Long COVID Emotion Analyzer

Gan Joo Han[1], Vimala Balakrishnan[2]
[1]*Faculty of Computer Science and Information Technology, University of Malaya, 50603 Kuala Lumpur, Malaysia*
[2]*Faculty of Computer Science and Information Technology, University of Malaya, 50603 Kuala Lumpur, Malaysia*
wih190002@siswa.um.edu.my, vimala.balakrishnan@um.edu.my

***Abstract*—COVID-19 has spread around the world since 2019. In the past two years, countless people who have recovered from COVID-19 will still have sequelae. The sequelae are called as long-COVID. Although many people recover without treatment, they can experience long-COVID for weeks or even months. This paper aims to find out the emotions of people towards long-COVID topic using a machine learning approach based on the data obtained from one of the most popular social media – Twitter. Furthermore, I deployed the proposed model by developing a prototype named Long COVID Emotion Analyzer. Twitter data were collected using the hash tag Long COVID for a total of five months (i.e., May – September 2021), resulting in 97098 clean tweets. IBM Watson Tone Analyzer was used to label the emotion of the tweets. (i.e., sadness, joy, fear, anger, analytical, tentative, confident, neutral). Several machine learning algorithms were used to train and test the dataset. Results indicate Logistic Regression with Unigram of Bag of Words to be the best predicting model, with accuracy = 88%, F1-score = 88%, Area Under Curve = 97%.**

***Keywords*—Emotion; Machine learning; Twitter; Long COVID.**

1. INTRODUCTION

The COVID-19 pandemic, also known as coronavirus disease 2019, began in Wuhan, China, in late December 2019 and swiftly spread to other countries across the world by early March, causing the US to declare a national emergency on March 13, 2020. (World Health Organization, 2021) Although most infected persons with COVID-19 recover within weeks after being unwell, some people experience post-COVID symptoms. Post-COVID syndromes are a slew of new, recurring, or chronic health issues that can strike four weeks or more after a person is initially infected with the COVID-19 virus. Even patients who showed no signs of COVID-19 in the days or weeks after their infection may experience post-COVID complications. These illnesses might manifest themselves as a range of health issues throughout time. (Centers for Disease Control and Prevention, 2021).

Post-COVID-19 syndrome, long-haul COVID, post-acute COVID-19, long-term consequences of COVID, or chronic COVID are all terms used to describe long COVID syndrome. (Centers for Disease Control and Prevention, 2021) It was first identified in social support groups, then in scientific and medical circles. COVID-19 survivors of all ages, including younger individuals, children, and those who are not hospitalized, are afflicted by this long-term COVID illness, which is still little understood. The most common symptoms, according to several studies, are weariness and trouble breathing, which can linger for months after exposure to COVID-19. Long COVID can cause cognitive and mental deficits, chest and joint pains, palpitations, myalgia, smell and taste dysfunctions, cough, headache, and gastrointestinal and cardiac problems. (Yong, 2021)

The public became more concerned as news of the new coronavirus circulated throughout the globe. On the Internet, a growing number of public emotions are freely expressed. Through social media posts such as Tweets, Facebook posts, Instagram status, and comments on posts, people expressed their concerns about the rise of COVID and the long COVID symptoms, including long COVID patients. Despite the fact that individuals may convey their thoughts and feelings through audio and video files, the majority of people still communicate and engage in everyday life through text on social media platforms. When it comes to communicating their feelings about other people, events, or items, people still prefer words (Sailunaz, et al., 2018).

2. LITERATURE REVIEW

During recent years, the usage of social media during times of COVID-19 crisis has increased dramatically. Using Twitter APIs, several academics from all around the world are building COVID-19 datasets (Banda et al., 2021; Chen et al., 2020). Mostly of the tweets are related to keywords such as COVID, corona, pandemic, and quarantine. There is still has limited study which related to our research about long COVID.

Few would have predicted that the coronavirus disease 2019 (COVID-19) pandemic, which was declared by the World Health Organization (WHO) in March 2020, would be chronic. The new severe acute respiratory syndrome coronavirus 2 is the cause of COVID-19 (SARS-CoV-2) (Hu et al, 2021). This longer-lasting COVID-19 cases started gaining attention and the term long COVID or post COVID syndrome or long-haul COVID-19 started to gain recognition in the communities of medical and scientific (Callard & Perego, 2021). The possible persistent symptoms of long COVID include chest and joint aches, cognitive and mental impairments, palpitations, myalgia, smell and taste dysfunctions, cough, headache, cardiac difficulties, etc (Yong, 2021). There are still having some undiscover symptoms and more research is required to prove the factors cause long COVID.

To perform emotion analysis for this project, we have investigated a few emotion models to select the suitable model for our project. Emotion is a concept that describes personal or group feelings (Ortony, et al., 1990). Ekman (1992) had previously created an emotional model in which the main

elements were anger, disgust, fear, happiness, sadness, and surprise. Plutchik (1982) presented a three-dimensional hybrid model with eight fundamental complex emotions which are joy, sadness, anger, fear, trust, disgust, surprise, and anticipation as shown in Figure 1. He divided emotions into four categories: main, secondary, tertiary, and opposing emotions, like a color wheel. A popular and revered framework called the circumplex model which is developed by Posner (2005) as shown in Figure 2. Posner plotted out emotions on a 2D plane. On the x-axis is how pleasant emotion is while on the y-axis is the activation level of emotion. Figure 3 shows Scherer's circumplex continuum model (Scherer, 2005), which is based on the assumption that all emotions exist in a two-dimensional space continuum with four dimensions: valence (how positive or negative an emotion is), excitement (the degree of energy or excitement associated with the emotion), coping potential (feeling of control over an organism in certain situations), and goal attainment (ability to achieve a goal). The task of studying texts and determining emotion from them is fairly difficult. It is hard to detect emotion from the language and semantics of these texts.

Although there are many works available on tweets classification, there are only few attempts to classify emotion using COVID-19 related tweets. There is no related work on emotion detection related to long COVID tweets. Thus, some related to emotion analysis studies are reviewed. Kabir & Madria (2021) created a neural network model and trained it with manually labelled data to automatically recognise various emotions in COVID-19 tweets at fine-grained labelling. There were 10 labels which are neutral, optimistic, happy, sad, surprise, fear, anger, denial, joking, and pessimistic were being labelled. They also built a customized Q&A RoBERTa model to extract terms from tweets that are most responsible for the associated emotions. Their classification model outperformed the other systems, with a Jaccard score of 64.75% and an accuracy of 89.51%. The custom RoBERTa Q&A model surpasses other models with a Jaccard score of 78.65%. Kulai et al., (2021) fetched tweets using Twitter APIs for peak COVID months and perform emotion analysis by using Facebook's FastText. Facebook's FastText is an NLP library generally used for text classification and representation. It classifies tweets into 12 emotions namely anger, relief, boredom, happiness, hate, fun, love, surprise, worry, enthusiasm, sadness and empty. They also compared their proposed method with commonly used algorithms, and they found that FastText based approach performed better in terms of classifying the user sentiments or emotions.

Salam & Gupta (2018) focused on data gathered from Twitter, one of the most famous social media platforms, by evaluating live and archived feeds and extracting emotions. The twitter data required by them is English language and is converted into a vector eight emotions which are anticipation, enjoyment, sad, disgust, anger, surprise, fear, trust or no emotion. They used supervised machine learning techniques such as Naïve Bayes and SVM and unsupervised machine learning techniques such as K-means to perform classification. They found the SVC with 5-fold validation is the best performer with an accuracy of 66%. They also found that Naïve Bayes and SVM performed better than K-means and 5-fold cross validation performed better than 10-fold. Gupta et al., (2017) proposed a novel approach to detect emotions like happy, sad, angry and others in textual conversations using an LSTM based deep learning model. They used machine learning algorithms such as SVM, Gradient Boosted Decision Tree, Naïve Bayes as their baselines and they found that Sentiment and Semantic LSTM (SS-LSTM) gives the best performance on F1 score for each emotion class as well as on Average F1.

An et al. (2017) used Naïve Bayes algorithms to do the automatic classification of Chinese music emotions by analysing the lyrics. They evaluate the classifiers trained by four different datasets and the final accuracy is 68%. Hassan et al., (2017) presented a work that used social media data to research and detect emotions from text using emotion theories, machine learning, and NLP approaches to determine a person's depression degree. They employed three classifiers, SVM, Naive Bayes, and Maximum Entropy, to create binary and multi-class sentiment classification approaches at the sentence level. They concluded by using a voting model and feature selection approach to compare SVM, NB, and ME classifiers. Their findings showed that SVM performed better than NB and ME classifiers. SVM has a 91% accuracy, Naive Bayes has an 83% accuracy, and ME has an 80% accuracy.

Perikos & Hatzilygeroudis (2016) demonstrated a sentiment analysis system that uses an ensemble of classifiers to recognise emotions in text automatically. They used ensemble of Naive Bayes, maximum entropy learner, and knowledge-based tools and the best performance of a sole classifier is achieved by the Naïve Bayes classifier. Da Silva et al., (2014) developed a method that uses classifier ensembles and lexicons to automatically classify the sentiment of tweets. They used Multinomial Naive Bayes, Logistic Regression, Random Forest, Support Vector Machines (LibSVM), and ensembles and they found that ensembles showed better accuracy rates than single base classifiers for all the datasets.

Hasan et al., (2014) proposed a new approach, Emotex for automatically classifying text messages of individuals to infer their emotional states. They utilized the well-established Russell's Circumplex model by considering four major classes of emotions which are Happy-Active, Happy-Inactive, Unhappy-Active, and Unhappy-Inactive. For classifying different emotions categories of Twitter tweets, they examined the accuracy of numerous machine learning methods, including SVM, KNN, Decision Tree, and Naive Bayes. Their method achieved a high accuracy of over 90% and is robust across different learning algorithms. Desmet & Hoste (2013) who used natural languages processing techniques and machine learning methodology such as SVM for fine fine-grained emotion detection in suicide notes.

However, to our knowledge there is no available works on tweets emotion extraction related to long COVID. There is also lack of publicly available datasets. In this project, we scraped the raw tweets related to long COVID from Twitter and provided emotion label to the data. Then we used the supervised machine learning approach to train an emotion predictor model and deployed it as a data product which is a web application hosting on Heroku platform.

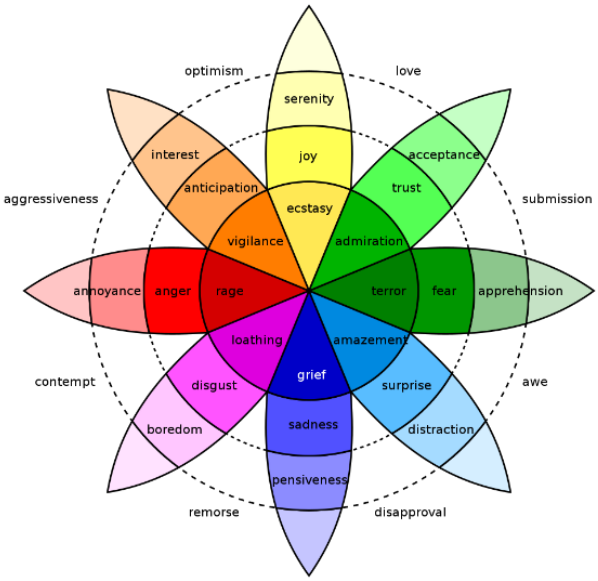

Figure 1: Plutchik's Wheel of Emotion (1982)

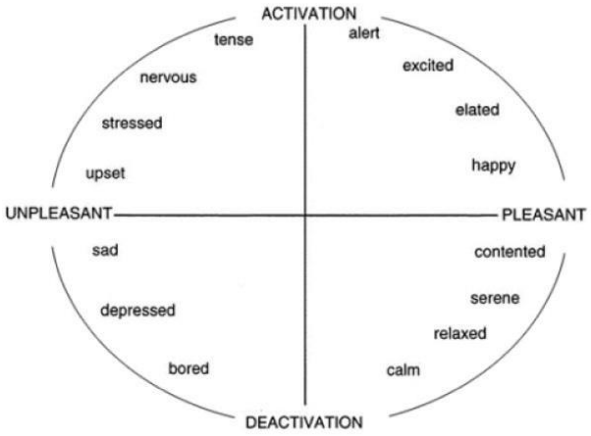

Figure 2: Posner's Circumplex Model (2005)

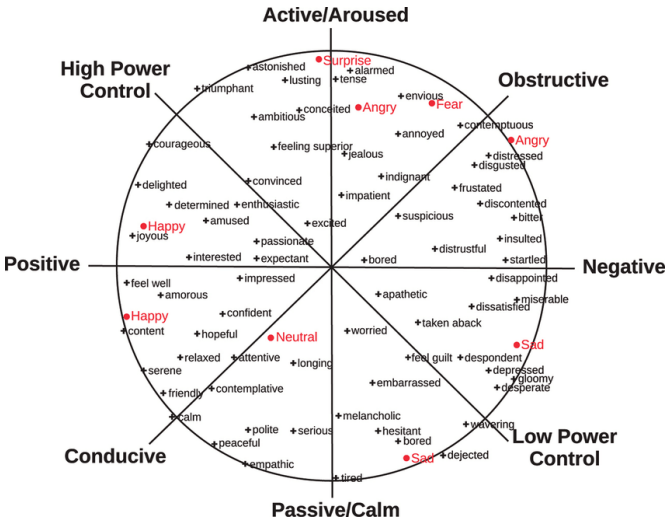

Figure 3: Scherer's circumplex model (2005)

3. PROBLEM FORMULATION

3.1 Problem Statement

Every area and country have instituted lockdown measures since the emergence of COVID-19, and people seldom venture out to mingle with others. This incident shows that some people, including long COVID patients, are unable to express their emotions face-to-face with their friends or family. This is because everyone is locked up at home, it is difficult to know what the emotions of patients with long COVID experience is when they suffer from long COVID symptoms. At the same time, determining what kinds of feelings people have about long COVID is tough. At this time, everyone is more inclined to vent their emotions on social media, including one of the biggest social networking platforms Twitter, which has 353 million active users. (Sailunaz, et al., 2018)

The goal of this project is to create a long COVID emotion analyzer for persons who are concerned about long COVID issues or who are long COVID patients. It is a web application which can analyze how people react to the long COVID and what emotions they have. The application can find out the trends of keywords used by people regarding the long COVID issues. The application can also analyse the emotion from the text.

3.2 Project Scope

The scope of the project includes searching for the dataset, training the model, and deploying a model on a web application for the users to view the tweets analysis results about long COVID and predict the emotion of text entered by users. At the end of the project, a web application is deployed and hosted online, and it can be accessible anywhere from a computer. It also can support other web browsing enabled devices which having smaller screens such as smartphones and tablets, but it is not recommended for a better visual experience.

3.3 Objective

The objectives of this project are:

1. To explore the trend of keywords across the different emotions
    - What are the keyword trends for long COVID topics?
2. To determine the emotions using supervised algorithms
    - What machine learning algorithms can be used to determine the emotions from texts?
3. To assess the effectiveness of the emotion model by using evaluation metrics
    - What are metrics can be used to evaluate the emotion model?
4. To develop a product for the emotion analyzer
    - How to share the insights of the data and the application of the emotion model to stakeholders?

4. METHODOLOGY

This section presents the methodology of the proposed study, including the overall long COVID emotion analyzer architecture, including data collection and pre-processing, emotion analysis, prediction models, the evaluation metrics, and web application deployment. The overview of overall development process is shown in the Figure 4.

4.1 Data collection

Tweets were crawled and collected using Twitter streaming API service between May 1, 2021 to September 30, 2021, based on one hashtag, namely, #LongCOVID. The initial number of tweets gathered were 130005. The dataset is stored in a CSV file.

4.2 Data pre-processing

Data collected from social media platforms, including Twitter, are known to contain a lot of noise. Thus, a series of pre-processing tasks were performed to improve the dataset's quality. Several common criteria were used, namely:

1. Tweets in any other language other than English were removed,
2. All URLS, emails, emoticons and emojis were removed,
3. Special characters such as @, #, & etc., multiple spaces, punctuations and digits were removed,
4. Tweets containing fewer than three words (e.g. good day, sleep)were removed,
5. User handles (e.g. @Maguire) and hashtags were removed,
6. All re-tweets and duplicate entries were removed,
7. Convert upper case to lower case, and
8. Stop words were removed.

By implementing these steps above, a total of 32,908 tweets were reduced, leaving 97,098 tweets for further analysis. After that, the lemmatization is applied to the clean tweets instead of stemming because stemming operates on a single word without knowledge of the context (e.g. change, changing, changes, changed, changer → chang) while lemmatization determines the lemma of a word based on its intended meaning. The process converts words into their basic forms. Unlike stemming, lemmatization depends on correctly identifying the intended part of speech and meaning of a word in a sentence (e.g. change, changing, changes, changed, changer → change; drive, drives, drove, driven → drive) (Müller, 2015).

4.3 Data Labelling

It is necessary to choose the emotion model before we explore the tweets emotion. In this project we decided to utilize the Scherer's circumplex model by considering 7 emotions and classify tweets as 'joy', 'sadness', 'anger', 'fear', 'analytical', 'tentative', 'confident' or no emotion which is 'neutral'. The circumplex model of this project is shown in Figure 5. The emotions were labelling using the IBM Watson Tone Analyzer which calculates the emotion score between a range of 0 to 1. The highest score in "Document-level" is picked. The example is shown in Table 1.

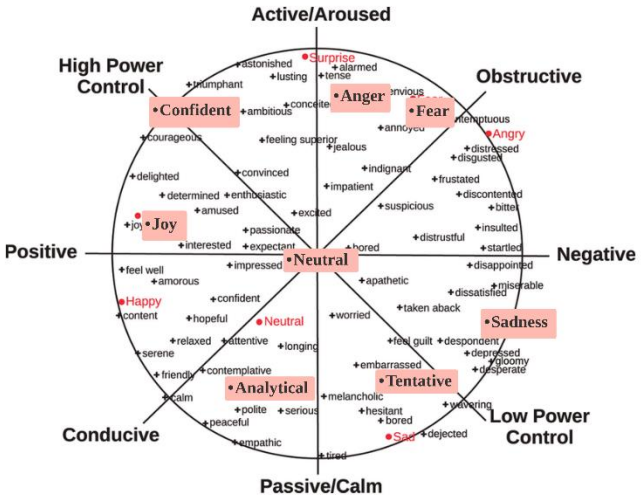

Figure 5: Circumplex model for this project

Table 1: Labelling the example tweets.

Example Tweets	Confidence Score
"migraine like headaches weeks week stopped amp havent come wonder weve covidish thing long covid knowing weve covid point pos antibody test remember ill"	Sadness: 0.65 Analytical: 0.62 Tentative: 0.61 ∴ Label: Sadness
"favorite aunt covid bad thing believed amp spread conspiracy theories vaccine pacemaker longterm effect vaccine wouldnt"	Sadness: 0.54 Analytical: 0.88 Tentative: 0.81 ∴ Label: Analytical
"stiking thinking antiopioid tardsdoctors prescribing opioids covid longhaulers raising addiction fears"	Fear: 0.92 Analytical: 0.65 ∴ Label: Fear
"rare bullshit friends w long covid hospitalized dont know cases cfs infections friends cfslike cases covid year"	Anger: 0.64 ∴ Label: Anger
"number young healthy people got case covid tiny people experiencing longterm symptoms tiny fraction laughable believes longer bc look obviously false"	No result ∴ Label: Neutral

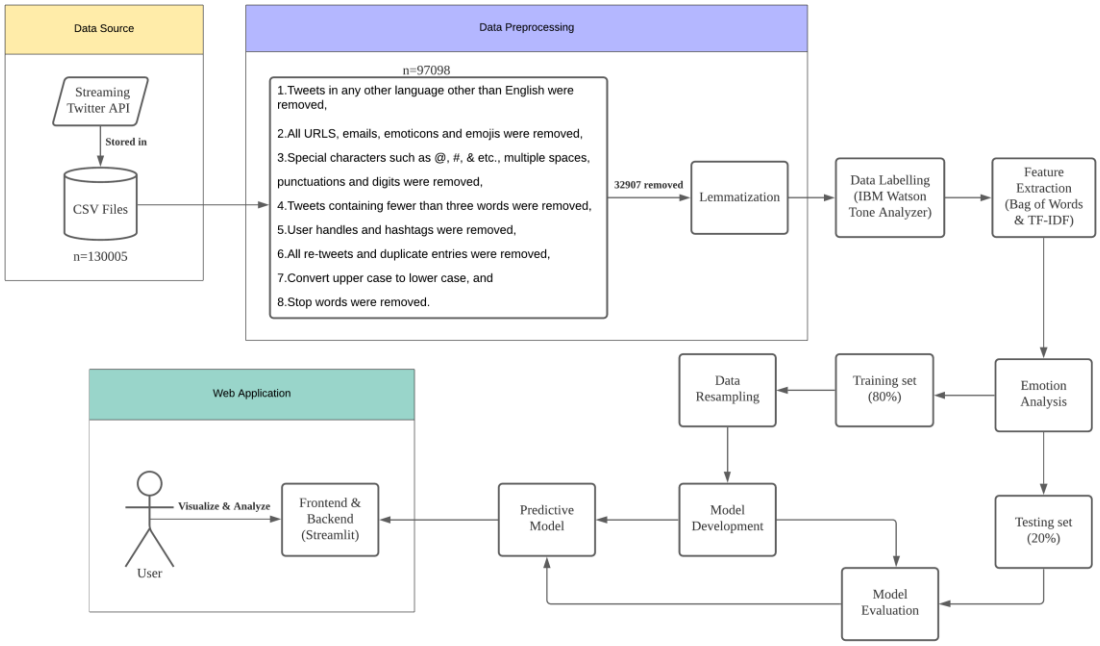

Figure 4: Overall Pipeline

4.4 Feature Extraction

The way of text is analyzed and how it is represented is important for the performance of a machine learning approach. In this project, bag-of-words (BOW) and term frequency-inverse document frequency (TF-IDF) representation technique are used for the representation of natural language text. BOW is the simplest form of text representation in numbers and is often used in text mining applications. BOW represents each text document as a numeric vector where each dimension is a specific word from the corpus and the value could be its frequency in the document, occurrence denoted by 1 or 0. TF-IDF is a numerical statistic that is intended to reflect how important a word is to a document in a collection or corpus. There might be some terms that occur frequently across all documents and these terms may to overshadow other terms in the feature set, especially words that do not occur as frequently, but might be more interesting and effective as features to identify specific emotion categories. In short, BOW creates a set of vectors containing the count of word occurrences in the text while TF-IDF contains information on the more important words and the less important ones as well. Besides that, N-grams feature extraction method also were experimented in the vectorization process. N-grams basically is a collection of word tokens from a text document such that these tokens are contiguous and occur in a sequence. In this project, BOW with unigram and bigram and TF-IDF with unigram and bigram are used. All the tasks were accomplished using Python.

4.5 Emotion Analysis

Exploratory data analysis is carried out to find some initial insights and more understand about the data. Result of the IBM Watson Tone Analyzer labelling revealed 36530 (37.62%) analytical tweets, 19822 (20.41%) sadness tweets, 18031 (18.57%) neutral tweets, 7959 (8.20%) tentative tweets, 5957 (6.14%) joy tweets, 4791 (4.93%) confident tweets, 2610 (2.69%) fear tweets, and 1398 (1.44%) anger tweets. The analytical category is the highest followed by sadness category while the anger category is the lowest. Thus, people have an analytical and sadness emotion towards the topic of long COVID. The emotions classes are imbalanced and needed to be balanced before passing into machine learning algorithms. This is because imbalanced dataset will cause the model biased to the major category class. Besides that, we also find out some of the most common words in the dataset which can be considered as stop words and need to be removed before proceeding to model building part. This is because these common stop words exist in all categories of emotions and will not help in the model building, even will worsen the performance of model. Round 2 of data cleaning step is taken to remove these common stop words such as 'covid', 'long', 'vaccine', 'know', 'people', 'amp', 'time', 'need', 'like', 'year', 'term', 'risk', 'vaccinate', 'symptom', 'work', etc. The others visualization results will be shown in *5 RESULT* section.

4.6 Machine Learning Model

4.6.1 Train Test Splitting

The dataset is split into 80% of training set and 20% of testing set. The training set is used for model training while the testing set is used for model evaluation.

4.6.2 Data Resampling

From the previous emotion analysis, the problem of imbalanced data is found. In this step, the training set is resampled to become balance before passing into machine learning algorithms for training. Majority classes such as analytical, sadness, and neutral are downsampled while minority classes such as tentative, joy, confident, fear, and anger are upsampled.

4.6.3 *Model Development*

The task is to build a model that could predict the emotion from the user's input text which is a multiclassification problem. There are a variety of algorithms to choose from, each with its own set of advantages and disadvantages. Exploration of the various algorithms is essential to determine the best algorithms to be used to build the model for web application. Thus, six machine learning algorithms were used in this project, namely, Multinomial Naïve Bayes, Logistic Regression, Random Forest Classifier, K-Nearest Neighbors, Support Vector Machine, and Decision Tree.

Below are the parameters for each algorithm used in this experiment:

- Multinomial Naïve Bayes: alpha = 0.315789, fit_prior=True;
- Logistic Regression: C=10000, solver='saga', n_jobs=-1;
- Random Forest Classifier: n_estimators=200, n_jobs=-1
- K-Nearest Neighbors: Default
- Support Vector Machine: loss='log', penalty='l2',alpha=1e-3, random_state=42, max_iter=5, tol=None
- Decision Tree: criterion='gini', max_depth=None

4.6.4 *Model Evaluation*

Performance evaluation is a crucial step in analysing a model's performance by comparing the model's predictions to the actual labels. The evaluation metrics which are used for the model evaluation are accuracy, F1-score (macro), Precision (macro), Recall (macro) and Area Under Curve (AUC).

4.6.5 *Predictive Model*

The best performing model which is logistic regression with unigram is used as the predictive model of the web application.

4.6.6 *Web Application*

The architecture of the web application being used is Streamlit framework. Streamlit is an open-source app framework which all the code are in Python. No front-end and back-end experience required to build a powerful web application using Streamlit. After that, the web application is hosted by using Heroku, a website that provided free web hosting services. Heroku is chosen because it provided free web hosting services to host the web application. Heroku is also user-friendly for developer to deploy their application. The only limitation of Heroku is when nobody accesses to the server, the server will shut down. When user accesses to the web application again, it takes time for the server to boosting on again.

5 RESULTS

5.1 *Emotion Analysis*

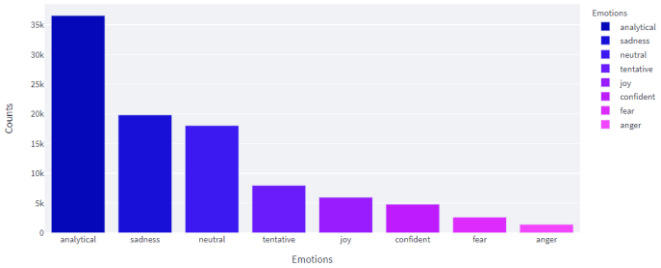

Figure 6: Distribution of emotion in the dataset

Figure 6 shows that the distribution of emotion in the dataset after labelling. From the results, we know that the highest emotion is analytical, followed by sadness and neutral, while anger is the lowest. Thus, we can know that people are having an analytical emotion towards the long COVID topic. This maybe because of majority of the people are not very knowledgeable about long COVID, so they maintain an analytical mood to look at and study about long COVID. Apparently, people do not have enough information about the long COVID situation. However, we could see that sadness category is the second highest emotion in this dataset. This means there are still have people who feel sad because of long COVID. The reasons could be they are the victims who suffering from long COVID, or their family, parents, partner, siblings, or friends are the victims of long COVID

Word cloud is used to visualize the most popular words used in each category of emotion.

Figure 7: Word cloud of analytical category

Figure 7 shows that words such as "effect", "think", "longterm", etc. are most frequently used in analytical category. Words like "fact", "think", "study", "understand", "cause", "impact" can indicate people's reasoning and analytical attitude about long COVID topic.

Figure 8: Word cloud of sadness category

Figure 8 shows that words such as “die”, “death”, “suffer”, “bad”, “lose”, etc. are most frequently used in sadness category. These words can indicate people’s feeling of loss and disadvantage.

Figure 9: Word cloud of neutral category

Figure 9 shows that words such as “new”, “case”, “test”, “health”, “day”, etc. are most frequently used in neutral category. These words can indicate people feel indifferent and nothing in particular.

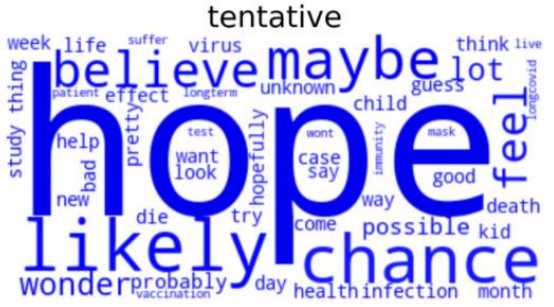

Figure 10: Word cloud of tentative category

Figure 10 shows that words such as “hope”, “likely”, “chance”, “believe”, “maybe”, “wonder”, “probably”, “hopefully”, “possible”, “unknown”, “guess”, etc. are most frequently used in tentative category. These words can indicate a person feels hesitant or unsure about something, might be questionable, doubtful, or debatable.

Figure 11: Word cloud of joy category

Figure 11 shows that words such as “good”, “great”, “thank”, “love”, “happy”, etc. are most frequently used in joy category. These words can indicate a person has shades of enjoyment, satisfaction, and pleasure. Joy brings a sense of well-being, inner peace, love, safety, and contentment.

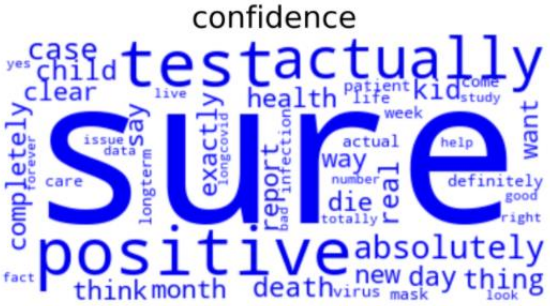

Figure 12: Word cloud of confidence category

Figure 12 shows that words such as “sure”, “actually”, “absolutely”, “exactly”, “completely”, etc. are most frequently used in confidence category. A person who feels confidence might be perceived as assured, collected, hopeful, or egotistical.

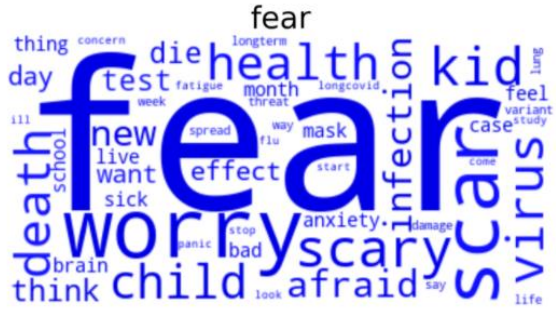

Figure 13: Word cloud of fear category

Figure 13 shows that words such as “fear”, “worry”, “scary”, “afraid”, “scar”, etc. are most used in fear category. These words can indicate a person has an unpleasant emotion caused by the threat such as danger, pain, or harm. It may be a mild caution or an extreme phobia.

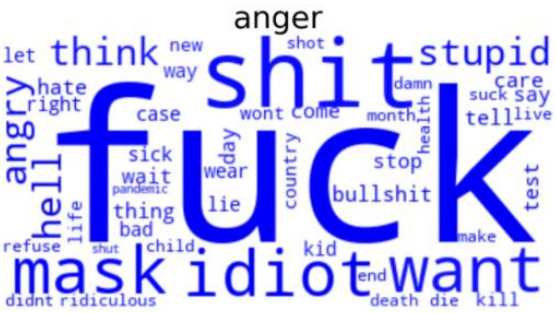

Figure 14: Word cloud of anger category

Figure 14 shows that words such as “fuck”, “idiot”, “shit”, “angry”, “hate”, “bullshit”, “stupid”, etc. are most frequently used in anger category. These words can indicate a person has a strong feeling of annoyance, displeasure, or hostility. This emotion may evoke due to injustice, conflict, humiliation, negligence, or betrayal. If anger is active, the person attacks the target, verbally or physically. If anger is passive, the person silently sulks and feels tension and hostility.

N-grams are widely used in text mining and natural language processing. It is a continuous sequence of N elements from a given sample of text. The N value must be an integer and is set to be a unigram n=1 (one or characters), a bigram n=2 (two words or characters), or a trigram n=3 (three words

or characters) (Majumder et al., 2022). In this project, we used N-grams to explore the top words in the dataset.

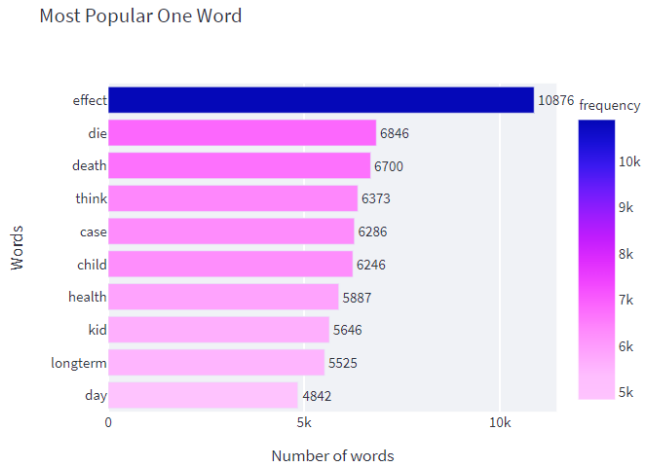

Figure 15: Top 10 most popular one word

Figure 15 shows the top 10 most popular one word in the dataset. Words like "effect", "die", "death", "think", and "case" are the most often used single word.

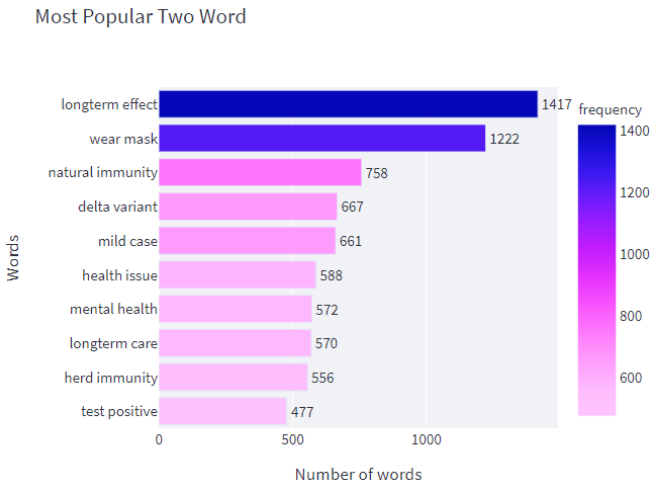

Figure 16: Top 10 most popular two words

Figure 16 shows the top 10 most popular two word in the dataset. Words like "longterm effect", "wear mask", "natural immunity", "delta variant", and "mild case" are the most often used two words.

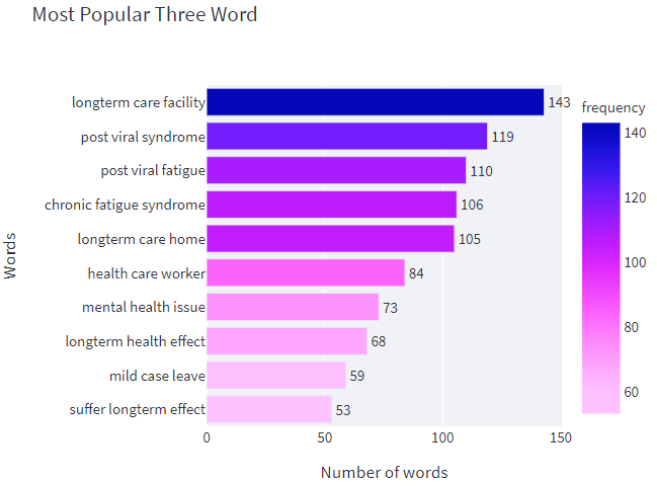

Figure 17: Top 10 most popular three words

Figure 17 shows the top 10 most popular three words in the dataset. Words like "longterm care facility", "post viral syndrome", "post viral fatigue", "chronic fatigue syndrome", and "longterm care home" are the most often used three words.

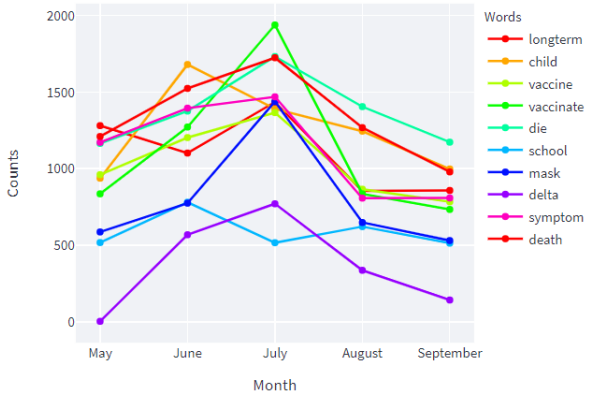

Figure 18: Trendy words across timeline

Figure 18 shows the trendy words across timeline. The line chart is plotted to highlight the changes in perception. The trendy words are extracted from the tweet text for all tweets in the dataset. The trendy words with emerging popularity are estimated based on the usage counts. From Figure 18, we can see that there is a small leap for "mask", "delta", "vaccinate". This can be related to the event happened in July 2021 which Delta variant continues to spread across every country, Brisbane and Australia announced to lockdown again as Delta variant continues to spread on July 31, Israel announced to give a third dose of COVID-19 vaccine to people over 60, becoming the first country to do so (Infoplease Staff, 2021).

5.2 Model Performance

In this section, a comparison of all classifiers is presented. This section will also show the web application which is the final product of this project. In Table 3,4,5,6 the comparison of all classifiers is based on the results of the accuracy, F1-score, precision, recall and Area Under Curve (AUC) model evaluation results. These results were achieved using the grid's best parameter. The dataset was split once it was collected and pre-processed: 80% of the data was allocated for training, while 20% was allocated for testing.

Table 2: Abbreviations Used in Table 3-6

MNB	Multinomial Naïve Bayes
LR	Logistic Regression
RF	Random Forest
KNN	K-Nearest Neighbors
SVM	Support Vector Machines
DT	Decision Tree

Table 3: Experiment result of Bag of Words (Unigram)

Bag of Words (Unigram)					
Model	Accuracy	F1-score	Precision	Recall	AUC
MNB	72.93	72.07	72.18	72.87	94.59
LR	**87.79**	**87.55**	**87.52**	**87.72**	97.69
RF	84.95	84.41	84.51	84.82	**98.04**
KNN	57.40	57.01	62.69	57.11	87.22
SVM	72.11	71.90	72.24	72.02	93.90
DT	81.40	80.72	80.58	81.27	89.35

Table 4: Experiment result of Bag of Words (Bigram)

Bag of Words (Bigram)					
Model	Accuracy	F1-score	Precision	Recall	AUC
MNB	75.38	74.49	74.58	75.30	95.68
LR	**77.89**	**78.19**	79.98	**77.73**	**96.16**
RF	76.37	76.78	**80.55**	76.15	94.88
KNN	52.07	51.78	63.18	51.81	82.61
SVM	70.06	69.79	70.79	69.93	91.20
DT	75.71	75.92	78.66	75.51	89.74

Table 5: Experiment result of TF-IDF (Unigram)

TF-IDF (Unigram)					
Model	Accuracy	F1-score	Precision	Recall	AUC
MNB	72.57	71.58	72.01	72.52	94.49
LR	**86.39**	**86.10**	**86.01**	**86.32**	97.43
RF	85.48	85.01	85.04	85.37	**98.20**
KNN	43.43	42.48	78.72	43.00	84.37
SVM	64.33	63.86	64.77	64.30	91.25
DT	81.63	80.96	80.81	81.49	89.45

Table 6: Experiment result of TF-IDF (Bigram)

TF-IDF (Bigram)					
Model	Accuracy	F1-score	Precision	Recall	AUC
MNB	75.39	74.34	74.60	75.31	95.61
LR	**77.76**	**77.57**	77.63	**77.65**	**96.13**
RF	76.27	76.42	**79.10**	76.07	94.64
KNN	52.09	51.38	61.95	51.80	82.26
SVM	68.11	67.91	74.46	68.21	91.68
DT	75.04	74.95	77.23	74.84	89.01

Only the confusion metric report and ROC curves of the best performing classifier which is Logistic Regression with BOW and unigram will be displayed:

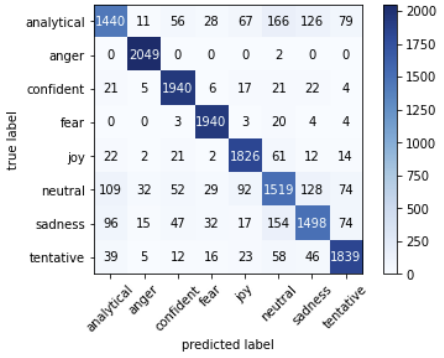

Figure 19: Confusion metric of LR with BOW and unigram

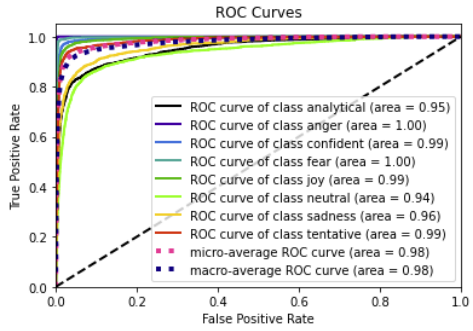

Figure 20: ROC curves of LR with BOW and unigram

By comparing the performance of the six classifiers through the use of N-grams, BOW and TF-IDF, Logistic regression with BOW and unigram obtained the highest accuracy rate at 87.79%, and scored 87.55%, 87.52%, 87.72% and 97.69% for F1-score, precision, recall, and AUC, respectively. The classifier that showed the least accuracy was KNN with TF-IDF and unigram, with a rate of 43.43%. Thus, the best classifier which is Logistic regression with BOW and unigram is deployed in the web application.

5.3 Web Application

The frontend and backend of the web application is built using Streamlit framework and hosted on Heroku platform. The link of the web application is https://long-COVID-emotion-analyzer.herokuapp.com/ The web application consists of few pages and each of the pages provide different functionalities:

Home Page: The landing page for users who access to the website is shown in the Figure 21. This page consists of an emotion analyzer which is the trained machine learning model, Logistic Regression with BOW and unigram. The emotion analyzer allows users to input their text and predict the emotion.

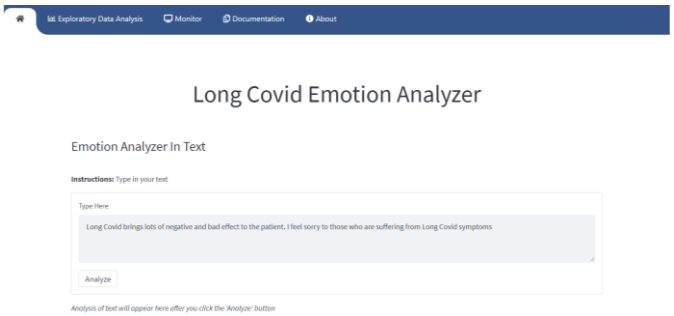

Figure 21: Home page

Exploratory Data Analysis Page: The page that consists of a table which shows the dataset, an interactive view of sample tweets for each emotion categories, and a few interactive visualizations that could help in understanding more about the dataset. The page is shown in Figure 22.

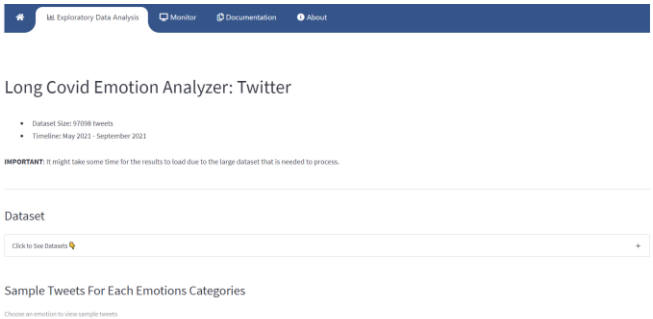

Figure 22: Exploratory Data Analysis page

Monitor Page: The page that collects the inputs text data in emotion analyzer from user and display the past analyzed text entered by the user in "Home" page and the results as shown in Figure 23.

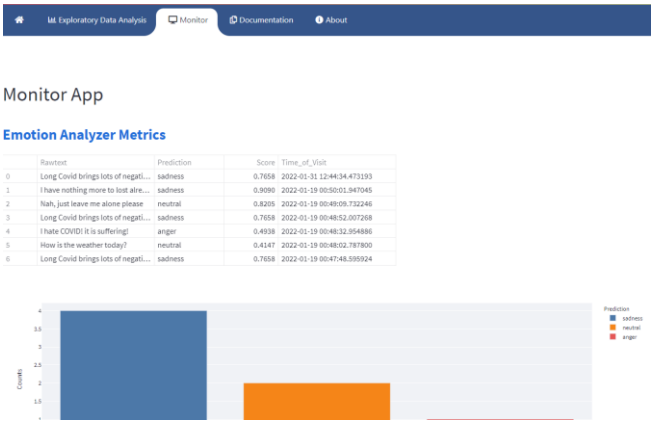

Figure 23: Monitor page

Documentation Page: The page that consists of documentation and details of each page as shown in Figure 24.

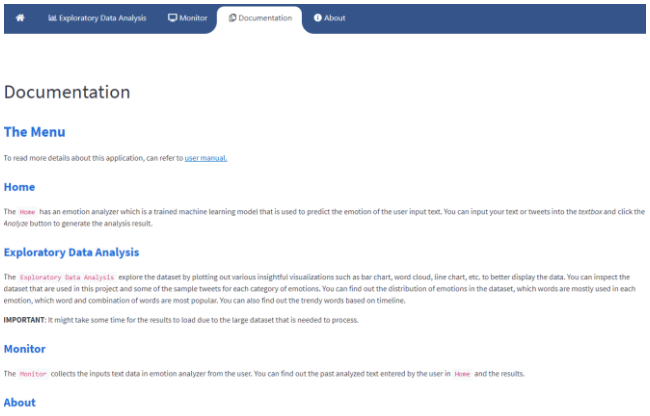

Figure 24: Documentation Page

About Page: The page that consists of the information and details about this web application as shown in Figure 25.

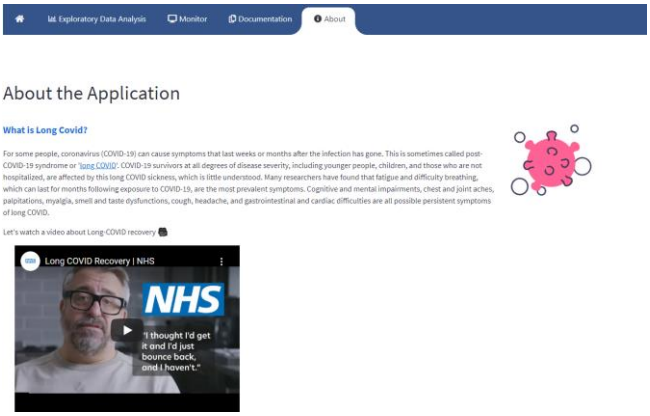

Figure 25: About page

6 DISCUSSION

The findings of emotional analysis and all supervised machine learning models can be seen in previous mentioned figures and tables. The machine learning classifiers perform greatly in terms of accuracy, F1-score, precision, recall, and AUC, except k-Nearest Neighbors which always perform the worst in all result. Logistic Regression outperformed the other classifiers on all methods. Overall, Logistic Regression with Bag of Words and Unigram is the best performing models with accuracy, F1-score, precision, recall, and AUC of 87.79%, 87.55%, 87.52%, 87.72% and 97.69%. The second-best model is Logistic Regression with TF-IDF and unigram with accuracy, F1-score, precision, recall, and AUC of 86.39%, 86.10%, 86.01%, 86.32%, and 97.43%. While the third-best model is Random Forest with TF-IDF and unigram with accuracy, F1-score, precision, recall, and AUC of 85.84%, 85.01%, 85.04%, 85.37% and 98.20%. Besides that, we found that bi-grams of both word embedding methods which are BOW and TF-IDF performed poorly compared to unigram for this dataset.

The most common challenge of emotion analysis is that we cannot ignore the importance of the processing of natural languages. The granularity of the data set, which is created after NLP in the case of emotion analysis, is directly related to the experiment's accuracy and performance (Gupta, 2020). Many irregularities, diversity, and subjectivity in the dataset are needed to be tackled while dealing with the natural language. There is also has limitation which the data generated on Twitter keeps evolving with new abbreviations and emojis. For example, “tbh” which is frequently used lately means “To Be Honest”, “fyi” means “For your information”, “wfh” means “work from home”. There is also some noisy words which is hard to clean such as “haaapppppyyyy” which means “happy”. Although we already reclean the dataset for second round by removing some new customize stop words, there are still have some underlying stop words which can be removed before model training to improve the model performance. We do not consider the emoticons and hashtags of the tweets because it can obstruct the efficiency of the model. For example, users can post sadness tweets with happy emoticons or hashtags.

Since Twitter become one of the biggest platforms for people to express their emotions, sarcasm has become the most popular way of expression now. Another limitation of the model is it unable to detect the emotion underlying the sarcastic tweets, resulting in predict the wrong emotion. Positive sentences with terrible feelings, or vice versa, are examples of sarcastic viewpoints. Training a model with some of the most often used sarcastic words might lead to improved emotion analysis in the future.

7 CONCLUSION

In conclusion, based on the results discussed, people are having an analytical emotion towards the long COVID topic. After pre-processing the dataset, six different supervised machine learning text classifiers were used which are Multinomial Naïve Bayes, Logistic Regression, Random Forest Classifier, K-Nearest Neighbors, Support Vector Machine, and Decision Tree. These algorithms were used to analyse and predict the emotion from the tweets. Furthermore, the models have been evaluated in the aspects of accuracy, F1-score, precision, recall, and AUC. The greatest performance of this project was obtained by LR with BOW and unigram with an accuracy of 87.79%. In terms of future work, the performance of the machine learning model might be improved by increasing the size of the collected dataset, adding new customize stop words, testing more classifiers or deep learning, and testing different types of feature extraction such as Word2Vec and GloVe.

ACKNOWLEDGEMENTS

I hope that this project will benefit the community by sharing the useful information. I hope that this project can also inspire people to expand the long COVID emotion analysis research scope, especially in computer science field, because there is limited study about long COVID topic in computer science. Lastly, I would like to thank Dr. Vimala Balakrishnan for being my supervisor in this project. She has given her support and guidance to me in completing this project.

REFERENCES

An, Y., Sun, S., & Wang, S. (2017). Naive Bayes classifiers for music emotion classification based on lyrics. Paper presented at the 2017 IEEE/ACIS 16th International Conference on Computer and Information Science (ICIS).

Banda, J. M., Tekumalla, R., Wang, G., Yu, J., Liu, T., Ding, Y., . . . Chowell, G. (2021). A large-scale COVID-19 Twitter chatter dataset for open scientific research—an international collaboration. Epidemiologia, 2(3), 315-324.

Callard, F., & Perego, E. (2021). How and why patients made Long COVID. Social Science & Medicine, 268, 113426.

Centers for Disease Control and Prevention. (2021, September 16). Post-COVID Conditions. Retrieved from https://www.cdc.gov/coronavirus/2019-ncov/long-term-effects/index.html

Chen, E., Lerman, K., & Ferrara, E. (2020). COVID-19: the first public coronavirus Twitter dataset. arXiv e-prints. arXiv preprint arXiv:2003.07372.

Da Silva, N. F., Hruschka, E. R., & Hruschka Jr, E. R. (2014). Tweet sentiment analysis with classifier ensembles. Decision Support Systems, 66, 170-179.

Desmet, B., & Hoste, V. (2013). Emotion detection in suicide notes. Expert Systems with Applications, 40(16), 6351-6358.

Devgan, S. (2021, March 3). Twitter Statistics 2021. Retrieved from https://statusbrew.com/insights/social-media-statistics/#twitter-marketing-statistics

Ekman, P. (1992). An argument for basic emotions. Cognition & emotion, 6(3-4), 169-200.

Gupta, P., Kumar, S., Suman, R., & Kumar, V. (2020). Sentiment Analysis of Lockdown in India During COVID-19: A Case Study on Twitter. IEEE Transactions on Computational Social Systems.

Gupta, U., Chatterjee, A., Srikanth, R., & Agrawal, P. (2017). A sentiment-and-semantics-based approach for emotion detection in textual conversations. arXiv preprint arXiv:1707.06996.

Hasan, M., Rundensteiner, E., & Agu, E. (2014). Emotex: Detecting emotions in twitter messages.

Hassan, A. U., Hussain, J., Hussain, M., Sadiq, M., & Lee, S. (2017). Sentiment analysis of social networking sites (SNS) data using machine learning approach for the measurement of depression. Paper presented at the 2017 International Conference on Information and Communication Technology Convergence (ICTC).

Hu, B., Guo, H., Zhou, P., & Shi, Z.-L. (2021). Characteristics of SARS-CoV-2 and COVID-19. Nature Reviews Microbiology, 19(3), 141-154.

Infoplease Staff. (2021, July 7). July 2021 Current Events: World News. Retrieved from https://www.infoplease.com/july-2021-current-events-world-news#:~:text=Haiti%20President%20Assassinated,Lebanon%2C%20South%20Africa%2C%20and%20Cuba

Kabir, M. Y., & Madria, S. (2021). EMOCOV: Machine learning for emotion detection, analysis and visualization using COVID-19 tweets. Online Social Networks and Media, 23, 100135.

Kulai, A., Sankhe, M., Anglekar, S., & Halbe, A. (2021). Emotion Analysis of COVID Tweets using FastText Supervised Classifier model. Paper presented at the 2021 International Conference on Communication information and Computing Technology (ICCICT).

Majumder, P., Mitra, M., & Chaudhuri, B. (2002). N-gram: a language independent approach to IR and NLP. Paper presented at the International conference on universal knowledge and language.

Müller, T., Cotterell, R., Fraser, A., & Schütze, H. (2015). Joint lemmatization and morphological tagging with lemming. Paper presented at the Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing.

Ortony, A., Clore, G. L., & Collins, A. (1990). The cognitive structure of emotions: Cambridge university press.

Perikos, I., & Hatzilygeroudis, I. (2016). Recognizing emotions in text using ensemble of classifiers. Engineering Applications of Artificial Intelligence, 51, 191-201.

Plutchik, R. (1982). A psychoevolutionary theory of emotions. In. New York: Harper and Row.

Posner, J., Russell, J. A., & Peterson, B. S. (2005). The circumplex model of affect: An integrative approach to affective neuroscience, cognitive development, and psychopathology. Development and psychopathology, 17(3), 715-734.

Sailunaz, K., Dhaliwal, M., Rokne, J., & Alhajj, R. (2018). Emotion detection from text and speech: a survey. Social Network Analysis and Mining, 8(1), 1-26. doi:10.1007/s13278-018-0505-2

Salam, S. A., & Gupta, R. (2018). Emotion Detection and Recognition from Text using Machine Learning.

Scherer, K. R. (2005). What are emotions? And how can they be measured? Social science information, 44(4), 695-729.

World Health Organization. (2021). Coronavirus disease (COVID-19) pandemic. Retrieved from https://www.who.int/emergencies/diseases/novel-coronavirus-2019

Yong, S. J. (2021). Long COVID or post-COVID-19 syndrome: putative pathophysiology, risk factors, and treatments. Infectious Diseases, 1-18. doi:10.1080/23744235.2021.1924397.