values in column

The `loc_analysis` creates a donut chart using `plot` to visualize the distribution of location counts from a Data Frame. It first prepares the data by selecting the top 15 locations and their counts. The chart is arranged to show labels, percentages and names on hover with a hole in the center to create a donut chart. Layout specifies the title of the chart and places the legend horizontally above the chart. Finally, the script creates a figure object with the prepared data and layout, centers the title,

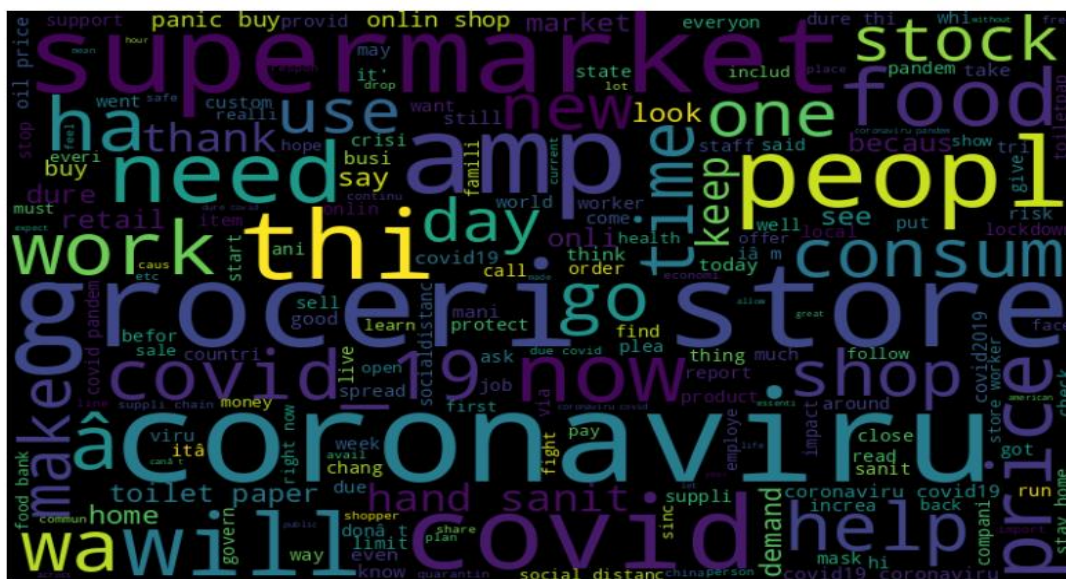


0.6 Piechart of Top 15 Locations

and displays the chart. This visualization helps to understand the relative proportions of the top 15 positions in the dataset

4.4.5 Word Cloud Graph

The script creates and displays a cluster of words from the tweet data contained in the "Tweet" column of the data Frame. It starts by concatenating all the tweets into a single chain. It then generates a cluster of words using the Word Cloud library, specifying dimensions, randomness for reproducibility, and maximum font size. Finally, display the word cloud using matplotlib, turning off the axis labels for a clean presentation. This visualization helps identify the most common



0.7 Word Cloud Graph of tweet Columns

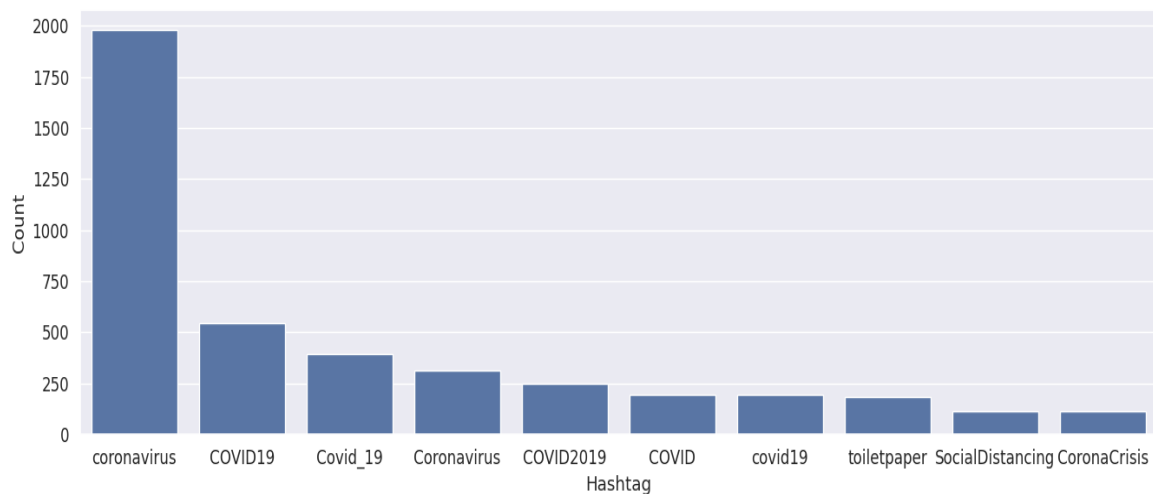
words in tweets and provides insight into common themes and topics being discussed. Different Word cloud graph are used in this project.

4.4.5 Understanding the impact of Hashtags on tweets sentiment

Collects tweets from Twitter using different hashtags and processes them for analysis. By specifying various hashtags, the script gathers tweets related to different topics or themes, enabling a comprehensive exploration of discussions on Twitter.

4.4.5.1 Extremely Positive

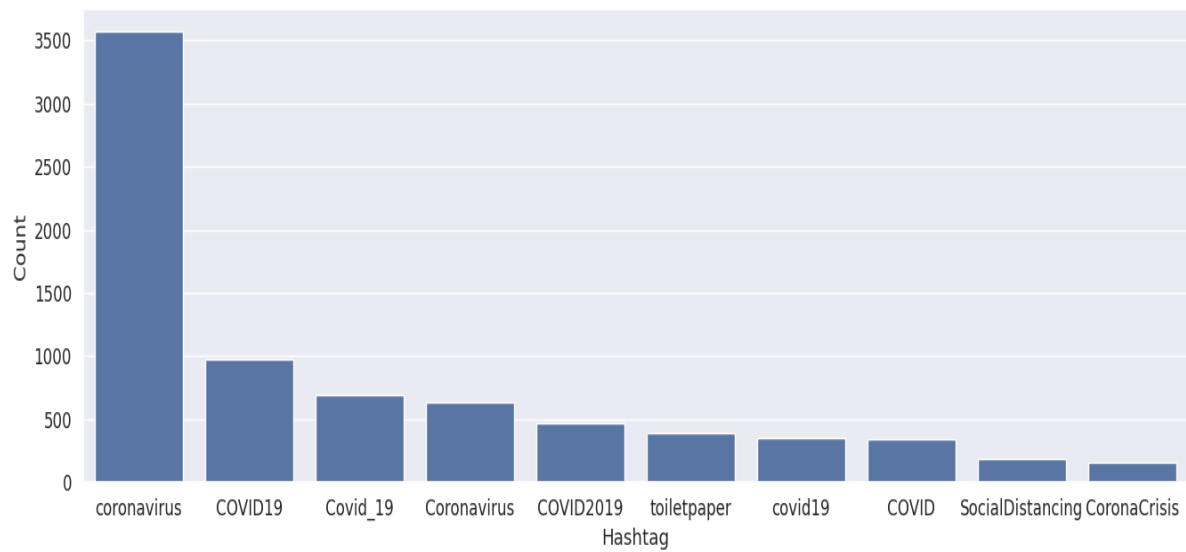
In this a bar graph showing the 10 most frequently occurring hashtags in tweets labeled "Extremely Positive". It uses the Natural Language Toolkit (NLTK) to calculate the frequency distribution of hashtags and pandas to organize the data for visualization. The resulting bar graph offers a look at the most shared hashtags associated with highly positive sentiment on Twitter.



0.8 Extremely Positive Graph according to hashtags

4.4.5.2 Positive

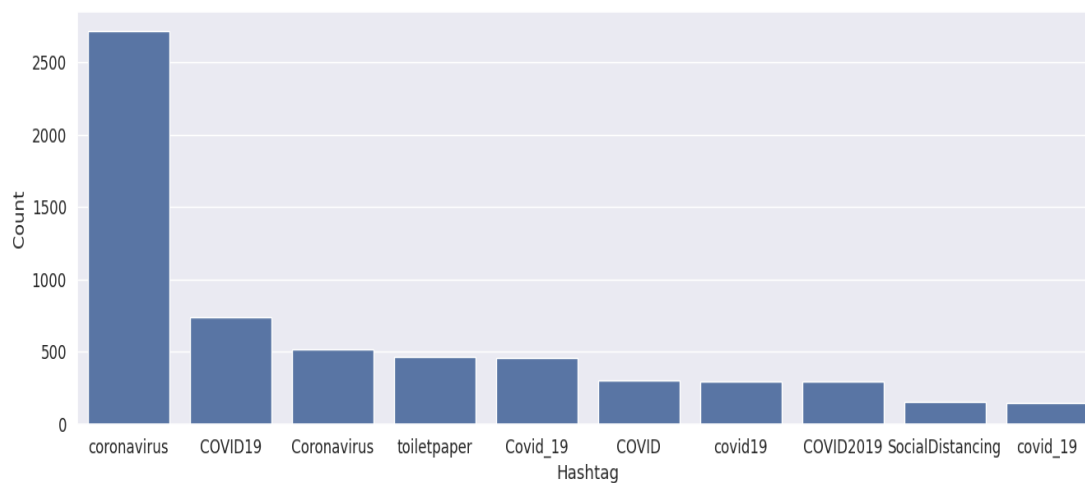
In this a bar graph showing the 10 most frequently occurring hashtags in tweets labeled "Positive". It uses the Natural Language Toolkit (NLTK) to calculate the frequency distribution of hashtags and pandas to organize the data for visualization. The resulting bar graph offers a look at the most shared hashtags associated with highly positive sentiment on Twitter.



0.9 Extremely Positive Graph according to hashtags

4.4.5.3 Neutral

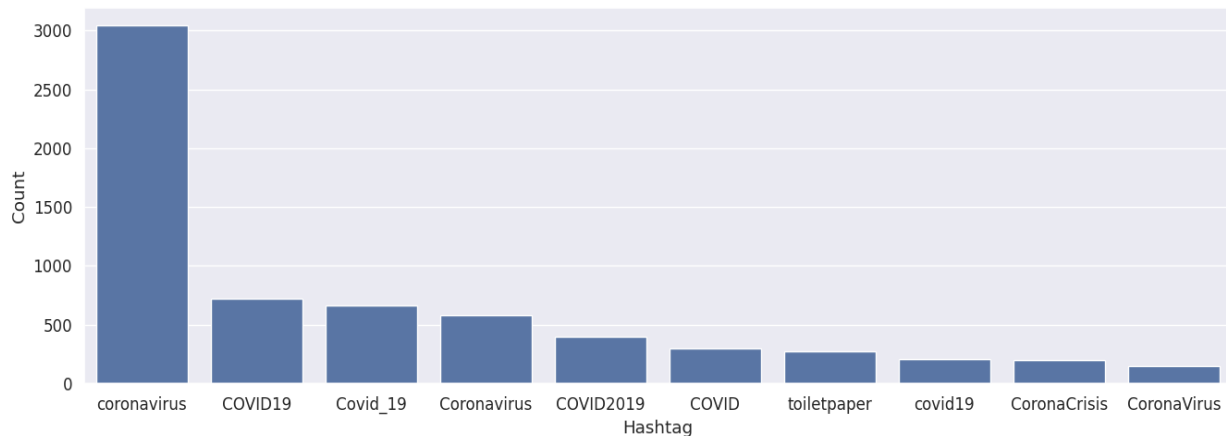
In this a bar graph showing the 10 most frequently occurring hashtags in tweets labeled "Neutral". It uses the Natural Language Toolkit (NLTK) to calculate the frequency distribution of hashtags and pandas to organize the data for visualization. The resulting bar graph offers a look at the most shared hashtags associated with highly positive sentiment on Twitter.



0.10 Neutral Graph according to hashtags

4.4.5.4 Negative

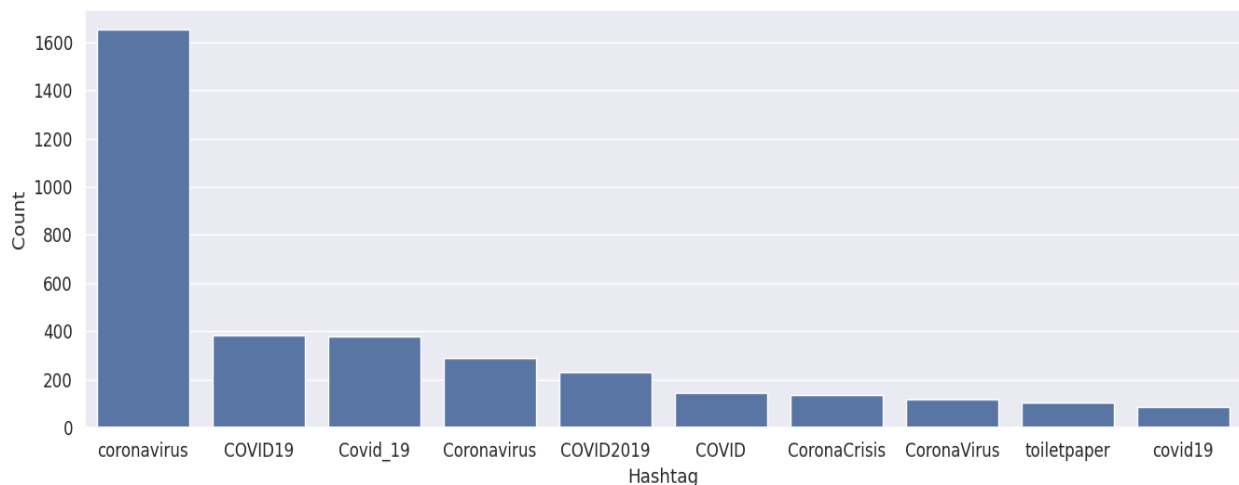
In this a bar graph showing the 10 most frequently occurring hashtags in tweets labeled "Negative". It uses the Natural Language Toolkit (NLTK) to calculate the frequency distribution of hashtags and pandas to organize the data for visualization. The resulting bar graph offers a look at the most shared hashtags associated with highly positive sentiment on Twitter.



0.11 Negative Graph according to hashtags

4.4.5.5 Extremely Negative

In this a bar graph showing the 10 most frequently occurring hashtags in tweets labeled "Extremely Negative". It uses the Natural Language Toolkit (NLTK) to calculate the frequency distribution of hashtags and pandas to organize the data for visualization. The resulting bar graph offers a look at the most shared hashtags associated with highly positive sentiment on Twitter.



0.12 Extremely Negative Graph according to hashtags

4.4.6 Spitting Our Dataset into Training and Testing Dataset

The dataset is split into training and test sets using the ``train_test_split`` function with a ratio of 80:20, where 80% of the data is allocated for training and 20% for testing. By reserving a portion of the data for testing, the model's performance can be assessed on unseen data to help measure its generalization capabilities. In addition, the ``stratify`` parameter ensures that the distribution of sentiment labels remains consistent across both sets and preserves the representation of different classes in both the training and test subsets. This approach is critical to maintaining the integrity of the dataset and building reliable machine learning models.

Table 1 Spitting Our Dataset into Training and Testing Dataset

Train shape	32925, 2
Testing shape	8232, 2

4.4.7 Use of Counter Vectorizer for Multi Class Classification

Table 2 Use of Counter Vectorizer for Multi Class Classification

<code>X_train.shape</code>	(32925, 35984)
<code>X_train.shape</code>	(8232, 35984)
<code>y_train.shape</code>	(32925,)
<code>y_valid.shape</code>	(8232,)

4.4.8 Machine Learning Algorithm

4.4.8.1 Naive Bayes Classifier for MULTICLASS Classification

Training Accuracy Score: 0.7187

Validation Accuracy Score: 0.4684

Precision: Accuracy is a measure of the accuracy provided by the model. Indicates the ratio of correctly predicted instances of a class to the total number of instances predicted as that class.

Recall: Recall, also known as sensitivity, measures the ability of a model to find all relevant cases within a data set. It is the ratio of correctly predicted instances of a class to the actual instances of that class in the data.

F1 Score: The F1 score is the harmonic mean of precision and recall. It provides a balance between accuracy and memorability.

Support: The number of actual occurrences of the class in the specified data set.

```

training accuracy Score      : 0.718663629460896
Validation accuracy Score    : 0.4684159378036929
      precision      recall  f1-score      support

Extremely Negative          0.33          0.58          0.42          614
Extremely Positive          0.39          0.59          0.47          870
      Negative            0.53          0.42          0.47         2487
      Neutral            0.31          0.68          0.42          697
      Positive            0.64          0.41          0.50         3564

      accuracy                                0.47         8232
      macro avg            0.44          0.54          0.46         8232
      weighted avg         0.53          0.47          0.47         8232

```

0.13 Result of Naive Bayes Classifier for MULTICLASS Classification

Extremely Negative:

Table 3 Extremely Negative of Naive Bayes Classifier for MULTICLASS Classification.

Precision	0.33
Recall	0.58
F1-score	0.42
Support	614

Extremely Positive:

Table 4 Extremely Positive of Naive Bayes Classifier for MULTICLASS Classification.

Precision	0.39
Recall	0.59
F1-score	0.47
Support	870

Negative:

Table 5 Negative of Naive Bayes Classifier for MULTICLASS Classification.

Precision	0.53
Recall	0.42
F1-score	0.47
Support	2487

Neutral:

Table 6 Neutral of Naive Bayes Classifier for MULTICLASS Classification

Precision	0.31
Recall	0.68
F1-score	0.42
Support	697

Positive:

Table 7 Positive of Naive Bayes Classifier for MULTICLASS Classification.

Precision	0.64
Recall	0.41
F1-score	0.50
Support	3564

Accuracy: The proportion of correctly classified instances.

Macro Avg: Average of the metrics (precision, recall, F1-score) for each class without considering class imbalance.

Weighted Avg: Average of the metrics, weighted by the support values for each class. It provides an overall score while considering class imbalance.

4.4.8.2 Stochastic Gradient Descent-SGD Classifier(MULTICLASS CLASSIFICATION)

Training Accuracy Score: 0.8568

Validation Accuracy Score: 0.5548

Precision: Indicates the proportion of true positive predictions out of all positive predictions made by the model.

Recall: Reflects the proportion of true positive predictions out of all true positive instances in the dataset.

F1-score: Harmonic average of precision and recall, providing a balance between the two metrics.

Support: Number of actual occurrences of each class in the dataset.

```

Training accuracy Score    : 0.856826119969628
Validation accuracy Score  : 0.5547862001943634
              precision      recall  f1-score   support

Extremely Negative        0.65      0.61      0.63      1169
Extremely Positive        0.68      0.62      0.65      1446
      Negative            0.42      0.49      0.45      1695
      Neutral            0.76      0.56      0.65      2089
      Positive            0.41      0.52      0.46      1833

      accuracy
macro avg          0.59      0.56      0.57      8232
weighted avg       0.58      0.55      0.56      8232

```

0.14 Result of Stochastic Gradient Descent-SGD Classifier

Extremely Negative:

Table 8 Extremely Negative of SGD

Precision	0.65
Recall	0.61
F1-score	0.63
Support	1169

Extremely Positive:

Table 9 Extremely Positive of SGD

Precision	0.68
Recall	0.62
F1-score	0.65
Support	1446

Negative:

Table 10 Negative of SGD

Precision	0.42
Recall	0.49
F1-score	0.45
Support	1695

Neutral:

Table 11 Neutral of SGD Classifier

Precision	0.76
Recall	0.56
F1-score	0.65
Support	2089

Positive:

Table 12 Extremely Positive of SGD

Precision	0.41
Recall	0.52
F1-score	0.46
Support	1833

Accuracy: 0.5548, indicating the proportion of correctly classified instances.

Macro Avg: Average of the metrics (precision, recall, F1-score) for each class without considering class imbalance.

Weighted Avg: Average of the metrics, weighted by the support values for each class. It provides an overall score while considering class imbalance.

4.4.8.3 RANDOM FOREST CLASSIFIER (For Multiclass Classification)

Training accuracy Score: 0.9962642369020501

Validation accuracy Score: 0.5444606413994169

```
Training accuracy Score    : 0.9962642369020501
Validation accuracy Score  : 0.5444606413994169
      precision    recall  f1-score   support

Extremely Negative      0.34      0.69      0.46       547
Extremely Positive      0.33      0.71      0.45       624
      Negative         0.53      0.50      0.51      2107
      Neutral          0.77      0.58      0.66      2040
      Positive         0.62      0.49      0.55      2914

      accuracy                    0.54      8232
      macro avg              0.52      0.59      0.53      8232
      weighted avg           0.60      0.54      0.55      8232
```

0.15 Result of RANDOM FOREST CLASSIFIER

Extremely Negative:

Table 13 Extremely Negative of RANDOM FOREST CLASSIFIER

Precision	0.34
Recall	0.69
F1-score	0.46
Support	547

Extremely Positive:

Table 14 Extremely Positive of RANDOM FOREST CLASSIFIER

Precision	0.33
Recall	0.71
F1-score	0.45
Support	624

Negative:

Table 15 Negative of RANDOM FOREST CLASSIFIER

Precision	0.53
Recall	0.50
F1-score	0.51
Support	2107

Neutral:

Table 16 Neutral of RANDOM FOREST CLASSIFIER

Precision	0.77
Recall	0.58
F1-score	0.66
Support	2040

Positive:

Table 17 Positive of RANDOM FOREST CLASSIFIER

Precision	0.62
Recall	0.49
F1-score	0.55
Support	2914

Accuracy: 0.54, indicating the proportion of correctly classified instances.

Macro Avg: Average of the metrics (precision, recall, F1-score) for each class without considering class imbalance.

Weighted Avg: Average of the metrics, weighted by the support values for each class. It provides an overall score while considering class imbalance.

4.4.8.4 Extreme Gradient Boosting (For Multiclass Classification)

Training accuracy Score: 0.696826119969628

Validation accuracy Score: 0.5738581146744413

```
Training accuracy Score      : 0.696826119969628
Validation accuracy Score : 0.5738581146744413
      precision    recall  f1-score   support

     0           0.64       0.49       0.56         1096
     1           0.71       0.54       0.61         1325
     2           0.54       0.45       0.49         1983
     3           0.55       0.79       0.65         1543
     4           0.54       0.59       0.56         2285

 accuracy
macro avg           0.60       0.57       0.58         8232
weighted avg        0.58       0.57       0.57         8232
```

0.16 Result of Extreme Gradient Boosting

Class0:

Table 18 Class 0 of Extreme Gradient Boosting

Precision	0.64
Recall	0.49
F1-score	0.56
Support	1096

Class 1:

Table 19 Class 1 of Extreme Gradient Boosting

Precision	0.71
Recall	0.54
F1-score	0.61
Support	1325

Class 2:

Table 20 Class 2 of Extreme Gradient Boosting

Precision	0.54
Recall	0.45
F1-score	0.49
Support	1983

Class 3:

Table 21 Class 3 of Extreme Gradient Boosting

Precision	0.55
Recall	0.79
F1-score	0.65
Support	1543

Class 4:

Table 22 Class 4 of Extreme Gradient Boosting

Precision	0.54
Recall	0.59
F1-score	0.56
Support	2285

Accuracy: 0.57, indicating the proportion of correctly classified instances.

Macro Avg: Average of the metrics (precision, recall, F1-score) for each class without considering class imbalance.

Weighted Avg: Average of the metrics, weighted by the support values for each class. It provides an overall score while considering class imbalance.

4.4.8.5 Support vector machine (For Multiclass Classification)

Training accuracy Score: 0.9044191343963554

Validation accuracy Score: 0.5947521865889213

```
Training accuracy Score    : 0.9044191343963554
Validation accuracy Score  : 0.5947521865889213
      precision    recall  f1-score   support

Extremely Negative      0.45      0.72      0.56        684
Extremely Positive      0.52      0.79      0.62        866
      Negative          0.56      0.53      0.54       2117
      Neutral           0.71      0.62      0.66       1750
      Positive          0.66      0.54      0.59       2815

      accuracy                    0.59       8232
      macro avg              0.58      0.64      0.60       8232
      weighted avg           0.61      0.59      0.60       8232
```

0.17 Result of Support vector machine

Extremely Negative:

Table 23 extremely negative of Support vector machine

Precision	0.45
Recall	0.72
F1-score	0.56
Support	684

Extremely Positive:*Table 24 extremely positive of Support vector machine*

Precision	0.52
Recall	0.79
F1-score	0.62
Support	866

Negative:*Table 25 Negative of Support vector machine*

Precision	0.56
Recall	0.53
F1-score	0.54
Support	2117

Neutral:*Table 26 Neutral of Support vector machine*

Precision	0.71
Recall	0.62
F1-score	0.66
Support	1750

Positive:*Table 27 positive of Support vector machine*

Precision	0.66
Recall	0.54
F1-score	0.59
Support	2815

4.4.8.6 Logistic Regression (For Multiclass Classification)**Training accuracy Score:** 0.9151404707668944**Validation accuracy Score** 0.6010689990281827

```

Training accuracy Score    : 0.9151404707668944
Validation accuracy Score  : 0.6010689990281827
      precision    recall  f1-score   support

Extremely Negative      0.61      0.66      0.63      1009
Extremely Positive      0.61      0.70      0.65      1166
      Negative         0.52      0.54      0.53      1910
      Neutral          0.71      0.62      0.66      1774
      Positive         0.58      0.56      0.57      2373

      accuracy
macro avg      0.61      0.62      0.61      8232
weighted avg   0.60      0.60      0.60      8232

```

*0.18 Result of Logistic Regression***Extremely Negative:***Table 28 Extremely negative of Logistic Regression*

Precision	0.61
Recall	0.66
F1-score	0.63
Support	1009

Extremely Positive:*Table 29 Extremely Positive of Logistic Regression*

Precision	0.61
Recall	0.70
F1-score	0.65
Support	1166

Negative:

Table 30 Negative of Logistic Regression

Precision	0.52
Recall	0.54
F1-score	0.53
Support	1910

Neutral:

Table 31 Neutral of Logistic Regression

Precision	0.71
Recall	0.62
F1-score	0.66
Support	1774

Positive:

Table 32 Positive of Logistic Regression

Precision	0.58
Recall	0.56
F1-score	0.57
Support	2373

4.4.8.7 CATBOOST MODEL (For Multiclass Classification)

Training accuracy Score: 0.6570387243735764

Validation accuracy Score: 0.609572400388727

```
Training accuracy Score    : 0.6570387243735764
Validation accuracy Score  : 0.609572400388727
      precision    recall  f1-score   support

Extremely Negative      0.52      0.70      0.60       820
Extremely Positive      0.55      0.76      0.64       971
      Negative          0.51      0.57      0.54      1784
      Neutral           0.81      0.59      0.68      2118
      Positive          0.63      0.57      0.60      2539

      accuracy
macro avg      0.61      0.64      0.61      8232
weighted avg   0.63      0.61      0.61      8232
```

0.19 Result of CATBOOST MODEL

Extremely Negative:*Table 33 Extremely Negative of Catboost model*

Precision	0.52
Recall	0.70
F1-score	0.60
Support	820

Extremely Positive:*Table 34 Extremely Positive of Catboost model*

Precision	0.55
Recall	0.76
F1-score	0.64
Support	971

Negative:*Table 35 Negative of Catboost model*

Precision	0.51
Recall	0.57
F1-score	0.54
Support	1784

Neutral:*Table 36 Neutral of Catboost model*

Precision	0.81
Recall	0.59
F1-score	0.68
Support	2118

Positive:*Table 37 Positive of Catboost model*

Precision	0.63
Recall	0.57
F1-score	0.60
Support	2539

4.4.8.8 All the multiclass models test accuracy in descending order:

CatBoost: 0.6096

Logistic Regression: 0.6011

Support Vector Machines: 0.5948

XGBoost: 0.5739

Stochastic Gradient Descent: 0.5548

Random Forest: 0.5445

Naive Bayes: 0.4684

These accuracies represent the performance of each model on the test dataset.

	Model	Test accuracy
6	CatBoost	0.609572
1	Logistic Regression	0.601069
0	Support Vector Machines	0.594752
5	XGBoost	0.573858
4	Stochastic Gradient Decent	0.554786
2	Random Forest	0.544461
3	Naive Bayes	0.468416

0.20All the multiclass models test accuracy in descending order:

4.4.9 Machine Learning Algorithm for Binary Classification

4.4.9.1 NAIVE BAYES CLASSIFIER FOR BINARY CLASSIFICATION.

training accuracy Score: 0.8706454062262718

Validation accuracy Score: 0.7879008746355685

```

training accuracy Score      : 0.8706454062262718
Validation accuracy Score : 0.7879008746355685
      precision    recall  f1-score   support

     0       0.68       0.74       0.70       2836
     1       0.85       0.82       0.83       5396

 accuracy
macro avg       0.77       0.78       0.77       8232
weighted avg       0.79       0.79       0.79       8232

```

0.21 Result of NAIVE BAYES

Class 0:

Table 38 class 0 of NAIVE BAYES

Precision	0.68
Recall	0.74
F1-score	0.70
Support	2836

Class 1:

Table 39 class 1 of NAIVE BAYES

Precision	0.85
Recall	0.82
F1-score	0.83
Support	5396

4.4.9.2 RANDOM FOREST CLASSIFIER FOR BINARY CLASSIFICATION

Training accuracy Score: 0.9987243735763098

Validation accuracy Score: 0.8196064139941691

```
Training accuracy Score      : 0.9987243735763098
Validation accuracy Score    : 0.8196064139941691
      precision      recall  f1-score      support

      0      0.67      0.82      0.73      2515
      1      0.91      0.82      0.86      5717

    accuracy                    0.82      8232
  macro avg      0.79      0.82      0.80      8232
weighted avg      0.84      0.82      0.82      8232
```

0.22 Result of RANDOM FOREST CLASSIFIER

Class 0:

Table 40 class 0 of RANDOM FOREST

Precision	0.67
Recall	0.82
F1-score	0.73
Support	2515

Class 1:

Table 41 class 1 of RANDOM FOREST

Precision	0.91
Recall	0.82
F1-score	0.86
Support	5717

4.4.9.3 LOGISTIC REGRESSION (BINARY CLASSIFICATION)

Training accuracy Score: 0.9490356871678056

Validation accuracy Score: 0.8544703595724004

Training accuracy Score : 0.9490356871678056					
Validation accuracy Score : 0.8544703595724004					
	precision	recall	f1-score	support	
0	0.76	0.84	0.80	2776	
1	0.91	0.86	0.89	5456	
accuracy			0.85	8232	
macro avg	0.83	0.85	0.84	8232	
weighted avg	0.86	0.85	0.86	8232	

0.23 Result of LOGISTIC REGRESSION

Class 0:

Table 42 class 0 of LOGISTIC REGRESSION

Precision	0.76
Recall	0.84
F1-score	0.80
Support	2776

Class 1:

Table 43 class 1 of LOGISTIC REGRESSION

Precision	0.91
Recall	0.86
F1-score	0.89
Support	5456

4.4.9.4 CATBOOST ALGORITHM FOR BINARY CLASSIFICATION

Training accuracy Score: 0.8777828397873956

Validation accuracy Score: 0.8435374149659864

```
Training accuracy Score      : 0.8777828397873956
Validation accuracy Score    : 0.8435374149659864
      precision      recall  f1-score      support
      0      0.70      0.86      0.77      2520
      1      0.93      0.84      0.88      5712

      accuracy
macro avg      0.81      0.85      0.83      8232
weighted avg   0.86      0.84      0.85      8232
```

0.24Result of CATBOOST ALGORITHM

Class 0:

Table 44 class 0 of CATBOOST ALGORITHM

Precision	0.70
Recall	0.86
F1-score	0.77
Support	2520

Class 1:

Table 45 Class 1 CatBoost Model

Precision	0.93
Recall	0.84
F1-score	0.88
Support	5712

4.4.9.5 XG BOOST(BINARY CLASSIFICATION)

Training accuracy Score: 0.8452847380410022

Validation accuracy Score: 0.814625850340136

```

Training accuracy Score      : 0.8452847380410022
Validation accuracy Score    : 0.814625850340136
      precision      recall  f1-score   support

      0         0.63         0.83         0.72        2342
      1         0.92         0.81         0.86        5890

   accuracy
  macro avg         0.78         0.82         0.79        8232
 weighted avg         0.84         0.81         0.82        8232

```

0.25 Result of XG BOOST

Class 0:

Table 46 class 0 of XG BOOST

Precision	0.63
Recall	0.83
F1-score	0.72
Support	2342

Class 1:

Table 47 class 1 of XG BOOST

Precision	0.92
Recall	0.81
F1-score	0.86
Support	5890

4.4.9.6 SUPPORT VECTOR MACHINE (BINARY CLASSIFICATION)

Training accuracy Score: 0.9576917236142749

Validation accuracy Score: 0.8367346938775511

```

Training accuracy Score      : 0.9576917236142749
Validation accuracy Score    : 0.8367346938775511
      precision      recall  f1-score   support

      0         0.67         0.86         0.76        2414
      1         0.93         0.83         0.88        5818

   accuracy
  macro avg         0.80         0.84         0.82        8232
 weighted avg         0.86         0.84         0.84        8232

```

0.26 Result of SVM

Class 0:*Table 48 class 0 of SVM*

Precision	0.67
Recall	0.86
F1-score	0.76
Support	2414

Class 1:*Table 49 class 1 of SVM*

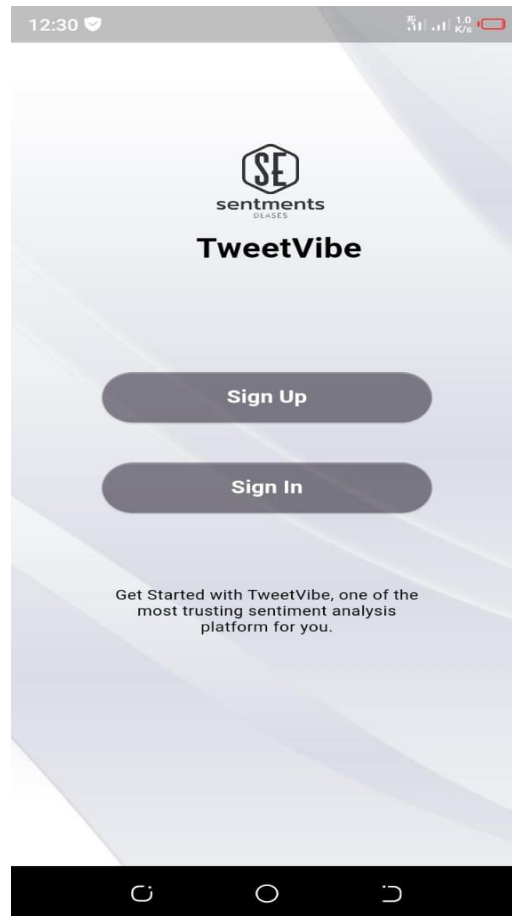
Precision	0.93
Recall	0.83
F1-score	0.88
Support	5818

4.4.10 Mobile Application for Sentiment

4.4.10.1 Start Screen

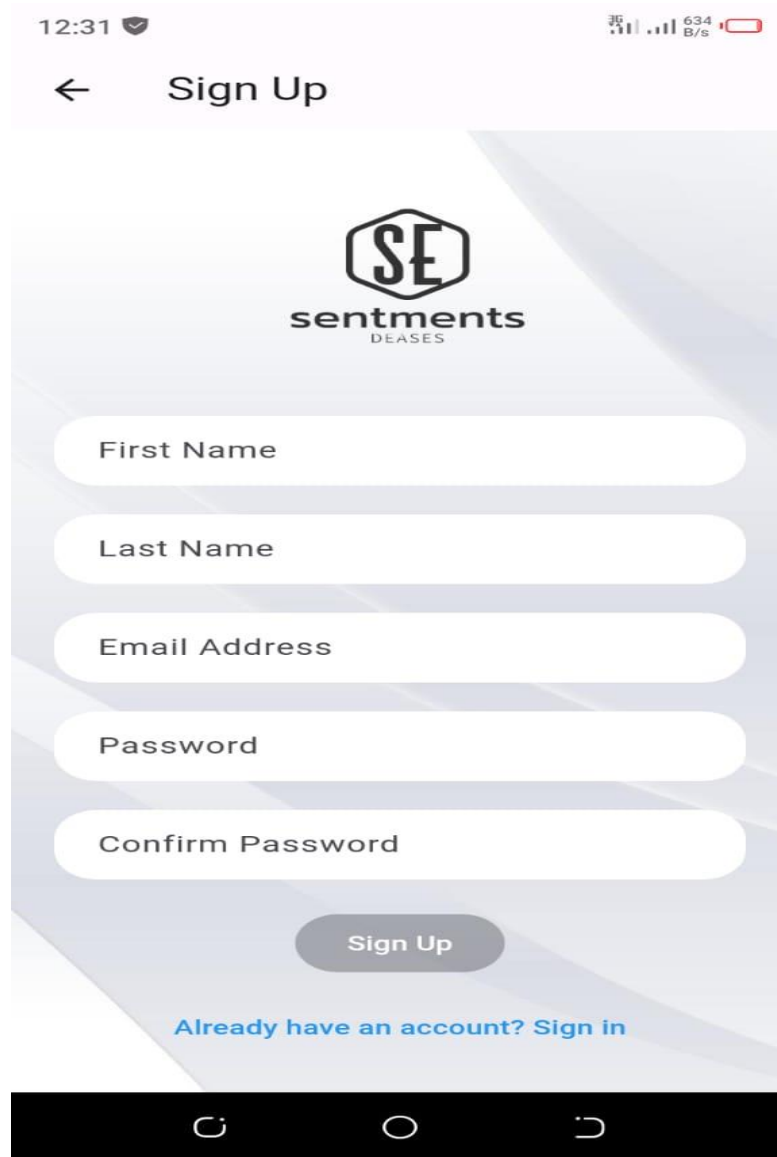
Sign Up: Create a new account to access all features.

Sign In: Log in to your existing account and continue where you left off.



0.27 Start Screen

4.4.10.2 Sign Up Screen



The image shows a mobile application sign-up screen. At the top, the status bar displays the time 12:31, a shield icon, signal strength bars, the text '634 B/s', and a battery icon. Below the status bar is a header with a back arrow and the text 'Sign Up'. The main content area features the SE sentiments DEASES logo, which consists of a hexagon containing 'SE' above the word 'sentiments' and 'DEASES' in smaller text. Below the logo are five rounded rectangular input fields labeled 'First Name', 'Last Name', 'Email Address', 'Password', and 'Confirm Password'. A dark grey 'Sign Up' button is positioned below the input fields. At the bottom of the form area, there is a link that reads 'Already have an account? Sign in'. The entire screen is set against a light blue background with abstract geometric shapes. At the very bottom, there is a black navigation bar with three white icons: a home button, a circle, and a back button.

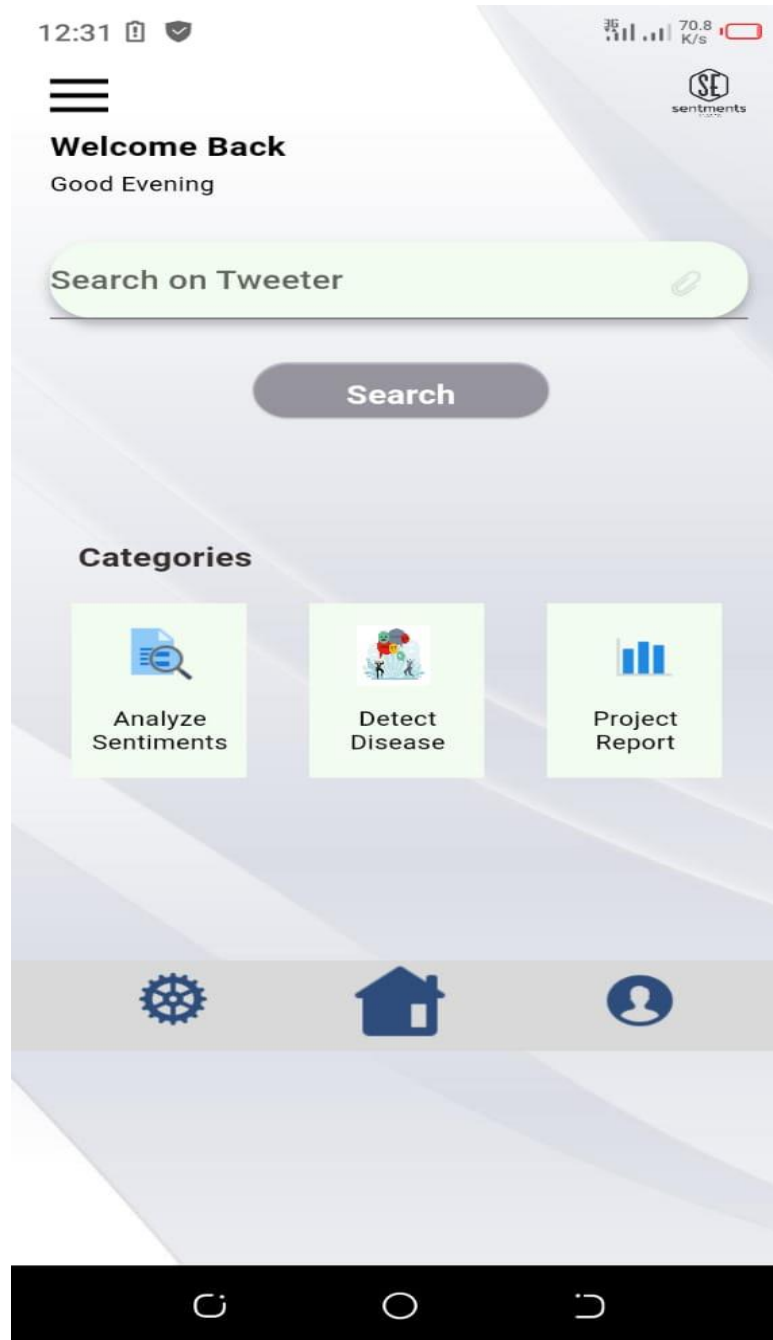
0.28 SignUp Screen

4.4.10.3 Login Screen



0.29 Login Screen

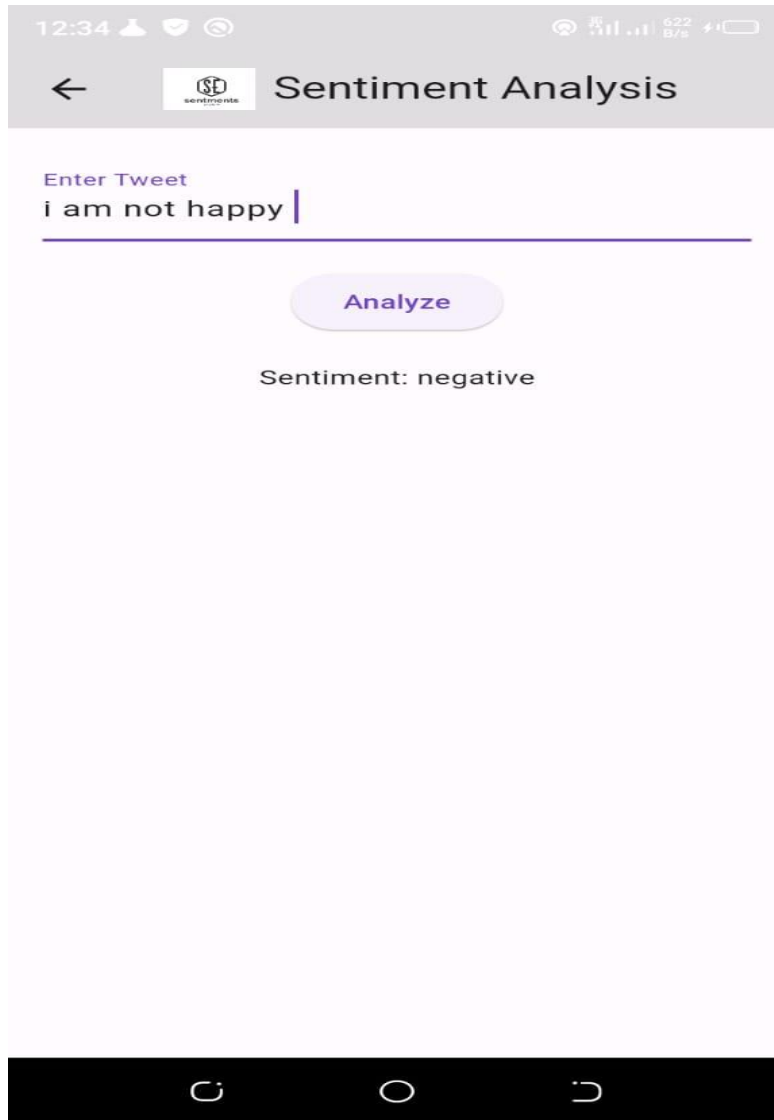
4.4.10.4 Start Screen



0.30 Start Screen

4.4.10.5 Sentiment Result

Analyze a tweet for sentiment: Enter a tweet to see if it's positive, negative, or neutral. Get instant feedback and insights into the sentiment behind the words.



0.31 Sentiment Result

Chapter 5

Conclusion & Future Work

5.1 Discussion and Conclusion

5.1.1 Discussion

Applying sentiment analysis using machine learning and natural language processing (NLP) to predict disease outbreaks represents a cutting-edge approach in public health surveillance. This method uses the vast amount of text data generated daily on social media platforms, news sites and other digital communication channels. By capturing and analyzing the sentiments of this data, we can identify public anxiety and fear related to health issues that often precede official reports of disease outbreaks.

One of the key strengths of this approach lies in its ability to provide real-time information. Traditional epidemiological methods, although reliable, often suffer from delays due to the time required for data collection, processing and analysis. Sentiment analysis, on the other hand, can quickly detect changes in public sentiment and potentially serve as an early warning system for emerging health threats. For example, a sudden increase in tweets expressing concern about flu symptoms in a certain area could prompt health authorities to investigate further and take preventive measures.

Despite its potential, several challenges and limitations must be considered. The accuracy of sentiment analysis models can be affected by the quality and representativeness of the data. For example, social media data may not be representative of the entire population as it tends to be biased towards younger demographics and those with internet access. Furthermore, the interpretation of sentiment in textual data is inherently complex due to nuances such as sarcasm, idioms, and cultural differences. Improving NLP algorithms to better understand these subtleties is an ongoing area of research.

Another critical aspect is the integration of sentiment analysis with other data sources. Combining sentiment data with epidemiological data, mobility models and environmental factors can provide a more comprehensive view of disease dynamics. This multimodal approach can improve the robustness of predictions and help mitigate false alarms.

In summary, while sentiment analysis using machine learning and NLP offers a promising tool for disease outbreak prediction, it is not without challenges. Future work should focus on improving model accuracy, expanding data sources, and addressing ethical issues to fully exploit the potential of this innovative approach. Collaboration between technologists, public health experts, and policy

makers will be essential to effectively integrate these tools into public health strategies and improve global health outcomes.

5.1.2 Conclusion

Sentiment analysis using machine learning and natural language processing (NLP) has shown significant promise in predicting disease outbreaks. By analyzing social media posts, newspaper articles, and other sources of textual data, we can detect early signs of public concern and emerging health issues. Machine learning models, especially those using advanced NLP techniques, can effectively classify and interpret sentiments expressed in large amounts of unstructured text. These insights can complement traditional epidemiological methods and offer a more timely and detailed understanding of potential disease outbreaks. Integrating sentiment analysis into public health surveillance systems can improve the ability to respond quickly to emerging threats and potentially mitigate the spread of disease.

5.2 Future Work

Future work should focus on improving the accuracy and reliability of sentiment analysis models in the context of disease outbreak forecasting. This includes the development of more sophisticated NLP algorithms that can better handle context, sarcasm and multilingual data. Additionally, integrating sentiment analysis with other data sources such as search engine queries, mobility data, and health records could provide a more comprehensive view of disease dynamics. Expanding the geographic and demographic scope of data collection will also be key to building models that can be generalized across different populations and regions. Collaboration between data scientists, epidemiologists, and public health officials will be necessary to validate these models and ensure their practical applicability in real-world scenarios. By advancing in these areas, we can improve the predictive capabilities of sentiment analysis and contribute to more effective disease surveillance and response systems.

References

A wide range of web resources were used during the development of this project figure some of the references we used:

1. <https://doi.org/10.2196/22590>
2. <https://doi.org/10.1371/journal.pone.0231924>
3. <https://doi.org/10.4018/978-1-7998-6825-5.ch026>
4. <https://doi.org/10.1016/j.scitotenv.2020.138882>
5. Abd-Alrazaq A, Alhuwail D, Househ M, Hamdi M, Shah Z. Top concerns of tweeters during the COVID-19 pandemic: infoveillance study. *J Med Internet Res.* (2020) 22:e19016. doi: 10.2196/19016
6. Barkur G, Vibha GBK. Sentiment analysis of nationwide lockdown due to COVID 19 outbreak: evidence from India. *Asian J Psychiatr.* (2020) 51:102089. doi: 10.1016/j.ajp.2020.102089
7. Pedrosa AL, Bitencourt L, Fróes ACF, Cazumbá MLB, Campos RGB, de Brito SBCS, et al. Emotional, behavioral, and psychological impact of the COVID-19 pandemic. *Front Psychol.* (2020) 11:566212. doi: 10.3389/fpsyg.2020.566212
8. Fernandes N. Economic Effects of Coronavirus Outbreak (COVID-19) on the World Economy. Available at SSRN 3557504 (2020).
9. Abbasi A, Javed AR, Chakraborty C, Nebhen J, Zehra W, Jalil Z. ElStream: an ensemble learning approach for concept drift detection in dynamic social big data stream learning. *IEEE Access.* (2021) 9:66408–19. doi: 10.1109/ACCESS.2021.3076264
10. Samuel J, Ali G, Rahman M, Esawi E, Samuel Y, et al. Covid-19 public sentiment insights and machine learning for tweets classification. *Information.* (2020) 11:314. doi: 10.3390/info11060314
11. Mittal A, Patidar S. Sentiment analysis on twitter data: a survey. In: *Proceedings of the 2019 7th International Conference on Computer and Communications Management.* Bangkok (2019). p. 91–5.
12. . M.Z. Asghar, A. Khan, F.M. Kundi, M. Qasim, F. Khan, R. Ullah, I.U. Nawaz, Medical sentiment analysis lexicon: an incremental model for mining health reviews. *Int. J. Acad. Res.* 6(1), 295–302 (2014)

13. H. Iyer, M. Gandhi, S. Nair, Sentiment analysis for visuals using natural language processing. *Int. J. Comput. Appl.* 128(6), 31–35 (2015)
14. M.T. Khan, S. Khalid, Sentiment analysis for health care, in *Big Data: Concepts, Methodologies, Tools, and Applications*, IGI Global (2016), pp. 676–689
15. M.M. Mostafa, N.R. Nebot, Sentiment analysis of spanish words of arabic origin related to islam: a social network analysis. *J. Lang. Teach. Res.* 8(6), 1041–1049 (2017)
16. A. Shoukry, A. Rafea, Sentence-level Arabic sentiment analysis, in *2012 International Conference on Collaboration Technologies and Systems (CTS)*, IEEE (2012, May), pp. 546–550
17. M. Korayem, D. Crandall, M. Abdul-Mageed, Subjectivity and sentiment analysis of arabic: a survey, in *International Conference on Advanced Machine Learning Technologies and Applications* (Springer, Heidelberg, 2012, December), pp. 128–139
18. Goodfellow I., Bengio Y., Courville A. *Deep Learning*. MIT Press; Cambridge, MA, USA: 2016. [Google Scholar]
19. Russell S., Norvig P. *Artificial Intelligence: A Modern Approach*. Prentice Hall; Hoboken, NJ, USA: 2002. [Google Scholar]
20. Wild C.P. The exposome: From concept to utility. *Int. J. Epidemiol.* 2012;41:24–32. doi: 10.1093/ije/dyr236. [PubMed] [CrossRef] [Google Scholar]