الجامعة البريطانية في دبي

The British University in Dubai

Sentiment Analysis for opinion leaders on Twitter: A Case Study of COVID-19

تحليل المشاعر للقادة المؤثرين في الرأي علي تويتر: دراسة عن كوفيد-19

by

REEM SAJID MIR

Dissertation submitted in partial fulfilment

of the requirements for the degree of

MSc INFORMATICS

at

The British University in Dubai

November 2022

DECLARATION

I warrant that the content of this research is the direct result of my own work and that any use made in it of published or unpublished copyright material falls within the limits permitted by international copyright conventions.

I understand that a copy of my research will be deposited in the University Library for permanent retention.

I hereby agree that the material mentioned above for which I am author and copyright holder may be copied and distributed by The British University in Dubai for the purposes of research, private study, or education and that The British University in Dubai may recover from purchasers the costs incurred in such copying and distribution, where appropriate.

I understand that The British University in Dubai may make a digital copy available in the institutional repository.

I understand that I may apply to the University to retain the right to withhold or to restrict access to my thesis for a period which shall not normally exceed four calendar years from the congregation at which the degree is conferred, the length of the period to be specified in the application, together with the precise reasons for making that application.

Signature of the student

COPYRIGHT AND INFORMATION TO USERS

ABSTRACT

The coronavirus or COVID-19 is an ongoing global problem where a pandemic was implemented early in 2020 during the outbreak. Social media platforms were used during the pandemic to share views and exchange information. This study aims to provide a framework for sentiment analysis of opinion leaders on Twitter. The experiments were conducted by aiming COVID-19 specific tweets from four opinion leaders by applying machine learning models. The dataset collected uses covid hashtags and tweets posted in English. Sentiment analysis are then performed on these tweets for analysis. The tweets are then preprocessed to prepare it for evaluation. This research provides findings from these tweets using sentiment analysis on machine learning models where the logistic regression model provided the best accuracy results followed by the Multi-layer perceptron model, Support vector machine, Convolutional neural network, and Decision tree. As the tweets directly affect people's thoughts, the purpose of these results was to know about the tweet's sentiments from diverse public opinion leaders around the world during COVID-19.

نبذة مختصرة

يعد فيروس كورونا أو COVID-19 مشكلة عالمية مستمرة حيث تم تنفيذ جائحة في وقت مبكر من عام 2020 أثناء تفشي المرض. تم استخدام منصات التواصل الاجتماعي أثناء الوباء لتبادل الآراء والمعلومات. تهدف هذه الدراسة إلى توفير إطار عمل لتحليل آراء قادة الرأي على تويتر. أجريت التجارب من خلال توجيه تغريدات محددة لـ COVID-19 من أربعة قادة رأي من خلال تطبيق نماذج التعلم الآلي. تستخدم مجموعة البيانات التي تم جمعها علامات التجزئة والتغريدات المنشورة باللغة الإنجليزية. ثم يتم إجراء تحليل المشاعر على هذه التغريدات لتحليلها. ثم تتم معالجة التغريدات مسبقًا لإعدادها للتقييم. يقدم هذا البحث نتائج هذه التغريدات باستخدام تحليل المشاعر على نماذج التعلم الآلي حيث قدم نموذج الانحدار اللوجستي أفضل نتائج الدقة متبوعًا بنموذج الإدراك متعدد الطبقات وآلة ناقلات الدعم والشبكة العصبية التلافيفية وشجرة القرار. نظرًا لأن التغريدات تؤثر بشكل مباشر على أفكار الناس ، كان الغرض من هذه النتائج هو معرفة مشاعر التغريدات من قادة -COVID. 19خلال الرأي العام المتنوعين حول العالم

ACKNOWLEDGEMENTS

Firstly, I would like to thank Allah for giving me the capability, opportunity, and knowledge to take this research and complete it with satisfaction. Then, I would like to extend the appreciation to my supervisor Dr Manar Alkhatib and Dr Khaled Shaalan who supported me through my dissertation, from giving their precious time to guiding me throughout. I have learned so much from them in terms of sentiment analysis and research. I also would like to thank all my Master's lecturers for their constant support and for helping me to complete my Master's degree. I would also like to thank my family for their motivation and provision to continue my studies.

.

Table of Contents

1. Introduction

2. Literature Review

3. Methodology

4. Results and Discussion

List of Figures

List of Tables

List of Abbreviations

Abbreviation	Description
COVID	Corona Virus Disease
LR	Logistic Regression
ML	Machine Learning
DT	Decision tree
SVM	Support Vector Machine
MLP	Multi-layer Perceptron
CNN	Convolutional Neural Network
API	Application Programming Interface

List of keywords

- Sentiment Analysis
- Opinion Leaders
- COVID-19
- Coronavirus
- Users
- Tweets
- Machine learning models

1. CHAPTER ONE: INTRODUCTION

The first case of the virus was detected in 2019 in China, Wuhan (WHO, 2020) where later it was name as coronavirus or COVID-19. Coronavirus was declared as a global emergency in 2020. During the global emergency, millions of people went to social media platform to share their thoughts. (H. Manguri et al., 2020)Various social media platforms were used to give reviews on covid global emergency by public opinion leaders.

Opinion leaders are individuals that have a lot of authority within the community while possessing the strength to shape the views of others they are linked to. Among the two-way mode of understanding, content is sent from a single opinion leader to a broader public and so on. During COVID-19, opinion leaders are addressed more widely for the broadcasting of info. An opinion leader identified in one social network may have accounts in another social platforms. (Parau, 2017)

We are specifically targeting the twitter social media to evaluate the sentiment analysis of public from renowned public leader's posts. In our research, we selected four famous public opinion leaders that are Donald trump, Emmanuel macron, Justin Trudeau, and Boris Johnson.

According to WHO (World health organization), till we are drafting this study, 634,522,052 confirmed cases have contacted COVID worldwide and 6,599,100 deaths were recorded. (WHO, 2022)From this data, various professionals and researchers are finding ways to evaluate people emotions and the environmental impact on people. As social media platform twitter played an important role of communication during the virus outbreak,(Zhu et al., 2016) tweets from public opinion leaders are used to identify the sentiments.(el Barachi et al., 2021) researched that positive words have a powerful impact on people with hope and willingness to survive during the global outbreak whereas negative words cause depression and lack of hope with attempt to suicidal emotions and bad mental health. Keeping this in mind, we are conducting sentiment analysis on these tweets to categorize the tweets and further investigate finding of the output using machine learning models.

1.1 Problem Statement

Covid was declared as a global pandemic in 2020 as the number of cases increased. These public opinion leaders gave comments and reviews on diverse covid restrictions and situations. The public opinion leaders have a huge number of followers where some of the followers agree to the posts, and some disagree. The people sentiments are affected (Kausar et al., 2021)and will be discussed in this research based on the posts by public opinion leaders. We have used machine learning algorithms to analyze and evaluated the sentiments.

1.2 Related work

There are numerous researchers on sentiment analysis with diverse social media platforms. This part includes few pieces of research that is utilized as reference to learn on sentiments analysis during covid using machine learning models. (Hatami et al., 2021) investigated by applying network analysis to find evidence for opinion leaders in covid tweets by categorizing them as tweets for diverse purposes.(Wang et al., 2021) studied national leaders and their dialogue with the world though tweets. (Wang et al., 2021) compares national leaders with different Natural language processing models to study the behavior. (Kausar et al., 2021) studied public sentiments from eleven affected counties in Twitter during global pandemic using sentiment analysis algorithms. Researchers examine the feelings and categorize them. For this purpose, Twitter API was used in our research to obtain COVID tweets and use sentiment analysis to specify as positive, negative, and neutral to be used in our different machine learning models by various visualizations.

1.3 Research Questions

In our research, we mainly address the following questions and discussed them below.

- Does opinion leader’s tweet on COVID-19 affected the public sentiments?
- How are the machine learning models used to conduct sentiment analysis of public opinion leader’s tweet?

- What are the results of the public leader’s tweets on COVID-19 based on sentiment analysis?

1.4 Contribution

In this research, various machine learning models are applied using python programming language to study about the sentimental analysis of COVID-19 tweets by different public leaders. The brief contribution on this dissertation is

- Using Twitter API to extract the tweets of public opinion leaders on COVID-19
- Using Machine learning models to find sentiment of the tweets.
- Applying selected machine learning g models as decision tree, logistic regression, convolution neural network, multi-layer perceptron and support vector machine to conduct sentiment analysis.
- Evaluate the performance of the models with respect to the outcomes to identify the accuracy of the sentiment from the models with the actual outcomes.
- The results are used to study about the tweet’s sentiment affecting the public on COVID-19.

1.5 Tools

This research has been done using software and programming languages. We have mainly used python programming language and Anaconda software with Jupyter environment. The data has been collected from Twitter API using Twitter elevated developer account.

1.6 Scope

There are various public opinion leaders that have diverse reviews on COVID-19. Some researchers used national leaders to evaluate the responses on Twitter during covid.(Wang et al., 2021) .We are specifically targeting renowned leaders and the tweets to focus on the sentiments and analyze the effect on their followers and the public. We will use popular machine learning models to deliver research results.

1.7 Organization of Thesis

The remaining section of this research is organized using the other chapters. Chapter 2 gives an overview of the literature review on Twitter sentiment analysis and a detailed description on the machine learning models used. Chapter 3 demonstrates the methodology of sentiment analysis framework on data collection and cleaning, preprocessing, feature extraction and applying machine learning models. Chapter 4 provides a detailed results and discussion after implementing machine learning models followed by chapter 5 of conclusion and future work required for this research. Chapter 6 shows references used in this research.

2. CHAPTER TWO: LITERATURE REVIEW

2.1 Outline

This research provides an understanding of the tweets from public opinion leaders during covid. It uses various machine learning algorithms using sentiment analysis to determine if the tweets are positive or negative. During the global pandemic, the tweets from public leaders played an important role in impacting people.(Hatami et al., 2021) Tweets spread positivity or negativity among people and their reactions from the tweets affected overall situation.

2.2 Twitter API

To use the dataset from Twitter API, initially twitter sign up was used to create an account. This account is then used to apply for a twitter developer account. (Twitter, 2022) There are several questions to be responded from Twitter in order to approve the access. After numerous questions, queries, and emails, Twitter developer account access was given with Twitter API v2 elevated account. API v2 has a high level of access to tweets and features.

2.3 Twitter Sentiment analysis

Sentiment analysis is a method of natural language processing to evaluate the textual data based on sentiments. (H. Manguri et al., 2020)The data is first collected through Twitter API to identify and calculate sentiments in the dataset. Sentiment analysis follows a certain process through polarity and subjectivity to analyze the tweet. (Yue et al., 2019) The tweets are then categorized as positive, negative, or neutral.(Samuel et al., 2020) The positive tweets have an overall positive impact on the public, giving confidence to people while negative tweets have an adverse influence on the people leading them to depression and harm mental health.

2.4 Machine Learning Models

Machine learning is a branch of computer science that used various algorithms to build machine learning models. The models of machine learning are trained to different patterns to develop machine learning models (Bi et al., 2019). Machine Learning models are usually categorized as supervised or unsupervised. In our research, we will be using five machine learning models, where details are discussed below.

2.4.1 Logistic Regression (LR)

Logistic regression is a linear classifier using function. We have used logistic regression in our text classification of tweets to calculate the observations. We have used binominal classification which is categorized as negative and positive or 0's and 1's. It uses the following function to calculate probability close to 0's or 1's (Mirko Stojiljković, 2022)

$$f(\mathbf{x}): p(\mathbf{x}) = 1 / (1 + \exp(-f(\mathbf{x}))$$

Where p(x) is the probability predicted for given x as 1, whereas 1 – p (x) is the probability predicted as 0. Logistic regression has built in packages for ease of use (Ansari et al., 2017) .In our research, we have used python libraries to extract libraries and packages of logistic regression. We have train and test the model to evaluate results.

2.4.2 Decision Tree (DT)

We have used decision tree model as it is one of the well-known supervised classifiers. It has an assemble that also supports in decision making with conception as flowchart. It can effectively handle dimensional data. Decision tree have GINI index to generate split points using the formula (DataCamp, 2018)

$$Gini(t) = 1 - \sum_{i=1}^{j} P(i|t)^2$$

From the above formula, J represented the class, P represented the ratio of class. We have used several libraries in python to import decision tree methods. The amount of records and number of attributes in the given data determine the time complexity of trees. We have used 80 percent of the data for the training decision tree model and 20 percent of the data for testing. The training period of decision tree is slightly faster than the neural network model.

2.4.3 Support Vector Machine (SVM)

Support vector machines (SVM) are a collection of supervised learning techniques which work well at large-scale environments. It’s likewise is memory effective as SVM only uses an insignificant segment of training points for the function of decision identified as support vectors. SVM uses a variety of mathematical formulas and parameter sets.(scikit-learn developers, 2022c)

In our research, we have used SVM.SVC classifier with probability as true for arrays and labels where SVC is defined as support vector classification. This was implemented using various libraries and functions. Accuracy, precision and recall were calculated using datasets divided between testing and training. The output were created using a prediction of digits up to three decimals. The outputs were shown and discussed in the results section below.

2.4.4 Multi-Layer Perceptron (MLP)

Multi-Layer Perceptron (MLP) is a type of neural network with supervised learning algorithm. MLP uses various parameters to return the values. MLP can use various hidden layers among input and output layers. In our research we have used one hidden layer of MLP shown in Figure 1 below. Since the estimated values of the function in relation to model parameters were also calculated on every stage i.e., to upgrade the variables, MLP Classifier is trained repeatedly. (scikit-learn developers, 2022b)Data supplied as numpy arrays of floating variable values are compatible with this approach.

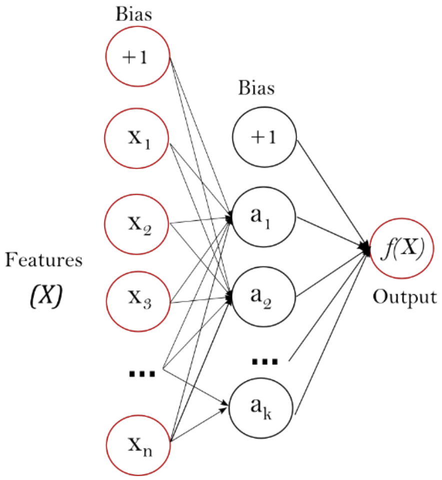
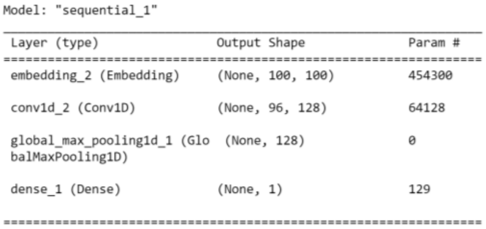


Figure 1 One-layer MLP (scikit-learn developers, 2022b)

The input layers include neurons whereas the final output receives the input from the last layers. We have used MLP Classifier for our sentiment analysis using parameters as, random_state=1, alpha=1e-5, solver='lbfgs'.

2.4.5 Convolutional Neural Network (CNN)

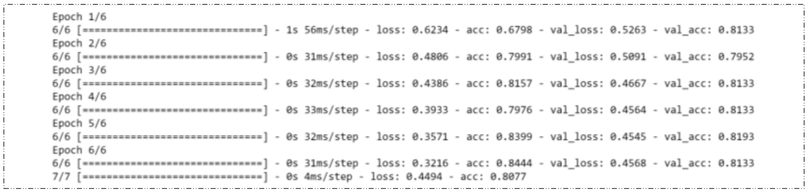
CNN is a type of neural network consisting of layers named as convolutional layers. (Vijay Choubey, 2018). We have utilized a pre-made embedded library. For instance, available are diverse embedded libraries that are accessible such as Glove and Word2Vec. (Amin & Nadeem, 2018). As we use text vectors, relationships among words for a given class can be found easily.



```
Model: "sequential_1"
_________________________________________________________________
 Layer (type)                Output Shape              Param #
=================================================================
 embedding_2 (Embedding)     (None, 100, 100)          454300

 conv1d_2 (Conv1D)           (None, 96, 128)           64128

 global_max_pooling1d_1 (Glo  (None, 128)              0
 balMaxPooling1D)

 dense_1 (Dense)             (None, 1)                 129

=================================================================
```

Figure 2 CNN model layers

Figure 2 shows the summary of our model. We have used sequential model with one Dimensional (1D) convolutional layer and 1Max Pooling layer using sequential method. We have used glove embedded library and get words using embedded dictionary. The outputs are shown and discussed in the section below.



```
Epoch 1/6
6/6 [==============================] - 1s 56ms/step - loss: 0.6234 - acc: 0.6798 - val_loss: 0.5263 - val_acc: 0.8133
Epoch 2/6
6/6 [==============================] - 0s 31ms/step - loss: 0.4806 - acc: 0.7991 - val_loss: 0.5091 - val_acc: 0.7952
Epoch 3/6
6/6 [==============================] - 0s 32ms/step - loss: 0.4386 - acc: 0.8157 - val_loss: 0.4667 - val_acc: 0.8133
Epoch 4/6
6/6 [==============================] - 0s 33ms/step - loss: 0.3933 - acc: 0.7976 - val_loss: 0.4564 - val_acc: 0.8133
Epoch 5/6
6/6 [==============================] - 0s 32ms/step - loss: 0.3571 - acc: 0.8399 - val_loss: 0.4545 - val_acc: 0.8193
Epoch 6/6
6/6 [==============================] - 0s 31ms/step - loss: 0.3216 - acc: 0.8444 - val_loss: 0.4568 - val_acc: 0.8133
7/7 [==============================] - 0s 4ms/step - loss: 0.4494 - acc: 0.8077
```

Figure 3 CNN model Epoch history

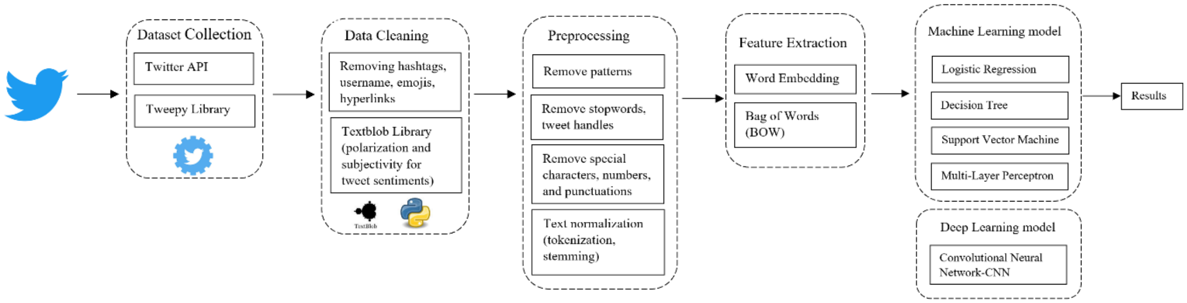
Figure 3 shows the epoch history of the parameters used with calculates loss and accuracy with every epoch. An overall of 6 epochs was used with a validation split of 20% as testing and 80% as training data.

3. CHAPTER THREE: METHODOLOGY

In our research, we have used various methods and models to find the results of our dataset. Below is the explained analysis of our sentiment analysis approach.

3.1 Sentiment Analysis Framework

The framework used in this research is shown in figure 4. The framework has focused on using the best process to collect findings on the sentiment analysis of selected public opinion leader’s tweets during COVID. This framework uses chosen models to classify the tweets and sentiments. In the section below, we have discussed each process in detail.



Dataset Collection
Twitter API
Tweepy Library
Data Cleaning
Removing hashtags, username, emojis, hyperlinks
Textblob Library (polarization and subjectivity for tweet sentiments)
TextBlob
Preprocessing
Remove patterns
Remove stopwords, tweet handles
Remove special characters, numbers, and punctuations
Text normalization (tokenization, stemming)
Feature Extraction
Word Embedding
Bag of Words (BOW)
Machine Learning model
Logistic Regression
Decision Tree
Support Vector Machine
Multi-Layer Perceptron
Deep Learning model
Convolutional Neural Network-CNN
Results

Figure 4 Twitter Sentiment Analysis Framework

3.2 Data Collection

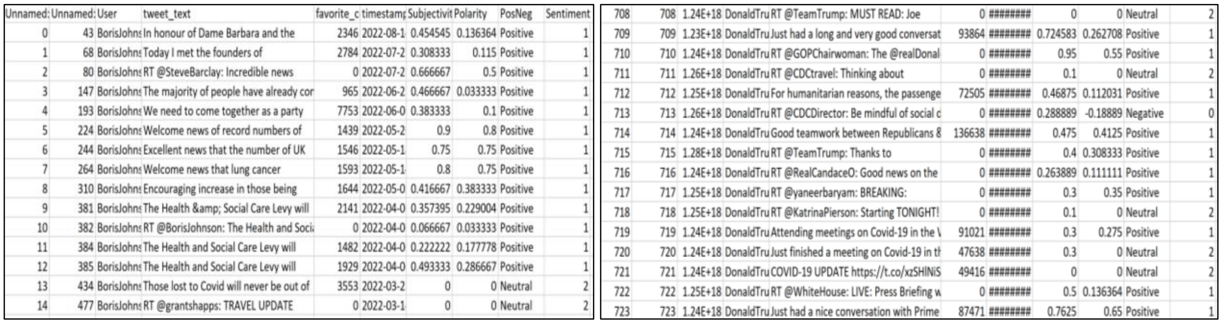
In this research, Twitter API was used to collect 22,000 tweets from four renowned world leaders. Twitter API is used by researchers to collect the data using keywords and hashtags. We have used tweepy, Twitter API to connect with data streams. All of the data was collected using a recent timeframe. Hashtags were used to collect the data as #COVID-19 , #covid, #coronavirus, #covid19. The data was collected using the public leader’s personal account. On 8 January, 2021 Twitter announced that former president Donald trump account is permanently suspended. Due to the suspended account all the available public tweets disappeared. Hence, Donald trump data was collected using an archived tweets dataset(*The Trump Archive*, 2020). This archive has thousands of tweets available from trump's account. The tweets collected were 13,000. Whereas 9000 tweets were collected using twitter API from President Emmanuel macron, Justin Trudeau and Boris Johnson accounts. Some of the irrelevant or languages other than English tweets were eliminated from the dataset.Some of the tweets contain data other than the selected topic i.e. covid. These tweets are then cleaned and removed from the dataset.

Based on the best practices, 80 percent is utilized for training data and 20 percent of the data for testing. Where 1 is represented as positive and 0 as negative. Table 1 below refers to the data extraction strategies used in this research.

Search Keywords	"COVID-19", "Coronavirus", "Pandemic"
Number of tweets	22,000
Year of tweets	2019 & 2020
Topic area	Covid Sentiment Analysis
Language	English
Query string platform	Twitter API- tweepy
Programming Language	Python
Search date	October 2022

Table 1 Data Extraction approach

The Table figure below shows the list of records from dataset used in our research of four public opinion leaders on covid tweets.



Unnamed:	Unnamed:	User	tweet_text	favorite_c	timestamp	Subjectivit	Polarity	PosNeg	Sentiment
0	43	BorisJohns	In honour of Dame Barbara and the	2346	2022-08-1	0.454545	0.136364	Positive	1
1	68	BorisJohns	Today I met the founders of	2784	2022-07-2	0.308333	0.115	Positive	1
2	80	BorisJohns	RT @SteveBarclay: Incredible news	0	2022-07-2	0.666667	0.5	Positive	1
3	147	BorisJohns	The majority of people have already cor	965	2022-06-2	0.466667	0.033333	Positive	1
4	193	BorisJohns	We need to come together as a party	7753	2022-06-0	0.383333	0.1	Positive	1
5	224	BorisJohns	Welcome news of record numbers of	1439	2022-05-2	0.9	0.8	Positive	1
6	244	BorisJohns	Excellent news that the number of UK	1546	2022-05-1	0.75	0.75	Positive	1
7	264	BorisJohns	Welcome news that lung cancer	1593	2022-05-1	0.8	0.75	Positive	1
8	310	BorisJohns	Encouraging increase in those being	1644	2022-05-0	0.416667	0.383333	Positive	1
9	381	BorisJohns	The Health &amp; Social Care Levy will	2141	2022-04-0	0.357395	0.229004	Positive	1
10	382	BorisJohns	RT @BorisJohnson: The Health and Soci	0	2022-04-0	0.066667	0.033333	Positive	1
11	384	BorisJohns	The Health and Social Care Levy will	1482	2022-04-0	0.222222	0.177778	Positive	1
12	385	BorisJohns	The Health and Social Care Levy will	1929	2022-04-0	0.493333	0.286667	Positive	1
13	434	BorisJohns	Those lost to Covid will never be out of	3553	2022-03-2	0	0	Neutral	2
14	477	BorisJohns	RT @grantshapps: TRAVEL UPDATE	0	2022-03-1	0	0	Neutral	2

708	708	1.24E+18	DonaldTru	RT @TeamTrump: MUST READ: Joe	0	########	0	0	Neutral	2
709	709	1.23E+18	DonaldTru	Just had a long and very good conversat	93864	########	0.724583	0.262708	Positive	1
710	710	1.24E+18	DonaldTru	RT @GOPChairwoman: The @realDonal	0	########	0.95	0.55	Positive	1
711	711	1.26E+18	DonaldTru	RT @CDCtravel: Thinking about	0	########	0.1	0	Neutral	2
712	712	1.25E+18	DonaldTru	For humanitarian reasons, the passenge	72505	########	0.46875	0.112031	Positive	1
713	713	1.26E+18	DonaldTru	RT @CDCDirector: Be mindful of social c	0	########	0.288889	-0.18889	Negative	0
714	714	1.24E+18	DonaldTru	Good teamwork between Republicans 8	136638	########	0.475	0.4125	Positive	1
715	715	1.28E+18	DonaldTru	RT @TeamTrump: Thanks to	0	########	0.4	0.308333	Positive	1
716	716	1.24E+18	DonaldTru	RT @RealCandaceO: Good news on the	0	########	0.263889	0.111111	Positive	1
717	717	1.25E+18	DonaldTru	RT @yaneerbaryam: BREAKING:	0	########	0.3	0.35	Positive	1
718	718	1.25E+18	DonaldTru	RT @KatrinaPierson: Starting TONIGHT!	0	########	0.1	0	Neutral	2
719	719	1.24E+18	DonaldTru	Attending meetings on Covid-19 in the \	91021	########	0.3	0.275	Positive	1
720	720	1.24E+18	DonaldTru	Just finished a meeting on Covid-19 in tl	47638	########	0.3	0	Neutral	2
721	721	1.24E+18	DonaldTru	COVID-19 UPDATE https://t.co/xzSHlNiS	49416	########	0	0	Neutral	2
722	722	1.25E+18	DonaldTru	RT @WhiteHouse: LIVE: Press Briefing w	0	########	0.5	0.136364	Positive	1
723	723	1.24E+18	DonaldTru	Just had a nice conversation with Prime	87471	########	0.7625	0.65	Positive	1

Table 2 Dataset Collected tweets

3.3 Data Cleaning

Data collected in the previous step was then cleaned using the python libraries. Initially, the dataset includes several punctuations, hashtags, emojis and hyperlinks which was eliminated. The output of this dataset is then used to get the polarity and subjectivity of every tweet using python textblob library. This library is used to process textual data for various tasks. We have used textblob library for sentiment analysis to define the score of the tweet as positive, negative, or neutral. It does sentiment analysis, as well as visualization on output, to aid us in better comprehending the information we were dealing. Because textblob library exclusively understands English language, if tweets were with terms from other languages, it will provide erroneous results. The output of this library is saved to analyze data for preprocessing. In our research, we are only aiming on positive or negative tweets, hence the neutral tweets are then eliminated from the dataset. The neutral tweets are not clear to identify as people satisfy or dissatisfy with the public opinion leader's tweets.

3.4 Data Preprocessing

In order to prepare the data for analysis, data preprocessing is performed on dataset to eliminate irrelevant information and preprocess the data. Several libraries are installed inside the python for our dataset requirement. Initially, the panda library is used to read the file. Only three columns are extracted 1.e. tweet id, label (as 0 or 1), and tweet. Tweets are taken as an object datatype while tweet id and label columns are taken as integer.

Then, removing the pattern method is used in the input text of dataset. All non-overlapping matches of the patterns would be returned in a single component by the function after iterating over every row inside the dataset.

Vectorize function is used to eliminate the twitter handles refereeing to @xyz @user etc. The special characters, punctuations and numbers are then eliminated from the tweets text.

Lambda library is used to clean the tweets by removing short words from the text e.g. ‘In, I, we’ are removed from the text to target the actual text. Tokenization function is then used from ntlk library where a word is taken as a token, and a phrase is taken as a token in a sentence or paragraph since a token is a portion of a whole. The technique of dividing a word into then a collection of tokens is called as tokenization.

The words are then stemmed using the ntlk libraries. We have used porter stemmer as it is one of the common libraries used. This stemmer that is primarily used for feature extraction simply involves breaking down word to its base elements. Since it only supports the English language in its apps, it is based on the assumption that English suffixes were composed of a mixture of smaller and more straightforward adjectives. This stemmer has a lower mistake probability than other related stemmers while delivering the best results.

The words are then combined in a single sentence to be tokenized again as tweets and saved as output using the join function. The frequency distribution function is used in our data preprocessing from python library to count the number of probabilities of keywords either positive or negative, appear in a text.

3.5 Feature Extraction

This is also known as text vectorization. A significant area of use for ML techniques is feature extraction. In this research, word embedding (word2vec) and Bag of words technique is used to convert tweets into frequency based count. The magnitude of every phrase is represented in a data visualization approach of expressing text information, shows their occurrence and relevance. To use a word cloud, relevant text elements may well be emphasized. These were frequently used for network data analysis.(Datacamp wordcloud, 2019)

The occurrence or significance of every phrase is represented by a cloud that is frequently packed with several phrases of various ranges. Tag clouds and phrase clouds are terms for this. It helps us to examine text data and enhance the visual appeal in our research. By importing the word cloud from the library, we have used parameters of word cloud with width and height as 800 x 500 and random state as 42 with font size as 100 to generate all words.

Moreover, because most algorithms require raw textual content of varying length instead of numeric vectorization having defined data sizes, it is not possible to feed the original data, series of characters, straight towards the algorithms itself. (scikit-learn developers, 2022a)

Scikit-learn offers tools for most popular methods of extracting numeric characteristics from textual data. It includes tokenizing texts and designating a numeric id for each and every potential word, for example through utilizing punctuation and blank space for credential spacers. determining how many times every text has a certain token. Standardizing with a decreasing relevance that are found in a large number of texts.(scikit-learn developers, 2022a)

In our research, we refer to the broad method for converting a group of textual information to numeric extracted features as "vectorization." The Bag of Words depiction is the name given to the particular method of tokenization and normalizing. By fully disregarding the comparative data of the words within data file, data file is characterized by word frequency. We have used bag of words vector for count vector function with parameters.

4. CHAPTER FOUR: RESULTS AND DISCUSSION

We will discuss various outcomes and findings of machine learning models implemented in the dataset.

4.1 Tweets visualization

In our research, we have analyzed our dataset through a seaborn library in python. This library plot graph as bar chart shown in Figure 5 below. The graph uses two variables of count and label for sentiment distribution. The count variable is used for tweets and label variables as 0 and 1 where 0 is represented as negative tweets and 1 as positive tweet. The count of positive covid tweets from opinion leaders are more than negative covid tweets in our dataset.

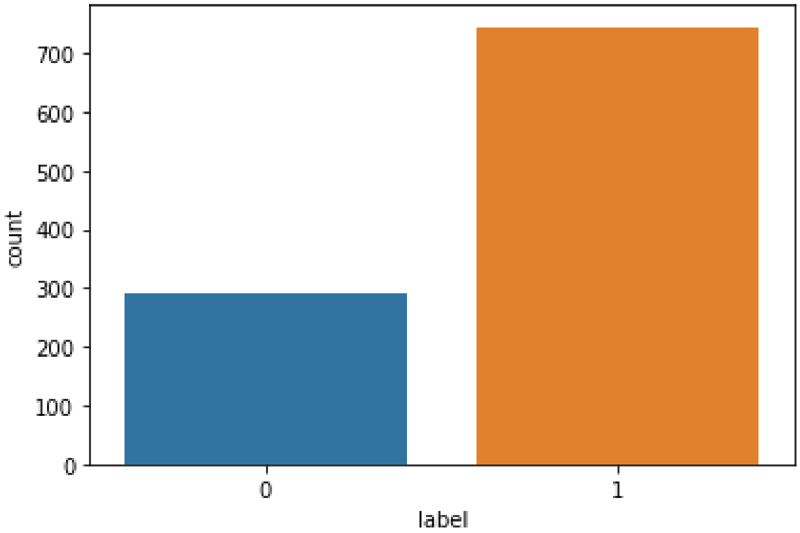


Figure 5 Sentiment Distribution of tweets

Above barchar displays the review of sentiments on overall data. We have individually analyzed every individual opinion leader. The individual bar chart plot results of covid tweets are shown below.

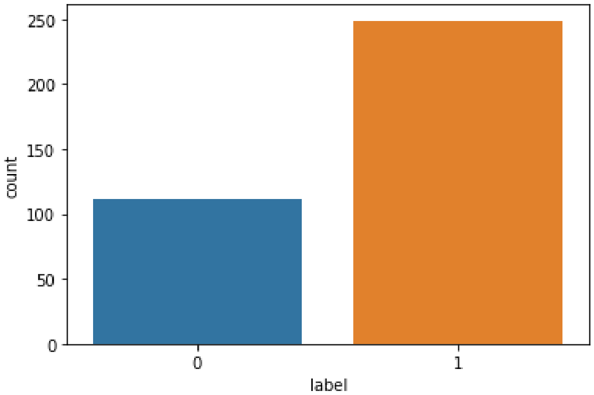


Figure 5.1 Sentiment Distribution of tweets-Boris Johnson

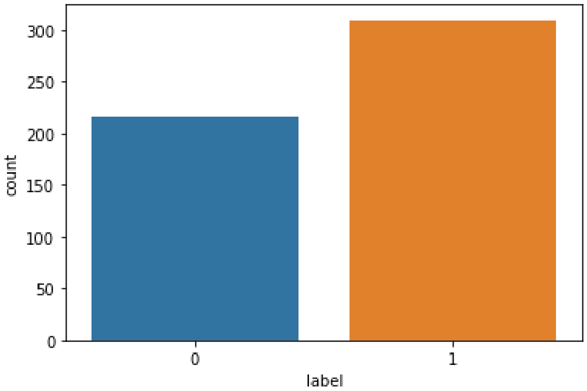


Figure 5.2 Sentiment Distribution of tweets-Donald Trumph

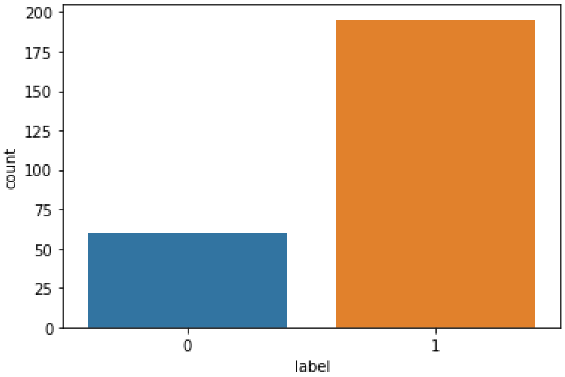


Figure 5.3 Sentiment Distribution of tweets-Justin Trudeau

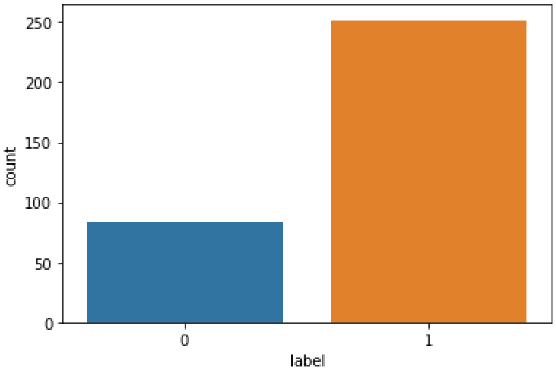


Figure 5.4 Sentiment Distribution of tweets-Emmanuel Macron

4.2 Word cloud of Tweets

Before implementing and showing output of the machine learning models, data is first visualized using the frequent words in the dataset. (Flores-Ruiz et al., 2021) Figures displayes the words cloud from the dataset that includes covid tweets by opinion leaders.

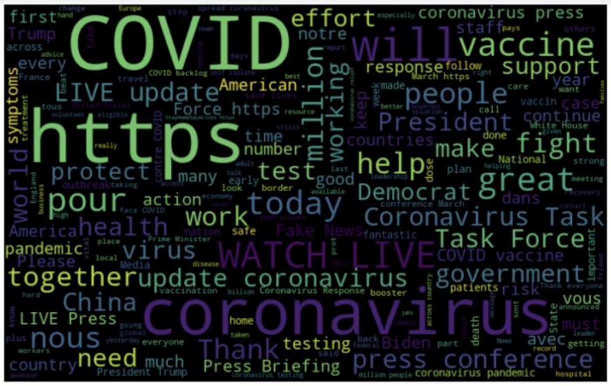


Figure 6 Word cloud of preprocessed tweets

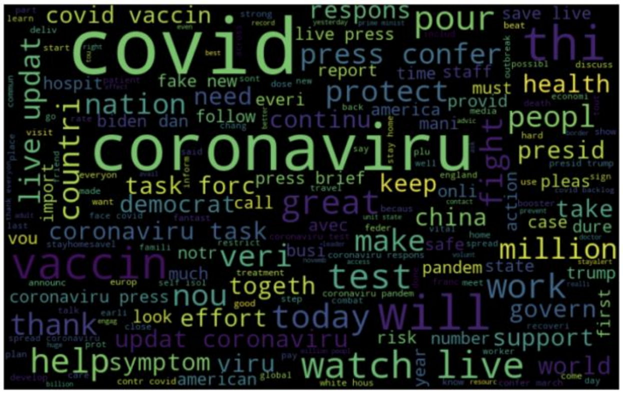


Figure 7 Word cloud of preprocessed tweets by Stemming

In figure 6, it includes a group of words highlighting main words as ‘COVID’ and ‘Coronavirus’. It also includes other words as ‘vaccine’, ‘President’, ‘people’, ‘effort’, ‘fight’ etc. It shows a group of words mostly appear in the dataset and targeting the main words that is focused by most of the public leaders. In figure 7, the word cloud of covid tweets is displayed using stemming. In our research, we have used porter stemmer method that is renowned for the efficiency & ease. It was frequently helpful in our environment, as data extraction for quick recollection or obtaining of targeted keywords were collected. Context texts are typically vectorized comprising words or concepts. Phrases with same root was signified as a same thing. It used suffix to stem the words or reduce it to phrases understandable. It has English libraries that is used through a lookup table to stem words by also applying various algorithmic guidelines.

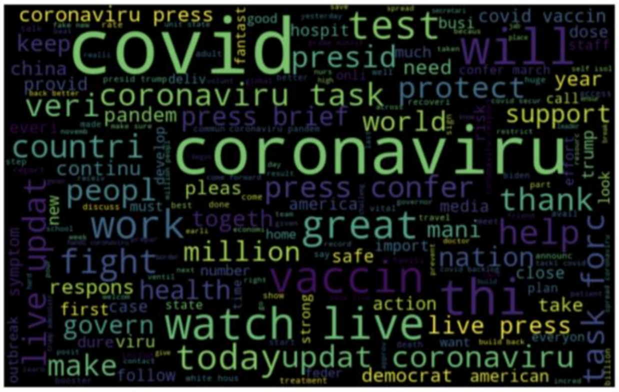


Figure 8 Word cloud of frequent positive words

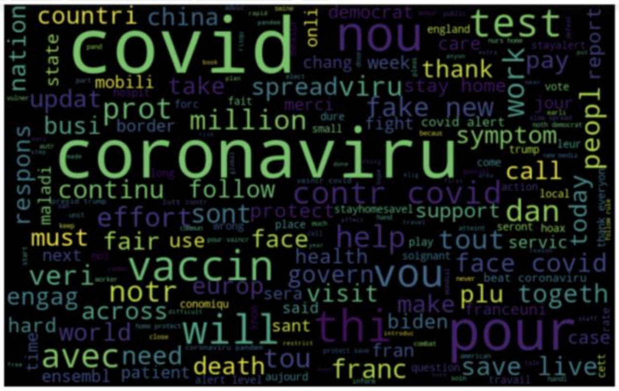


Figure 9 Word cloud of frequent negative words

Figure 8 shows the word cloud of positive words used by public leaders in tweets. The ‘covid’ and ‘coronavirus’ stood out the most in positive and negative tweets. In the volume of all positive tweets, the most common terms were ‘vaccine’, ‘health’, ‘protect’, ‘will’, ‘together’, ‘action’, ‘live’ was used to show the stronger attitude of these tweets and spread positivity during the hard times of global pandemic. However, in figure 9, tweets words such as ‘spread’, ‘help’, ‘death’, ‘save’, ‘hard’, ‘fight’, ‘effort’, and ‘need’ shows the weak emotions of people during the global pandemic. It shows the sadness and fear in these tweets.

4.3 Tweet Hashtags

Figure 10 below shows the bar chart of the top six most common hashtags used by the public leaders in the tweets. Among these #coronavirus has the highest count of tweets that involved this hashtag, followed by #covid, #stayhomes, #franceuni, #stayalert and #getboost. This tweets also show the hashtags that were used involving universities to research about covid or various students that were affected during the pandemic. Hence the #stayalert tweet was used by opinion leaders to keep the public attentive on new updates related to the pandemic and covid. As the pandemic arrived, there was no vaccine and only came out by later, thus #getboost hashtag had the lowest number of tweets from opinion leaders motivating people to get vaccinated against covid to limit the spread of the virus.

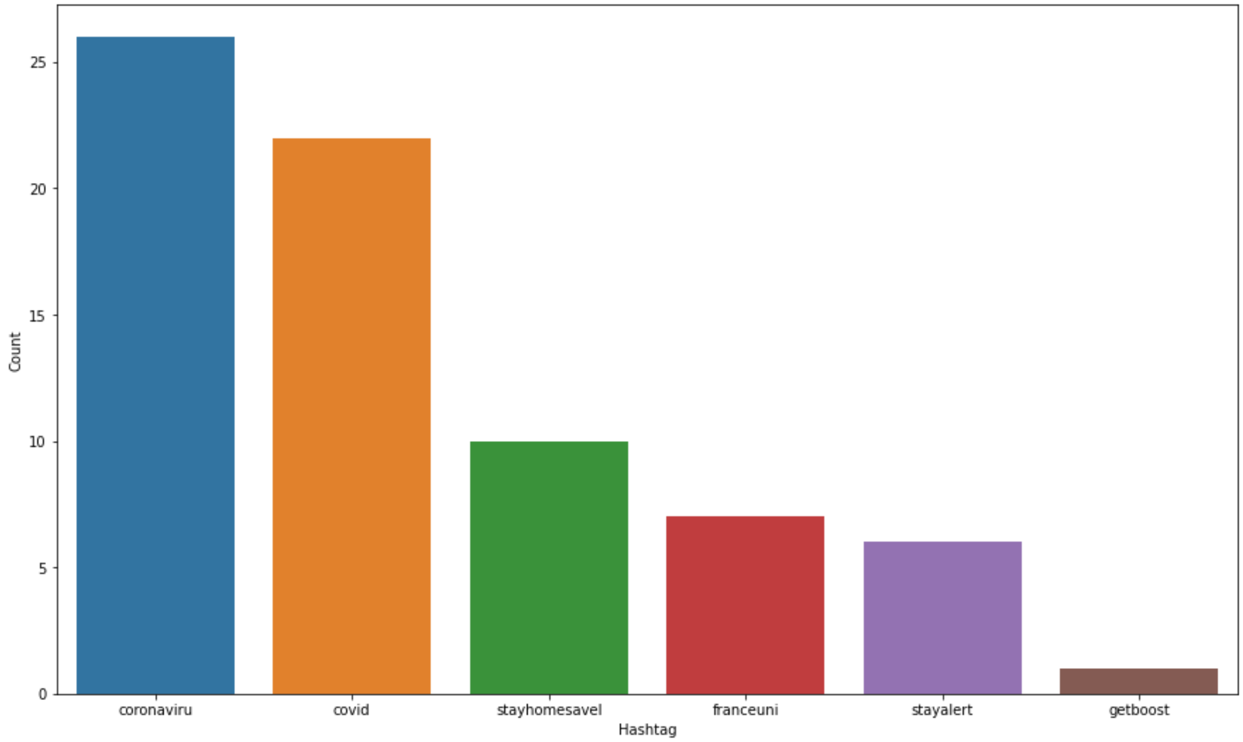
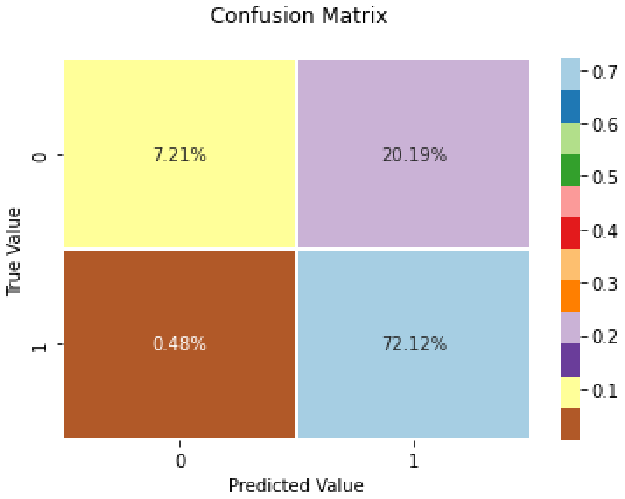
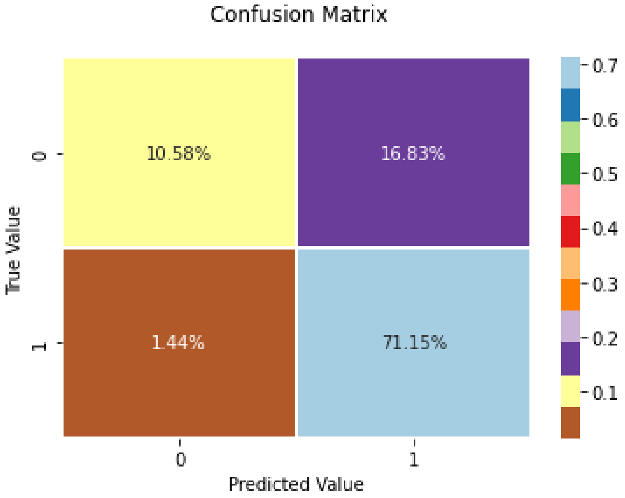


Figure 10 Bar chart of frequent Hashtags

4.4 Performance Evaluation of Machine learning models

The most fundamental classifying measure is the confusion matrix. That is a tabular representation of the modeling projections and truth figures. Every row would show all occurrences of either an actual class while every row would reflect the occurrence inside a predicted category (Aniruddha Bhandari, 2020) Although it isn't mainly a performance measure, the confusion matrix would serve as the basis for those other measures that seem to be necessary to deliver good outcomes. True positive, true negative, false positive, and false negative were the four categories under which the confusion matrix fall.

Table 4 shows the confusion matrix of machine learning models with accuracy displaying the highest by the logistic regression model. Closer examination shows that the true predicted value i.e., column 1, and true actual value i.e., column 1 are maximum for almost all the models.

Machine Learning Models	ML models Confusion Matrix	Model Overall accuracy %
Decision Tree (DT)		79.3
Support Vector Machine (SVM)		81.2

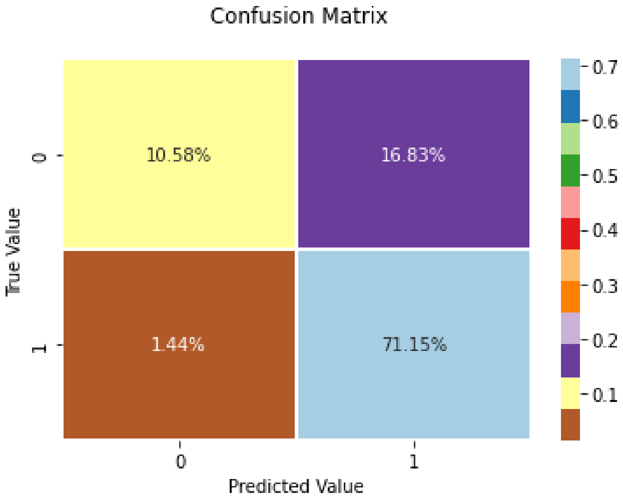
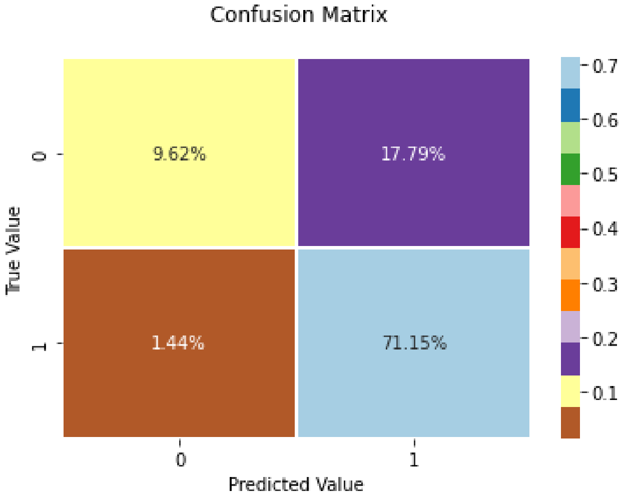
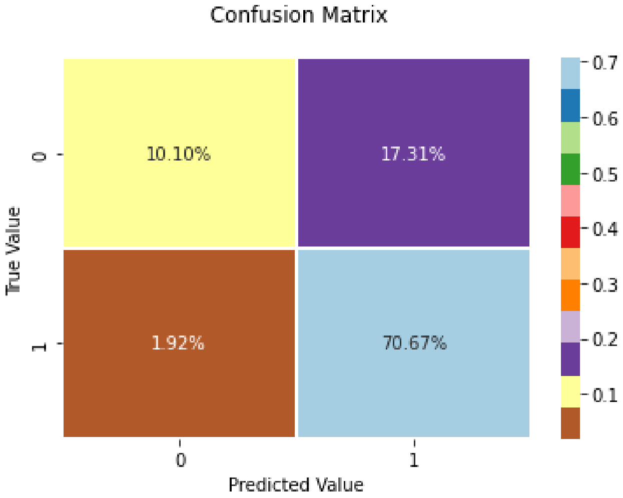
Multi-layer Perceptron (MLP)		83.1
Logistic Regression (LR)		84.6
Convolutional Neural Network (CNN)		80.2

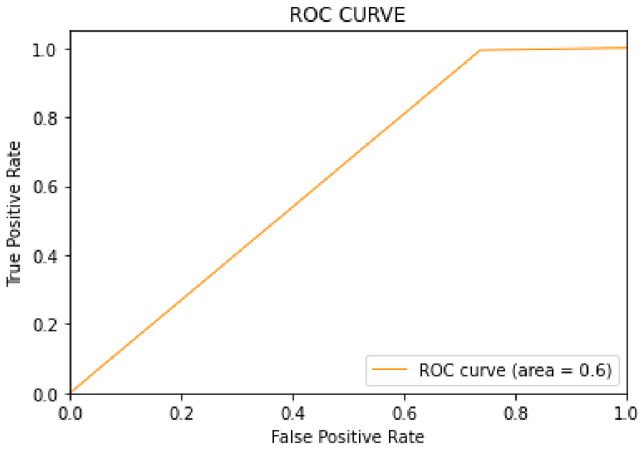
Table 3 Confusion matrix of Machine learning models

From the above table of confusion matrix, in the decision tree model, 72.12% of tweets are positive for actual and predicted values. Where 20.19% of tweets are predicted as positive, but they were classified as negative ones. In the support vector machine model, 16.83% were mislabeled as positive while the actual values were negative tweets. In the multi-layer perceptron model, 10.58% of tweets were predicted as negative while they were actually negative followed by just 1.44% predicted as negative while they were actual positive tweets.

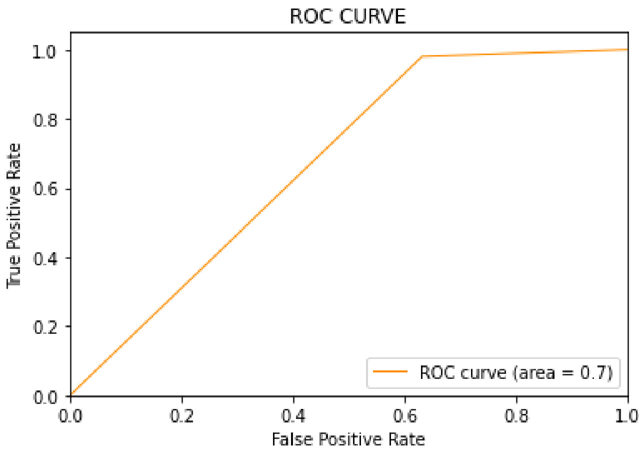
Only 17.79% of negative values were classified as positive in a logistic regression model with 71.15% of values were truly predicted as original tweets in the dataset. Similarly, with a Convolutional neural network using an embedded library, 70.67% of values were accurately predicted from actual values with 10.10% were precisely predicted as negative with real data tweets as negative too.

4.5 Graph Evaluation of Machine learning models

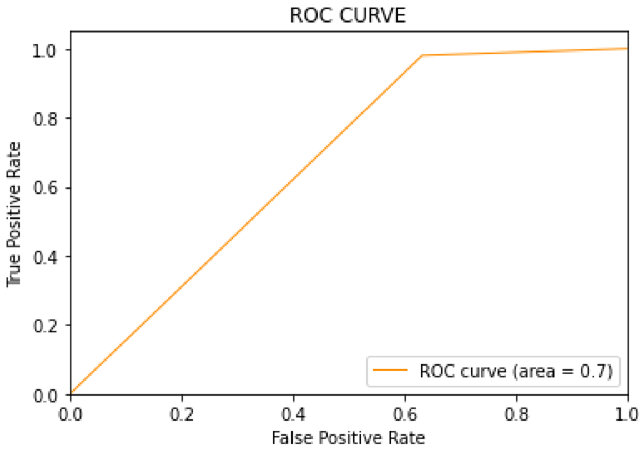
In our research, we have used ROC known as the Receiver Operating Characteristic graph to evaluate classification models at the different thresholds. (developers google, 2022) The graph represents a visual display. It demonstrates key relationships among each potential cut-off and responsiveness for the test run. The ROC curves could indeed show the results of several experiments together.

Machine Learning Models	ML models ROC
Decision Tree (DT)	

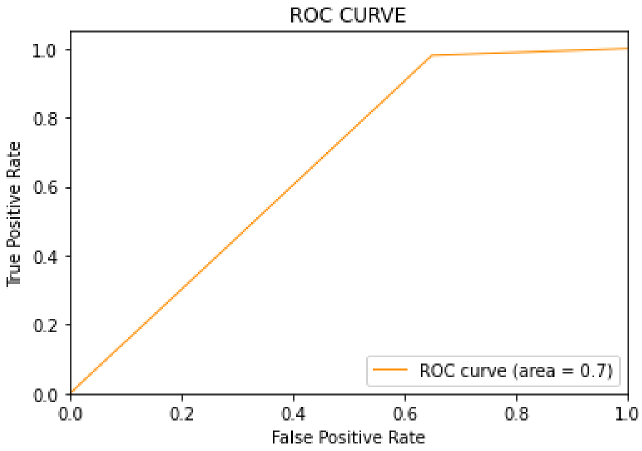
Support Vector Machine (SVM)

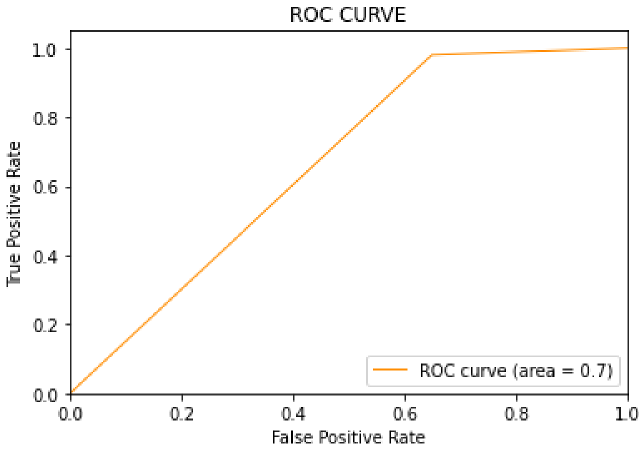


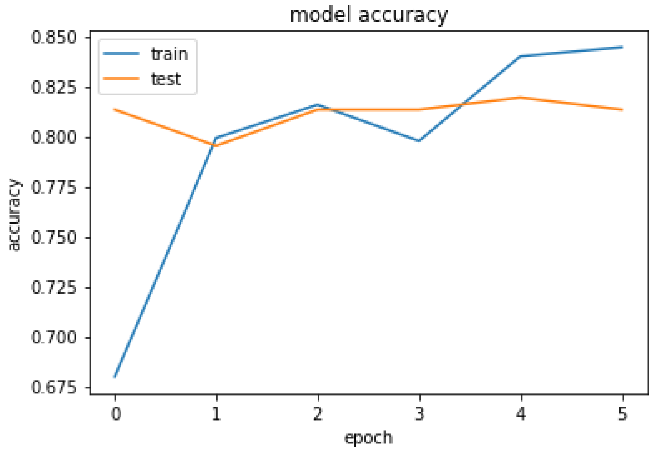
Multi-layer Perceptron (MLP)



Logistic Regression (LR)



<table>
<tr>
<td>Convolutional Neural Network (CNN)</td>
<td>




</td>
</tr>
</table>

Table 4 Graphical plot of Machine learning models

Table 5 shows the ROC curve of ML models where the Support Vector Machine, Logistic Regression, Multi-Layer Perceptron, and Convolutional Neural Network have a ROC area of 0.7 which is defined as a good value among the true positive rate and false positive rate by diverse classification thresholds. As Decision Tree displays the ROC area of 0.6 displaying a lower rate by the classification model. In CNN, two graphs were displayed, where the second graph named as model accuracy demonstrates the accuracy of the epoch utilized in the model history of CNN parameters. It shows the train and test models overlapping to show the similarities and comparable skills between both, the train and test datasets.

4.6 Comparison of Machine Learning models

The metrics of the confusion matrix above are utilized to estimate the accuracy, precision, and recall metrics. In our research, we have used scikit-learn library functions to calculate the precision, accuracy, and recall. Accuracy is calculated by the model's precise predictions of the data. Precision scores are calculated using the formulas (Jeremy Jordan, 2017) as
Precision = True Positives / (True Positives + False Positives)
Where Recall is calculated using the formula (Jeremy Jordan, 2017) in the library as
Recall = True Positives / (True Positives + False Negatives)

Table 6 below shows the accuracy, precision, and recall of all models. The highest accuracy of the logistic regression model as 86.61% followed by MLP at 83.17% proceeded with SVM, CNN, and DT models. The highest precision was recorded for the support vector machine and multi-layer perceptron model as 80.11% proceeded with LR and CNN model with least percentage of decision tree model at 78.12 %. The maximum recall was noted for DT as 99.34% with SVM and MLP at 98.68% followed by the least for LR and CNN at 98.01.

Machine Learning Models	Accuracy	Precision	Recall
Decision Tree (DT)	79.32%	78.12%	99.34%
Support Vector Machine (SVM)	81.25%	80.11%	98.68%
Multi-layer Perceptron (MLP)	83.17%	80.11%	98.68%
Logistic Regression (LR)	84.61%	80.00%	98.01%
Convolutional Neural Network (CNN)	80.28%	79.57%	98.01%

Table 5 Accuracy, Precision, and Recall of Machine Learning Models

5. CHAPTER FIVE: CONCLUSION AND FUTURE WORK

In conclusion, our research studied sentiment analysis of COVID data on Twitter by means of various Machine learning models. We analyzed that the public opinion leaders have a more positive manner during the COVID-19 global outbreak on Twitter.

A deeper look at the datasets can have potential analytical problems. Despite the outcome, the observations demonstrate that all opinion leaders have a fairly positive framework. The confusion matrix and model accuracy discussed above indicates a more optimistic context that shapes public opinion and behavior. Using Machine learning models, it is observed that the ML model's performed effectively with the dataset by good observations recorded. The sentiment analysis on tweets during covid by opinion leaders from different countries shows that they promote a more safer environment to protect people and guide them to '#getboost' vaccinated in order to decrease the risk of covid. The sentiment analysis case study on covid using machine learning models also reveals the negative emotions during the tweets that subject to sadness and fear among people. This analysis gives us the opportunity to specify and eliminates the tweets while considering the strong negative impact on people causing them or leading to depression and other issues.

Due to the lack of time, various parts have been left for the future. it would be useful if future research should expand by specifying tweets with related to covid and evaluating people's comments on every tweet in support or opposition of the specific tweet by public leaders. This research was only conducted for English tweets and can be extended to include other languages. This can also be expanded by implementing other machine learning algorithms and combining two machine learning models as one algorithm while testing the dataset. The research is only aimed at a single social media environment as twitter, future research can be done using a related famous platform like Instagram and Facebook. Moreover, it is crucial to explore further social media platform with respect to sentiment analysis using machine learning models.

6. REFERENCES

Amin, M. Z., & Nadeem, N. (2018). *Convolutional Neural Network: Text Classification Model for Open Domain Question Answering System*.

Aniruddha Bhandari. (2020). *Confusion Matrix*. https://www.analyticsvidhya.com/blog/2020/04/confusion-matrix-machine-learning/

Ansari, A. F., Seenivasan, A., Anandan, A., & Lakshmanan, R. (2017). *CS5228 Project 2 Twitter Sentiment Analysis Group No. : 29*.

Bi, Q., Goodman, K. E., Kaminsky, J., & Lessler, J. (2019). What is machine learning? A primer for the epidemiologist. *American Journal of Epidemiology*, *188*(12), 2222–2239. https://doi.org/10.1093/aje/kwz189

DataCamp. (2018). *DT*. https://www.datacamp.com/tutorial/decision-tree-classification-python

Datacamp wordcloud. (2019). *Datacamp wordcloud*. https://www.datacamp.com/tutorial/wordcloud-python

developers google. (2022). *ROC graph*. https://developers.google.com/machine-learning/crash-course/classification/roc-and-auc

el Barachi, M., AlKhatib, M., Mathew, S., & Oroumchian, F. (2021). A novel sentiment analysis framework for monitoring the evolving public opinion in real-time: Case study on climate change. *Journal of Cleaner Production*, *312*. https://doi.org/10.1016/j.jclepro.2021.127820

Flores-Ruiz, D., Elizondo-Salto, A., & Barroso-González, M. D. L. O. (2021). Using social media in tourist sentiment analysis: A case study of andalusia during the Covid-19 pandemic. *Sustainability (Switzerland)*, *13*(7). https://doi.org/10.3390/su13073836

H. Manguri, K., N. Ramadhan, R., & R. Mohammed Amin, P. (2020). Twitter Sentiment Analysis on Worldwide COVID-19 Outbreaks. *Kurdistan Journal of Applied Research*, 54–65. https://doi.org/10.24017/covid.8

Hatami, Z., Hall, M., & Thorne, N. (2021). Identifying Early Opinion Leaders on COVID-19 on Twitter. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *13094 LNCS*, 280–297. https://doi.org/10.1007/978-3-030-90238-4_20

Jeremy Jordan. (2017). *Metrics Confusion Matrix*. https://www.jeremyjordan.me/evaluating-a-machine-learning-model/

Kausar, M. A., Soosaimanickam, A., & Nasar, M. (2021). Public Sentiment Analysis on Twitter Data during COVID-19 Outbreak. In *IJACSA) International Journal of Advanced Computer Science and Applications* (Vol. 12, Issue 2). www.ijacsa.thesai.org

Mirko Stojiljković. (2022). *Logistic Regression*. https://realpython.com/logistic-regression-python/#:~:text=The%20logistic%20regression%20function%20%F0%9D%91%9D(%F0%9D%90%B1)%20is%20the%20sigmoid%20function,that%20the%20output%20is%200.

Parau, P. (2017). *Opinion Leader Opinion Leader Detection*.

Samuel, J., Ali, G. G. M. N., Rahman, M. M., Esawi, E., & Samuel, Y. (2020). COVID-19 public sentiment insights and machine learning for tweets classification. *Information (Switzerland)*, *11*(6). https://doi.org/10.3390/info11060314

scikit-learn developers. (2022a). *Feature Extraction*. https://scikit-learn.org/stable/modules/feature_extraction.html

scikit-learn developers. (2022b). *MLP*. https://scikit-learn.org/stable/modules/generated/sklearn.neural_network.MLPClassifier.html

scikit-learn developers. (2022c). *SVM*. https://scikit-learn.org/stable/modules/svm.html

The trump archive. (2020). https://www.thetrumparchive.com/insights/frequency

Twitter. (2022). *Twitter*. https://developer.twitter.com/en

Vijay Choubey. (2018). *CNN*. https://medium.com/voice-tech-podcast/text-classification-using-cnn-9ade8155dfb9

Wang, Y., Croucher, S. M., & Pearson, E. (2021). National Leaders' Usage of Twitter in Response to COVID-19: A Sentiment Analysis. *Frontiers in Communication*, *6*. https://doi.org/10.3389/fcomm.2021.732399

WHO. (2020). *Archived: WHO Timeline - COVID-19* . WHO. https://www.who.int/news/item/27-04-2020-who-timeline---covid-19

WHO. (2022). *WHO Coronavirus (COVID-19) Dashboard*. WHO Coronavirus (COVID-19) Dashboard. https://covid19.who.int/

Yue, L., Chen, W., Li, X., Zuo, W., & Yin, M. (2019). A survey of sentiment analysis in social media. *Knowledge and Information Systems*, *60*(2), 617–663. https://doi.org/10.1007/s10115-018-1236-4

Zhu, M., Lin, X., Lu, T., & Wang, H. (2016). *Identification of Opinion Leaders in Social Networks Based on Sentiment Analysis: Evidence from an Automotive Forum*.