SENTIMENT ANALYSIS OF TWITTER DATA TO ANALYZE THE EFFECT OF COVID-19

by

Adeola Ayandeyi

A project report submitted in conformity with the requirements
for the degree of Master of Science, Information Technology

Department of Mathematical and Physical Sciences
Faculty of Graduate Studies
Concordia University of Edmonton

SENTIMENT ANALYSIS OF TWITTER DATA TO ANALYZE THE EFFECT OF COVID-19

Adeola Ayandeyi

Approved:

Supervisor Date

Committee Member Date

Dean of Graduate Studies: Rorritza Marinova, Ph.D. Date

SENTIMENT ANALYSIS OF TWITTER DATA TO ANALYZE THE EFFECT OF COVID-19

Adeola Ayandeyi
Master of Science, Information Technology
Department of Mathematical and Physical Sciences
Concordia University of Edmonton
2021

Abstract

Coronavirus pandemic has caused major changes in peoples' personal and social lives. The psychological effects have been substantial because it has affected the ways people live, work, and even socialize. It has also become major discussions on social media platforms as people showcase their opinions and the effect of the virus on their mental health particularly. This pandemic is the first of its kind as humans has never encountered anything like this virus and this is due to its airborne characteristics which leads to social distancing. Before the virus surfaced, some countries of the world were dealing with mental health cases, with over 40 percent of adults in the USA reported experiencing mental health challenges, including anxiety and depression. Social media has become one of the major sources of information due to information sharing on a very large scale. Peoples' perception and emotions are also portrayed through their conversations. In this research work, the interaction and conversation of people on social media most especially Twitter as it relates with the pandemic and its effect on mental health, will be analyzed using machine learning tools and algorithms. This analysis will help suggest the area of concentration to medical practitioners in order to speed up the diagnosis and recovery process and reduce the mental health issues which has escalated due to the virus.

Keywords:Coronavirus, COVID-19, mental health, Sentiment analysis, Twitter data, Data Preprocessing, machine learning classifiers, Supervised learning, Semi-supervised learning, performance metrics

Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Background

Coronavirus, a unique virus took the world by surprise with the first case noted in November 2019 in Wuhan, China and successfully spread to all nations of the world. Since its inception, the virus created a pandemic which has been a major conversation on akmost all microblogging sites. The negative effect of this virus on every aspect of human existence was enormous and it became a major issue to the social interaction of human beings due to its air borne nature. The virus has claimed millions of lives in different countries, United states of America topping the chart with 621293 deaths as at 4th of July 2021, with Brazil taking the second place with 524,417 deaths with increase in death of 718. India makes the third place with 402,757 deaths with increasing daily deaths of 742 [1]. The total number of confirmed cases worldwide as at this date is 184,498,556 with total recovered being 168,785.185. The total deaths recorded is about 3,991,905. These numbers throw the world into chaos as humans has never seen or experienced a virus like this. There were a lot of isolation, school closure, job losses, illness, mental breakdown, loss of income and depression. There was complete shutdown with no movement of some nations in order to understand the nature of this virus and be able to provide adequate solution to eliminate or minimize the spread. There were a lot on restrictions put in place restricting people from travelling and connecting whether for business, networking or to meet with loved ones. Although the increase in confirmed cases and deaths are no longer at its peak, its effect on human life and activities has taken a drastic turn. Social distancing has been inculcated into daily activities, unemployment increased drastically, school closure is still very paramount in most nations and the economy of most nations are suffering. All these have in more ways than one, has had adverse effect on human beings mental health. Humans are social beings and therefore the effect of this virus has led to

loneliness, depression and increase in mental health cases. The recent survey carried out by Census Bureau and the Centers for Disease Control and Prevention shows that coronavirus is associated with rapid rises in psychological distress across many nations most especially among women, the less educated and some minority ethnic groups like black Americans [2]. The fear of the virus has also triggered new mental illnesses which means that the measurable impact is greater than the actual number of casualties. Between April 2020 and March 2021, the google trend on Covid and its relationship with mental health shows how consistent the trend is and how worried people are. From the trend, the top five countries are North American countries, South Africa, Ireland and some parts of Pakistan.

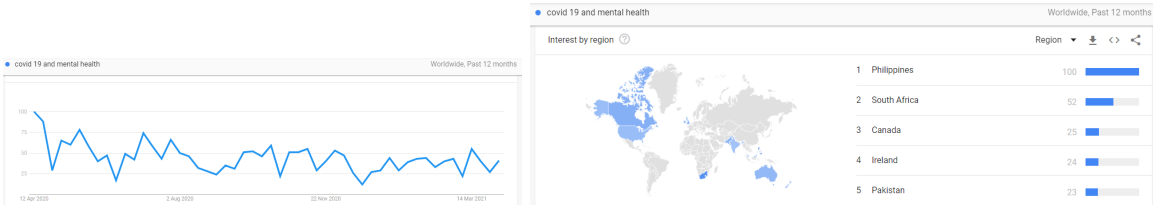


Figure 1.1: Google Trends on COVID-19 and in connection with Mental Health from Apr 2020 till Mar 2021

According to Canadian Mental Health Association, 42% of Albertans report isolation throughout the pandemic to be the top mental health concern for themselves and their community. 44% of rural Albertans report that a lack of socialization throughout the pandemic has become a mental health challenge. 57% of Albertans say staying connected with loved ones has helped their mental health and wellbeing during the pandemic [3]. A lot of conversations and emotional expressions on the virus and its effect on peoples' lives has becoming a very popular topic on social media. In this research work, we will be analyzing these conversions and expressions in order to determine the areas where people are most affected by the virus and help the medical practitioners and the government narrow down to the specific areas to look into in order to reduce the effect of the virus on peoples' lives.

To immune human beings from this virus, different companies came up with vaccines and some of the approved vaccines by World Health Organization are AstraZeneca manufactured by Oxford, BioNTech manufactured by Pfizer, Beijing manufactured by Sinopharm and Sputnik V [4]. About two hundred and twelve (212) countries commenced the vaccination of people against coronavirus. For a person to be considered fully vaccinated, It is required that 2 shot of the vaccine is administered with six weeks intervals. The first 10 countries with Canada ranked 15th with the most administered doses (at least one dosage) of approved vaccines [5] as at displayed in the Table 1.1.

Table 1.1: Vaccinaction against COVID 19 in 10 Countries

Country	Population	Total doses administered	Vaccines	Last Observation Date
China	1.44 billion	1.24 billion	CanSino, Sinopharm/Beijing, Sinopharm/Wuhan, Sinovac V	July 3, 2021
India	1.38 billion	329.16 million	Covaxin, Oxford/AstraZeneca, Sputnik V	July 3, 2021
USA	331 million	326.52 million	Johnson&Johnson,Moderna, Pfizer/BioNTech	July 3, 2021
Brazil	212.5 million	101.48 million	Oxford/AstraZeneca, Pfizer/BioNTech, SinovacV	July 1, 2021
United Kingdom	67.9 million	77.91 million	Moderna, Oxford/AstraZeneca, Pfizer/BioNTech	Jul. 2, 2021
Germany	83.7 million	74.87 million	Johnson&Johnson,Moderna, Oxford/AstraZeneca, Pfizer/BioNTech	July 1, 2021
France	67.5 million	53.79 million	Johnson&Johnson,Moderna, Oxford/AstraZeneca, Pfizer/BioNTech	June 30, 2021
Italy	60.4 million	51.58 million	Johnson&Johnson,Moderna, Oxford/AstraZeneca, Pfizer/BioNTech	July 3, 2021
Turkey	84.3 million	50.61 million	Pfizer/BioNTech, Sinovac	July 3, 2021
Canada	37.7 million	37.38 million	Moderna, Oxford/AstraZeneca, Pfizer/BioNTech	July 3, 2021

The complete table can be found **here** [5]. With the introduction of vaccination, we see the google trends on covid 19 with the effect on mental health starts to decrease from June 2021 as shown in Fig 1.2

Figure 1.2: Google Trends on COVID-19 and in connection with Mental Health from Jul 2020 till June 2021

As people get vaccinated, the world begins to open. Restrictions are being lifted and more people are getting their lives back. Isolation has reduced drastically, more schools are back on campus and more jobs are opening up for people.

1.2 Problem Statement

With this virus, there has been an increase in mental health issues. The various problems that came along with the virus, such as job loss, isolation, depression, loss of income, homeschooling, and a lot more have had quite a lot of mental stress and this illness has increased drastically. The government does carry out surveys from time to time to understand people's perception of the virus as it relates to mental health but this method is tedious, time-consuming, and expensive. Even with the survey, only a minute percentage of the population can be reached and this sample is too small to predict and help health care practitioners make proper diagnoses and create solutions for the eradication of this pandemic. With Twitter data, there is access to free large data. Although Twitter data is not reliable as there may be false tweets, false tweets can be detected and government can rely on this data from initial findings. These initial findings can help the government takes quick steps to avoid social and economic damage which could be enormous if time and adequate data are not put into perspective.
This research work takes into account people's opinions on Twitter (as a large community is created on social media which is limitless to race, region, or country) as people err their views and opinions on the effect of covid 19 on mental health. With machining learning, this research work tables different models to analyze peoples' sentiments and deduced the different avenues that health care workers can utilize for more accurate diagnosis and cure which eventually leads to a decrease in mental health.

1.3 Contribution of the thesis

- The research work involves the collection, cleaning, and analysis of data for the extraction of useful insights/information. The contribution to this project work involved the collection of Twitter data from January 2020 till May 2021. The data collected is about 30 gigabytes which runs into millions of rows. The features of the data were limited to the date, the id of the Twitter user, and the tweets(represented as text).
- The collected data is preprocessed which involved the removal of special characters, missing values, stopwords, reducing the data to about 10 million rows of clean data.
- With machine learning algorithm, feature selection was performed on the data to utilized only features (columns) that give more accurate analysis to the data. Not

all features of the data are relevant for data analysis. Adding irrelevant features to the analysis may result in less accurate analysis.

- Machine Learning classifers were performed on a portion of the data, the train data, to create a model that can be used to test the remaining data. About seventeen (17) classifiers were performed to create the models and the accuracies of each model was tested to determine which model has the best performance. The classifiers utilized includes: KNeighborsClassifier,Support Vector Machine, Gaussian Process Classifier, Decision Tree Classifier, Random Forest Classifier, MLP Classifier, Ada Boost Classifier, Linear Discriminant Analysis, Quadratic Discriminant Analysis, Gradient Boosting Classifier, Logistic Regression, Multinomial NB, SDG Classifier, LGBM Classifier and XGB Classisifer.

- For each classifier, the accuracy and precision were calculated for both supervised learning and semi-supervised learning. For supervised learning, the target feature is known which helps to train the remaining data. For semi-supervised learning, some of the data have the target variable while some don't. The data with the target variable serves as the train data with which the model is created. The model is tested on the test data, which automatically generated the target for the test (unlabelled) data.

1.4 Organization of the thesis

For this research work, there are six (6) chapters. Chapter 1 is the Introduction which provides a detailed background of the research work, explaining the reasons for research work, the problem statement, the prediction of Covid 19 data in Canada, and how the introduction of vaccines had helped calm the mental health of people. Chapter 2 carries the Literature Review and is sectioned into three. The first section discusses related works of sentimental analysis of Twitter data. Section two discusses the related work of sentimental analysis of other social media data asides from Twitter data while section three discusses related works that pertain to the sentimental analysis of Covid-19 Twitter data. Chapter 3, which is the methodology discusses in detail the three main methods for the research work which include feature selection, supervised learning, and semi-supervised learning. Chapter 4 is the Data collection and preprocessing which discussed in detail how data was collected, preprocessed to ensure it is cleaned enough for machine learning analysis. Chapter 5 talked about the results and discussions of the analysis. Every step of the analysis is broken down here to understand different insights that were extracted from the data. Chapter 6 concludes this research work

and also discussed the future works for further analysis.

Chapter 2

Literature Review

There are quite several researches that focuses on social media analysis. People's opinions and perspectives on social media are quite important due to the volume of information that can be extracted and analyzed to help provide relevant insights. Twitter data has been analyzed for different purposes which includes live traffic updates, real-time notifications such as large-scale fire emergencies and downtime on services provided by content providers [6]. In Business, analysing twitter data helps companies understand their competitors and consumers and can strategically position themselve to serve their consumers better than their competitor. The Twitter data analysis helps business understands how their present and perspective consumers think and can create the prducts to better meet the needs of their consumers in real time. In this reseach work, related researches are classified based on sentiment analysis of twitter data which are not related to COVID-19, Analysis of data from other social media as it relate to COVID-19 and the analysis od twiitter data relating to COVID-19. The following sections will discuss the Sentimental analysis of Twitter data, data from other social media platforms and Covid-19 twitter data

2.1 Twitter Data Sentiments

Agarwal et al acquired their data with distance learning technique and separated the data into positive and negative sentiments based on the emotions of the twitter user. They build models using different machine learning models and reported that Support Vector Machine (SVM) outperforms other classifiers. With feature selection, they utilized Unigram, and Bigram model in conjunction with parts-of-speech (POS) features. Their report shows that unigram model outperforms all other feature selection models [7]. With Twitter data HJ Do et al investigated people's emotional responses during the 2015 Middle East Respiratory Syndrome (MERS) outbreak in

South Korea. The result of their research in relation to MERS outbreak assisted with the understanding of the human behaviour and the characteristics of sociocultural system [8].

2.2 Other Social Media Data Sentiments

Li et al collected data from Weibo, China second largest social media platform and then used Natural Language Processing (NLP) on the datasets to classify the information into seven different types, such as help seek, emotional support, donations, cautions. This type of information helps governments, healthcare personnel, and individuals to diagnose and respond appropriately to the mental and personal health issues relating to the pandemic [9].
Wang et al carried out emotional analysis method on COVID-19 data collected from Sina Weibo ,a microblogging site in China using machining learning models such as Support Vector Machine, naïve Bayes and Random Forest classifiers. Extracting reviews of 184 hospitals from mouthshut.com, Ankita Bansal et al analysed the sentiment of these hospital reviews to provide relevant and important information about the operating conditions and current status of hospitals to the general public. Sentiment analysis applied to patient's reviews to quantify the direction and/or magnitude of the emotive content. Patient comments were segregated into different sections and analysed to quantify the positive or negative aspects of the reviews. With this analysis, the overall rating of the hospital, based on key parameters will help people to understand the hospital's current condition. This analysis births information which provides great value to the patients, giving them the power to make the best options of available hospitals [10].

2.3 Covid-19 Tweets Sentiments

With a lot of researches carried out on twitter data analysis, only a few of these reseaches focused on sentimental analysis of social media datasets on the COVID-19 pandemic. Koustuv Saha et al carried out a research on the psychosocial effects of the COVID-19 crisis by using social media data (Twitter) from 2020. In their research, they found out that people's mental health symptomatic and support expressions increased significantly during the COVID-19 period as compared to similar data from 2019 [11]. Jelodar et al with NLP automated the extraction of COVID-19–related discussions from social media to discover various issues related to COVID-19 from public opinions. They also investigated the use the DL-based Long Short-Term

Memory (LSTM) approach for sentiment classification of COVID-19 comments and discovered this approach produces better results than other well-known ML approaches [12]. Lyu, Chen, Wang, and Luo analyze millions of twitter data to identify two groups of users who use term "coronavirus" and the other terms "Chinese virus" or "Wuhan virus", and predict the number of users who are more likely to use later words than the former using machine learning models [13]. Amrita Mathur et al extracted information from twitter and classified the tweets into positive and negative sentiments. They further classified these tweets into six basic emotions given by Ekman i.e joy, sadness, anger, fear, disgust and surprise [14]. While analyzing twitter data, Man Hung et al identified different themes that relate to COVID-19 which are health care environment, emotional support, business economy, social change, and psychological stress. These themes dictate the effects of the pandemic on human mental and overall health [15]. Sohini Sengupta et al carried out an overview on the discussions about mental health as at June 2020 and the impact of covid-19 on the health issue. They also carried out an overal sentiment analysis to further understand the emotions of twitter users [16]. Jim Samuel et al present textual analyses of Twitter data to identify public sentiment, specifically, tracking the progress of fear, which has been associated with the rapid spread of Coronavirus and COVID-19 infections. Their work outlines a methodological approach to analyzing Twitter data specifically for identification of sentiments, key words associations and trends for crisis scenarios in relation to the current COVID-19 pandemic. Their disucssion centres around the search for insights with descriptive textual analytics and data visualization [17]. Anna Kruspe et al analyze Twitter messages (tweets) restricted to only European countries that were collected during the first months of the COVID-19 pandemic. These data were analyzed with a neural network for sentiment analysis using multilingual sentence embeddings. The results were separated by country of origin, and their temporal development were correlated with the events in those countries. The moods of the citizens were studied and they found out that the lockdown announcements correlate with a deterioration of mood in almost all surveyed countries, which recovers within a short time span [18]. Furqan Rustam et al conducted a research on the performance of various machine learning classifiers using proposed feature set which was created by concatenating the bag-of-words and term frequency-inverse document frequency. Twitter data were classified as positive, neutral, or negative sentiment. Performance of classifiers was evaluated on the accuracy, precision, recall, and F1 score. Further investigation was made on the dataset using the Long Short-Term Memory (LSTM) architecture of the deep learning model. With comparison with the machine learning classifiers, LSTM achieved lower accucary as Extra Trees Classifiers outperform all other models by

achieving a 0.93 accuracy score [19]. Harleen Kaur et al conducted a reseach work by designing an algorithm called Hybrid Heterogeneous Support Vector Machine (H-SVM) which was used to perform sentimental analysis of twitter data. This twitter data was collected based on hashtag keywords, including COVID-19, coronavirus, deaths, new case, recovered.The sentiment was classified into positive, negative and neutral sentiment scores. They compared the performance of the proposed algorithm on certain parameters like precision, recall, F1 score and accuracy with Recurrent Neural Network (RNN) and Support Vector Machine (SVM) [20].

Chapter 3

Mental Health Prediction

Analyzing data to extract using insights requires some methods and this research work is not an exception. The methodologies utilized can have a huge impact on the outcome or performance of the project work. The proposed architecture for mental health prediction with machine learning that depicts these methods is shown in Figure 3.1

3.1 Text Features

The data are usually in text data types that are not compactible with machine learning models. Machine learning models are only compactible with numeric values. There are techniques to convert these text data into numeric features without losing the meaning of the data. Some of the feature extraction techniques are mentioned below:

3.1.1 N-gram

N-gram is a sequence of words in a sentence. N-gram is probably the simplest concept in machine learning. There are quite some varieties of the usefulness of N-gram. It can be used for autocorrection of words, auto spell check, and also grammar checks [21]. It also helps to check the relationship between words especially when one is trying to figure out what someone is more likely to say to determine the emotions or sentiments through the word said. N-grams are the combination of words that are used together. Unigrams are N-grams with N = 1. For N = 2, these are referred to as bigrams and trigrams for N = 3. N-grams capture the structure of the language helping to determine which word is likely to follow a given word.

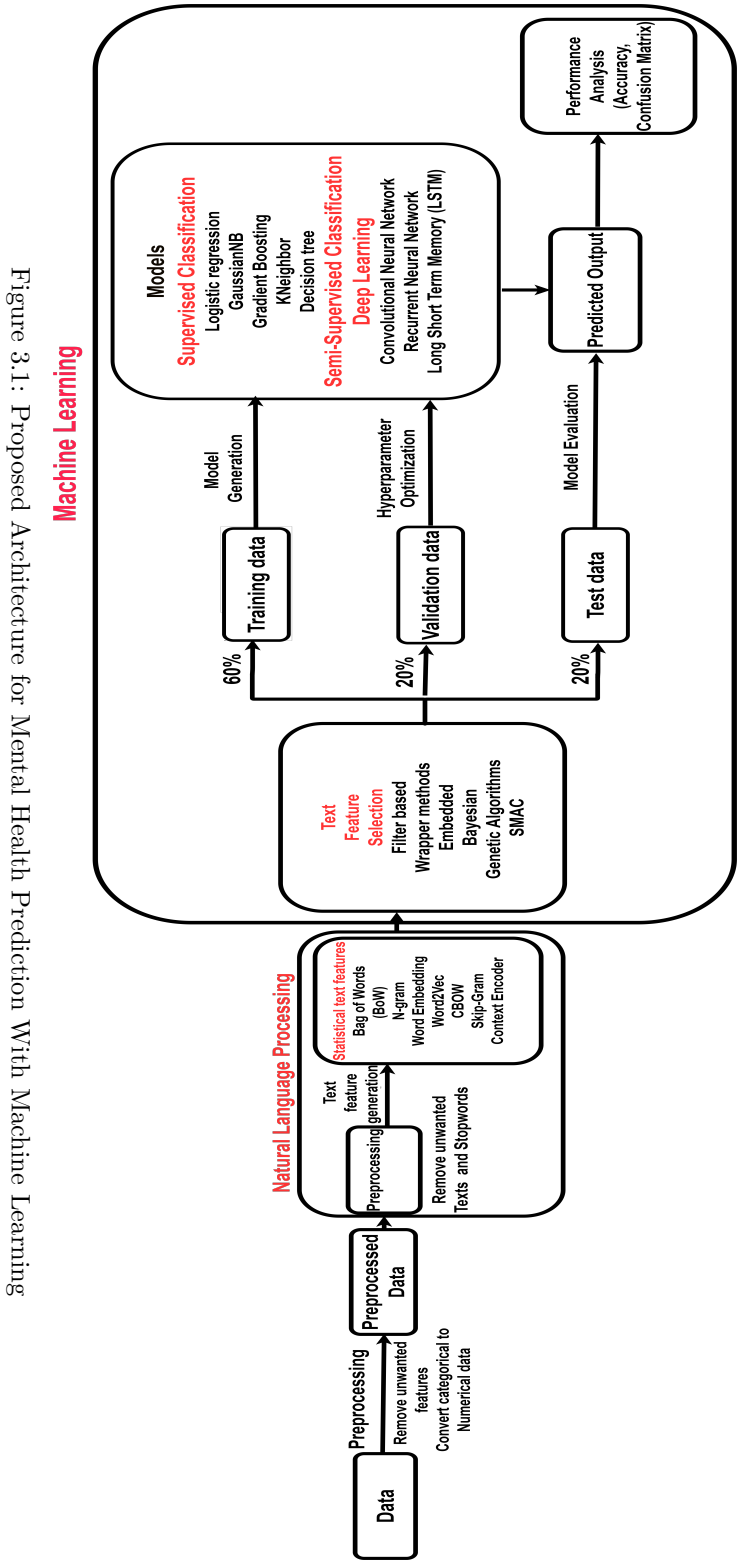


Figure 3.1: Proposed Architecture for Mental Health Prediction With Machine Learning

3.1.2 The Word2Vec Model

Word2vec is a method to efficiently create word embeddings. Created by Google in 2013, it is a predictive deep learning-based model to compute and generate high quality, distributed and continuous dense vector representations of words, which capture contextual and semantic similarity [22]. It is a type of unsupervised model that takes a large corpus of words, creates a vocabulary of words from it, and generates dense word embeddings for each word. The words are transformed into vectors to allow machine learning algorithms to perform algebra operations on numbers as against words. This transformation is referred to as word embedding [23]. With distributed Hypothesis in Word2Vec, the lexicon for a word is found in its neighboring words. A word can be predicted by looking at these close words.

3.1.2.1 Continuous Bag of Words

In the CBOW model, the architecture tries to predict the current target word (usually the word in the middle) based on the source context words which are the surrounding words [22], [24]. The corpus is built such that extraction of each unique word from the dictionary can be done and mapped to a unique numeric identifier. The major python module used is Kera preprocessing. After this, the CBOW generator is built with two variables: context and target. Keras and Tensorflow are used to build the deep learning architecture for the CBOW model. This model tends to do better with smaller datasets. It is very fast to train the model and provide better accuracy.

3.1.2.2 Continuous Skip-Gram Model

This is the inverse of CBOW as it predicts the surrounding words from the current target words. This model does better with a larger dataset. The objective is to predict the contexts of a given word. To implement this model, the corpus dictionary is built such that each unique word can be extracted from the dictionary and assigned a unique identifier. The mappings to transform words to and from their unique identifiers are also maintained. The skip-gram generator is built next which will provide the pair of words and their relevance. To build the skip-gram model, Keras on top of TensorFlow is taken advantage of to build it. The model is trained and the embedded words are retrieved [22].

3.1.3 Bag of Words

Bag of words is a type of feature extraction or feature encoding used to extract features from the text. It is called Bag of words because the order with which the text appears

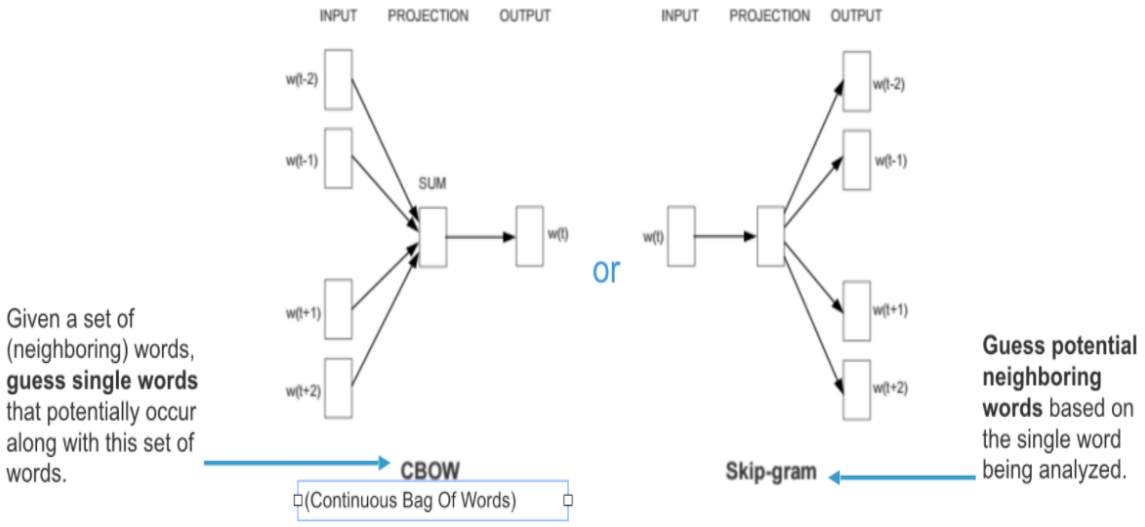


Figure 3.2: CBOW and Continuous Skip-Gram Model Architectures [23]

is irrelevant and discarded. It is only concerned with whether or not the known word occurs in the text, not the position in the text. Each document of free text is converted into a vector which can be used as input or output for a machine learning model. In a python programming language, the module for BOW is referred to as CountVectorizer. It transforms a given text into a vector-based on the frequency (count) of each word that occurs in the entire text. See documentation **here**. Before generating vector representation of words, CountVectorizer pre-processes the text data. CountVectorizer creates a matrix in which each unique word is represented by a column of the matrix, and each text sample from the document is a row in the matrix [25].

3.2 Supervised Learning

This is a type of Machine learning where a dependent variable can be predicted based on one or more independent variables. For example, suppose we want to predict whether a bank customer will pay a loan or not (which is represented as a variable called status) based on its loan amount, duration, and demography, the status variable here is the dependent variable whilst loan amount, duration and demography are the independent variables. The independent variables, usually represented with X are also known as input variables while the dependent variable, usually represented by Y is also known as output variables. As an illustration, there is a dataset with different animals which are dogs, cats, horses, and lions. A subset of the data that are properly labeled is given to the machine learning model to understand. After

this, the remaining data (test data) is given to the model to sort. Because the model already understands the features of the different animals, it can sort out which animal is which using the test data accurately. Based on the independent variables, supervised machine learning can predict the dependent variable. Supervised learning is generally a classification problem where the target variable (also the dependent variable) is a categorical data type to determine which category a data is. Some of the algorithms to create supervised learning models under supervised machine learning are LogisticRegression, Random Forest, Decision Tree, KNN, Linear Support Vector Machines, Non-Linear Support Vector Machines, Naive Bayes Theorem, and many more.

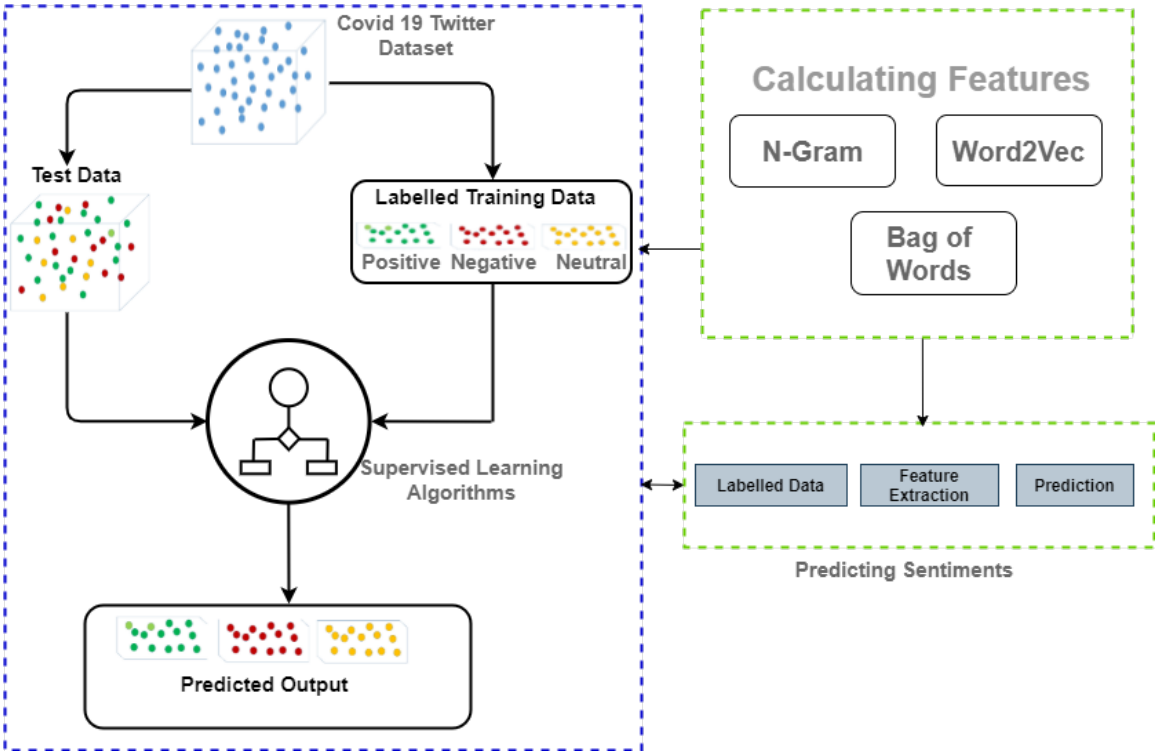


Figure 3.3: Supervised Learning

3.3 Semi-supervised Learning

Semi-Supervised Machine learning is a type of machine learning where a large portion of the data is unlabelled (input data) and a small portion of the data are labeled (output data). This data is used to train the model. This type of machine learning falls between unsupervised and supervised machine learning. Most real-world data falls into this category. It trains its model with pseudo labeling which can combine many neural network models and training methods [26]. The model is trained with

the small portion of the data that is labeled just like the case of supervised learning until a good result is obtained. The unlabelled training data (which are pseudo labels) are used to predict the outputs. This may not be accurate. The labels of the labeled training data are linked with the pseudo labels. Their inputs of both the labeled training data and the unlabelled data are linked as well. The model is re-trained again as earlier to decrease errors and improve the accuracy of the model. Figure 3.4 depicts the Semi-Supervised learning illustration.

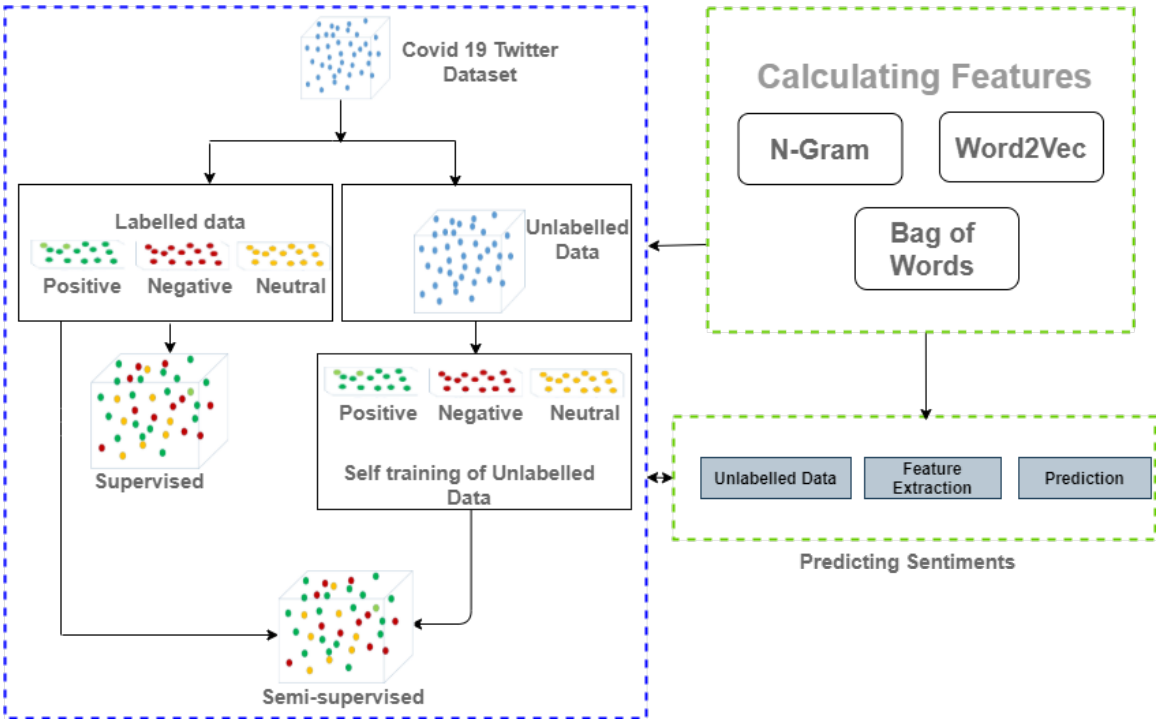


Figure 3.4: Semi-Supervised Learning

Chapter 4

Data Collection and Preprocessing

4.1 Data Collection

In recent times, human interaction is centered around the web and as such, individuals tend to communicate how they feel over social media such as Facebook, Twitter, TikTok, and other social platforms. Twitter information tweeted by individuals pertaining to the pandemic was extracted for this research work to dissect the feelings of each individual.

Several types of research were done on the collection of Twitter data. To access Twitter data, the user must have a Twitter developer account. This is done by applying for this account and provide the required information for approval. Upon approval, which usually takes twenty-four to forty-eight hours, the user's profile is created, and API access is given. From the Twitter developer account profile, the user can retrieve the access token, access token secret key, consumer key, and consumer secret key. Without these keys, it is impossible to access Twitter data. The data extracted is usually in a JSON file. The JSON file is usually a mix of 'root-level attributes and child objects. Some of the root-level attributes are stated in the table below:

Other attributes can be found in this **link**.

There are several ways of retrieving Twitter data. For this research work, the focus is on python modules that can be used for tweets extraction some of which are:

4.1.1 Tweepy

With this module, tweets can be retrieved with a StremListener. The API is accessed with tweepy's OAuthHandler. The consumer key and consumer secret key are supplied to the OAuthHandler to allow Twitter to give authorization for data access. Here, specific information that the user desire to access can be stated, and only the tweets with this information are retrieved. For example, if a user desires to retrieve only

Table 4.1: The Root-level and Child Atrributes of Twitter Data

Attribute	Type	Description
created_at	String	UTC time when this Tweet was created. Example: "created_at": "Wed Oct 10 20:19:24 +0000 2018"
Id	Int64	The integer representation of the unique identifier for the Tweet. Using a signed 64 bit integer for storing this identifier is safe as some programming languages may find it difficult to detect it.
text	String	This is the text or conversation by the twitter user.
User	User Object	The user who posted the Tweet. User data dictionary has the complete list of attributes see **link** for details
retweet_count	Int	Number of times this tweet has been retweeted
lang	String	This detect the language of the tweet

information about covid-19, information such as coronavirus, corona, covid19, covid, social distancing, etc. can be added to a list to filter only the tweets with the information. The documentation for tweepy can be found here (www.tweepy.org). To use tweepy in python, install with the line of code below:

'pip install tweepy'.

The disadvantage of tweepy is that data can only be extracted for the last seven days from the date of extraction. Other twitter modules for scraping twitter data are twitter-scaper. For this research work, data was extracted with a python module called twarc.

4.1.2 Twarc

This is a command tool and a python module for retrieving and archiving Twitter data as JSON files. Its installation is with the line of code stated below:

'pip install twarc'

. As a command tool, it can be executed on PowerShell subject to the prior installation of python 2.7 or a higher version. Once installed, it must be configured. To configure twarc, the line of code below is typed in Powershell or command prompt.

'twarc configure'

The secret keys are requested for authorization. Once 'Authorize app is selected, the Twitter app is connected to twarc to utilize the APIs for tweets extraction. After

all the instructions have been followed, the tweets extraction can be done. For this research work, tweet ids of Twitter users that converse on the pandemic and its relationship with mental health were accessed from the Zenodo database. The tweet ids are categorized monthly so that the data are grouped in such a way as well. With the tweet ids, the conversations (tweets) by these ids are extracted with the command line stated below:

'twarc hydrate tweet_ids.txt > tweets_hydrated. jsonl'.

All the root level and child attributes of each tweet are extracted, and the data extracts all information as long as the tweet id is active on Twitter. To extract the required columns from the file, which are the id, text, date, and location, a git bash command is opened, and the line of code below is executed.

$$\text{awk -F "," "}\{\text{print \$1 \$2 \$4 \$16}\}\text{" > output.txt}$$

where $1,$2, $4 and $16 are the positions of text, id, date and location.
The JSON file is converted to text file. The text file is read into python to create a dataframe the id, text, date and location are separated into different columns. The python code named 'extract combine and convert to dataframe' is found **here**. The dataframe with three columns (id, text and date) is saved as a text file and this represents that unstructured data that is preprocessed in the next section.

4.2 Data Preprocessing

Data Preprocessing is the most important step of data mining which deals with the transformation and preparation of datasets for knowledge extraction [reference 2 data preprocessing]. There are several techniques involving in preprocessing. Some of them are cleaning, integrating, transforming, and reducing the dataset. This results in structured/clean data that are useful for modeling. Data collected or extracted from different sources are usually in their raw format which is not feasible for analysis, therefore, the raw data must first be cleaned before analysis. For all analysis projects, data cleaning takes about 70% of project work. It is cumbersome but extremely unavoidable. For this project work, data was extracted from January 2020 to April 2021. The data extraction was limited to the tweet id, the date of tweets, and the text. The focus is on the text as that is the conversations to be analyzed. The uncleaned data was preprocessed to ensure the data is clean enough for model acceptance. The data contains three columns; the id, text, and the month of the tweet. A new column is created where the preprocessed data is stored named tidy_text. The preprocessing

is done of column text. Some of the preprocessing done on our raw data are stated below:

4.2.1 Remove the urls from the text

To remove the urls from a text, a function as shown below is written in python and applied to the text.
def remove_url(row):
txt = str(row['tidy_text']).split('https')[0]
return txt
data['tidy_text'] = data.apply(remove_url, axis = 1)

4.2.2 Remove the special characters

Some regular expressions were expressed to remove the special characters as stated in the code below
data['tidy_text'] = data.tidy_text.str.replace("?!,\ &:;%()", " ", regex=True)

4.2.3 Remove all usernames with @

The following line of code can be used to remove and replace usernames.
data['tidy_text'] = data['text'].str.replace('@[\w:]*',")

4.2.4 Convert to lowercase

All the characters are converted to lowercase.The line of code below convert the characters into lowercase to avoid dulpication of words with different cases.
data['tidy_text'] = data['tidy_text'].apply(lambda x: x.lower())

4.2.5 Remove numbers from characters

Here, all numeric values are removed with the regular expression (regex) pattern \d+. The addition sign ensures multiple numbers such as 10 are interpreted as such and not interpreted as two separate numbers. The line of code below removed all numeric values from the text.
data['tidy_text'] = data['tidy_text'].str.replace('\d+',"").

4.2.6 Drop Null Values

Some unstructured/raw data, there may be missing values. These values are sometimes called null values. They are usually filled up with the most common words or are

dropped completely.If null values are not dealt with, it will affect our models are machine models don't accept null values. In our preprocessing, we dropped the null values with the line of code stated below:

data['tidy_text] = data['tidy_text].dropna(inplace = True)

4.2.7 Removal of stopwords

Stopwords are words that do not add meaning to a sentence and therefore can be ignored or removed without tampering with the meaning of the sentence.[reference stopwords]. Stopwords are found in most languages but for the purpose of this project work, stopwords in English are utilized. For sentimental analysis, Stop words are to be removed from the data to keep only the root words. Common stopwords include [i,me,my,myself,we,our,ours,ourselves,you,your]. A list of common stopwords (in english) can be found **here**. With these data, stopwords are removed by importing stopwords from the corpus of nltk, the python module for natural language processing.

```
import nltk
from nltk.corpus import stopwords
stop_words = set(stopwords.words('english'))
```

Using the lambda function, apply the stopwords to the text to remove them. Also, extract words that have 4 characters and above in order to filter words that are meaningful to the research work.

```
data['tidy_text'] = data['tidy_text'].apply(lambda x: ' '.join([w for w in x.split() if w not in stop_words]))
data['tidy_text'] = data['tidy_text'].apply(lambda x: ' '.join([w for w in x.split() if len(w)> 4]))
```

Below is the top 10 data showing the id, month, the uncleaned text and the tidy text

The complete python code for data preprocessing can be accessed **here**.

Out[29]:

	id	text	month	tidy_text
0	1381308761533452291	Le #Portugal a d\u00e9cid\u00e9 de suspendre ...	Apr	portugal decide suspendre provenance bresiln c...
1	1381308765841002501	Tommorrow is an exciting day as the Island mo...	Apr	tommorrow exciting island moves stage reconnec...
2	1381308766482751493	Nos hacemos acopio de esta noticia publicada ...	Apr	hacemos acopio noticia publicada informacifn i...
3	1381308768764370945	401 people	Apr	people
4	1381308768898576387	@mybmc UK\u00a0Clinical Trial Confirms SaNOti...	Apr	ukaclinical trial confirms sanotizes breakthro...
5	1381308770404466688	Th\u00eam 1 ca\u00a0COVID-19	Apr	theam caacovid-
6	1381308773944414209	Matthew Hancock MP	Apr	matthew hancock
7	1381308773998936074	@xotep my co-worker got a cert saying he got ...	Apr	co-worker saying vaccine conformed covid-
8	1381308775118766087	Stay Safe Mumbai! Masks on and battle against...	Apr	mumbai masks battle covid support government m...

Figure 4.1: The Unstructured and Pre-processed data

Chapter 5

Results and Discussions

The extraction of data from Twitter is carried out and the data is preprocessed as discussed in Chapter 4. The preprocessed data is fed into different types of natural language processing models for statistical text features which are described in chapter 3. The data is further fed into different text feature selection methods to select the best features from the datasets that will provide relevant and accurate results. Feature Selection is the process of filtering irrelevant from the dataset. If the right subset is chosen, it improves the accuracy of the model and helps the model train its data faster. Some of the feature selection methods utilized are Filter Methods, Wrapper Methods, and Embedded Methods.

5.1 Data Exploration

Data was collected from January 2020 till May 2021. After cleaning, the total row of data is 12.5 million rows.

The frequency of the tweets is visualized on a monthly basis to see how much discussion is made on Covid-19 and its effect on mental health. Figure 4.2 displayed below shows the dynamics of the conversations. The conversation is quite steady between January 2020 and April 2020. Quite a lot of conversation was going on during this period but seem to be centered around major effects of the pandemic which includes loss of jobs, homeschooling, and isolation. With the invention of vaccines, the conversation increased drastically.

The conversation increased drastically in February 2021 as more countries are opened to the vaccination and quite a lot of incentives were introduced especially in North American countries to further encouraged their citizen to receive the vaccine so as to keep people safe and more importantly open the countries to international communities. The conversation during these periods was centered around covid-19,

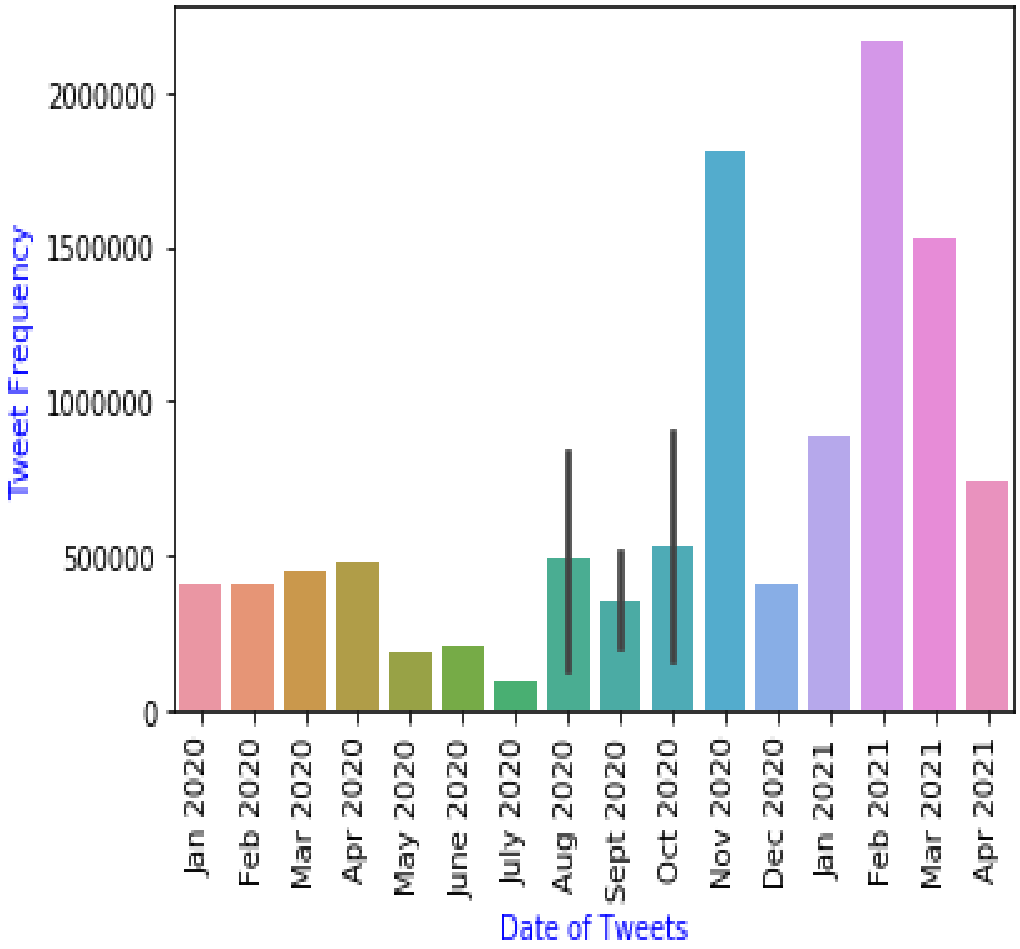
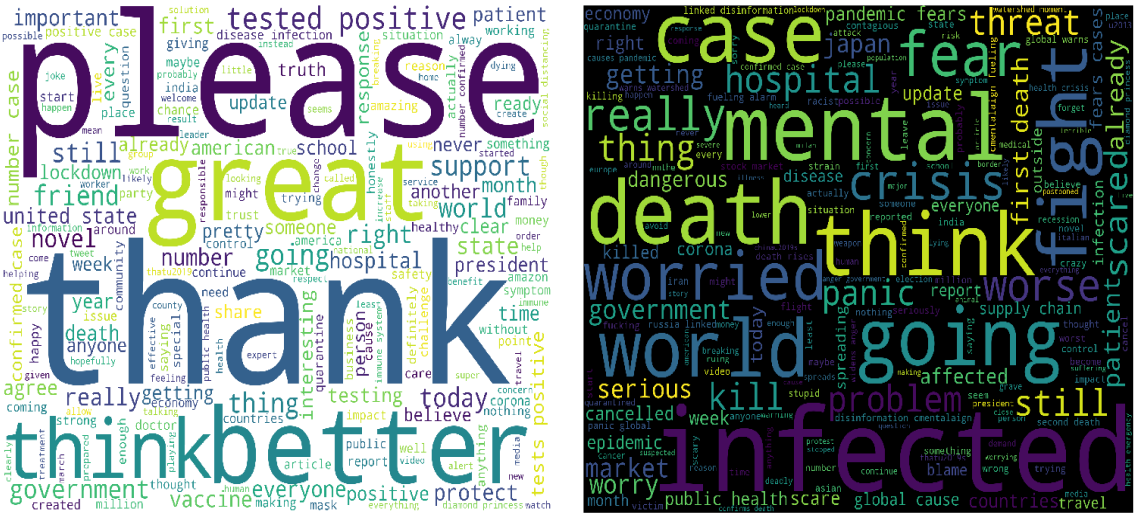


Figure 5.1: Tweet Frequency month by month



(a) Positive sentiments

(b) Negative sentiments

Figure 5.2: The Wordcloud of frequent words classified as Positive and Negative Sentiments

mental health, vaccination, and different types of approved vaccines. There was a lot of conversation on the improvement of the situation as more jobs are being created and the standard of living of people is improving. The fear of isolation is also reducing as fully vaccinated people can go and relate with people without the fear of contracting the virus. Most countries have also opened up as the isolation has reduced drastically.

5.2 Text Feature Selection

5.2.1 Filter Methods

There are univariate metrics with which the filter method ranks the features of a dataset. After ranking, it selects the features with the highest rank. Some of the filter methods for text feature selection utilized in this research work are discussed below

5.2.1.1 Information Gain

From python, sklearn module is the feature selection package from where the information gain package is imported from. It determines how the independent variable is utilized to predict the target variable. From our experiment, the mutual information function was run on our train data and the figure below shows the word with the most sentiments. The bar chart shows the combination of words that has determines the sentiments of words in the dataset.

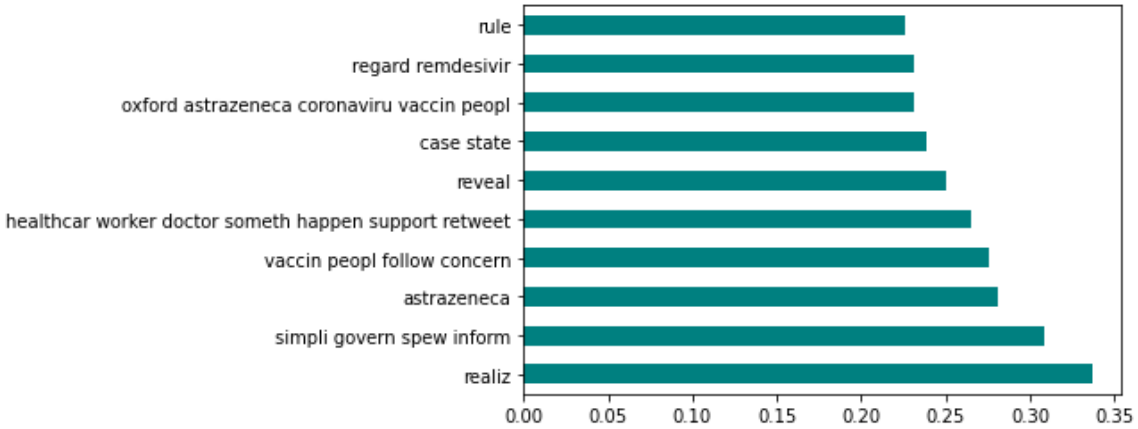


Figure 5.3: Text Feature Selection Utilizing Information Gain

5.2.1.2 Fisher's Score

With this type of Text Feature Selection, features are ranked according to their importance. All these words in the chart below carry the same ranking based on their

importance. Under the fisher criterion, the features are selected independently based on their scores. This leads to a suboptimal subset of features.

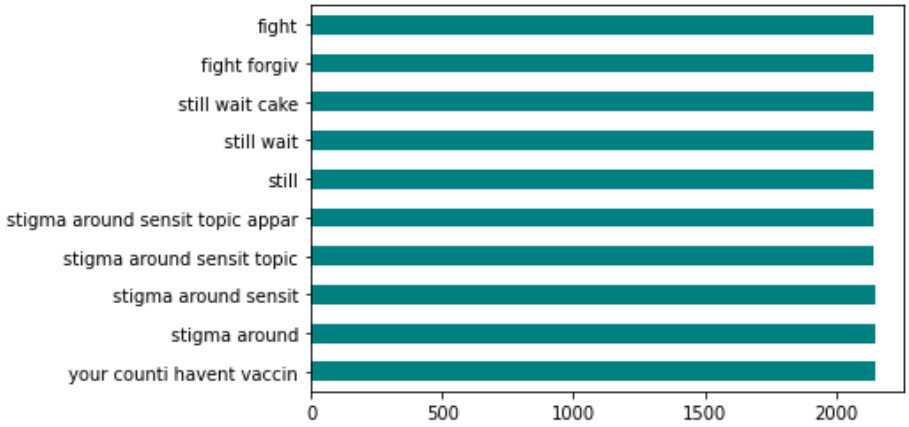


Figure 5.4: Text Feature Selection Utilizing Fisher's Score

5.2.1.3 Chi-Square Test

The relationship between the features is tested and see how one feature deviates from the other. This function eliminates the features that are most likely to be independent and are therefore irrelevant for classification. Utilizing this test on our sample data, the chart below is generated.

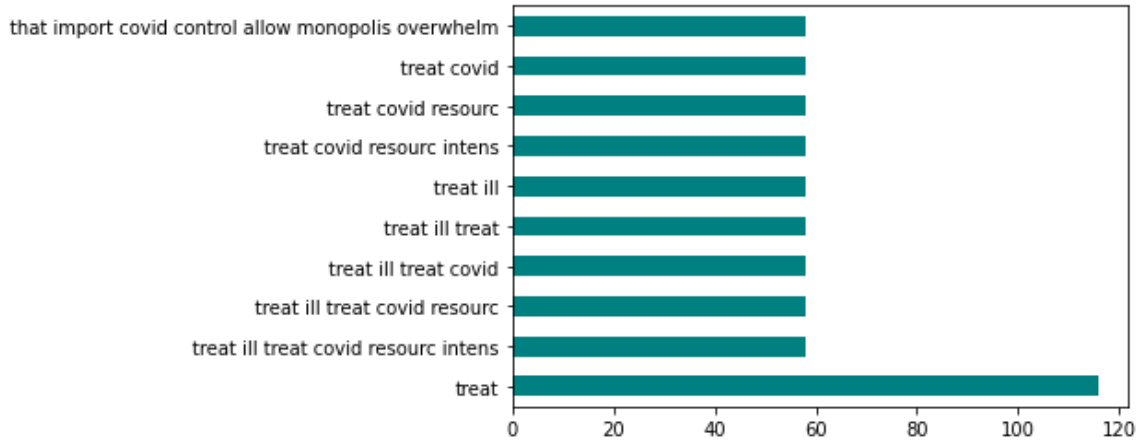


Figure 5.5: Text Feature Selection Utilizing Chi-Square Test

5.2.1.4 Variance Threshold

This method of text feature selection eliminate features with low importance, that is, feature with not so much relevant information. Utilizing this feature on our sample data, the relevant information is stated.

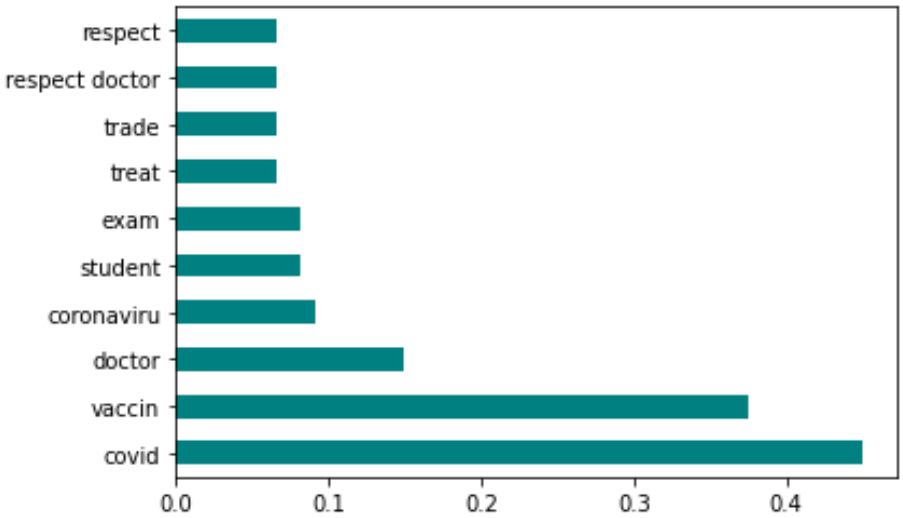


Figure 5.6: Text Feature Selection Utilizing Variance Threshold

5.2.1.5 Mean Absolute Difference (MAD)

This method is similar to the variance threshold method, just that it does not include and square. This is a scaled variant that calculates the mean absolute difference from the mean value for a given feature. This is depicted in Figure 5.1.5 where this method is utilized on the data sample.

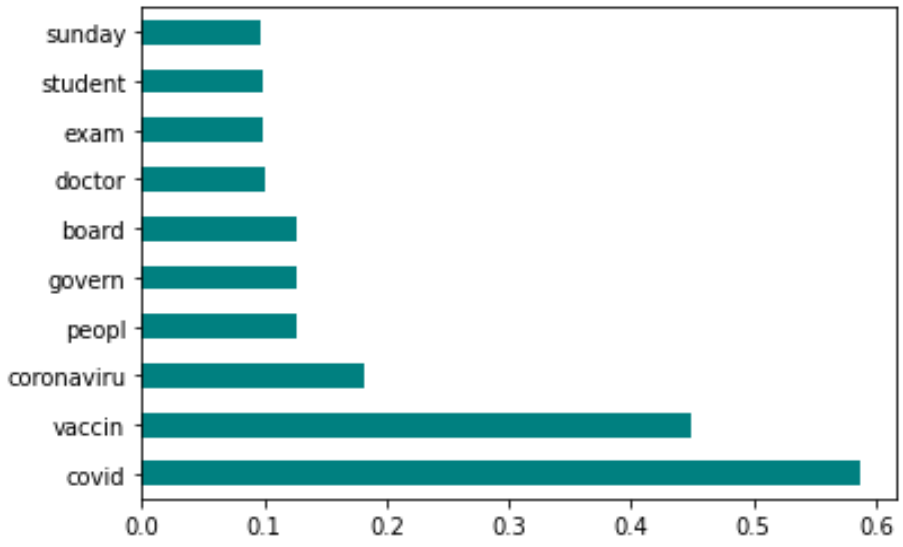


Figure 5.7: Text Feature Selection Utilizing Mean Absolute Difference

5.2.1.6 Dispersion Ratio

Higher dispersion corresponds to more relevant features. The dispersion ratio is calculated by dividing the arithmetic mean and the geometric mean for the given feature. Utilizing this method on our sample data, the word with the highest ratio is "covid" as shown on the graph.

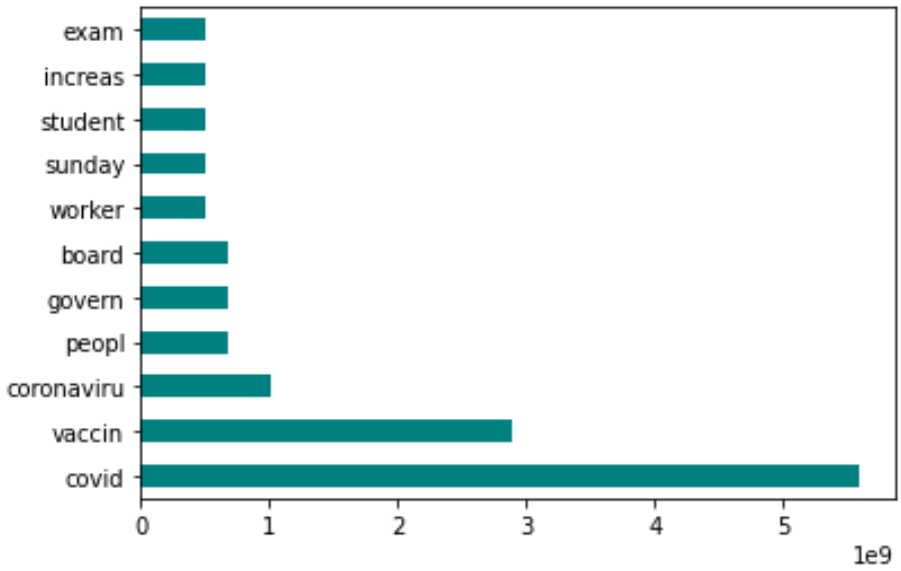


Figure 5.8: Text Feature Selection Utilizing Dispersion Ratio

5.2.2 Wrapper methods

Here, a subset of the features is utilized to train the algorithm in an iterative manner [27]. All the features are considered for selection and passed into the algorithm. The model is repeated until the performance of the algorithm is well based on the user of the model. After the conclusion of the model, the best features are selected. Some of the wrapper methods utilized are Recursive Feature Elimination (RFE) with random forest, Forward Feature Selection

5.2.2.1 Recursive Feature Elimination (RFE) with Random Forest

RFE is one of the most used features selection algorithms due to its ease of configuration, use, and effectiveness at selecting the features that are most relevant for the prediction of the target variable. The two most important attributes of RFE are the selection of the relevant columns in the datasets and also the algorithm for choosing these columns.

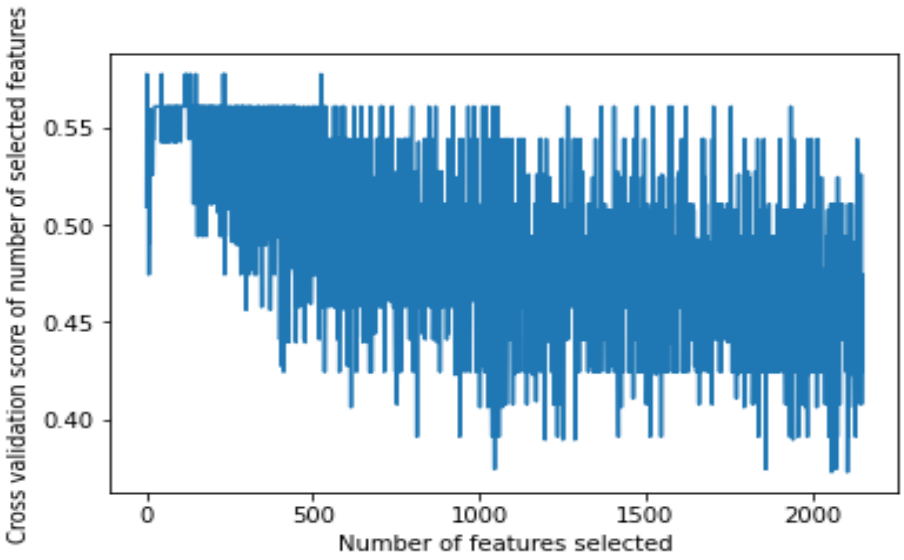


Figure 5.9: Recursive Feature Elimination (RFE) with Random Forest

5.2.2.2 Forward Feature Selection

The algorithm commences the feature selection process with empty features and then gradually adds features that improve the model. It continues this process for every iteration until the addition of a new feature does not improve the performance of the model.

Other techniques here are Backward elimination, bi-directional elimination, exhaustive selection, and recursive elimination.

5.2.3 Embedded methods

This feature selection method merges the advantages of both the filter method and the wrapper method. The algorithm here has its own built-in feature selection methods and its considers a combination of features. These methods are fast like the filter methods but produce more accurate results than the filter methods. Some techniques utilized here are the regularization techniques (Lasso (L1 regularization) and Elastic nets (L1 and L2 regularization)) and the tree-based method which uses feature importance as a basis for feature selection.

5.3 Feature Engineering

After selecting the relevant features from the datasets, some feature engineering are performed to transform the features of the datasets into **Vectors** and also creating new features from the dataset Some of these engineering features are discussed below:

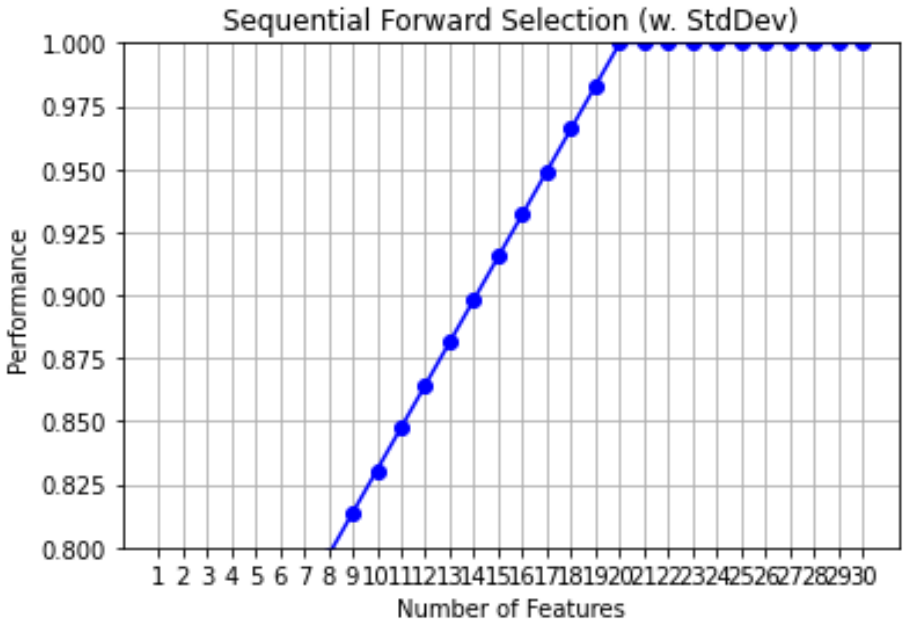


Figure 5.10: Forward Feature Selection

5.3.1 Count Vectors as features

Text data are represented with numbers 1 or 0 depending on the position of the text. If specific text is in a feature, it is represented by 1, else 0. Every time the word is encountered again, the count is increased and leaving a 0 everywhere else. This is also known as One-Hot Encoding. Human efforts with this will be cumbersome, hence a python package name CountVectorizer under Scikit Learning module to help utilize this feature engineering method.

5.3.2 TF-IDF Vectors as features

This is utilized to rank the importance of a vector or feature. TF which means Term Frequency is defined by the number of times a word w appears in a document divided by the number of terms in the document whilst IDF which means Inverse Document Frequency is the logarithm of the number of documents divided by the number of documents with the word w in it [28]. TF-IDF can be generated under word-level, character levels, and n-gram levels.

5.3.3 Text / NLP based features

Here, some extra features are created in order to improve the text classification model. Some of the features that are created from our dataset are the word count,

the character count, and the word density which is the average length of words in the dataset.

5.3.4 Word Embeddings

This is a Natural Language Processing Approach where the encoding of words in vector formats is performed such that the word closer to that vector should have a similar meaning. For instance, considering the pair of words fruit: pineapple and man: apple. Due to the understanding our the language, it is easy to same fruits: pineapple are similar and not man: apple. Word embedding is to allow the machine to also have a similar understanding instantly about words closest to them.

To calculate the accuracies of this method with different classification models, below is the table that shows the result for each feature engineering method to determine the most suitable for our dataset.

Table 5.1: Featuring Engineering Accuracies with Different Classifiers

Model	Count Vectors	Word Level TF-IDF	Char. Level TF-IDF	Char. Level Vector
Naive Bayes	0.41	0.545	0.5	0.45
Linear Classifier	0.45	0.45	0.5	0.5
Boosting	0.409	0.182	0.409	0.45
GradientBoost	0.5	0.454	0.591	0.272
AdaBoost	0.454	0.409	0.045	0.454
DecisionTree	0.5	0.454	0.136	0.318

5.4 Sentiment Classification

Twitter data has been extracted and the features relevant to this research work are the tweet id, the date of the tweets, and the conversation which is represented by the text. The dataset has been preprocessed. Stop words, special characters and irrelevant words have been removed with natural language processing. Text feature extraction has been carried out to select the most relevant features which will produce an accurate performance of the models. Feature Engineering has also be carried out to convert the features into vectors.

The performance evaluation was carried out on the accuracy, the F1 score, precision, recall, and kappa parameters of the scikit learn metrics. About 13 models were created to test these parameters. The result of metrics for the models are displayed in Table 5.2

Table 5.2: Covid-19 Twitter Text Classification Results

Classifier	Accuracy	Precision	Recall	F1	Kappa
Nearest Neighbors	0.71	0.69	0.71	0.69	0.24
LogisticRegression	0.73	0.62	0.73	0.64	0.05
Decision Tree	0.74	0.67	0.74	0.68	0.16
Random Forest	0.74	0.57	0.74	0.63	0.00
MLP	0.70	0.65	0.70	0.66	0.16
AdaBoost	0.74	0.66	0.74	0.67	0.12
GaussianNB	0.13	0.52	0.13	0.03	0.00
LinearDiscriminant	0.73	0.60	0.73	0.63	0.01
GradientBoost	0.76	0.71	0.76	0.69	0.19
MultinomialNB	0.66	0.65	0.66	0.65	0.17
SGD	0.51	0.62	0.51	0.55	0.08
LGBM	0.77	0.73	0.77	0.72	0.26
XGB	0.76	0.71	0.76	0.68	0.16

5.4.1 Performance Evaluation with Supervised Learning Classification

Supervised Learning algorithms are utilized to create models with labeled data. The labeled part of the dataset was utilized here for the creation of the model. The model classified and predicted the train and the test score and compared the prediction with the actual scores. Table 5.3 shows the predicted scores with supervised learning classifiers. From the table, LGBM Classifier has the best prediction for both train and test scores.

Table 5.3: Performance Evaluation for Supervised Learning

Classifier	Train Score	Test Score
Nearest Neighbors	0.83	0.71
LogisticRegression	0.73	0.73
Decision Tree	0.75	0.74
Random Forest	0.74	0.74
MLP	0.75	0.70
AdaBoost	0.74	0.66
GaussianNB	0.13	0.13
LinearDiscriminant	0.74	0.73
GradientBoost	0.79	0.76
MultinomialNB	0.67	0.66
SGD	0.54	0.54
LGBM	**0.85**	**0.77**
XGB	0.78	0.76

From the performance evaluation of the supervised learning models, LGBM has the best performance of 0.77. This means the model was able to predicted corrected 77% of the test score.

5.4.2 Performance Evaluation with Semi-Supervised Classification

With a dataset that has a combination of labeled and unlabelled data like our dataset, semi-supervised learning is the appropriate machine learning to create the appropriate label predictions. Different semi-supervised learning classifiers were utilized to create different models for accurate predictions. The performance evaluation for the different models is displayed in Table 5.3.

Table 5.4: Performance Evaluation for Semi-Supervised Learning

Classifier	Train Score	Test Score
Nearest Neighbors	0.89	0.93
LogisticRegression	**0.84**	**0.99**
Decision Tree	**0.85**	**0.99**
Random Forest	**0.85**	**1.00**
MLP	0.80	0.86
AdaBoost	0.84	0.97
GaussianNB	0.47	0.98
LinearDiscriminant	**0.84**	**0.99**
GradientBoost	0.86	0.97
MultinomialNB	0.78	0.96
SGD	0.69	0.69
LGBM	0.89	0.96
XGB	0.85	0.97

From the performance evaluation of the semi-supervised learning models, three(3) different models have the same result of 0.99 while the Random Forest model has the best performance of 1.0. Comparing the performance results of both supervised learning and semi-supervised learning, it is obvious that semi-supervised learning has better performance evaluation. Semi-supervised learning is appropriate for data with a combination of some labeled data and large unlabeled data.

Chapter 6

Conclusion and Future Works

Data has evolved and major insights are being drawn from data for effective decision making. This research work focuses on Twitter data because enormous information can be retrieved from this social media platform. Research has shown that there are about 200 million active Twitter users daily. During the pandemic, quite a lot of users voiced their opinions about the situation and hence a lot of information was retrieved, analyzed with machine learning in order to derive insights to help provide long-term solutions especially on the effect on mental health. Leveraging machine learning, twitter data were analyzed to retrieve the sentiments of people's opinions. These sentiments were categorized into positive and negative sentiments. With machine learning, different models were designed under semi-supervised and supervised learning. Different metrics were utilized to determine the model with the best performance. Also, we concluded that semi-supervised models outperformed supervised models and the best performance under the supervised model predicted 0.77 for the best model performance and semi-supervised prediction for the same model was 0.96.

From the study and analysis of Twitter data using Natural Language Processing and Scikit machine learning algorithm to create AI models, we deduced that the mental health-related issues increased between January 2020 and April 2020 but gradually reduced as people tends to manage the situation. The trend of Covid 19 and its effect on mental health kept fluctuating as the months go by but still within minimized range. The effect of this pandemic on mental health is more centered around personal challenges such as unemployment, isolation from loved ones, travel restrictions, and other unforeseen circumstances. The conversation between Jan 2021 and April 2021 was more positive with the introduction of vaccination. Quite a lot of compensation was introduced to encourage people to get vaccinated. The effectiveness of the vaccines reduced isolation, more jobs were created and people with job loss were able to get jobs. Schools resume physically and people are beginning to generate income. This

study will assist medical professionals to focus on the specific challenges that triggers the mental issues in patients and help with resolutions as quickly as possible. The research is still in progress as we keep analyzing the effect of this virus on mental health using social media data so as to bring about the last solution to mental health issues.

Bibliography

[1] *Coronavirus cases*, 2021. [Online]. Available: https://www.worldometers.info/coronavirus/?utm%5C_campaign=homeAdvegas1?%5C#countries.

[2] medrxiv.org. [Online]. Available: https://www.medrxiv.org/content/10.1101/2020.08.07.20170548v1.

[3] *Canadian mental health association on reconciliation and mental health*, 2021. [Online]. Available: https://cmha.ca/news/canadian-mental-health-association-on-reconciliation-and-mental-health.

[4] *Covid-19 vaccinations*, 2021. [Online]. Available: https://ourworldindata.org/covid-vaccinations.

[5] *Vaccine ranking of countries per doses administered*, 2021. [Online]. Available: https://www.atlas-mag.net/en/article/covid-19-vaccine-ranking-of-countries-per-doses-administered.

[6] M. Motoyama, B. Meeder, K. Levchenko, G. M. Voelker, and S. Savage, "Measuring online service availability using twitter.," *WOSN*, vol. 10, pp. 13–13, 2010.

[7] A. Agarwal, B. Xie, I. Vovsha, O. Rambow, and R. Passonneau, "Sentiment analysis of twitter data," *Proceedings of the Workshop on Languages in Social Media*, Jan. 2011.

[8] D. H. Jin, L. Chae-Gyun, K. Y. Jin, and C. H. Jin, "Analyzing emotions in twitter during a crisis: A case study of the 2015 middle east respiratory syndrome outbreak in korea," in *2016 international conference on big data and smart computing (BigComp)*, IEEE, 2016, pp. 415–418.

[9] L. Lifang, Z. Qingpeng, W. Xiao, *et al.*, "Characterizing the propagation of situational information in social media during covid-19 epidemic: A case study on weibo," *IEEE Transactions on Computational Social Systems*, vol. 7, no. 2, pp. 556–562, 2020.

[10] B. Ankita, M. Manoj, K. Niranjan, and T. Siddharth, "Analysis of hospital reviews through sentiment analysis: An approach to aid patients in the times of covid-19 pandemic," 2021.

[11] S. Koustuv, T. John, C. E. D, and D. C. Munmun, "Social media reveals psychosocial effects of the covid-19 pandemic," *medRxiv*, 2020.

[12] J. Hamed, W. Yongli, O. Rita, and H. Shucheng, "Deep sentiment classification and topic discovery on novel coronavirus or covid-19 online discussions: Nlp using lstm recurrent neural network approach," *IEEE Journal of Biomedical and Health Informatics*, vol. 24, no. 10, pp. 2733–2742, 2020.

[13] L. Hanjia, C. Long, W. Yu, and L. Jiebo, "Sense and sensibility: Characterizing social media users regarding the use of controversial terms for covid-19," *IEEE Transactions on Big Data*, 2020.

[14] M. Amrita, K. Purnima, and V. Sonali, "Emotional analysis using twitter data during pandemic situation: Covid-19," in *2020 5th International Conference on Communication and Electronics Systems (ICCES)*, IEEE, 2020, pp. 845–848.

[15] H. Man, L. Evelyn, H. E. S, *et al.*, "Social network analysis of covid-19 sentiments: Application of artificial intelligence," *Journal of medical Internet research*, vol. 22, no. 8, e22590, 2020.

[16] S. Sohini, M. Sareeta, and S. Garima, "An exploration of impact of covid 19 on mental health-analysis of tweets using natural language processing techniques," *medRxiv*, 2020.

[17] S. Jim, A. GG, R. Md, E. Ek, S. Yana, *et al.*, "Covid-19 public sentiment insights and machine learning for tweets classification," *Information*, vol. 11, no. 6, p. 314, 2020.

[18] K. Anna, H. berle Matthias, K. Iona, and Z. X. Xiang, "Cross-language sentiment analysis of european twitter messages duringthe covid-19 pandemic," *arXiv preprint arXiv:2008.12172*, 2020.

[19] R. Furqan, K. Madiha, A. Waqar, R. Vaibhav, M. Arif, and C. G. Sang, "A performance comparison of supervised machine learning models for covid-19 tweets sentiment analysis," *Plos one*, vol. 16, no. 2, e0245909, 2021.

[20] K. Harleen, A. S. Ul, A. Bhavya, and C. Victor, "A proposed sentiment analysis deep learning algorithm for analyzing covid-19 tweets," *Information Systems Frontiers*, pp. 1–13, 2021.

[21] *Understanding n-grams.* [Online]. Available: `https://towardsdatascience.com/understanding-word-n-grams-and-n-gram-probability-in-natural-language-processing`.

[22] *Implementing deep learning methods.* [Online]. Available: `https://www.kdnuggets.com/2018/04/implementing-deep-learning-methods-feature-engineering-text-data-cbow.html`.

[23] *An implementation guide to word2vec using numpy and google sheets*, 2018. [Online]. Available: `https://towardsdatascience.com/an-implementation-guide-to-word2vec-using-numpy-and-google-sheets-13445eebd281`.

[24] *Implementing deep learning methods.* [Online]. Available: `https://towardsdatascience.com/nlp-101-word2vec-skip-gram-and-cbow`.

[25] *Using countvectorizer to extracting features from text*, 2020. [Online]. Available: `https://www.geeksforgeeks.org/using-countvectorizer-to-extracting-features-from-text`.

[26] *Semi-supervised learning*, 2020. [Online]. Available: `https://algorithmia.com/blog/semi-supervised-learning`.

[27] *Semi-supervised learning*, 2021. [Online]. Available: `https://www.geeksforgeeks.org/feature-selection-techniques-in-machine-learning/`.

[28] N. Abbasi, *What is tf-idf?* [Online]. Available: `https://www.educative.io/edpresso/what-is-tf-idf`.