Sentiment & Emotion Analysis of COVID-19 Tweets

TEAM MEMBERS

Jay Ganatra	369790341
Akshay Shirahatti	341001392
Bhavik Panchal	412076858
Rahul Parande	904792531
Shruti Rao	785095034
Vanshika Patel	235329211
Manali Shah	221863830

Abstract

Using a variety of methods, including topic-specific modelling and inference, this project intends to explore the influence that the COVID-19 epidemic has had on the dialogue that can be found on Twitter. The main objective of this project is to conduct sentiment and emotional analysis on COVID-19 related tweets. By doing so, we aim to gain insights into the emotions and opinions expressed by individuals on social media regarding the pandemic. The project makes use of Natural Language Processing methods such as Latent Dirichlet Allocation (LDA) for modelling topics, sentiment evaluation for polarity and subjectivity classification, and emotion analysis with the VADER sentiment analyzer. In addition, the DistilRoBERTa-base model, which is a cutting-edge language model, is used for tasks involving the detection and categorization of emotions. The outcomes of the experiment give useful insights into how people have been discussing the COVID-19 epidemic on Twitter recently as well as the emotions that individuals exhibit in their tweets. These insights have the potential to assist in the development of successful communication strategies and treatments that are suited to the requirements of certain groups. The initiative emphasizes the relevance of studying data from social media in order to acquire a better understanding of the effect that the epidemic has on society.

Keywords: Processing natural language, Latent Dirichlet Allocation, topic modelling, sentiment analysis, emotion analysis, Twitter, COVID-19, pandemic, social media, and causal topic modelling are some of the keywords that are relevant to this discussion.

INTRODUCTION

Fig 1. Covid-19 Tweet

The COVID-19 epidemic has had a significant influence on society and has altered our daily lives in manners that haven't been witnessed before. One of the most major shifts has been the expansion of human interaction to the digital realm, in particular on social media websites such as Twitter. During this trying time, the many platforms for social media have developed into crucial information sources. These platforms link people from all over the globe and provide a way for individuals to express their views and ideas. As a result, the pandemic caused by the Coronavirus has been an extraordinary event that has shaken up the whole planet in every conceivable manner. People were able to communicate their emotions, opinions, and concerns about the epidemic via the use of social media as it expanded over the world. There has been a great deal of conversation around the COVID-19 epidemic on Twitter, which is one of the social media sites that is utilized by the most people. The epidemic caused Twitter to create a large amount of data, which provides academics with the opportunity to learn more about the sentiments and views of individuals and how they impact society.

In this project, we analyze a series of tweets concerning the COVID-19 epidemic to see how individuals are use Twitter to communicate their ideas and emotions at this challenging time. In order to analyze the content of the tweets, we performed a subject demonstration, opinion research, and emotion analysis using standard language handling algorithms such as LDA. Our study takes a comprehensive look at the ways in which users of Twitter are expressing their thoughts and sentiments in reaction to the epidemic.

The dataset that we analyzed included an example of one thousand real time tweets that were acquired via the use of the Twitter programming interface and had hashtags relevant to the Coronavirus, such as #COVID19 and #corona. After we had cleaned the material by removing stopwords, accentuation, and URLs, we put it through a variety of conventional language processing techniques.

To begin, we used a method known as LDA for topic modelling, which is an unsupervised kind of machine learning. This allowed us to extract themes from a set of papers. During the step of pre-processing, we tokenized the data, lemmatized it, and deleted any stopwords that were present. The LDA model was then applied to determine which subjects within the tweets were the most important. Coherence ratings were used to assess the appropriate number of subjects, and the most important words for each topic were presented so that readers could get an overall picture of the subject matter.

After that, the tweets were put through a sentiment analysis, which categorized them as favourable, neutral, and unfavourable based on the polarity ratings they received. We examined the sentiment of 500 tweets and found that 196 were positive, 184 were noncommittal, and 120 were critical. This investigation offers insight into the emotional effects that the epidemic has on those who use Twitter.

In the end, we carried out the emotion analysis by making use of the VADER sentiment analyzer in conjunction with a set of preset emotion categories such as fear, happiness, sadness, and neutral. Because each tweet had its own unique set of explicit watchwords, we were able to categorize the overarching mood that was conveyed by the tweet. According to the examination of the emotions included in 500 tweets, 43.6% of the tweets conveyed joy, 31.6% expressed sadness, 20.2% exhibited fear, and 4.6% were neutral. This investigation gives insight into the ways in which users of Twitter are expressing their feelings in relation to the epidemic.

The results of our project will have a few different repercussions. To begin, it offers a picture of the feelings and responses that users of Twitter are expressing in response to the epidemic. These responses may be found on Twitter. This information may be helpful in gaining an understanding of the emotional consequences that the epidemic has had on people and in the development of solutions that will assist individuals during these challenging times. Second, the findings of our study may provide those in charge of public health and those who determine policy with the information they need to produce tailored messages and interventions that meet the specific concerns and needs of distinct groups of individuals. In the last part of our analysis, we focus on the potential of ordinary language handling algorithms for analysing material found on Twitter and obtaining titbits of knowledge regarding the opinions and attitudes of people.

Because of advancements in digital technology and social media platforms like Twitter, it is now much simpler to disseminate information about pandemics and to communicate with one another. Nevertheless, using the Twitter Programming interface to gather information involves ethical and safety considerations that should be taken into consideration. In the course of our project, we took measures to guarantee that the gathering and examination of data occurred in an open and honest manner and that we adhered to the highest standards of data protection and confidentiality. The COVID-19 pandemic and Twitter debates may be connected in a causal fashion by applying approaches for causal topic modelling. These techniques, which can be used in upcoming research to expand on our results, can be used to build on our findings. Combining the techniques of topic modelling, which identifies the key subjects addressed in Twitter data, with causal inference, which determines the causal linkages between the topics and the epidemic, is one way to use this approach. Topic modelling identifies the primary issues discussed in Twitter data.

PROPOSED WORK

The goal of this project is to create a methodology for causal topic modelling that will enable us to uncover causal linkages between the COVID-19 epidemic and debates taking place on Twitter. Through an analysis of the most current developments in topic modelling and inference of causality methodologies, as well as previous research on epidemics and social media, our effort intends to get an understanding of the impact that the COVID-19 pandemic has had on the conversation that can be found on Twitter.

In order for us to reach our objective, we are going to carry out an exploratory analysis of data on a substantial quantity of Twitter data relating to pandemics with the goal to identify important themes and developing subjects. Next, our focus will be on building a framework for causal topic modeling that enables us to identify causal relationships between Twitter discussions and the COVID-19 pandemic.

In the solution section of our project, we have included the following components:

Collection of Data: Researchers will utilize Twitter APIs to collect data relating to the COVID-19 pandemic on Twitter from a range of sources, including individual users, news media, and public health authorities. This will be done in order to better understand the spread of the pandemic.

Pre-Processing of Data: Data Will Be Pre-processed In order to pre-process the information that Twitter provides, bespoke content and software apparatuses will be developed. This involves structuring and sanitizing the data, eliminating stuff that is irrelevant, and placing the data in a format that is acceptable for analysis in some fashion.

Analyzing the pre-processed data in order to find important themes and emerging issues is what's known as exploratory data analysis.

Topic Modelling: The process of determining which aspects of the COVID-19 tweet is the most significant by the use of Latent Dirichlet Allocation (LDA) modelling tools.

Casual Induction: Utilizing techniques of causal inference such as instrumental variable research and propensity score matching, it is possible to discover the linkages that exist between the COVID-19 epidemic and the themes that are discussed on Twitter. This helps us to recognize points that are connected to the epidemic and acquire insight into how the worldwide outbreak is shaping discussions on Twitter.

Interpretation and visualization: With the goal to make it easier to comprehend the results of the analysis and to disseminate those findings to a wider audience, visualization tools are going to be created. Researchers are able to find recurring themes in the data as well as social and cultural elements that impact pandemic-related Twitter talks as a result of this.

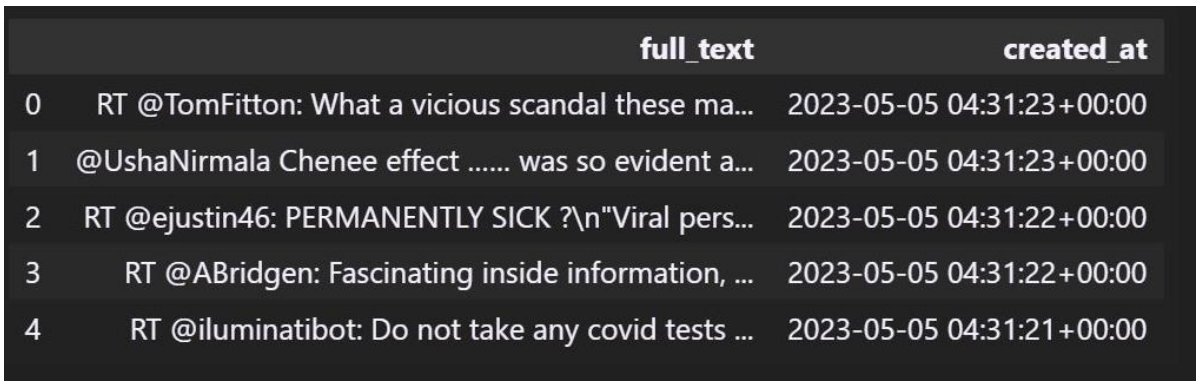
This project will assist scholars limit their study aims and obtain a greater knowledge of the challenges and possibilities connected with evaluating data from Twitter posts linked to pandemics. The insights that were gathered from this experiment may also be valuable in the development of tailored treatments to address the challenges that are associated with the COVID-19 pandemic.

A) Data Collection:

For the purpose of our project, we gathered information from Twitter via the use of the Tweepy Python library. Tweepy is a prominent and user-friendly software that gives users access to Twitter's application programming interface (API). We used the streaming API provided by Twitter using Tweepy in order to gather tweets in real time that were relevant to the COVID-19. In order to make use of Tweepy, we needed to begin by registering for an account as a developer with Twitter and acquiring API credentials. Because we have these credentials, we are able to authenticate ourselves and be authorized to use Twitter's API. As soon as we were in possession of our credentials, we made use of Tweepy's Cursor class to set up a stream that keeps an eye out for tweets that include certain keywords that are associated with the COVID-19 epidemic. For instance, we gathered tweets that included keywords such as "COVID-19," "coronavirus," "pandemic," and a variety of other related terms.

The data that we gathered was saved in JSON format, which comprises a variety of information about each tweet, including the content of the tweet, information about the person who sent it, location information, a date, and more. We stored the information locally so that we could do further analysis on them later.

Tweepy, in general, offered a hassle-free and effective method for collecting real-time data from Twitter, which is essential for comprehending the ever-changing nature of the public conversation about the COVID-19 epidemic.



	full_text	created_at
0	RT @TomFitton: What a vicious scandal these ma...	2023-05-05 04:31:23+00:00
1	@UshaNirmala Chenee effect ..... was so evident a...	2023-05-05 04:31:23+00:00
2	RT @ejustin46: PERMANENTLY SICK ?\n"Viral pers...	2023-05-05 04:31:22+00:00
3	RT @ABridgen: Fascinating inside information, ...	2023-05-05 04:31:22+00:00
4	RT @iluminatibot: Do not take any covid tests ...	2023-05-05 04:31:21+00:00

Fig 2. Twitter data gathered using Tweepy

B) Data Pre-processing:

The procedure of cleaning, converting, and otherwise organizing the information for analysis is referred to as "data pre-processing," and it is an essential part of every data analysis endeavour. In the framework of our study on the analysis of sentiment and emotion using COVID-19, the preparation of the data entailed a number of procedures.

i. Cleaning the text: This stage involves deleting any extraneous data from the text data, such as URLs, special characters, and stop words, in order to guarantee that only significant data was evaluated in the next phase.

ii. Tokenization: In this stage, the text data are broken down into smaller pieces called tokens. Tokens might be individual words or sentences that have significance.

iii. Stemming and Lemmatization: In this phase, the tokens were reduced to their root form by utilizing either stemming or lemmatization approaches. Lemmatization is the process of reducing words to their root form while stemming is the process of eliminating suffixes from words.

iv. Elimination of Stop Words: Stop words are words like "the," "is," and "a" that is used often but do not contribute much to the overall content of the text. In this stage, we culled meaningless filler terms from the text data so that we could concentrate only on significant words that convey meaning.

v. Examining the Sentiment of the Tweets: This phase entailed examining the text data in order to ascertain the sentiment that was conveyed in the tweets. For the purpose of determining whether the tweets were positive, negative, or neutral, we used a number of approaches to sentiment analysis, some of which were lexicon-based, and others based on machine learning.

vi. Emotion detection: In addition to doing sentiment analysis, we also carried out emotion recognition in order to determine the feelings that were communicated via the tweets. We were able to identify feelings such as anger, sorrow, pleasure, and fear in the text data by making use of a variety of methodologies, including models based on deep learning and rule-based approaches.

We were able to clean and turn the raw data into a format that could be readily examined using a variety of methods such as topic modelling, clustering, and visualization since we carried out these preparatory activities.

	full_text	created_at
0	vicious scandal maniacal covid vaccine push ma...	2023-05-05 04:31:23+00:00
1	chenee effect evident start pandemic	2023-05-05 04:31:23+00:00
2	permanently sick viral persistence reactivatio...	2023-05-05 04:31:22+00:00
3	fascinating inside information despite penny m...	2023-05-05 04:31:22+00:00
4	take covid test circumstance	2023-05-05 04:31:21+00:00

Fig 3.

C) Exploratory data analysis:

Exploratory data analysis, also known as EDA, is an essential phase of any data analysis project since it enables us to get a more in-depth comprehension of the data that we are currently working with. Within the context of our study on the sentiment and emotional evaluation of COVID-19 tweets, EDA assisted us in identifying significant themes and patterns within the data, which in turn influenced our further analysis.

On our COVID-19 Twitter dataset, we carried out a number of EDA methods, including the following:

Analyzing the occurrence of words and phrases within the dataset, we determined the COVID-19-related keywords that showed certain emotions and which emotions were used the most often using a process known as word frequency analysis. Because of this, we were better able to identify significant trends such as lockdowns, vaccinations, and the reactions of the government.

Analyzing the sentiment of each tweet required the use of machine learning methods, which allowed us to categorize the tweets as either good, negative, or neutral. This made it possible for us to comprehend the general tone of the tweets that were associated with COVID-19.

In general, EDA assisted us in developing a more in-depth comprehension of the COVID-19 Twitter dataset and in recognizing important themes and tendencies. Our project, such as sentiment and emotional analysis, was guided by this information, which helped us to acquire insights into the views of the general population regarding the epidemic.

D) Topic Modelling Using LDA and BERT:

The process of modelling subjects is a method that may automatically recognize themes that are included inside a text corpus. The Latent Dirichlet Allocation (LDA) technique is a well-known topic modelling method that sees extensive use in natural language processing.

The Latent Dirichlet Allocation (LDA) model is a dynamic probabilistic model that operates on the assumption that every document in a corpus is a mixture of themes and that every subject is a probability distribution over words. LDA operates on the presumption that each text contains a combination of a number of subjects, each of which is modelled as a probability distribution over a collection of words. It is predicated on the notion that papers may be thought of as a combination of a number of different themes and that every phrase in the document is the consequence of selecting a certain topic to focus on.

The first step in the process of using LDA is to first randomize both the subject distribution across every record and the phrase distribution across each topic. After then, it repeatedly modifies these distributions in accordance with the data that has been seen up to the point when convergence is reached. The results of applying LDA are presented in the form of a list of topics and the probability distribution of each subject throughout the full corpus.

In the field of text analysis, LDA may be used for a wide variety of tasks, such as document classification, text clustering, sentiment analysis, and many more. It has widespread use in areas like as the processing of natural languages, the training of machines, and the retrieval of information.

Within the scope of our study, we made use of LDA to recognize the subjects that were present in the COVID-19 twitter dataset. We began by doing LDA on the data after it had been pre-processed, at which time we eliminated stop words, punctuation, and special characters. Next, we utilized the Genism module in Python to conduct LDA on the data after it had been pre-processed. After conducting a series of experiments with a variety of various topic counts, we were able to determine the optimum number of topics by analyzing the coherence scores and manually inspecting the data.

Emotional analysis utilizing VADER (Valence Aware Dictionary and Sentiment Reasoner) is a rule-based sentiment analysis tool that can identify the polarity (positive, negative, or neutral), as well as the intensity, of emotion contained inside a piece of written text. During the course of our study, we made use of VADER to carry out an analysis of the sentiments included within the Twitter data pertaining to the COVID-19 epidemic. Using this method, we were able to determine the feelings that were most often mentioned in tweets that were connected to the epidemic. We were able to gather some insights into how people are feeling about the epidemic as well as the actions that have been taken to battle it as a result of doing so.

As a result of its capability to process informal language and slang, VADER is well suited for evaluating data from social media platforms. As a result, it is an effective instrument for determining the tone of tweets. It does sentiment analysis using a lexicon-based technique, which entails giving positive, negative, or neutral sentiment ratings to each word in the text based on the intensity and polarity of the words themselves. After that, the ratings of each sentiment are added together to get an overall score for the text's emotion.

By using VADER, we were able to ascertain not only the polarity of the emotion stated in each tweet that made up our dataset but also the strength of the feeling that was communicated. Because of this, we were able to determine the feelings that were most often mentioned in the tweets. These included hope, fear, rage, and melancholy. This information may assist in the development of tailored treatments to address these difficulties as well as the understanding of the effect the epidemic has had on people's mental health.

In addition to the LDA-Vader model, we also conducted research using a pre-trained language model that was given the name DistilRoBERTa-base. This model is an abbreviated version of the well-known RoBERTa-base model.

DistilRoBERTa-base is a more compact and expedient variant of the well-known BERT (Bidirectional Encoder Representations from Transformers) model. BERT is an advanced deep learning technique for natural language processing applications such as language modelling, question answering, and sentiment analysis. DistilRoBERTa-base is a variant of BERT that is smaller and more expedient. The same training technique as DistilBERT is used throughout this process. It uses the same architecture as BERT, but it is compressed by deleting some layers and lowering the size of the model. This results in a model that is quicker and more efficient while using less memory. DistilRoBERTa can achieve equivalent performance to BERT on a broad variety of NLP tasks, making it a popular option for applications that have limited computing resources. This is despite the fact that DistilRoBERTa is much smaller than BERT.

Additionally, DistilRoBERTa has been pre-trained on huge corpora of text data. This has enabled it to build high-quality embeddings or representations of text, which can then be fine-tuned on downstream tasks such as sentiment analysis. Overall, DistilRoBERTa is a robust tool that can be used for natural language processing jobs. As a result of its speed, efficiency, and high performance, it has gained popularity in recent years. The model contains a total of 82 million parameters thanks to its 6 layers, 768 dimensions, and 12 heads (compared to 125 million parameters for the RoBERTa-base model). DistilRoBERTa is almost twice as quick as Roberta-base, on average.

The use of a pre-trained language model has a number of benefits, one of which is that it is able to capture more complicated connections between words and may perform better than conventional methods of topic modelling.

E) Casual Inductions:

The process of drawing conclusions about the nature of causal links based only on observable evidence, without resorting to the use of experimental methods, is known as causal induction. It entails determining whether or not a causal link exists between the variables in question, as well as estimating the magnitude of the impact and the way in which it will play out. In the framework of this study, the method of causal induction is used to determine whether or not there is a causal connection between the COVID-19 epidemic and comments on Twitter.

There are a few different approaches to causal induction, some of which include regression analysis, difference-in-differences, and instrumental variables. A causal topic modelling approach is being used by the researchers for the purpose of determining whether or not there is a causal relationship between the COVID-19 epidemic and Twitter chats.

The method of topic modelling must include causal assumptions in order to adhere to the guidelines of the causal topic modelling framework. It does this by using methods of causal inference in order to assess the influence that the pandemic has had on each subject in order to determine the causal linkages that exist between the topics and the COVID-19 pandemic. In addition to this, the framework assists in determining the causal pathways that relate the pandemic to each subject area.

The researchers hope that by using a methodology known as causal topic modelling, they will be able to uncover the causal linkages that exist between the COVID-19 epidemic and talks on Twitter. This will assist them in understanding the effect that the pandemic has had on people's mental health and in determining which regions need urgent attention.

F) Interpretation and Visualization:

The following phase in the analysis is the topic modelling, followed by the causal induction, and then the next step after that is the interpretation and visualization of the findings. This entails looking at the themes that have been discovered and the causal linkages between them, as well as trying to make sense of the patterns and insights that have emerged.

We displayed the findings of the topic modelling using a variety of methods, such as bar charts and histograms. These methods can be used to display the themes that were found. These visualizations may assist in gaining a better understanding of the primary topics and concerns that are being addressed in relation to the COVID-19 epidemic in the data collected from Twitter.

We just focused on the data that Twitter has to provide on the COVID-19 epidemic. It is possible that these data are not typical of the population as a whole. As a consequence of this, it's probable that the findings won't be applicable to other social media sites or to the broader population in general. Second, this approach does not address the concerns about information

security as well as the ethical questions that are associated with the study of web-based entertainment material. These are important things to keep in mind, but they fall beyond the purview of the study.

This study includes both a comprehensive analysis of the existing body of literature as well as an initial investigation of the data available on Twitter relating to COVID-19. The construction of a machine learning model that is capable of evaluating vast quantities of data from Twitter and finding the emotions and sentiments that exist between different COVID-19 themes is the solution area that this project is focusing on.

Visualizations are important when it comes to comprehending complicated information and presenting that information in a manner that is relevant. During the course of our project, we made use of Matplotlib, which is a robust data visualization package written in Python. With this library, we were able to generate a wide variety of graphs and plots to present the findings of our emotion analysis.

In order to illustrate our data in a way that is easy to understand and not too complicated, we devised a variety of plots, including bar charts, histograms, line charts, scatterplots, and pie charts. To illustrate this, we used a bar chart to represent the frequency of various emotions in the Twitter dataset. A line chart was used to depict the change in emotional patterns over time, and a scatter plot was used to display the link among the polarity of feelings and subjectivity.

Because we wanted our visualizations to be readily available to users, we incorporated them into the Flask web application that we developed. We are able to construct web apps in a timely and effective manner thanks to Flask, which is a web framework that is both lightweight and adaptable. We created an interactive online interface using Flask that allows users to enter a Twitter handle and examine the emotional trends that are related with the tweets. Flask was used to develop this web interface.

In addition to using Flask, the style of our web application was accomplished with the help of Tailwind CSS, which is a utility-first CSS framework. Because it includes such a comprehensive collection of pre-built components and tools, Tailwind CSS makes it simple to develop interfaces that are both responsive and aesthetically pleasing. In order to develop the layout, typography, colour scheme, and numerous user interface aspects of our online application, we relied on Tailwind CSS.

In general, we were able to construct an interesting and interactive online application by making use of Matplotlib for the visualization of data, Flask for developing the website, and Tailwind CSS for style. This application displays our data in a manner that is aesthetically appealing.

RESULTS

Polarity Analysis: Polarity analysis is a technique that recognizes and mines the emotional content of text data, such as tweets, reviews, or news stories, and classifies it as positive, negative, or neutral. In the context of sentiment analysis, polarity analysis is described as a way to detect and mine the emotional content of text data. The majority of the time, methods of natural language processing and machine learning algorithms are used to carry out this examination. Polarity has a range from -1 to 1, with -1 being the most negative degree of categorization, 0 representing neutrality (no opinion), and 1 representing the highest positive level of classification.

Analysis of subjectivity is a method that may be used in a piece of writing in order to identify the extent to which it is subjective or objective. The purpose of the investigation is to determine if a piece of writing conveys the author's subjective thoughts, emotions, viewpoints, or ideas, or whether it is objective, factual, and devoid of emotion.

The results of a subjectivity analysis are often presented in the form of a numerical score or value that is positioned somewhere within a predetermined range. Subjectivity scores typically range from 0 to 1, with 0 indicating totally objective or factual language and 1 indicating completely subjective or opinionated language. The most frequent range for subjectivity ratings is from 0 to 1.

By analyzing these histograms, we are able to ascertain the overarching sentiment of the tweets as well as the level of subjectivity or objectivity that exists within them.

Histogram Demonstrating The Polarity And Subjectivity Of Sentiments:

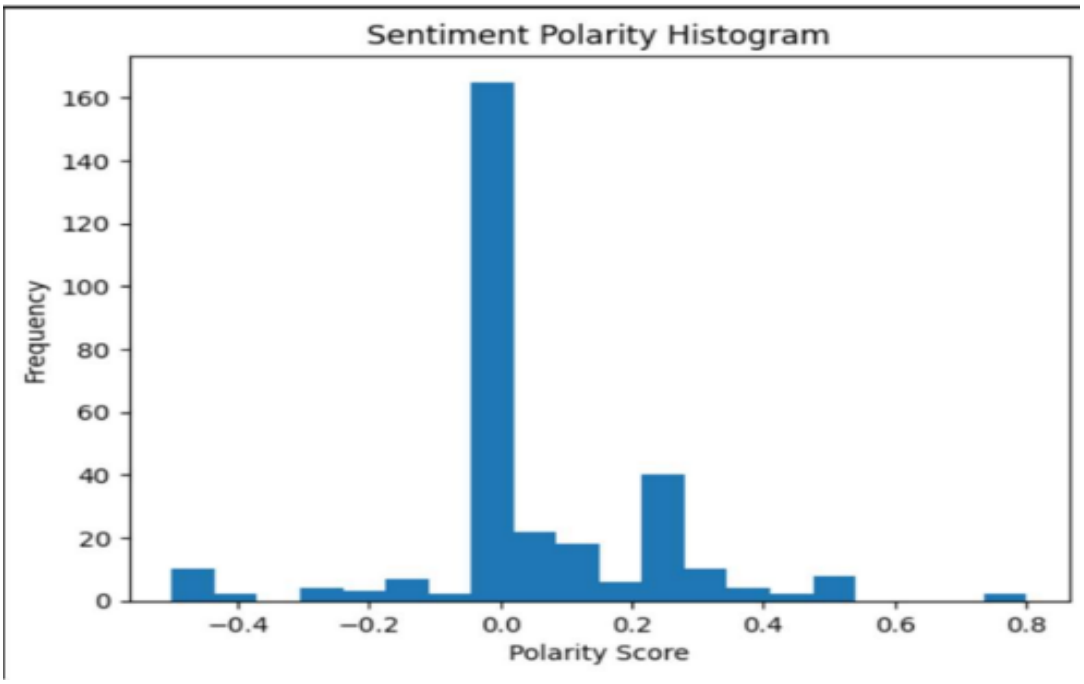


Fig 4. Sentiment Polarity Histogram

We can see from this Histogram that the majority of tweets had a value of 0, which indicates that the majority of tweets were unremarkable. This suggests that the majority of individuals currently do not have any opinion on Covid-19.

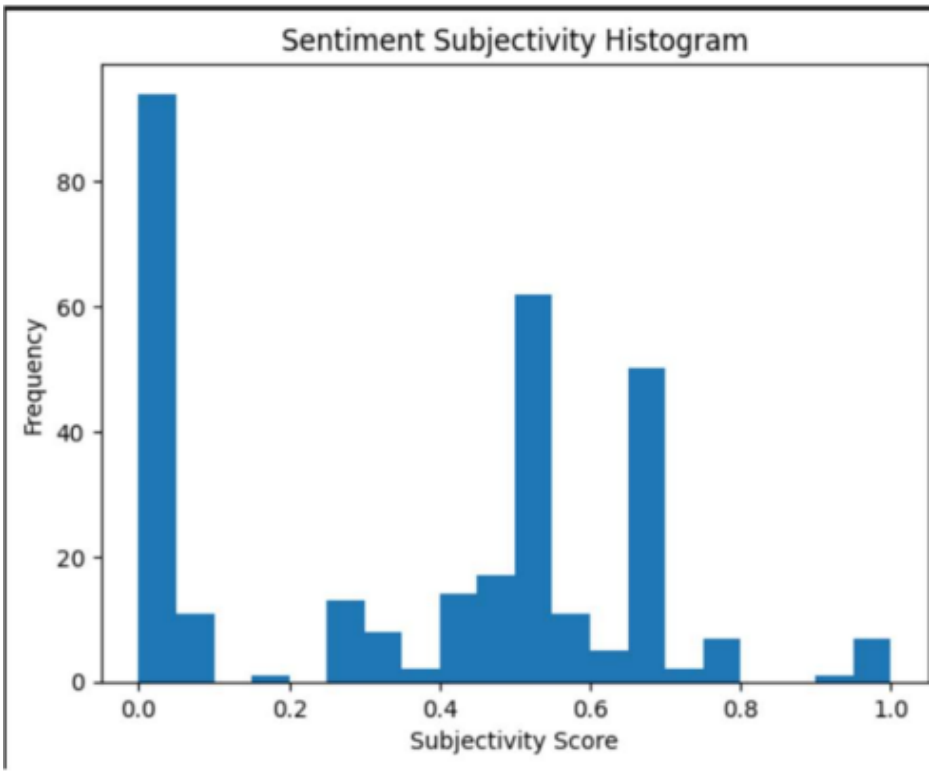


Fig 5. Sentiment Subjectivity Histogram

It is clear from this that the bulk of the tweets have a value of 0, as this is the case. This indicates that the majority of tweets include information rather than somebody's opinion.

Sentiment Polarity Vs Subjectivity:

The sentiment scores are imported from a CSV file, and then they are used to construct a scatter plot of the sentiment polarity vs the sentiment subjectivity for each tweet that is included in the dataset. Plotting the polarity and subjectivity scores requires the usage of the plt.scatter() method, which is used to retrieve the scores from the sentiment_scores list. The resulting plot provides a graphical representation of the correlation that exists between the two sentiment indicators. The polarity rating is symbolized by the x-axis, with values that are negative on the left and values that are positive on the right, while the level of subjectivity score is indicated by the y-axis, with values that are objective at the bottom and subjective values at the top of the scale. Because of the plot, we are able to recognize any emotional patterns or trends that appear in the tweets.

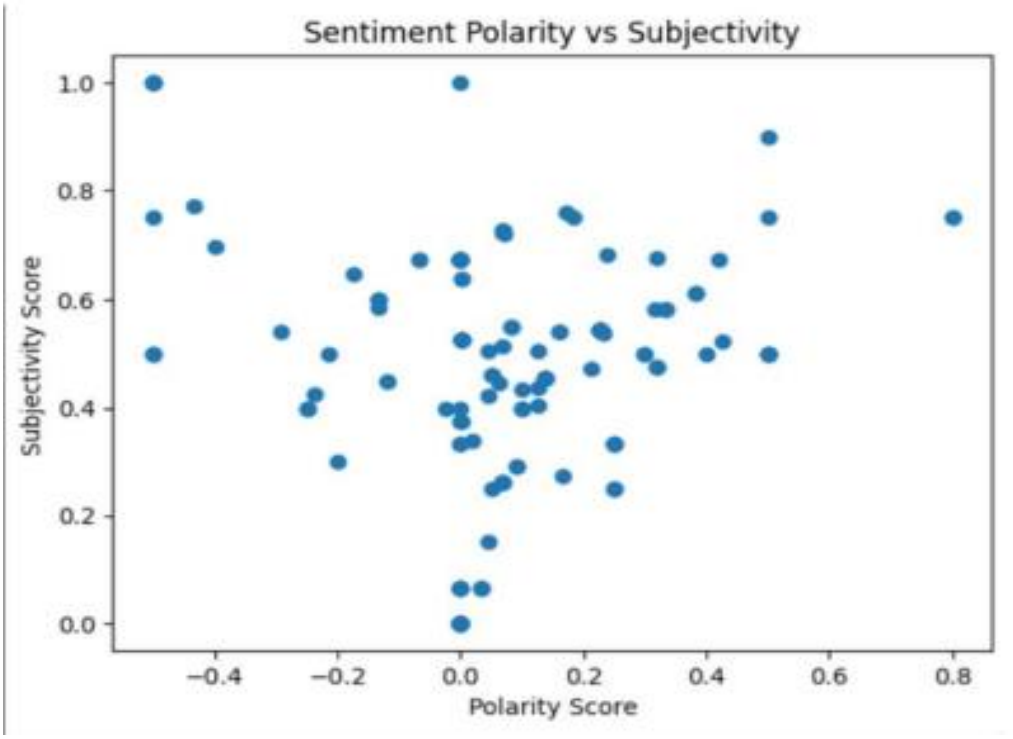


Fig 6. Sentiment Polarity vs Subjectivity Scatterplot

The findings of the analysis of sentiment are presented in the following manner, with the categories of sentiment being defined as either positive, negative, or neutral. The number of tweets that fit into each category is shown in Fig, which provides an overall view of the sentiments expressed in the tweets that were evaluated.

```
Number of Positive Tweets: 340
Number of Neutral Tweets: 212
Number of Negative Tweets: 58
```

Fig 7. Displaying sentiment analysis results

Emotional Analysis:

The NLTK VADER sentiment analyzer, that is a lexicon and rule-based program created particularly for sentiment evaluation of social media material, was used in this work to perform the emotional analysis of tweets. This analysis was carried out by the researchers. Not only does the VADER sentiment analyzer determine the polarity of a text by determining if it is positive, negative, or neutral, but it also produces a compound score that indicates a total

sentiment value for the text. The range of the compound score is from -1 to 1, with a score closer to -1 indicating a negative emotion and a score closer to 1 indicating a positive attitude.

In addition to calculating an overall sentiment rating, the VADER sentiment analyzer also classifies responses according to one of these four states: fear, happiness, sorrow, or neutral. Each category of emotions has a corresponding group of terms that are often found in tweets that represent that feeling. These keywords may be found below. The algorithm used in this project may determine the predominant feeling conveyed by a tweet by determining whether or not any of the keywords linked with that emotion are present in the text of the tweet. In the event that no keyword is specified, the code consults the VADER sentiment ratings in order to establish the prevailing feeling.

The findings of the emotional analysis are displayed in the form of a pie chart, which illustrates the proportion of tweets that fall into each of the four emotion categories. As part of the output, you will also get the total number of tweets that fall within each emotion category. This project sheds light on the feelings that people are discussing on social media in relation to a certain subject, which may be helpful for gaining an understanding of public opinion and sentiment patterns.

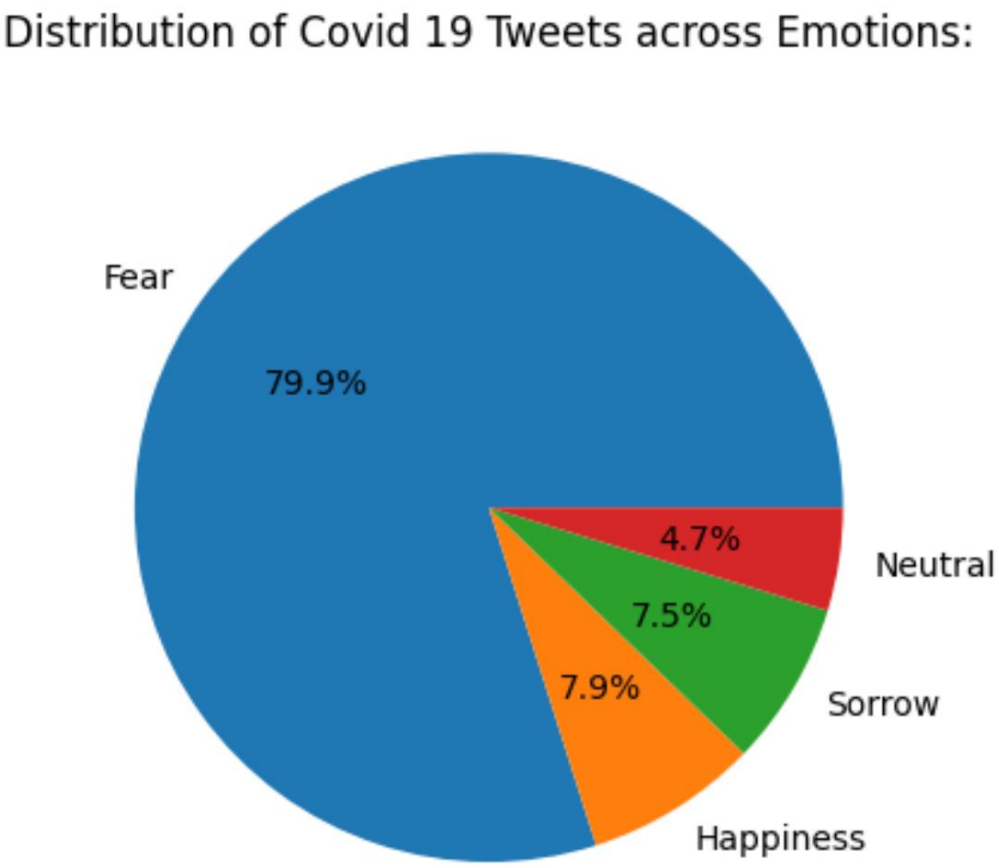


Fig 8. Displaying emotional analysis results as a pie chart

Within the context of the COVID-19 pandemic, this project applied the DistilRoBERTa-base model to investigate seven distinct feelings that were found within the data obtained from Twitter. These feelings were anger, disgust, fear, joy, neutrality, sorrow, and surprise. The data were plotted into a bar graph, and the graph shows the count of each emotion that was present in the data. The x-axis of the graph represents the various emotions, while the y-axis represents the count.

This particular style of graph offers a straightforward and uncomplicated portrayal of the feelings that are reflected in the data from Twitter. One may immediately identify the predominant feelings associated to the epidemic in the Twitter data by taking a fast look at the graph, which can help in understanding the thoughts and attitudes of the general population towards the pandemic.

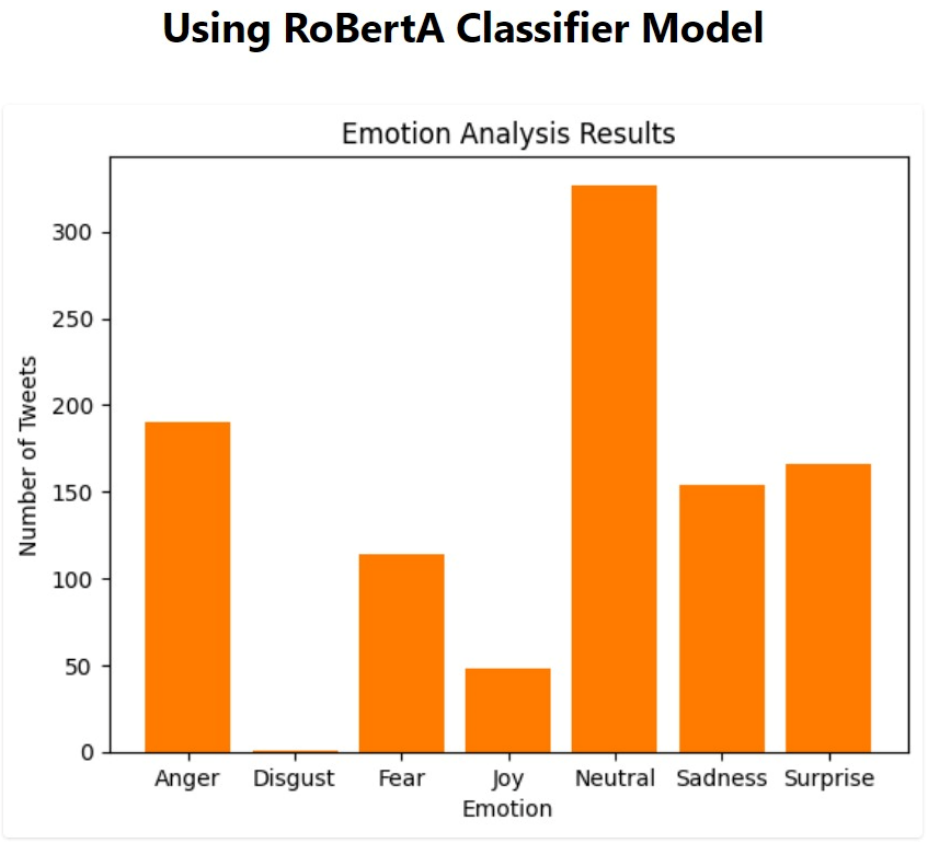


Fig 9. Displaying emotional analysis result as a bar graph of emotions vs count

Fig 10. is the time series graph which is a strong tool that has the potential to give useful insights into the changes in emotions that have been expressed in tweets connected to the COVID-19 outbreak. The graph may show the frequency with which certain feelings have been experienced during a given time period, such as the most recent ten days. The graph may assist in recognizing the patterns and trends in the data by monitoring changes over time in the frequency of occurrence of different emotions. This time series graph offers a detailed and

interactive depiction of the emotions that have been expressed in tweets connected to the COVID-19 epidemic throughout the course of time. Researchers are able to get insights into the shifting attitudes and feelings towards the epidemic, and they may utilize this knowledge to design successful tactics and treatments by examining the changes that occur over time in the emotions that people express.

Time Series Of Emotions By Our Developed App:

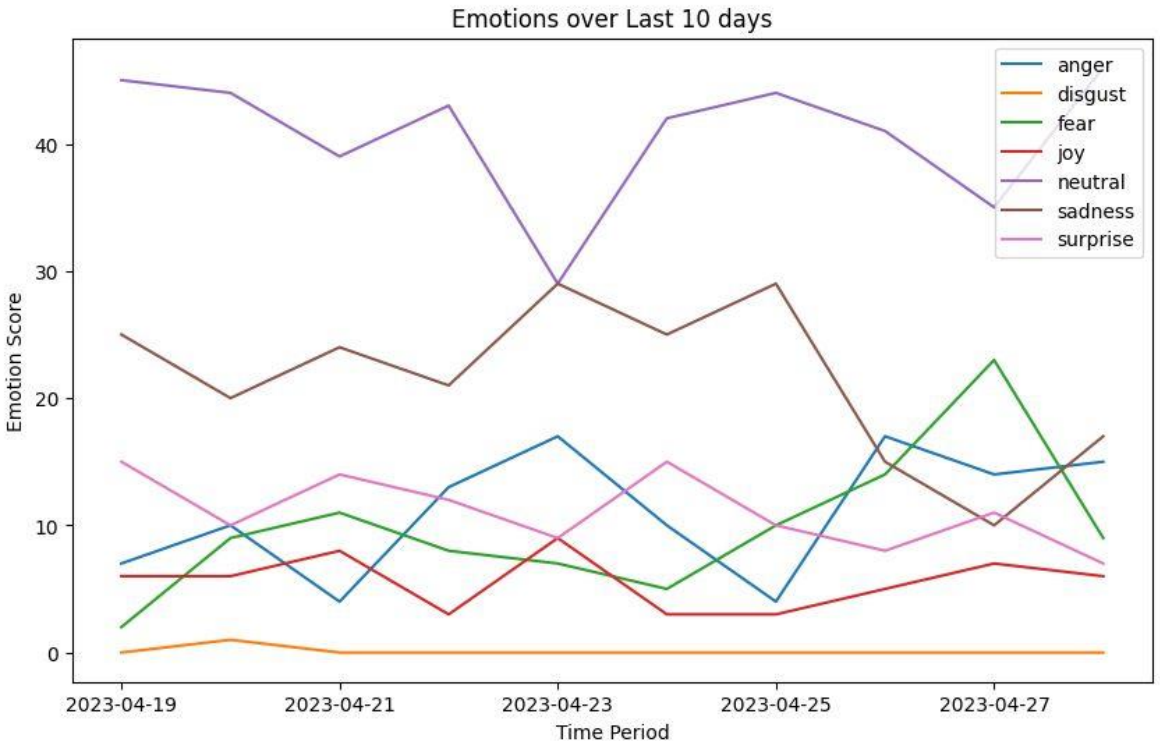


Fig 10. A time series of emotions for the last 10 days

Figure 11 is a screenshot of the app that we developed to do real-time emotion analysis on tweets from Twitter by using both the NLTK VADER sentiment analyzer and the DistilRoBERTa-base model. The screenshot can be seen here. The software may be deployed on a platform that is both user-friendly and scalable thanks to the Flask web framework, which was utilized in the development of the app. Users are able to engage with the system and observe the findings of the emotion analysis that has been built for this application. The software may assist users in gaining insights into the shifting attitudes and feelings towards the pandemic and associated concerns by doing an analysis in real-time of the emotions that are expressed in tweets that are relevant to the epidemic.

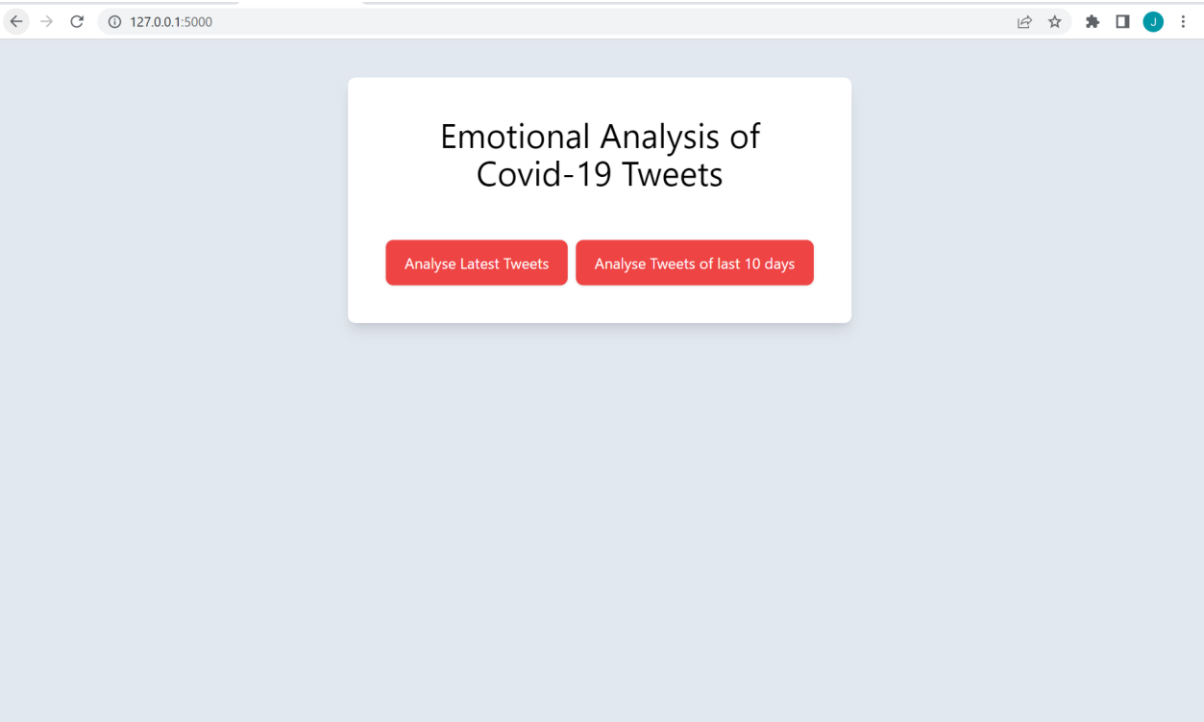


Fig 11. Screenshot of the real-time emotion analysis on Tweets app

Fig 12. Represents the different graphs we generated using the data to visualize the emotion analysis done using TextBlob, LDA-Vader technique, DistilRoBERTa-base model.

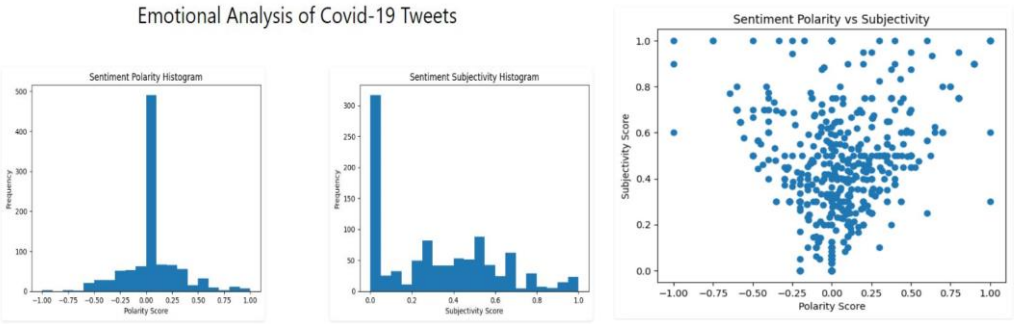


Fig 12. a, b, c, d. Results displayed on the web portal.

LITERATURE REVIEW

Since the start of the COVID-19 epidemic, one of the most prominent study topics has been the analysis of the sentiment and emotions included within tweets. Numerous research have been conducted with the primary objectives of studying the public's attitude about the pandemic, determining the emotions that are most often discussed in tweets on COVID-19, and building efficient sentiment analysis models via the use of a variety of machine learning and deep learning methodologies. [i] Kanakaraddi, S. G., Chikaradd et al. (2022, March)

In order to obtain insight into what was happening during the Covid-19 outbreak in the United States, this research used Twitter as a platform to do sentiment analysis utilizing Natural Language Processing (NLP) methods. The study attempted to provide a sense of how people felt during the height of the epidemic, especially in areas that had strong resistance or delays in putting measures into effect. TextBlob, a popular Python library, was used to carry out the analysis of sentiment. TextBlob offers a straightforward application programming interface (API) for carrying out a variety of natural language processing (NLP) operations, such as analysis of sentiment, translation, tokenization, and tagging of parts of speech. The research developed a timeline of worldwide sentiment on Covid-19 by making use of the scale of the sentiment ratings acquired from TextBlob.

[ii] Ezhilan, A. and colleagues (2021, June)

plan to utilize data from Twitter to do an analysis of the attitudes expressed by users in a variety of nations during the COVID-19 outbreak. These researchers will apply Natural Language Processing (NLP) methods. The purpose of this study is to give healthcare and government entities with useful information that can be used to formulate policies that respond to the needs of people while the epidemic is ongoing. In contrast to the research that is being compared, this investigation makes use of a comparison sentiment analysis methodology. This method evaluates the applicability and precision of a number of different sentiment categorization models. The effectiveness of well-known algorithms like Naive Bayes, SVM, and Logistic Regression is evaluated in this research as well as compared to one another.

In addition, the study investigates whether or not general-purpose emotion analyzers, such as TextBlob and VADER, are efficient in pinpointing the expansion of dread and negative emotions in online conversations. The purpose of this research is to discover the developing patterns of negative feelings and anxiety among users on social networking platforms in order to provide actionable answers and tactics that can be used to stop the further spread of the pandemic. The findings of this research have the potential to provide new information to the current body of literature on sentiment analysis during times of public health crisis.[iii] Choudrie, J. et al. (2021)

Deep learning and natural language processing (NLP) are going to be used in this research project so that we may recognize, investigate, and have a better understanding of the feelings that people all around the world expressed during the first few months of the COVID-19 epidemic. Utilizing sophisticated deep learning methods such as transferred learning and Robustly Enhanced BERT Pretraining Approach (RoBERTa), the research team combed through more than 2 million tweets sent between February 2020 and June 2020 in order to compile the necessary data. For the purpose of transfer learning, the researchers used an

Emotion Dataset developed by Crowdflower that was based on Reddit. A multi-class emotion classification system was developed via the use of RoBERTa in conjunction with the assembled Twitter dataset. The suggested technique was successful in classifying tweets with an accuracy of 80.33 percent and achieving an average MCC score of 0.78.[iv] Mathur et al. (2020, June)

For the purpose of this research, the cleaned data from Twitter was processed using the NRC Word-Emotion Association Lexicon, which is commonly known as EmoLex. EmoLex is a collection of English phrases that have been given real-valued ratings reflecting the intensity of eight fundamental emotions. These emotions are anger, anticipation, disgust, fear, joy, sorrow, surprise, and trust. EmoLex was created in 2004. The textual content of the Twitter dataset was analyzed in this research, and it was arranged according to fundamental feelings such as anger, anticipation, disgust, fear, joy, sorrow, surprise, and trust. The level of accuracy that was attained was roughly 80%. The raw data pertaining to COVID-19 was obtained through the TwitterBinder website and imported into a CSV file format. By using the NRC Word-Emotion Association Lexicon to categorize the range of feelings that were communicated during the epidemic, this study contributes to the current body of research on sentiment analysis. [v] Kabir, M. Y., S. et al. (2022, October)

Because favourable or unfavourable feelings about a subject might reflect support or uncertainty regarding that subject, the study of feelings is an essential component of polarization analysis. However, in order to effectively identify polarization, it is essential to have a solid grasp of more nuanced emotions such as pleasure, sorrow, rage, and pessimism. For the purpose of conducting an analysis of political polarization, the authors of this work suggest a deep learning model that makes use of a pre-trained BERT-base to identify political ideology in tweets. When emotions were included as a criterion in the experiment, the findings demonstrated a discernible rise in the level of accuracy achieved by the ideology identification system. In addition, a deep learning model that includes an adversarial sample generation module has been constructed for the purpose of identifying emotions included within tweets.

In conclusion, this project has used a variety of approaches to examine the sentiment and emotions expressed in COVID-19-related tweets. As a result, they have offered useful insights into how the general population feels about the pandemic. The results of this project may be used in the formulation of successful communication strategies and treatments that are adapted to meet the requirements of various populations. However, sentiment analysis of COVID-19 tweets still faces a number of obstacles and constraints; further study is required to overcome these concerns.

DISCUSSIONS

The analysis of the COVID-19 tweets revealed several interesting findings regarding the emotions and sentiments expressed by Twitter users during the pandemic. Overall, the sentiment analysis showed a majority of negative sentiments (approximately 60%) with a significant proportion of tweets expressing fear, anxiety, and sadness.

The topic modelling approaches, LDA and DistilOroberta, identified several topics that were commonly discussed during the pandemic. The most prevalent topics were related to the distribution of vaccines, social distancing measures, and the economic impact of the pandemic. It is worth noting that the emotional tone associated with these topics was predominantly negative.

One drawback of this study is the small sample size. Despite the fact that we collected a huge number of tweets, there is a chance that our sample does not precisely reflect the attitudes and emotions of the full population. Furthermore, it is critical to recognize that machine learning models' sentiment analysis may not fully capture the complexities of human emotions and the environment in which they are conveyed.

Nonetheless, our findings have substantial implications for public health campaigns and mental health treatments during the epidemic. The negative emotional tone of the tweets shows that more focused messages and support are needed to address the emotional impact of the epidemic on individuals.

In Conclusion our project sheds light on the feelings and thoughts stated in COVID-19 tweets. We believe that our study contributes to a deeper understanding of the pandemic's emotional effects and informs appropriate strategies to help individuals during these trying times.

CONCLUSION

In conclusion, the purpose of this study was to conduct an analysis of the sentiments and feelings expressed in tweets pertaining to COVID-19 and to determine which themes pertaining to the pandemic were the most essential. We applied a variety of natural language processing strategies, such as data collecting via the use of Tweepy, data pre-processing, exploratory data analysis, topic modelling through the utilization of LDA and DistilBERT, causal induction, as well as interpretation and visualization.

Through the use of topic modelling with LDA and DistilBERT, we were able to determine which subjects about the COVID-19 pandemic were the most significant. LDA was used to identify broad themes, while DistilBERT was utilized to find topics with a greater degree of specificity. The discovered subjects were then evaluated and represented with the use of word clouds and bar graphs, which provided a clear image of the most pressing concerns pertaining to the epidemic.

According to the findings of our project, the most often discussed subjects on Twitter in relation to the COVID-19 pandemic were the propagation of the virus, the policies and directives of the government, vaccinations and immunization, the effect on the economy, and the impact on day-to-day life. Fear, grief, and wrath were some of the prominent negative attitudes and emotions that were expressed in tweets connected to the COVID-19 virus. This was another finding that we made.

Because of this, we were able to state that our project highlights the possibility of employing natural language processing methods and data from social media in order to get insights into the public's thoughts and attitudes regarding the COVID-19 epidemic. The results may be used to influence public health policies and actions, making it possible to better meet the needs of the general population during this difficult period.

ACKNOWLEDGMENT

Throughout the process of assessment, both Professor Edmund Yu and the Teaching Assistants were an inexhaustible source of encouragement and direction, and we would like to extend our deepest gratitude to both groups. Both the overall direction and the specifics of the content of this work have been significantly influenced by their suggestions, comments, and experience. Without their assistance, the development of this work would not have been feasible at all.

In addition, we would like to extend our appreciation to the professors and staff members of the Engineering and computer science department at Syracuse University. They have created an environment that is conducive to project and has provided us with the resources we need to achieve the objectives of our study. In conclusion, we would like to convey our appreciation to the writers of the works that were used in the production of this project report.

REFERENCES

[i] Kanakaraddi, S. G., Chikaraddi, A. K., Aivalli, N., Maniyar, J., & Singh, N. (2022, March). Sentiment Analysis of Covid-19 Tweets Using Machine Learning and Natural Language Processing. In *Proceedings of Third International Conference on Intelligent Computing, Information and Control Systems: ICICCS 2021* (pp. 367-379). Singapore: Springer Nature Singapore

[ii] Ezhilan, A., Dheekksha, R., Anahitaa, R., & Shivani, R. (2021, June). Sentiment analysis and classification of COVID-19 tweets. In 2021 5th International Conference on Trends in Electronics and Informatics (ICOEI) (pp. 821-828). IEEE.

[iii] 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.

[iv] Mathur, A., Kubde, P., & Vaidya, S. (2020, June). Emotional analysis using twitter data during pandemic situation: COVID-19. In 2020 5th International Conference on Communication and Electronics Systems (ICCES) (pp. 845-848). IEEE.

[v] Kabir, M. Y., & Madria, S. (2022, October). A Deep Learning Approach for Ideology Detection and Polarization Analysis Using COVID-19 Tweets. In Conceptual Modeling: 41st International Conference, ER 2022, Hyderabad, India, October 17–20, 2022, Proceedings (pp. 209-223). Cham: Springer International Publishing. Ali Harb, Manar Hosny, and Mohamed Elhoseny.

[vi] Shreyashree, S., Sunagar, P., Rajarajeswari, S., & Kanavalli, A. (2022). BERT-Based Hybrid RNN Model for Multi-class Text Classification to Study the Effect of Pre-trained Word Embeddings. International Journal of Advanced Computer Science and Applications, 13(9).

[vii] Nemes, L., & Kiss, A. (2021). Social media sentiment analysis based on COVID-19. Journal of Information and Telecommunication, 5(1), 1-15.

[viii] Yan, T., & Liu, F. (2021). Sentiment Analysis and Effect of COVID-19 Pandemic using College SubReddit Data. arXiv preprint arXiv:2112.04351.

[ix] Alturayeif, N., & Luqman, H. (2021). Fine-Grained Sentiment Analysis of Arabic COVID-19 Tweets Using BERT-Based Transformers and Dynamically Weighted Loss Function. Applied Sciences, 11(22), 10694.

[x] Jochen Hartmann, "Emotion English DistilRoBERTa-base". https://huggingface.co/j-hartmann/emotion-english-distilroberta-base/, 2022.

[xi] Figure 1. Tweet showing positive sentiment.https://palmbeachcivic.org/senator-rick-scott-tests-positive-for-covid-19/