Sentiment Analysis during COVID-19 Pandemic using Social Media Data

Mushfiqul Alam

2016-1-60-040

Nishat Jahan Nishi

2016-1-60-015

Md. Mahadi Islam Mahadi

2016-1-60-061

A thesis submitted in partial fulfillment of the requirements for the degree of Bachelor of Science in Computer Science and Engineering

Department of Computer Science and Engineering

East West University

Dhaka-1212, Bangladesh

October, 2020

Declaration

We, hereby, declare that the work presented in this thesis is the outcome of the investigation performed by us under the supervision of Dr. Mohammad Rezwanul Huq, Associate Professor, Department of Computer Science and Engineering, East West University. We also declare that no part of this thesis hasbeen or is being submitted elsewhere for the award of any degree or diploma.

Countersigned

. . . . . .

(**Dr.Mohammad Rezwanul Huq**)

Supervisor

Signature

. . . . . .

(**Mushfiqul Alam)**

ID: 2016-1-60- 040)

Signature

. . . . . .

(**Md**. **Mahadi Islam Mahadi**)

(**ID: 2016-1-60-061**)

Signature

. . . . . .

(**Nishat Jahan Nishi**)

(**ID: 2016-1-60-015**)

Letter of Acceptance

This thesis report entitled “Sentiment Analysis during COVID-19 Pandemic using Social Media Data"submitted by Mushfiqul Alam (ID: 2016-1-60-040), Md. Mahadi Islam Mahadi (ID: 2016-1-60-061) and Nishat Jahan Nishi (ID: 2016-1-60-015) to the Department of Computer Science and Engineering, East West University is accepted by the department in partial fulfillment of requirements for the Award of the Degree of Bachelor of Science and Engineering on October, 2020.

Supervisor

. . . . . .

(Dr. Mohammad Rezwanul Huq)

Associate Professor,

Department of Computer Science and Engineering, East West University

Chairperson

. . . . . .

(Dr. Taskeed Jabid)

Chairperson and Associate Professor,

Department of Computer Science and Engineering, East West University

Acknowledgment

First we would like to express our deepest gratitude to the almighty for His blessings on us to perform the research successfully. After that we would like to express our sincere gratitude and appreciation to our honorable thesis supervisor, Dr. Mohammad Rezwanul Huq, Associate Professor, Department of Computer Science and Engineering, East West University, Bangladesh, who gave us the opportunity to integrate us into the sector "Sentiment Analysis during COVID-19 Pandemic using Social Media Data" and this work wouldn't have been possible without him. His guidance helped us in all the time research and writing of this thesis. His encouragement, visionaries and thoughtful comments and suggestions, unforgettable support at every stage of our B.Sc study were simply appreciating and essential.

Ultimately, we had the greatest debt to our dearest family for their patient, support, dedication, blessing, constant motivation and prayer. Several other individuals have already shown their Constant support and encouragement in different ways, explicitly or implicitly tied to our academic lives. We will hold them in our hearts and hope that they will be in the right place to acknowledge them in the future.

Mushfiqul Alam

October, 2020

Md. Mahadi Islam Mahadi

October, 2020

Nishat Jahan Nishi

October, 2020

Abstract

The latest outbreak of coronavirus disease in 2019 (COVID-19) has had a major effect on human life. In addition to overt physical and economic risks, the pandemic also indirectly impacts people's sentiments. The purpose of this research is to successfully predict people's sentiment during Covid-19 pandemic by using social media data. We build the Covid19-Tweet dataset by twitter API (twitterscraper) for the training purpose by 25570 English tweets and determining whether a piece of writing is positive, negative or neutral sentiments. With different combinations of different data models and machine learning algorithms, this research has been able to predict sentiments of twitter users with 89% accuracy. We analyze the accuracy to use Decision tree, Random Forest Classification (RFC) and LSTM recurrent neural network for sentiment classification of COVID-19 tweets**.** In this research, LSTM algorithm implemented on a dataset consisting responses has produced the best performance.

Table of Contents

List of Figures

List of Tables

List of Algorithms

Chapter 1

Introduction

1.1 Introduction

Mental health refers to psychological, behavioral, and emotional states of a person. According to WHO, mental well-being is a condition in which an person knows his or her own abilities, can cope with the normal stresses of life, can work productively and is able to do something for the betterment of the society. It shapes the way how a person thinks, feels and acts. Mental health is as important as physical health. At least 10% population of the world is affected by mental health conditions. WHO estimates that one in four people will be affected by mental health conditions at some time in their lives. There are many factors which are responsible for mental health problems including trauma, abuse, biological factors, family history, socioeconomic conditions, the extent of social participation of a person.

During the pandemic Covid-19 outbreak, it's adversely affecting the physical health and also the mental health of the peoples. Due to social distancing and lockdown, people are isolated. There are changes in their routine such as eating, sleeping pattern, worrying about their loved one's life and health, about their economic condition and these factors can lead a person to feel stress, depression, and anxiety which can affect mental health. So, we need to consider the mental states of the people due to covid19 to get organized in response to rising mental disorders. Due to lockdown everyone is at home and many of them shifted their workspace online and as a result of it people are spending more time online and on social media than usual and the intention of sharing personal feelings, thoughts, experiences through the internet is increasing day by day on social media including Facebook, Instagram, and Twitter. These online contents can be used to detect people's emotions or mental states using natural language processing (NLP) methods. In this work, we applied an LSTM text classifier, Random Forest, Decision Tree approach on the tweet dataset to predict the people's sentiment and find out the best model for this work.

1.2 Motivation

In leading a healthy, happy life, mental well-being is a must. It affects how a person handles stress or any situation or relates one thing with another. It influences a person's lifestyle and the others who are depending on him/ her. The foundation for individual welfare and the efficient functioning of a community is mental health. Being mentally healthy can increase productivity and proficiency in a person's daily activities. It enables an individual to accommodate changes in their lives and to cope with adversity.

So, one needs to give equal importance to mental health like physical health. According to WHO one in four people will be affected by mental health conditions at some time in their lives and this ratio can go higher. In this Covid-19 pandemic, peoples are isolated and this loneliness, fear, or the interrupt of a regular routine for a long time can cause stress, depression, anxiety, and many other mental health conditions. Before people have this tendency not to talk or share about their mental health condition even if it is spoken it confined within a small area like in families or whisper but now through the internet on social medias people are sharing their thoughts, ideas, feelings. From these data, we can identify the mental conditions of them using machine learning methods as it is an ever-growing field in the current world and predict how many people need a consultation to get stable mental health and a healthy society.

1.3 Challenges

We faced a few challenges while working on this work. On average 6000 tweets are tweeted every second and 500 million tweets per day. There are about 628 million tweets about Covid-19 and Coronavirus until now. But fetching the tweet data was difficult for the restricted availability of the APIs. Another challenge was there is lack of labeled dataset so we had to label the dataset on our own.

1.4 Objective

The main objective of our research is:

.

- To understand how people are reacting all over the world due to COVID – 19 pandemic.
- To find out the what is the major emotion is effecting peoples mental health.

1.5 Contribution

Here is our contribution to this research:

• We have created a dataset of our own by using twitter scrapper.

• By using Regular Expression, data preprocess was done.

• Three classification algorithms are applied to the selected features.

• We used Textblob to find out the sentiment polarity.

• We have observed which algorithm performs better by calculating accuracy, f1 measure, precision, recall.

1.6 Report Overview

Chapter 1: Chapter 1 introduces our motivation to predict the sentiment, the main objectives of our research, the contributions that we have made regarding the work.

Chapter 2: This chapter illustrates the background study.

Chapter 3: Chapter 3 gives a brief overview of the datasets and the data preprocessing techniques.

Chapter 4: Chapter 4 describes the proposed model, feature extraction technique, implementation process, algorithms that are used.

Chapter 5: Chapter 5 analyzes the results obtained from our proposed methods.

Chapter 6: Chapter 6 summarizes the overall work that we have done and our future plans.

Chapter 2

Background Study

There are many contributions made by many scholars and researchers regarding sentiment analysis.Due to the pandemic Covid-19 and its adverse impact on both people's mental and physical health,recently many researchers are working on the impact of covid-19 on mental health by natural language process methods or questionnaire classification which learns a mapping function from the labeled training dataset.

The work by Wolohan and Hamilton (2020) developed a LSTM text classifier [1] using fastText word embeddings at predicting user-level depression and used that classifier to estimate the population rate of depression. They used Reddit data to assess the possible effect of COVID-19 on depression and used a new sample of 20,000 Reddit responses during the first six months of 2018, 2019 and 2020 to quantify the population of depression during the COVID-19 pandemic. A comparative time-series study of three phases was carried out to assess the impact of the COVID-19 pandemic on language use. Their LSTM model achieved an AUC of 0.93 and an F1 score of 0.92 and indicates that the population rate of depression may increase by 50% in the first four months of 2020 relative to the first four months of 2019 and 2018.

The objectives of authors in [2] are to discover the polarity of public views published in a blog in Roman-Urdu with a mixture of English and Urdu languages. Naïve Bayesian, Decision Tree and KNN classification models were used to classify positive and negative emotions. The opinions written in Roman-Urdu are extracted from a blog containing public feedback on the "Facebook Usage Effect" using the Simple Web Extractor program. They used TF-IDF to apply weights to terms of relative significance in the corpus. They trained their machine with a dataset containing 150 positive and 150 negative opinions to build three models. They found that the Naïve Bayes algorithm worked well in terms of higher precision, higher accuracy, higher recall and higher F-value relative to the Decision Tree and KNN where Decision Tree was the fastest and KNN was the slowest classification technique.

A recent work in [3] where the authors build the EmoCT (Emotion-Covid-19-Tweet) dataset with BERT model for classifying COVID-19-related tweets into eight different emotions: anger, anticipation, disgust, fear, joy, sadness, surprise and trust using NLP. They suggested two models for both single-label and multi-label classifications, based on a multilingual BERT model with a learning rate of 0.00001, and obtained a promising result. To explain the reasons why the public may feel fear or sad, Attention Weight and POS tagging approaches were used to measure keywords correlations.

Another work in [4], where the authors presented a systematic framework based on NLP which is capable of extracting COVID-19 related comments on Reddit and proposed a deep learning model based on LSTM for sentiment classification. They collected 563,079 COVID-19 related comments from Reddit between January 20, 2020 and March 19, 2020 and for classifying the sentiment of the COVID-19 comments, they labeled each of the comment sentiment as very positive, positive, very negative, negative,

and neutral based on the sentiment score obtained using Sentistrength method. Their approach based on the LSTM model achieved 89.01% accuracy.

The authors in [5] introduced an approach to optimize word-embedding for classification based on tweet messages, with an emphasis on finding users suffering from depression. They used CLPsych2015 dataset for depression detection and Bell Let's Talk dataset for test generalization capability. The data was labeled as Control, Depressed, PTSD and also according to their gender and age. To evaluate the performance of depression detection, four neural network model were used where first three model uses CNN and the last model use RNN. The result shows that CNN based model performed better than RNN based model.

The authors in [6] presented textual analyses of Twitter data to identify public sentiment. They identified the fear sentiment and negative sentiment over time due to Covid-19. They showed the use of methods of exploratory and informative textual analysis and textual data visualization to discover early stage observations , such as grouping terms by levels of a particular non-text variable and provided a comparison of textual classification mechanisms used in artificial intelligence applications, and demonstrated their usefulness.

The authors in [7] introduced a model that applies the Random Forest algorithm, augmented by the AdaBoost algorithm in place of the conventional healthcare system to classify Covid-19 infected person. The dataset that was used in this work was accessed from Kaggle as "Novel Corona Virus 2019 Dataset" and it was further pre-processed. They compared between Decision Tree Classifier, Support Vector Classifier, Gaussian Naïve Bayes Classifier, and Boosted Random Forest Classifier and find out the best method for processing data. The analysis reveals that Boosted Random Forest performs better with an accuracy of 94% when predicting COVID-19 patients than other algorithms.

The authors in [8] analyzed the sentiments of people from Covid-19 tweets by using deep learning classifiers and proposed a Gaussian membership feature based on a fuzzy rule base to accurately classify feelings from tweets which gives an accuracy of 79%. In this work, two sets of datasets were used. Tweets were analyzed by unigrams and bigrams and they were labeled as positive, negative and neutral sentiments. Naïve Bayes Models, Ensemble models, Support Vector Machine Models, Linear Models, Multinomial Classifier, Bernoulli Classifier, AdaBoost Classifier, Linear SVC Classifier, Logistic Regression Classifier and Random Forest Classifier have been used as deep learning classifiers on bag-of-words model and Doc3Vec models for the two datasets where the accuracy yields up to 75% and 81%. The Susceptible Exposed Infectious Recovered (SIER) model was used to forecast the outbreak of the disease as the key idea was to measure the mortality and recovery rates using computational modelling.

Chapter 3

Dataset Overview

3.1 Data Types

We have used textual data collected from social media status for our work. People are increasingly using social media platforms to share their opinions and beliefs. As people are posting about everything in the context of everyday happenings and activities continuously, we can use these sentiments and expressions used in social posts to capture the emotion of a person behind the social media. So, we choose Twitter as the center of our textual data.

3.2 Data Collection

Twitter is known as one of the most popular websites for social networking or micro-blogging and there are about 628 million tweets related to covid-19. So, we applied Twitter API named twitterscraper to get Covid-19 related tweets. Twitter API stands for Application Programming Interface. API works as an intermediate system which delivers a request for certain data to the provider and returns the response to the requester.

Required API key fields which are begin_date, end_date, limits, and language was filled for collecting tweet data. We set our limit 6000 for every month and set the language as English. We collected the tweets by query_tweets() function. We have searched tweets with keywords like coronavirus, CoronavirusPandemic, covid19, mental health, isolation, anger, anxiety, loneliness, workfromhome, depression. We collected 25570 English tweets from December 2019 and April 2020 to August 2020.

3.3 Attributes of Dataset

There are a total number of 22 columns in our dataset. The feature of these columns are Unnamed, has_media, hashtags, img_urls, is_replied, is_reply_to, likes, links, parent_tweet_id, replies, reply_to_users, retweets, screen_name, text, text_html, timestamp, timestamp_epochs, tweet_id, tweet_url, user_id, video_url and username. The textual data that we have used is text which is in column 9. We have used also tweet_id to identify the texts which is in column 5.

3.4 Data Preprocessing

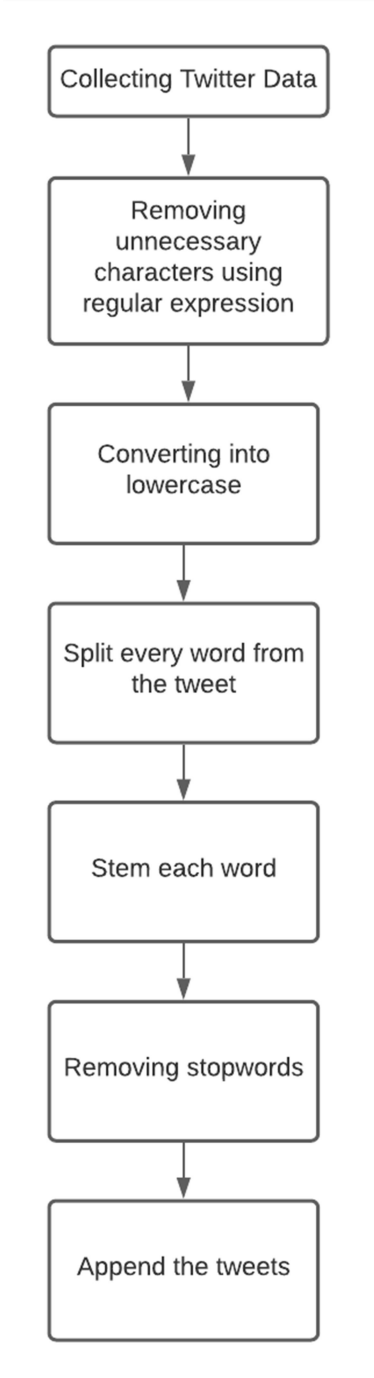

Figure 3.1: Preprocessing for Facebook status

Figure 3.1 shows the preprocessing steps performed for our dataset. The accuracy of the model could be very low because there could be lots of noise in the dataset. To increase the quality of data, preprocessed the text data. Regular expression is used for removing HTML tags, links, extra whitespaces, hash. After

that all alphabets was converted into lowercases. Then every tweet was splitted into single words. Stemming was done by using Natural Language Processing Toolkit (NLTK) SnowballStemmer.

Chapter 4

Proposed Model

4.1 Proposed model

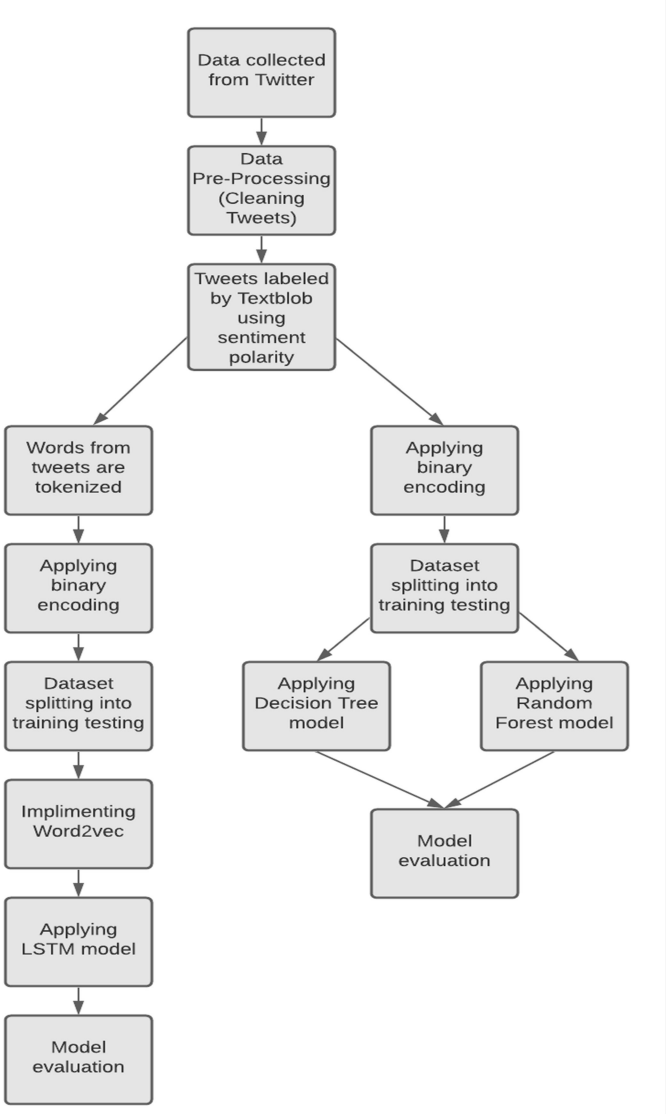

Figure 4.1: Proposed Model for the Experiment

Figure 4.1 demonstrate our proposed model. First of all, twitter data has been collected from twitter API named twitterscraper to get Covid-19 related tweets. Then clean the dataset by preprocessing and label the dataset by using texblob. We used TextBlob library and applied two function label1 and label2 on polarity. We labeled the data based on sentiment: positive, negative and neutral. Then we split the dataset into train (80%) and test (20%) set. After that, the model was built by applying various algorithms with machine learning classifiers. The following three algorithms are used for classification – Decision Tree, Random Forest Classifier (RFC) and LSTM (Long Short-Term Memory). Finally, evaluation matrices have been used to find the best model.

Algorithm	Parameters	Value
Random Forest	n_estimators	25
	criterion	Entropy
	min_samples_split	2
	bootstrap	True
Decision Tree	min_samples_split	7
	criterion	Entropy
	max_depth	None
LSTM	vocab_size	60000
	w2v_size	3
	w2v_window	7
	w2v_min_count	30
	Dense layer	3
	Hidden state	128
	Activation function	softmax
	Loss function	categorical_crossentropy
	Learning rate	0.0001

Table 4.1: Parameters of the algorithms

Table 4.1 shows the parameters and values of Random Forest, Decision Tree and LSTM Text classifier model. Here in the random forest model n_estimators is the number of the decision tree. The lowest number of samples required for splitting an internal vertex is min_sample_split, criterion is to measure the

quality of a spilled. A bootstrap sample is used for better performance. For the Decision Tree model, a criterion is used to measure the quality of a spilled, min_sample_spilt is the lowest number of sample which is required for splitting an internal vertex. Max_depth is the maximum depth of a tree.

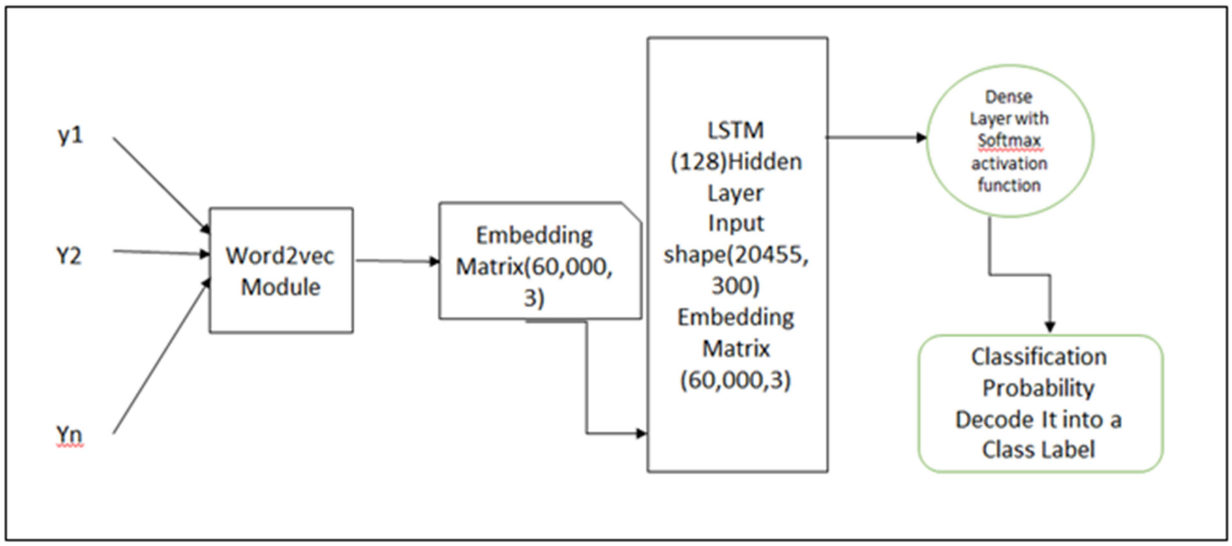

Figure 4.2: The architecture of the LSTM model

In figure 4.2 shows the architecture of the Long Short-Term Memory (LSTM) model.We have use sequential for generating the model where each layer has a weight. An embedding layer is employed to transform each word as a vector in a high dimensional space. The size of the embedding matrix is (60000,3) where 60000 is vocabulary size and each row has 3 values. We have used word2vec for creating the embedding. The word2vec size is the dimensionality of the each, window = 7 which defines the maximum distance within current and the predicted words, min_countignores all the words with the total frequency lower than this value, workers implemented for faster training of the model.

The LSTM model has 128 hidden state and the dense layer has the dimensionality of 3 which denotes the output of our model. The activation function of the model is "softmax" and the loss function is "categorical crossentropy". As an optimizer the model used Adam with the learning rate of 0.0001.

4.2 Feature Extraction Techniques

Word2vec is a two-layer neural network that handles text by vectoring words. The input is a corpus of text and the output is a vector collection. Although Word2vec isn't a deep neural network, it transforms text into a numerical form that can be interpreted by deep neural networks. Word2vec's applications do more than just parsing sentences. It can be applied to ascertained patterns like genes, code, social media graphs and other verbal or symbolic series [9].

The purpose and utility of Word2vec is to group in vector space the vectors of related terms together. Word2vec generates vectors which are distributed numerical representations of word characteristics, characteristics such as individual word meaning. Word2vec may make extremely precise guesses about the meaning of a word based on past appearances, provided enough evidence, use and contexts. Those guesses may be used to connect a word with other words or cluster documents and categorize them by

subject. Those clusters can form the basis of search, sentiment analysis and recommendations in many diverse fields [9]. To generate a distributed representation of words, Word2vec can use one of two model architectures, Skip-grams and Continuous Bag of Words model (CBOW). The most similar word of “covid19” are “base”, ”usatapaboca”, ”sit”, ”21”, ”salir”, ”absolut”, ”sometime”, ”da”,”apart” and “ficaemcasa”.

4.3 Algorithms

Implementing the proposed model, we have used three machine learning classifiers. They are - a) Decision Tree, b) Random Forest and c) Long short- term memory (LSTM) Classifier Here they are:

Decision Tree: The decision tree is to construct a training model that can be used to forecast the class or value of the target variable by studying basic decision rules obtained from previous data (training data). In Decision Trees, for predicting a class labels for a record it start from the root of the tree. A decision tree is a flowchart like tree structure. Each node of the tree denotes a test on an attribute, branches represents the outcome of the test and each leaf node represent a class label.

Random forest: Random forest is a supervised learning algorithm which is built from decision trees. Random forests merge the versatility of decision-making trees with flexibility resulting in a vast increase in accuracy. Each node in the decision tree uses a random subset of features to determine the output. To generate the final result the random forest combines the output of individual decision trees.

Long Short-Term Memory (LSTM): LSTM is a special form of recurrent neural network (RNN) that can learn long-term dependencies. It is beneficial for those modes of prediction that enable the network to maintain data over longer periods of time. Compared to conventional RNNs, LSTMs are designed to solve the vanishing gradient problem and allow them to maintain information for longer periods. LSTMs

can maintain a persistent error that enables them to continue learning over time and layers over multiple time-steps and back propagate. It is used on the basis of time series data for encoding, predicting and classifying [10]. LSTM has a chain structure comprising cells with four neural networks and multiple memory blocks. The cells hold information and the memory manipulations are performed by the gates [11].

4.4 Performance evaluation techniques

There are many criteria that you might use to judge how effective a classification model is. Precision, recall and F1 use positives and negatives to measure a model accurate when making predictions. We Used some performance evaluation techniques of machine learning and they are as follows.

- Accuracy
- Precision
- Recall
- F1

Accuracy: Accuracy is the ratio the number labeled correctly over the total number. It measures how much of the data labeled correctly classifies a data point out of all the total data points. Accuracy is calculated as shown in Equation as follows:

$$\text{Accuracy} = (TP + TN) / \text{Total Sample}$$

Here, TP, TN, FP, FN are respectively True Positive, True Negative, False Positive and False Negative. True positive is where the predicted class label is positive and the actual class is positive. True Negative is where the predicted class label is negative and the actual class is negative. False Positive is where the predicted class value is positive and the actual class is negative. False negative is where the predicted class value is negative and the actual class is positive.

Precision: It shows what fraction of the positive class predictions is actually positive. Using the following formula to measure accuracy:

$$\text{Precision} = TP / (TP + FP)$$

Recall: Recall is equal to the ratio of the True Positive (TP) samples to the sum of True Positive (TP) and False Negative (FN) samples [7]. Recall is calculated as given follows:

$$\text{Recall} = TP / (TP + FN)$$

F1 score: F1 score is equal to the harmonic mean of Recall and Precision value. The F1 Score strikes the perfect balance between Precision and Recall [7]. This is the most significant measure that we will be using to evaluate the model. F1 Score can be calculated as follows:

F1 score = 2 * (precision * recall) / (precision + recall)

Chapter 5

Performance Evaluation and Discussion

In this section performance of all the considered approaches have been evaluated and analyzed using various performance measurements. To analysis the positive sentiment, negative sentiment and neutral sentiment during covid-19 pandemic from the twitter dataset we used to train several Machine learning classification models. The models included in this study are: Decision Tree, Random Forest Classifier (RFC) and LSTM (Long Short-Term Memory). The accuracy of these tree models are shown in table bellow:

Table 5.1: Accuracy for sentiment analysis model for different algorithms

Algorithm	Accuracy
Decision Tree	0.6774
Random Forest Classifier	0.7413
LSTM(Long Short-Term Memory)	0.8965

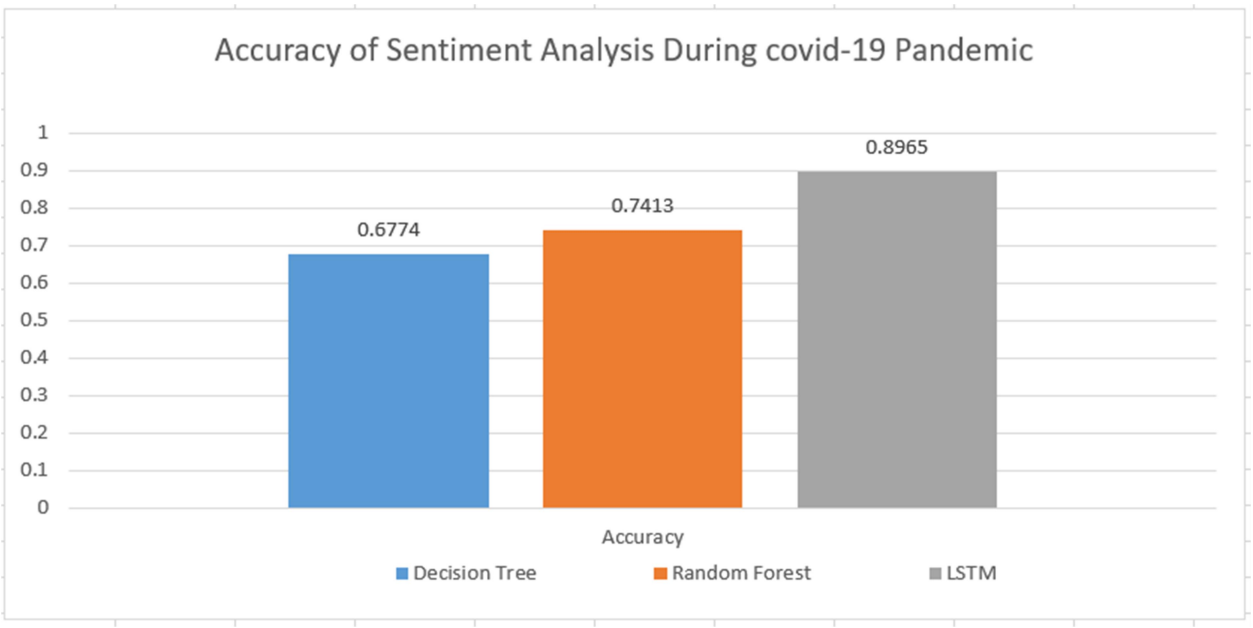

Figure 5.1: Accuracy scores for Data model

Table 5.1 and Figure 5.1 shows accuracy values are 67.74%, 74.13%, 89.65% respectively for decision tree, random forest and LSTM classifiers. Here shows that LSTM algorithm produce the best performance with 89.65% accuracy. LSTM give better accuracy where random forest (74.13%) gives 6.39% better accuracy than decision tree (67.74%).

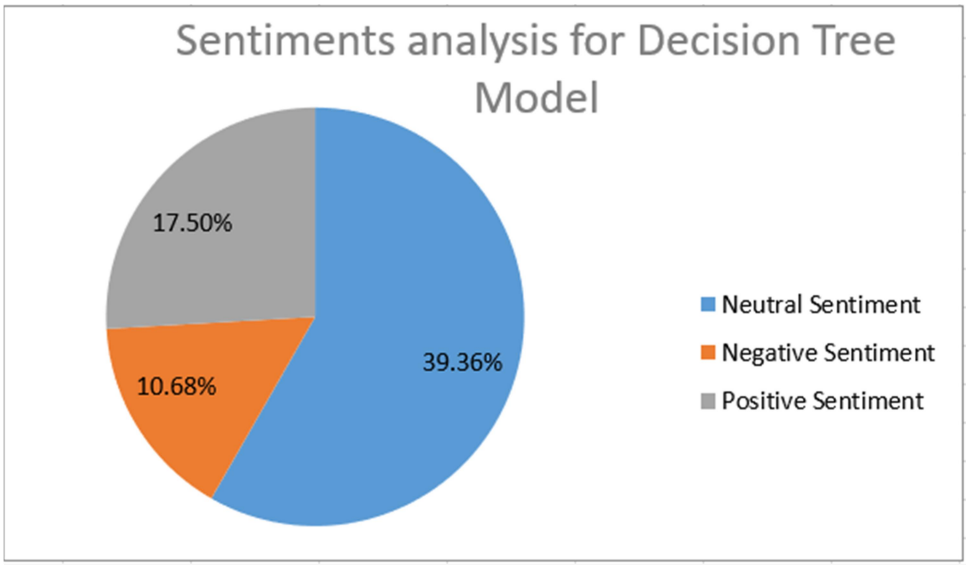

Figure 5.2: Ratio of positive, negative and neutral sentiment for Decision Tree model

In the figure 5.2 we can see that 39.36% (2013 tweets) showed neutral sentiments during COVID-19 Pandemic in their tweet status. Where respectively 10.68% (556 tweets) shows negative sentiment and 17.50% (895 tweets) positive sentiment for testing sample (5114 tweets).

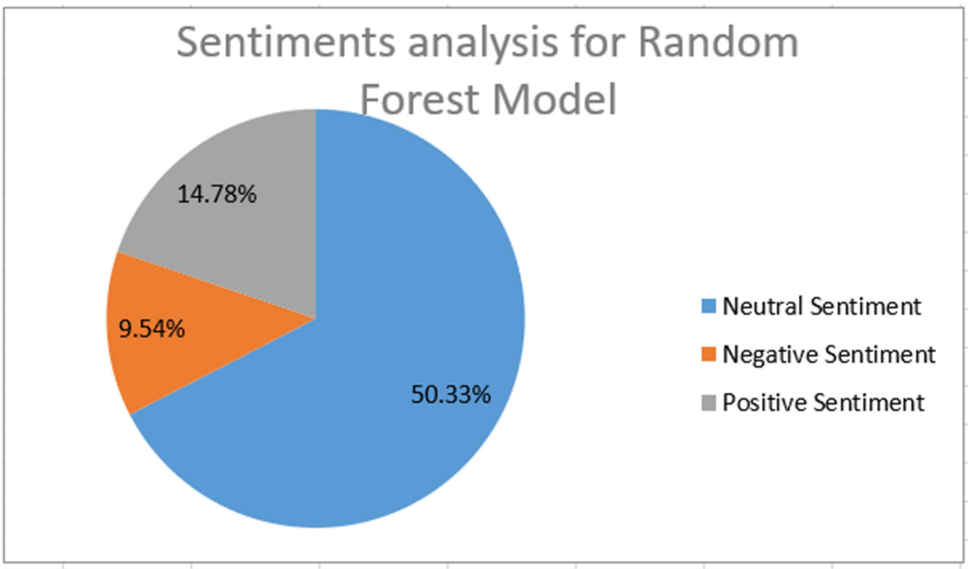

Figure 5.3: Ratio of positive, negative and neutral sentiment for Random Forest model

In the figure 5.3 we can see that 50.33% (2547 tweets) showed neutral sentiments during COVID-19 Pandemic in their tweet status. Where respectively 9.54% (488 tweets) shows negative sentiment and 14.78% (756 tweets) shows positive sentiment for testing samples (5114 tweets).

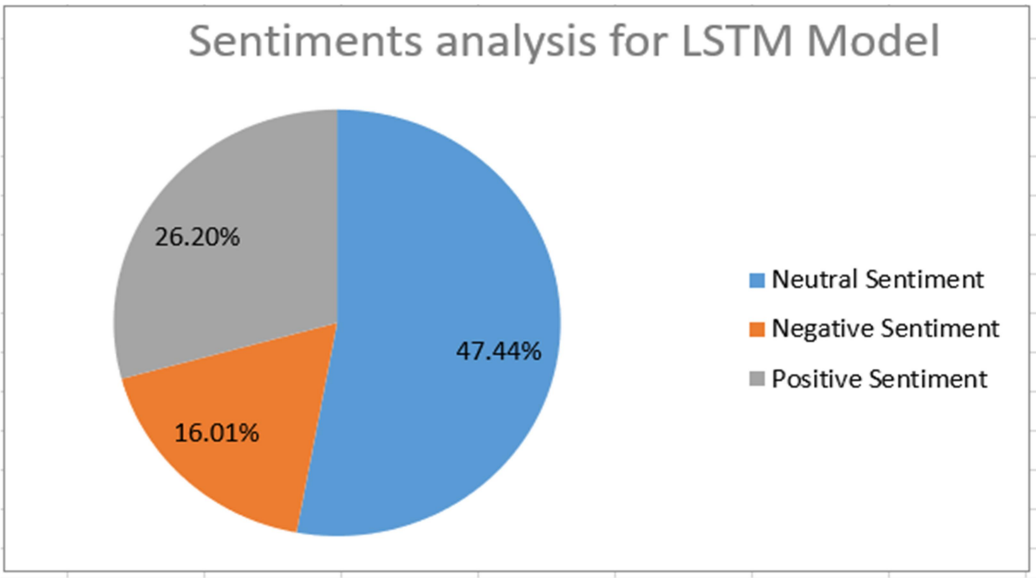

Figure 5.4: Ratio of positive, negative and neutral sentiment for LSTM model

In figure 5.4 shows, the LSTM model 47.44% (2426 tweets) showed neutral sentiments during COVID-19 Pandemic in their tweet status. Where respectively 16.01% (819 tweets) show negative sentiment and 26.20% (1340 tweets) show positive sentiment for testing samples (5114 tweets).

It is still easier to use the confusion matrix as the assessment criterion for the machine learning model. It gives a very simple and efficient performance measure for the model. Here are some of the most common performance measures use from the confusion matrix.

Table 5.2: Measuring precision, recall, f1-score for sentiment analysis using Decision Tree

People Sentiment	Precision	Recall	F1 Score
Neutral Sentiment	0.7730	0.7419	0.7572
Negative Sentiment	0.5577	0.5691	0.5633
Positive sentiment	0.5915	0.6285	0.6094

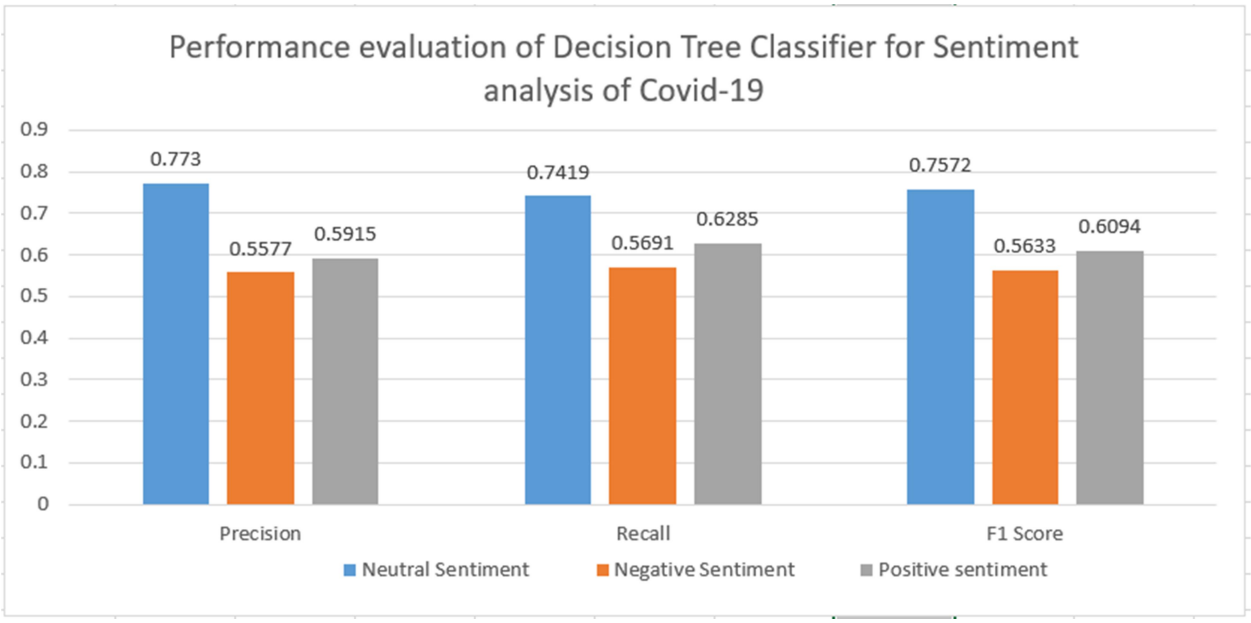

Figure 5.5: Measuring Precision, Recall and F1_score for sentiment analysis using Decision Tree algorithm.

Table 5.2 and figure 5.5 explain that the precision of the Decision Tree model for neutral, negative and positive sentiment is 0.7730 ,0.5577, 0.5915 respectively whereas recall for neutral, negative and positive sentiment is 0.7419, 0.5691, 0.6285 respectively. And the F1_score is 0.7572, 0.5633, 0.6094 respectively.

Table 5.3: Measuring precision, recall, f1_score for sentiment analysis using Random Forest

People Sentiment	Precision	Recall	F1 Score
Neutral Sentiment	0.9781	0.6785	0.8012
Negative Sentiment	0.4895	0.9779	0.6524
Positive sentiment	0.4997	0.8780	0.6369

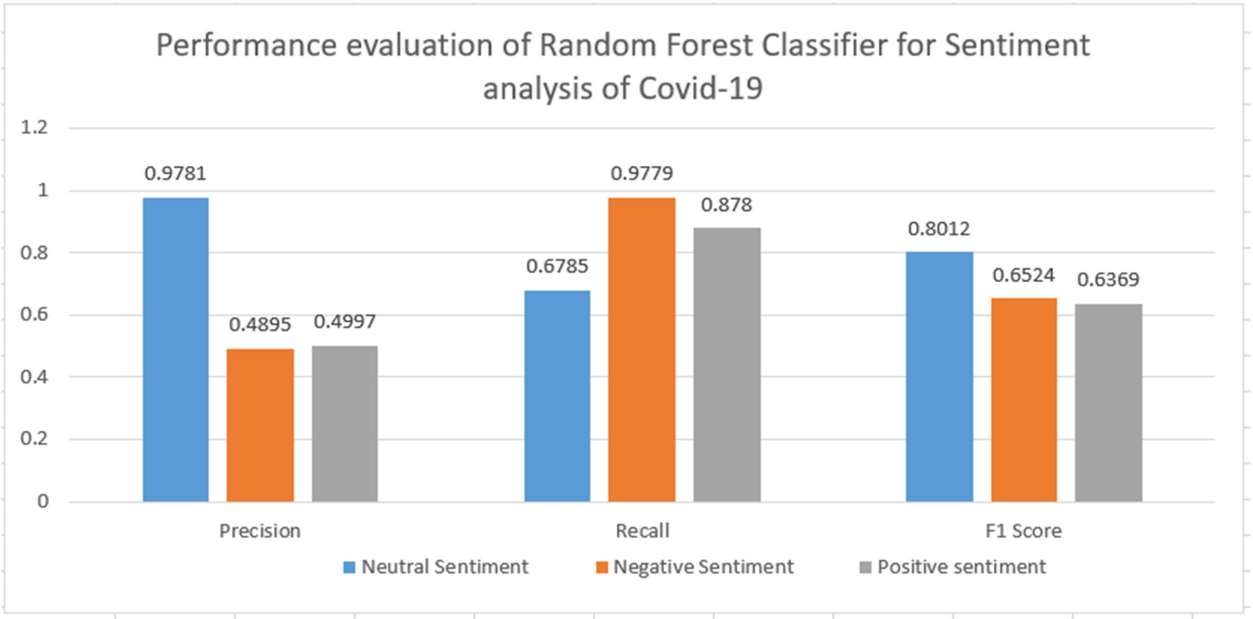

Figure 5.6: Measuring Precision, Recall and F1_score for sentiment analysis using Random Forest algorithm.

Table 5.3 and figure 5.6 explain that the precision of the Random Forest model for neutral, positive and negative sentiment is respectively 0.9781, 0.4895, 0.4997 whereas recall for neutral, negative and positive sentiment is 0.6785, 0.9779, 0.8780 respectively. And the F1_score is 0.8012, 0.6524 ,0.6369 respectively.

Table 5.4: Measuring precision, recall, f1_score for sentiment analysis using LSTM

People Sentiment	Precision	Recall	F1 Score
Neutral Sentiment	0.9316	0.9499	0.9407
Negative Sentiment	0.8215	0.7770	0.7986
Positive sentiment	0.8856	0.8898	0.8877

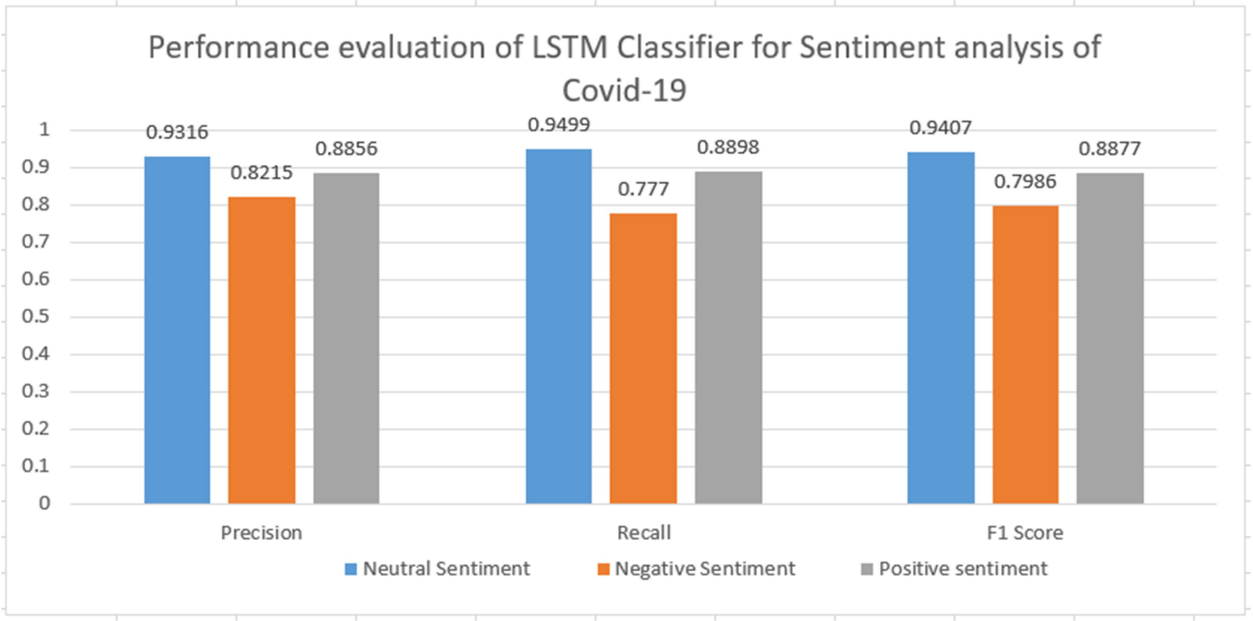

Figure 5.7: Measuring Precision, Recall and F1_score for sentiment analysis using LSTM model.

Table 5.4 and figure 5.7 explain that the precision of the LSTM model for neutral, positive and negative sentiment is respectively 0.9316 , 0.8215, 0.8856 whereas recall for neutral, negative and positive sentiment is 0.9499, 0.7770, 0.8898 respectively. And the F1_score is 0.9407, 0.7986, 0.8877 respectively.

Table 5.5: Precision, recall, F1 score values for data model

Algorithm	Precision	Recall	F1 Score
Decision Tree	0.6747	0.6773	0.6756
Random Forest	0.7959	0.7412	0.7235
LSTM	0.8984	0.8965	0.8973

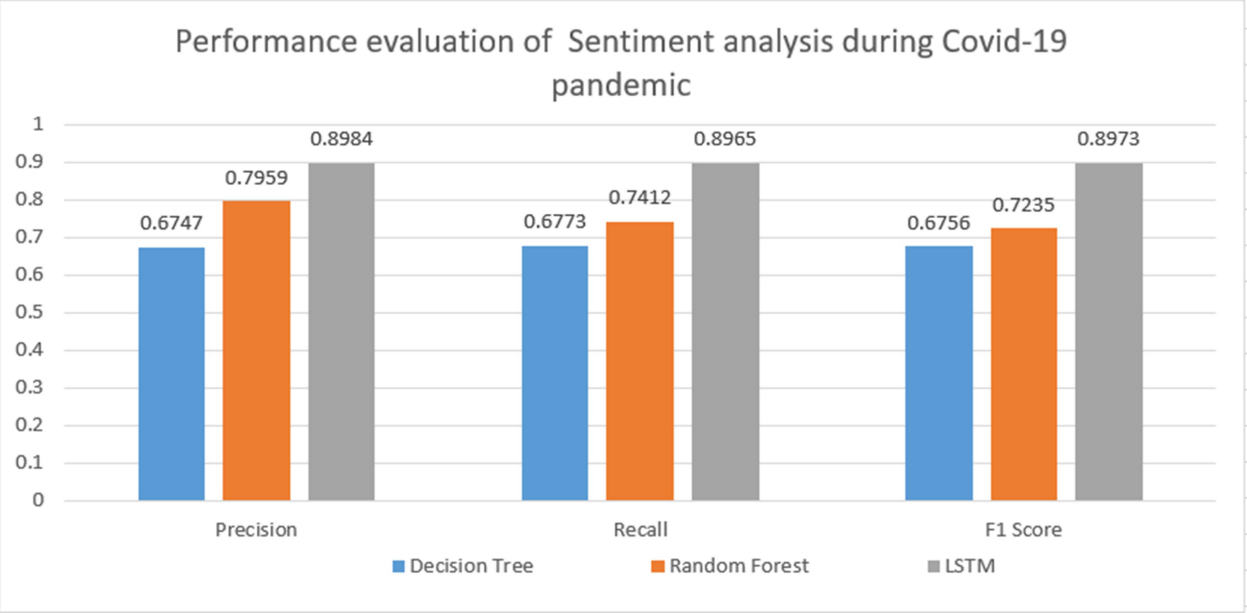

Figure 5.8: Precision, recall, F1 score values for data model

Table 5.5 and figure 5.8 shows the weighted precision, recall and f1 score values respectively for Decision Tree, Random forest classifier (RFC) and LSTM models. In a perfect world, we'd want a model that has a precision and recall value is 1. That means a F1-score of 1, which means 100% accuracy. But this is not possible. In our machine learning models we try to get a higher precision with a higher recall value. Here illustrated the precision for the Decision Tree, Random Forest Model (RFC) and LSTM model is 0.6747, 0.7959, 0.8984 respectively whereas recall for the Decision Tree, Random Forest Model (RFC) and LSTM model is 0.6773, 0.7412, 0.8965. And the F1-score for the Decision Tree, Random Forest Model (RFC) and LSTM model is 0.6756, 0.7235, 0.8973. Comparing three of the models, it can be said that LSTM model gives better precision, recall, f1 score with a higher accuracy than random forest and decision tree.

The purpose of the study is to accurately predict the outcome of people sentiments based on social media text data (Twitter). For this we label users sentiments as "Positive", "Negative" and "Neutral" classes. Figure 5.2, figure 5.3 and figure 5.4 shows that three of the models predict neutral emotions greater than positive and negative emotions. So, our model shows that peoples positive sentiments are greater than their negative sentiments. As a result of it our model couldn't classify negative sentiment accurately because of insufficient negative tweets in our dataset.

Chapter 6

Future Work and Conclusion

6.1 Future Work

In our work we labeled our dataset by using TextBlob library in positive, negative and neutral sentiments which didn't gave us proper polarity value for many texts and as a result of it we couldn't achieve a perfect sentiment rate of people in this Covid-19 pandemic. So, in our future work, we will label our dataset more accurately to get the perfect polarity value and by classifying the sentiments in depth, we can get more accurate sentiments. So, we are planning to work with individual sentiments like anger, sad, depression, loneliness and other sentiments rather than positive, negative and neutral sentiment. This analysis was done by limited data by using a larger dataset will try to increase the efficiency of our model by increasing the accuracy. By using a larger dataset, the accuracy of the model can be increased.

We will also try to create a dataset for mortality rate with geo-location so that we can find out the mortality rate of individual locations and establish a relation with a sentiment of the people of those areas for Covid-19 so that immediate measure can be taken to the most affected area.

6.2 Conclusion

The methods used in this study showed the ability to find out the sentiment from textual data that could be related to quickly evolving incidents, such as the COVID-19 pandemic. In this work, we build a Covid-19 tweet dataset for classifying Covid-19 related tweet sentiments and its effect on mental health. Our research was limited to English language text only. Consequently, the findings do not reflect remarks made in other languages. We addressed the positive, neutral, and negative sentiment of Twitter users for coronavirus and Covid-19. We demonstrated textual data analysis on three classifying algorithms: Decision tree, Random Forest, LSTM test classifier, and provided a comparison between them. For this approach, we used three classifying algorithms where the accuracy of Decision Tree is 67%, Random Forest 74% and LSTM gives 89%. So, we can see that LSTM performs better on classifying sentiments on textual data than Random Forest and Decision Tree.

References

[1] Wolohan, Jt and B. A. Hamilton. “Estimating the effect of COVID-19 on mental health: Linguistic indicators of depression during a global pandemic.” (2020). Available: https://www.aclweb.org/anthology/2020.nlpcovid19-acl.12

[2] Bilal, Muhammad & Israr, Huma & Shahid, Muhammad & Khan, Amin. (2015). Sentiment classification of Roman-Urdu opinions using Navie Baysian, Decision Tree and KNN classification techniques. Journal of King Saud University - Computer and Information Sciences. 28. https://doi.org/10.1016/j.jksuci.2015.11.003

[3] Li, I., Li, Y., Li, T., Álvarez-Napagao, S., & García, D. (2020). What are We Depressed about When We Talk. about COVID19: Mental Health Analysis on Tweets Using Natural Language Processing. ArXiv, abs/2004.10899

[4] Jelodar, Hamed & Wang, Yongli & Orji, Rita. (2020). Deep Sentiment Classification and Topic Discovery on Novel Coronavirus or COVID-19 Online Discussions: NLP Using LSTM Recurrent Neural Network Approach. 10.1101/2020.04.22.054973.

[5] Husseini Orabi, Ahmed & Buddhitha, Prasadith & Husseini Orabi, Mahmoud & Inkpen, Diana. (2018). Deep Learning for Depression Detection of Twitter Users. 88-97. 10.18653/v1/W18-0609.

[6] Samuel, Jim & Ali, G. G. Md. Nawaz & Rahman, Md. Mokhlesur & Esawi, Ek & Samuel, Yana. (2020). COVID-19 Public Sentiment Insights and Machine Learning for Tweets Classification. Information (Switzerland). 11. 1-22. 10.3390/info11060314.

[7] Iwendi C, Bashir AK, Peshkar A, Sujatha R, Chatterjee JM, Pasupuleti S, Mishra R, Pillai S, Jo O. COVID-19 Patient Health Prediction Using Boosted Random Forest Algorithm. Front Public Health. 2020 Jul 3;8:357. doi: 10.3389/fpubh.2020.00357. PMID: 32719767; PMCID: PMC7350612.

[8] K. Chakraborty, S. Bhatia, S. Bhattacharyya et al., Sentiment analysis of COVID-19 tweets by deep learning classifiers – A study to show how popularity is affecting accuracy in social media, Applied Soft Computing Journal (2020), doi: https://doi.org/10.1016/j.asoc.2020.106754

[9] **“**A Beginner's Guide to Word2Vec and Neural Word Embeddings” Available: https://wiki.pathmind.com/word2vec

[10] “Deep Learning Long Short-Term Memory (LSTM) Networks: What You Should Remember” Available: https://missinglink.ai/guides/neural-network-concepts/deep-learning-long-short-term-memory-lstm-networks-remember/

[11] “Deep Learning | Introduction to Long Short Term Memory” Available: https://www.geeksforgeeks.org/deep-learning-introduction-to-long-short-term-memory/