Heaven's Light is Our Guide

DEPARTMENT OF ELECTRONICS & TELECOMMUNICATION ENGINEERING

Rajshahi University of Engineering & Technology,
Bangladesh

A Natural Language Processing Approach on Negativity Detection on Social Media and Sentiment Analysis on COVID-19 related Public Sentiments

Author

Towrut Islam

1504023

Supervised By

Jannatul Robaiat Mou

Assistant Professor

RUET

December 2020

Acknowledgement

First and foremost, I would like to offer my gratitude to the Almighty for the blessings and ability to gain knowledge for completing this thesis work.

I would like to take express my sincere gratitude to my supervisor, **Jannatul Robaiat Mou**, Assistant Professor, Department of Electronics and Telecommunication Engineering, Rajshahi University of Engineering Technology for her immense and constant support. Without her knowledge, supervision, kind guideline, remarkable feedback, and encouragement, it was not possible to complete this research work. She has helped me to improve the lacking and kept me going with the finest improvement. She has enlightened me with valuable instructions from time to time that has helped me to reach the goal of this research. I consider myself lucky to be able to work and learn under her supervision.

I would also like to express my gratitude to my parents, siblings, family members, and friends who supported me in this journey thoroughly, who shared their views and opinions regarding the issues with mental health and COVID-19 for this research purpose, without which, this research work would be incomplete.

Towrut Islam

Heaven's Light is Our Guide

DEPARTMENT OF ELECTRONICS & TELECOMMUNICATION ENGINEERING

Rajshahi University of Engineering & Technology, Bangladesh

CERTIFICATE

This is to certify that the thesis entitled "A Natural Language Processing Approach on Negativity Detection on Social Media and Sentiment Analysis on COVID-19 related Public Sentiments" by Towrut Islam, Roll No. 1504023 has been carried out under my supervision. To the best of my knowledge, this thesis work is an original one and was not submitted anywhere for any degree or diploma.

Supervisor

....
(Jannatul Robaiat Mou)
Assistant Professor
Dept. of ETE
RUET

External

....

Head

....
(Dr. Md Munjure Mowla)
Ascociate Professor
Dept. of ETE
RUET

Abstract

Considering the increased level of negativity and cyberbullying on social media, we have taken a Natural Language Processing(NLP) approach using Deep Neural Network(DNN) to detect negativity on social media. In this study, additionally, we have applied Long Short Term Memory to classify the sentiment behind the opinion of people. We have trained a data set of six thousand comments and posts from different social sites and trained a dataset that contains 50 thousand opinions from Twitter on our DNN model. Our proposed model has provided a 0.93 accuracy and an 0.92 F-1 score for detecting negativity. For multi class classification of COVID-19 related sentiments, we have achieved an 0.67 accuracy score which is relatively better than the baseline accuracy(=0.61), and an F1- score of 0.65 with the application of Deep Neural Network. This result can be considered as a good performance for multi class classification. Undoubtedly, this model can perform on the real dataset to classify the sentiments in a good manner.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Intoduction

In modern times, social media is an inseparable part of life. From entertainment to marketing, writing to publishing expressions, social media is offering a huge platform in people's daily life.Social media can be simply defined as a virtual network through which people are connected to each other. There are over 3.6 billion people currently using social media [5]. People find their narrow escape from their monotonous life here every day. But everything has it's two sides. Misuse of social media has given birth of crimes like cyber-bullying, it is threatening privacy and security of personal life. We can see a lot of issues threatening mental health. Due to the popularity and use of social media, Sentiment analysis on social media is now one of the most enthusiastic research fields in the machine learning sector. Creating an autonomous system, that can detect sentiments or how people react to any issue is one of the most challenging work. The development of sentiment analysis has been able to detect people's sentiment based on the content or comments they provide. The year 2020 has been a year of a worldwide pandemic of novel Corona virus. This new type of SARS-COV-2 virus [6] attacks the respiratory system of the human body. Coronavirus is a contagious and fatal disease. Covid-19 was first detected in Wuhan, China in late 2019 [7] and reported the first death case in January 2020 [8]. But this became a global pandemic in 2020. Coronavirus spreads through human close contact. As immediately vaccine was not invented, World Health Organisation emphasized human distancing to prevent community spread. They suggested to lock down the area where the affected rate is high or at risk. Area lockdown leads to a nationwide lockdown slowly. China, the USA, England, Germany, Spain, India, Bangladesh, and many other countries went under lockdown for a long period [9]. As a precaution, WHO emphasized wearing a mask mandatory, to wash hands after touching anything outside, to use sanitizers [10]. This rules and precautionary steps brought a new change in regular life, people had to make new adjustments, many of them found this hard. Coronavirus has brought a huge change in both everyday life and the socio-economical environment. This has affected personal life, mental health, job life. Almost 36M people became unemployed [11]. All of these had mixed feelings and people were tweeting regularly about these factors on social media sites. It even somehow helped to spread panic and misinformation as well.

In this research work, we have proposed an autonomous model to detect the negativity on social media and classify COVID-19 related public sentiment.

1.2 Research Motivation

Human psychology or sentiment varies from one to another individuality and keeps responding differently to different circumstances throughout the whole life. From the different perspectives of life, when one person can accept changed behavior positively, some others might take that negatively. Public sentiment is an unavoidable weapon for the development of any issue. Social media is contributing to people with huge scope to express their opinion. But is ironic to find that instead of using the positive side, people these days are busy using this platform negatively. The level of negativity is spreading acutely which is directly affecting mental health. Safety security is threatened. Helpguide organization did a study on the adverse effects of social media and from the study result, we can see - loneliness, depression, anxiety, cyber bullying are affecting mental health negatively [12].To stop the spread of negativity is mandatory for mental health. But when there are numerous people using this, this is not possible to stop the use of abusive words or finding the hate words manually. But to create a system that can stop this is a crying need. People has spent more time than usual during the period of lockdown. A study in Harris Poll, conducted in

US, states that 46-51% adults used social media since the outbreak [13]. An online survey in China, states that they received 53.8% severe or moderated mental impact due to the pandemic. The social media contents were flooded with Covid-19 related sentiments, these also included fear, phobia, risk , myths, even wrong information [14], [15], [16]. while social media has helped a portion of people to readjust with all changes,it has also unfortunately contributed negatively to mislead a section of people.

The main motivation of this research has come to create an autonomous system that can automatically classify the sentiments from the social platform. So that the researchers can save their valuable time to do statistical research. This can also be helpful to study human nature for a psychology study.

1.3 Research Goal

In this research work, we have focused on a natural language processing approach to create an autonomous system that can detect negative words from the social media. So that, a default system or methodology can be improved to remove or warn while commenting with these words. Also, to create a model that can classify public sentiment about COVID-19 on this platform using deep neural network.

1.4 Research Outline

The research report consists of the following contents:

1. **Chapter 1** contains an overall introduction on social media,covid-19, and sentiment analysis.

2. **Chapter 2** contains a brief about background studies literature reviews related to this research

3. **Chapter 3** contains a detailed methodology result in detecting negativity on social media

4. **Chapter 4** contains a detailed procedure and result on Sentiment Classification of COVID-19 Related Public Opinion.

5. **Chapter 5** summarizes the work with a conclusion and some future work.

Chapter 2

Background Studies and Literature Review

Our research work has been accomplished with the help of natural language processing, machine learning algorithms, and deep neural network. To establish the model to detect negativity on social media and to analyze public sentiment on COVID-19, we have undergone a brief study on these. To detect negativity on social media and to classify COVID-19 related public sentiments, we have studied some related papers and journals to understand the working procedures in order to develop a furthermore better model.These background studies and literature review are presented as follows in a brief:

2.1 Background Studies

In this section, we represent the background studies that are inevitable for this research work. It mostly contains a brief on machine learning and its types, algorithms, deep neural network its layers, mandatory equations that helped to achieve the desired results.

2.1.1 Machine Learning

Machine learning is now one of the most trendy research topics in the field of artificial intelligence. Numerous numbers of research works are now carried in this field. Machine learning refers to the process of training a machine just like a human being. It enables the machine to identify the target from the input by following some instructions and methods.

Machine learning is highly being used in image recognition, voice recognition, social media platform, spam and malware detection, applications, and so on. Machine learning can be accomplished in two types.

i.Supervised Machine Learning Supervised learning is basically the machine learning type where we train a set of labeled data. In another way, we can say, in a supervised learning process, there is a desired output from the beginning. Each of the input data set is trained by supervised learning algorithms and this way, they can predict new untrained data.

ii.Unsupervised Machine Learning Unsupervised learning is just the opposite of supervised machine learning. Unsupervised machine learning deals with sets of unlabelled data.

2.1.2 Natural Language Processing

From the beginning of computer science, it only used to take human language as an input, then encoded it into a binary number, process the numbers, and before giving the output in human language by decoding this again. But with the development of science, the computer is now dealing with human sentiments. Natural language processing deals with making the computer or artificial intelligence understand the sentiments behind any word given as an input. Speech recognition, natural language understanding, natural language processing; are a few of the sectors scientists are trying to develop every single day. Text to speech processing, morphological analysis, syntactic analysis, lexical semantics, relational semantics, etc is a few of the tasks covered by NLP.

2.1.3 Algorithms

In order to complete this thesis work, we have gone through some basic algorithm studies. They are discussed here in a brief

1. **Logistic Regression**

Logistic Regression is one of the most used [17] machine learning algorithms. This algorithm deals with a discrete dataset. It has a massive impact on classification classes. Logistic Regression has been involved in brain tumor classification, sentiment analysis, online spam detection. There are two types of logistic regression. (1) Binary (2) Multi-Linear In terms of binary Logistic Regression, this predicts the probability and gives the output as 0/1. Till these days, the predicting limit for Logistic Regression is 0 to 1. From this, it is quite understandable that any limit having a greater value than 1 or less than 0, can not be predicted by Logistic Regression Algorithm. Logistic regression hypothesis expectation is given by

$$0 \le h_\theta(x) \le 1 \tag{2.1}$$

2. **Support Vector Machine (SVM)**

   Support vector machine is another useful algorithm in the machine learning field [18]. It detects N-number of features in N-dimensional space. It classifies distinct data points. Choosing a plane with a maximum margin is a key work to work with SVM properly. One positive side of SVM is, it takes less computation power. But gives better accuracy for the model. We can use SVM for both regression and classification problems. But this has a large impact on classification tasks.

3. **Stochastic Gradient Descent**

   Stochastic gradient descent is an iterative method,being largley used these days [19]. This is also applicable as an optimization algorithm. It offers the best suited smooth properties. While working in a high dimensional factor, the optimization problem occurs.And SGD can be a great rescue to solve this problem. This has the ability to train more than $10\hat{5}$ data and $10\hat{5}$ features. Though, this offers better optimization for a model, we cant get the desired result if the data set is not large in scale. So we need a huge dataset to apply this algorithm.

$$w \leftarrow w - \eta \left[\alpha \frac{\partial R(w)}{\partial w} + \frac{\partial L\left(w^T x_i + b, y_i\right)}{\partial w} \right] \tag{2.2}$$

4. **Naive-Bayes**

   Naïve Bayes classifies is a liberal predictor. In another word, while classifying or predicting,naïve Bayes does not relate one data with another. It does not even relate one condition to another. Naïve Bayes classifier gives better performance in comparison to logistic regression or SVM classifier. This has fast prediction in the training dataset.Naive Bayes is now engaged in many analyses [20]

   Bayes theorem is given by

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)} \tag{2.3}$$

   Bayes theorem can be rewritten as

$$P\left(y|X\right) = \frac{P\left(y|X\right)P\left(y\right)}{P(X)} \tag{2.4}$$

   where x is X=x$_1, x_2, x_3, ..., x_n$

Finally,

$$y = argmax_y P\left(_y\right) \prod_{i=1}^{n} P\left(x_i|y\right) \tag{2.5}$$

using this equation, it predicts the probability and obtain the class.

2.1.4 Deep Neural Network

A deep neural network can extract high-level functions from the information. This network consists of multiple layers like the input layer, hidden layers, and the output layer. designing the input layer, hidden layer, an output layer, choosing the activation function, and to optimize the performance, choosing the best-suited optimizer are the main important key works for designing the DNN.A simple deep neural network is drawn in figure 2.1

Input layers normally depend on padding length; the output layer is mostly the number of classes of expected classifications. Activation functions perform the functional mappings between the inputs and response variables. Relu, tanh, softmax, sigmoid is the mostly used activation function of DNN. A short description about this functions are provided as follows

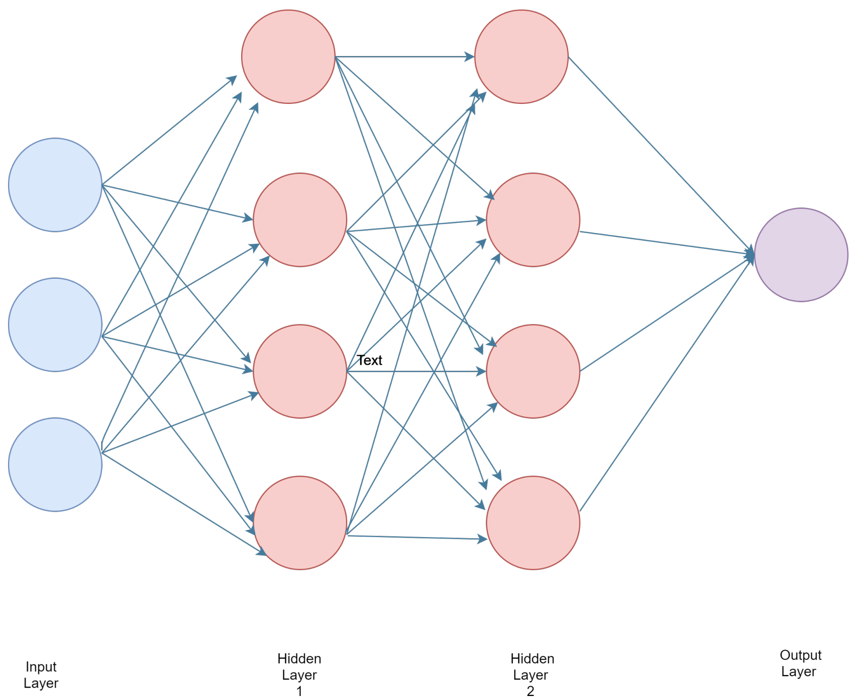


Figure 2.1: A simple Deep neural Network

1. **Rectified Linear Unit (ReLU)**

   ReLU stands for Rectified Linear Unit. This activation function is one of the most widely used activation function to output the input [21]. It indicates positive if the output is positive, or it defines the output as zero(0). ReLu is fast and easy to work on. This leaves a major impact on improving the performance of neural networks. The mathematical representation of ReLu is described further

   $$g(z) = max\{0, z\} \tag{2.6}$$

   A graphical representation od ReLU as an activation function is shown in figure 2.2

2. **Sigmoid**

   We can characterize the sigmoid function as differentiable with boundary. Sometimes a sigmoid function can also be named as a sigmoid curve but they stand for the same thing. This function normally has the limit from -infinity to + infinity. The mostly used range is also preferred as 0 to 1. While plotting the values, it originates an “S” shaped

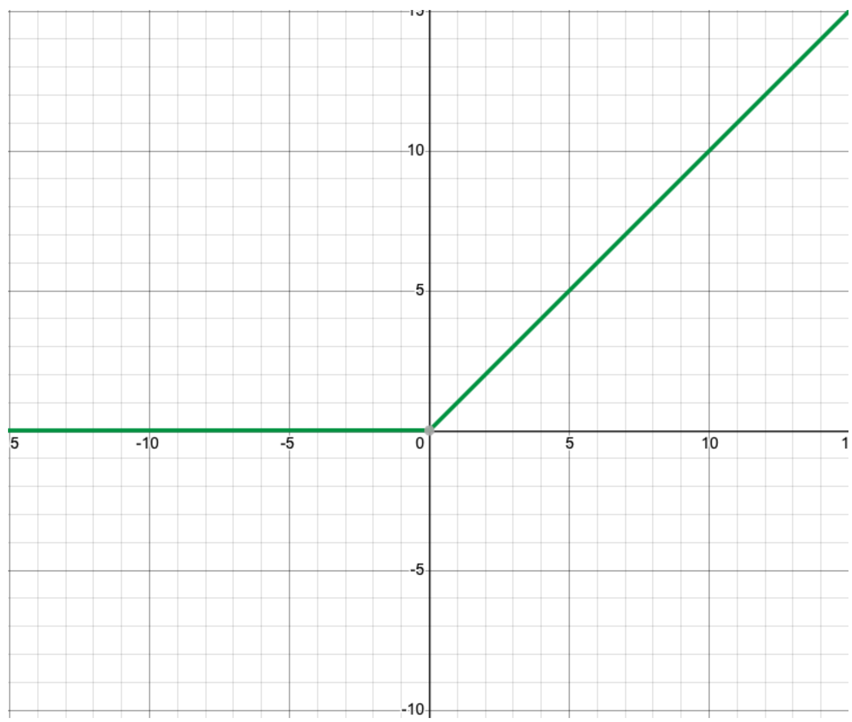


Figure 2.2: Graphical Representation of ReLU [1]

curve. The sigmoid curve is mostly used for predicting the probability of an output. For a case of yes or no, the sigmoid function is used based on its existing range.The mathematical representation of the Sigmoid activation function is. This is now one of the most used activation functions [22]

$$f(s) = \frac{1}{1+e^{-s}} \tag{2.7}$$

A graphical representation of sigmoid activation function is shown in figure 2.3

3. **Tanh**

   Tanh function is highly known as Tangent Hyperbolic Function. This has a value range from -1 to +1. It is non-linear in character. And it is widely being used in hidden layers of neural networks [22] The mathematical representation of Tanh is as follows:

$$f(x) = tanh(x) = \frac{2}{1+e^{-2x}} - 1 \tag{2.8}$$

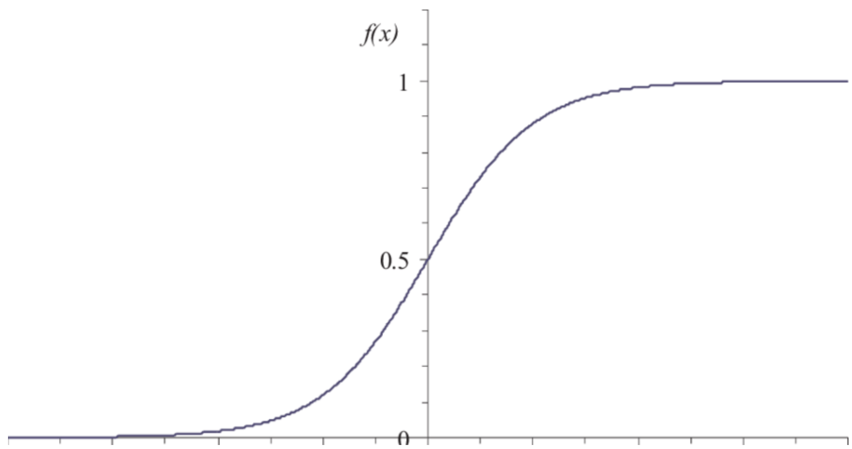


Figure 2.3: Graphical Representation of Sigmoid Functio [2]

A graphical representation of hyperbolic tangent function is represented in figure 2.4

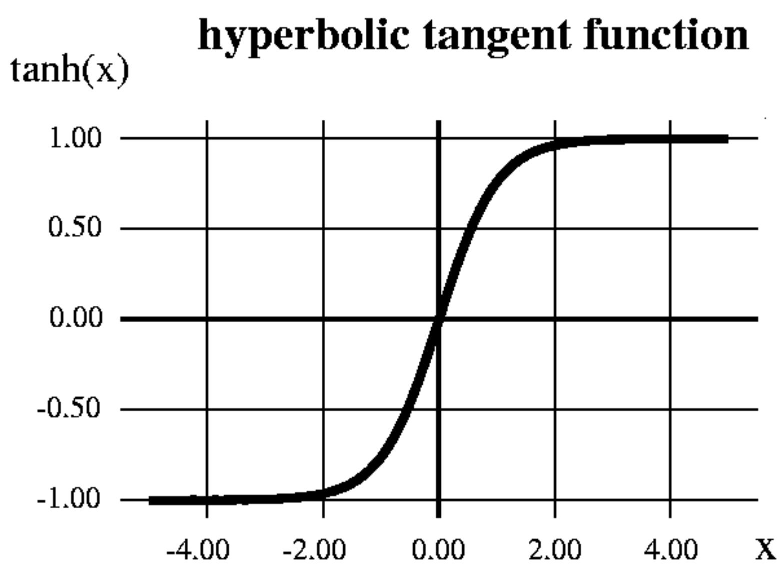


Figure 2.4: Graphical Representation of TanH Function [3]

(d) Softmax In natural language processing, Softmax activation function plays a vital role. Mostly, researchers show interest in classifying or detecting multiple sentiment labels in sentiment analysis work [23]. Softmax is best suited For classifying into multiple classes. This non-linear function is fed into the output layer of the neural network.The range is normally 0 to 1 The mathematical representation of the Softmax activation function is as follows

$$S(y_i) = \frac{e^{y_i}}{\sum_{j=1}^{i} e^{y_j}} \tag{2.9}$$

A graphical representation of Softmax activation function is shown in figure 2.5

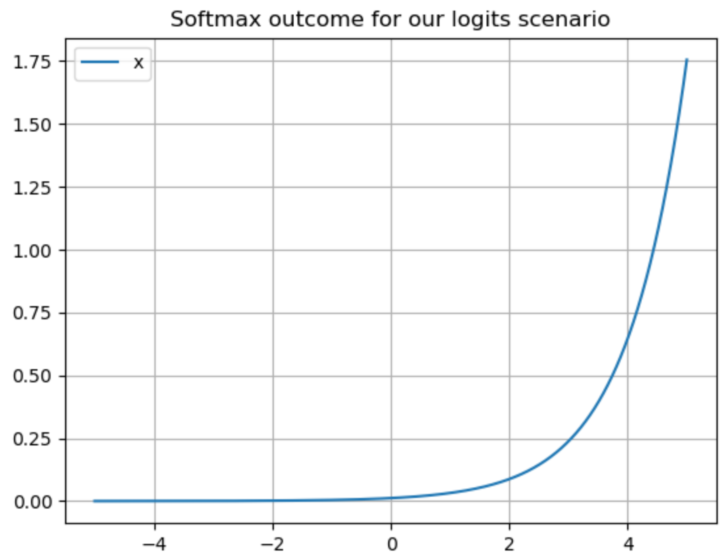


Figure 2.5: Graphical Representation of softmax [4]

2.1.5 Evaluation Matrices

Evaluation metrics are mainly used to evaluate the performance of any model or algorithm. Evaluation metrics is a policy that gives output in matrix form associated with prediction data with respect to actual data. A model might perform well when a labeled set of training data is fed into the network. But this does not define or indicate how well that model can perform with a set of valid or test data. Different evaluation matrices such as Precision, recall, F-1 score, classification accuracy; these scores can declare either the model can perform well in the real world or the model is not valid. In statistical analysis, the machine learning approach; evaluation matrices are making a noticeable contribution to analyze the performance of any proposed model.

1. **Accuracy**

The accuracy provides the ratio of the number of correct predictions to the total number of given data. The formula of accuracy is mentioned as follows:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{2.10}$$

2. **Precision**

Precision is the ratio of a true positive result to the predicted positive result. The equation of precision is

$$\text{Precision} = \frac{TP}{TP + FP} \tag{2.11}$$

3. **Recall**

Recall is the ratio of a correct positive result to all the relevant data. The mathematical formula of recall is given below

$$\text{Recall} = \frac{TP}{TP + FN} \tag{2.12}$$

4. **F1 Score**

F1 Score measures the accuracy of the test so that we can determine the performance of a model based on this score. The formula for

$$\text{F1-Score} = 2 * \frac{Recall * Precision}{Recall + Precision} \tag{2.13}$$

In all equation of Accuracy,Precision,Recall & F-1 Score,

TP= True Positive

TN= True Negative

FP= False Positive

FN= False Negative

2.1.6 Long Short Term Memory

Long Short Term Memory is a special kind of Recurrent Neural Network. It is capable of learning long-term dependencies. This can solve different NLP related tasks. LSTM includes different interacting layers of cell state, input gate layer, cell status update, output data. Lstm is very efficient for answering questions, sentiment analysis, image to text mapping, speech recognition, part of speech tagging, etc

2.2 Literature Review

Sentiment analysis is now one of the most demanding and important topics in the research platform of NLP. Considering the use of social media and its effect on our life, researchers are engaging more in developing this field and We have seen promising research progress by research enthusiasts so far.

Researchers worked on opinion mining using Machine Learning in [24]. They applied it to the dataset of restaurant reviews. After featuring extract with different N-gram techniques, they applied vectorization with a Hashing vectorizer, count vectorizer, and TF-IDF vectorizer. In order to classify the reviews, the scholars used SVM, DT LR and succeeded to achieve an accuracy of 75.58% with the SVM algorithm which was comparatively higher than the other two. Researchers worked on opinion mining using Machine Learning in [25]. They applied it to the dataset of restaurant reviews. After featuring extract with different N-gram techniques, they applied vectorization with a Hashing vectorizer, count vectorizer, and TF-IDF vectorizer. In order to classify the reviews, the scholars used SVM, DT LR and succeeded to achieve an accuracy of 75.58% with the SVM algorithm which was comparatively higher than the other two. In another publication [26], researchers worked on sentiment analysis on Bengali texts on online restaurant review using Multinominal Naive Bayes. They proposed a system that can classify the reviews in the positive and negative categories. And this model gained 80.48% accuracy using Multinominal naive Bayes. As the works kept on going, Moqsadur Rahman, aimed to identify categorize opinions expressed in the sentence [27]. They applied different deep learning algorithms i.e. CNN, multilayer perceptron, long short-term memory On a dataset of sports comments in different newspapers.

The F1 score for CNN was 0.4819 which was better than LSTM and DNN. A contribution in [28], To develop availability, researchers proposed a classification framework that could analyze sentiment from written texts. The model was generated by a neural network variance named CNN. this proposed model could perform with 99.87% accuracy.When Scholars continued their work on identification and classification, two research enthusiasts from India took a step to classify abusive comments in social media in [29]. They used Kaggle's toxic comment dataset to train a deep learning model to classify them into toxic, severe toxic, obscene, threat .insult, identity hate; these five categories. They used both LSTM CNN with or without GLoVe embeddings and GloVe pre-trained models. Where glove and CNN performed best for training and testing accuracy while glove Cnn LSTM did the worst.An investigation in [30], Researchers from ULAB worked on a dataset for sentiment analysis using deep recurrent models specifically LSTM. they used two different loss functions-binary cross-entropy and categorical cross-entropy. They expressed that the highest accuracy was achieved by the dataset with categorical cross-entropy loss and the rate was 70%. In an investigation on [31], authors did A different research work on the lexicon-based backtracking approach to analyze sentiment on song review. By collecting comments from a specific youtube channel, they implemented a backtracking algorithm. They made it possible to achieve up to 71.23% accuracy but this experiment failed to extract sentiment from the comment which could express both positive or negative sentiments.

This pandemic of COVID-19 has opened up a new section for the enthusiasts. Researchers also performing sentiment analysis on this important field. Hamed proposed a framework in [32] which combined with NLP and a deep learning method for the model. He applied LSTM for generating valuable comments on COVID. this proposed model was able to extract 81.15% info accurately. Gopalkrishna did a public sentiment analysis on [33] which is about nationwide lockdown in India. The author performed unsupervised learning and used the R Software to generate the sentiments; also created a word cloud. From almost, 24 thousand mixed data of all emotions, positivity stood out in the analysis. In this[34] paper, we saw an attempt to identify informative English tweets. They experimented by applying SVM, neural networks transformer models. They achieved the best f1 score of 90.96 on the hidden test data.

Chapter 3

Negativity Detection on Social Media

In chapter 3, we have discussed thoroughly the process applied to detect negativity on social media. At first, we have prescribed data collection, data pre-processing, and feature extraction process; then we have discussed the algorithms we have applied here; the architecture of Deep Neural Network; then we have shown how we have trained the model, evaluation matrices; we have showcased the result and analyzed the performance of our model in the end. In order to detect negativity from social media, we have represented our workflow for the research in figure 3.1

3.1 Dataset Collection & Description

To perform this experiment, we have collected over six thousand data from different platform of social media like Facebook, Instagram, Twitter, News Portals, etc. This dataset contains both English and Bangla language. Most of these comments or posts were made regarding real-life thoughts or random comments. Every data was labeled as negative or non-negative. Few examples from the dataset are illustrated in table 3.1

80% of the data has been used for training the model, 10% for validation and rest 10% is for test.

Data counts are given in table 3.2

Figure 3.1: Flow Diagram for Negativity Detection

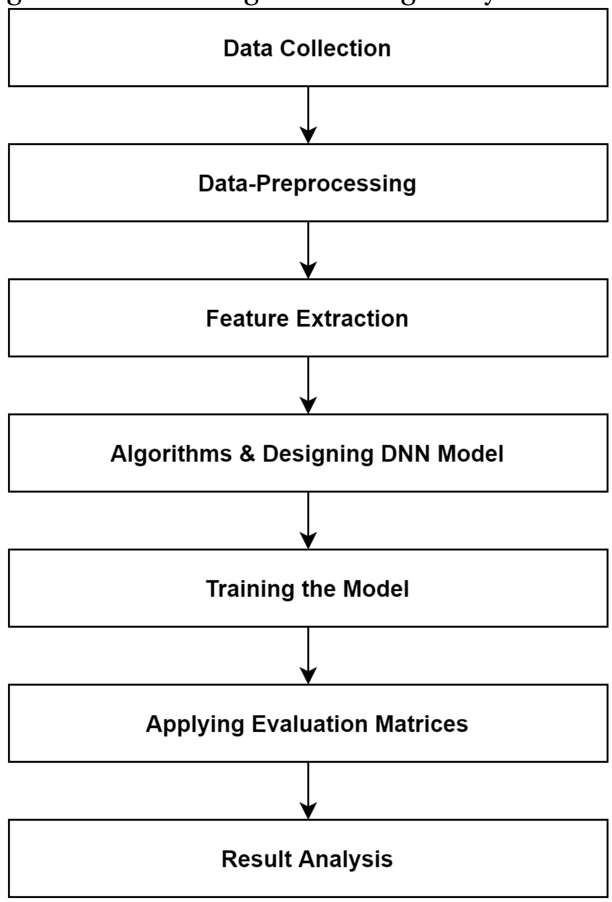


Table 3.1: Dataset Sample

Comments	**Labels**
I hate black skins!	Negative
Just what I needed atm!	Non-negative
life is getting out of control everyday!	Negative

Table 3.2: Data Counts

Comments	**Counts**
Negative	2763
Non-negative	3237

3.2 Data Pre-Processing

Data pre-processing involves the methods of removing any unwanted form of word or comments that do not contain any emotion. We have completed the data pre-processing into three parts as follows:

1. **Eliminating Stop Words**

   Stop words are mainly words that are not related to any sentiments. That's why we have eliminated all the stop words.

2. **Eliminating Punctuations**

   Every comment contained many punctuations like "," ";" "." "etc. these punctuations was not important for sentiment analysis, so the dataset has been made clear of punctuations.

3. **Eliminating Unnecessary Signs and Numbers**

   The collected data set had many unnecessary signs like "@", "", "$" and numbers. These were not helpful for the experiment. So we have removed all of these from the dataset. Finally we have a dataset which only contained data with sentiments.

3.3 Feature Extraction

The collected dataset was in human language and we directly can not put this into the neural network. Feature extraction helps with extracting the numeric features from the dataset. Tokenization, text to sequencing and padding, TF-IDF Vectorization; we have completed the feature extraction into this four-part.

1. **Tokenization**

   We have tokenized the whole dataset a created a new vocabulary list. We have found unique words from this.

2. **Text to Sequencing**

After tokenization, we have attempted text to sequencing. This assigns a numeric value to each of the contant of vocabulary set

3. **Padding**

   In order to maintain the length of each comment, we have done post padding that is adding <0 > after each word till it reached the length of 256.

4. **TF-IDF Vectorization**

   TF-IDF stands for Term Frequency Inverse Document Frequency. We used TF-IDF to measure how many times any word from the vocabulary list is repeated in the whole vocabulary set. The vectorizer counted the weights of the repeated words in the whole dataset.

3.4 Applied Algorithms

We have applied four algorithms as classifiers to detect negativity in this experiment.They are Logistic regression, Support Vector Machine (SVM), Stochastic Gradient Descent (SGD) Naive Bayes.After the application of four of these, we have fed this into Deep neural Network to improve the overall performance of the model even more.

3.4.1 Classifiers

The four classifiers which have been applied to carry on the detection process are described as follows

1. **Logistic Regression**

   We have applied logistic regression to predict probability. We had two data labels of negative and non-negative.The binominal logistic regression algorithm can predict the probability from two types of data. It can predict like yes/ no or even with a value of 0/1

2. **Support Vector Machine Classifier**

Support vector machine classifier is able to perform classifying and regression studies. In a supervised machine, learning input is fed and the desired output is labeled in the training set. SVM classifier is able to classify the train data efficiently and predict the valid data into its belonging class.

3. **Stochastic Gradient Descent (SGD)**

   We have chosen the stochastic gradient descent algorithm because it has the ability to perform efficiently. SGD also helps with optimization. It optimizes the performance of the proposed model. This is an efficient way to approach under SVM and logistic regression.

4. **Naïve-Bayes Classifier**

   Naïve Bayes classifies is a liberal predictor. In another word, while classifying or predicting,Naïve Bayes does not relate one data with another. It does not even relate one condition to another. Naïve Bayes classifier gives better performance in comparison to logistic regression or SVM classifier. This has fast prediction in the training dataset.

3.4.2 Deep Neural Network

We have applied a deep neural network to derive high-level functions. Designing the input layer, hidden layer, and output layer, choosing the best-suited activation functions and optimizer are the main steps for designing a DNN network.

1. **Input, Output, and Hidden Layers:**

   To proceed with DNN architecture, we have implemented 256 input layers. The reason for implementing the 256 input layer is we have 256 padding length. It's very important to maintain equality with padding length. We have taken one output layer for non-negative and negative output. And 3 hidden layers in the first hidden layer, there are 256 neurons and the rest two has 512 neurons in each shown in figure 4.2

2. **Activation Functions**

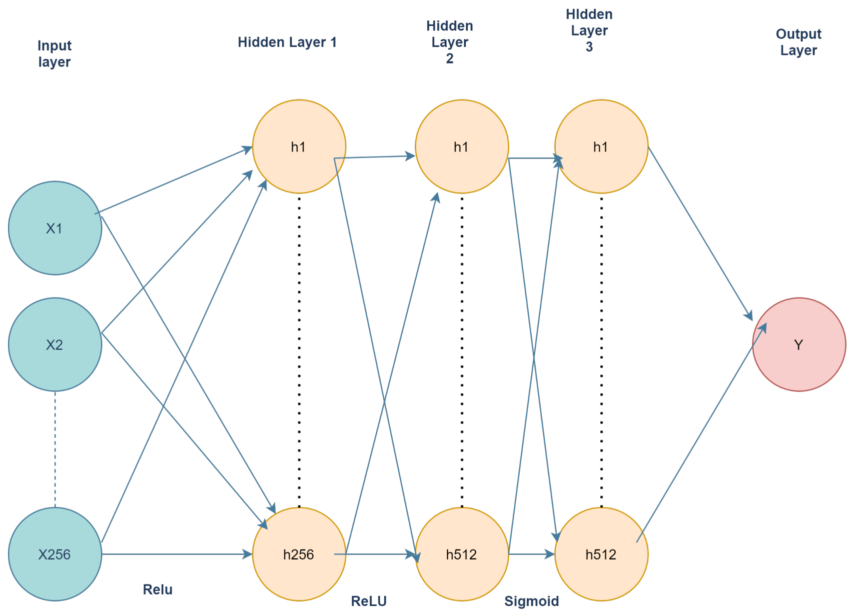


Figure 3.2: Training & validation accuracy across epochs

Activation functions perform the functional mappings between the inputs and response variables. It fundamentally changes over an input signal of a node into an output signal. That output signal is at that point utilized as an input within the next layer within the stack. Most of the actuation capacities are to make the systems nonlinear. In this study, we have utilized the following two activation functions.

ReLU:

Researchers are implementing ReLU largely in their respective research work as it is one of the largely used activation function in machine learning. We have extracted features from text and their range is between zero to infinity. We used ReLu as an activation function as we have a positive value as a feature. The mathematical expression of ReLu is:

$$g(z) = max\{0, z\} \tag{3.1}$$

Sigmoid:

We have chosen the sigmoid function here because we are working with two labels. The output probability can be counted as 0/1 or yes/no. Softmax works with multi-class outputs, so we applied the sigmoid activation function here. The graphical representation of this function can be described as 's' shape. The mathematical function is

$$f(s) = \frac{1}{1+e^{-s}} \tag{3.2}$$

3. **Optimization Algorithm**

   In neural networks, to shapes the proposed model into its most exact conceivable frame by altering the weights; we have applied 'Adam optimizer'. The memory requirement for Adam is relatively low and works with a little number of tuning parameters. We applied a 0.001 learning rate on our model. The Adam optimizer equation is noted as follows:

$$v_t = \beta_1 \text{Ö} v_{t-1} - (1-\beta_1)\text{Ö} g_t, \tag{3.3}$$

$$s_t = \beta_2 \text{Ö} s_{t-1} - (1-\beta_2)\text{Ö} g_t, \tag{3.4}$$

$$\Delta w_t = -\eta \frac{v_t}{\sqrt{s_t+\epsilon}} \text{Ö} g_t, \tag{3.5}$$

$$w_{t+1} = w_t + \Delta w_t. \tag{3.6}$$

   where
   η = Initial learning rate
   g_t = Gradient at time t along w_j
   v_t = Exponential average of gradients along w_j
   s_t = Exponential average of squares of gradients along w_j
   $\beta_1; \beta_2$= Hyperparameters

3.5 Evaluation Metrics

For validating our model performance, we have used different evaluation matrices specifically Accuracy, Precision, Recall, and F1 score. The accuracy provided us the ratio of the number of correct predictions to the total number of given data; precision provided the ratio of a true positive result to the predicted positive result; recall provided the ratio of a correct positive result to all the relevant data. F1 score helped to measure the accuracy of the test so that we could determine the performance of our model based on this score. We have generated the confusion matrix from the test dataset and evaluated the classification report by using the following equations:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{3.7}$$

$$Precision = \frac{TP}{TP + FP} \tag{3.8}$$

$$Recall = \frac{TP}{TP + FN} \tag{3.9}$$

$$F1 - Score = 2 * \frac{Recall * Precision}{Recall + Precision} \tag{3.10}$$

Here,

TP= True Positive

TN= True Negative

FP= False Positive

FN= False Negative

3.6 Result and Analysis

To evaluate the result of our experiment and analyze the performance of our model in the real world, we have separated this part into three segments. We have generated a training and validation curve, we have observed the performance of the confusion matrix and finally, we have analyzed the classification report.

3.6.1 Training-validation graph

we have plotted Training and validation accuracy across epochs shown in figure 3.3

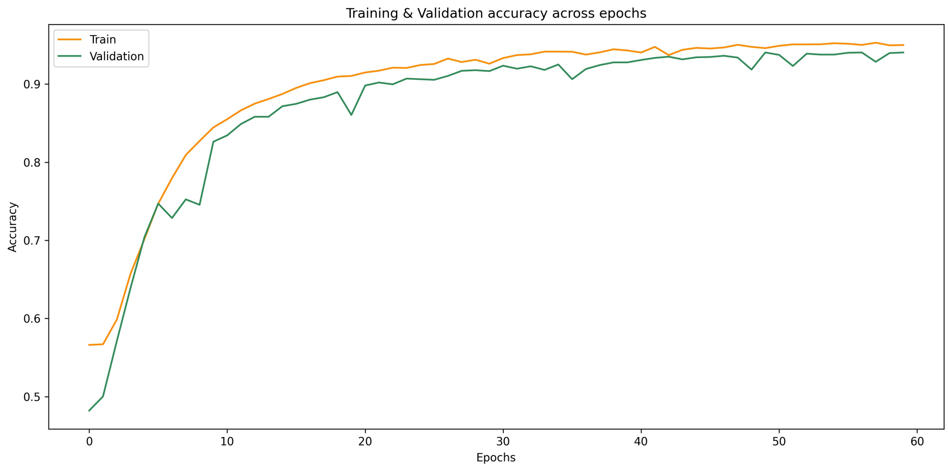


Figure 3.3: Training & validation accuracy across epochs

We have plotted up to 60 epochs and plotted up to 0.9 accuracies. From the graph, we can see that the training data has reached up to almost 0.95 and the validation data has reached up to 0.93. This is the indication that there is no overfitting issue in our model. This mode is realistic and can perform well in real data.

3.7 Confusion Matrix

We have generated the confusion matrix to evaluate the level of prediction of our proposed model. A sample of the prediction model is represented in figure 4.5

3.8 Classification Report

In our experiment, we have applied binominal logistic regression, SVM, SGD, naïve Bayes, and deep neural network. We have derived the result of precision, recall, f-1 score and accuracy for all of these algorithms. The output result of deep neural network for detecting negativity on social media stood out amongst all of the classifiers. It has been able to perform

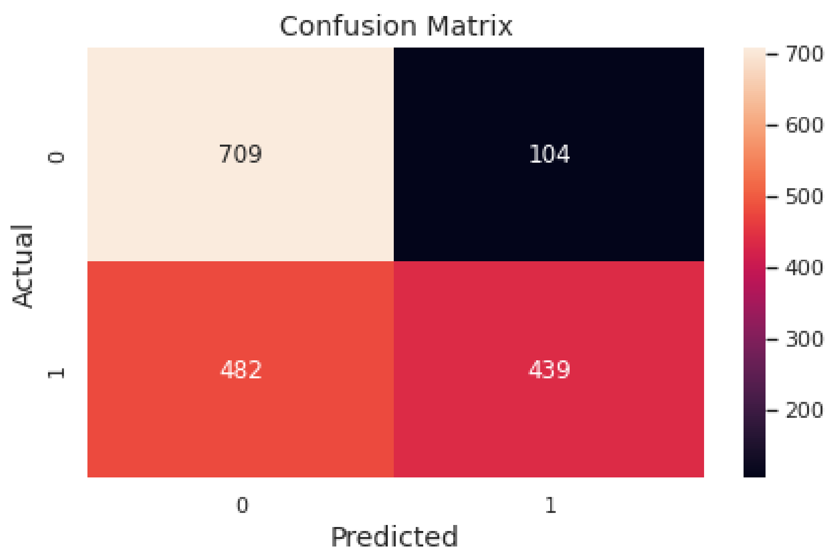


Figure 3.4: Confusion Matrix

the detection with highest precision, recall, and the f-1 score of 0.91, 0.87 & 0.92 respectively. The achieved results are represented in table 3.3

Table 3.3: Classification Result of Classifiers

Classifier	**Precision**	**Recall**	**F-1 score**
Logistic Regression	0.84	0.83	0.82
Support Vector Machine	0.71	0.73	0.76
Stochastic Gradient Descent	0.73	0.77	0.75
Naive Bayes	0.73	0.75	0.77
Deep neural Network	0.91	0.87	0.92

From the rest of the classifiers,logistic regression has stood out in achieving 0.84, 0.83 & 0.82 scores for precision, recall, and f1 respectively. Among the other three classifiers, SVM SGD achieved a score of 0.73 for precision; SGD achieved the highest score of 0.77 for recall, and SVM and Naive Bayes both showed a score of 0.77 for F-1 score.

Table 3.4 states the accuracy scores of all the classifiers. From the table, we can state that the performance of our model has hit the highest accuracy score of 0.93 when it is fed into deep neural network.

For the accuracy score, we can say that the logistic regression classifier is able to classify the sentiments up to a level of 0.82 whereas after applying a deep neural network, it is

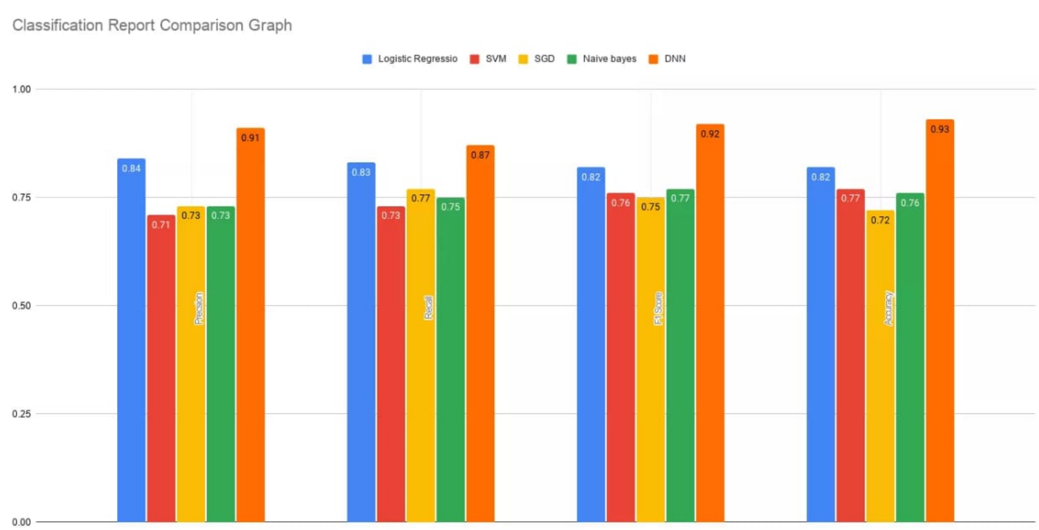
Table 3.4: Accuracy Score

Classifier	Accuracy
Logistic Regression	0.82
SVM	0.77
SGD	0.72
Naive Bayes	0.76
DNN	0.93

possible to perform accurately with an accuracy score of 0.93.

In figure 3.5 we have represented a comparison graph. This graph contains the value of precision, recall, F-1 score, and accuracy of all classifiers we have used in this research work. We have used Microsoft Excel to generate this graph.

Figure 3.5: Classification Report Comparison Graph



From the above analysis, we can say that our proposed model is able to perform promisingly in terms of detecting negativity from comments on social media in real life.

Chapter 4

Sentiment Analysis of Public Sentiment on COVID-19

In this chapter, we thoroughly discuss the research methodology for classifying public sentiment on coronavirus. The research work contains the following steps: data collection, data pre-processing, feature extraction, designing DNN Model, training data, applying evaluation matrices, and result analysis. The proposed workflow is represented in figure 4.1

4.1 Data Collection & Description

We have used a dataset from Kaggle containing 50000 tweets from twitter in English. All the data were related to coronavirus. All these tweets are labeled into five categories: 'positive', 'negative', 'neutral', 'extremely positive' ,and 'extremely negative'. The counts of the sentiments are given in Table 4.1

Table 4.1: Class distribution for Sentiment Classification

Class Labels	**Counts**
Positive	11422
Negative	10000
Extremely Positive	10000
Neutral	11578
Extremely Negative	7000

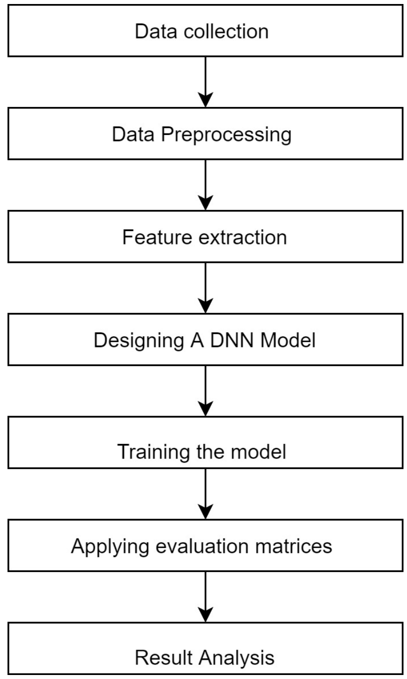


Figure 4.1: Flow Diagram of Research Procedure

From Table 1, this can be observed that the data-set is properly balanced for the Sentiment Classification. But in case of 'Extremely Negative' Class, the data is a bit less than the others.

4.2 Data Pre-processing

Usually, raw data contain a big number of unwanted elements that directly affect the performance of the machine learning or deep learning models. That is why, elements such as stop-words, punctuations, and unwanted characters have been removed from the dataset.

1. **Removing Stop-words**

   Stop-words are the words that do not contain any significant information, even sometimes decrease the performance of the model. So, all of the stop-words have been suppressed from the dataset.

2. **Removing Punctuations**

Punctuations do not carry any informative meaning for the classification tasks. That is why, to reduce the complexity,the punctuations have been removed from the data as well.

3. **Removing Unnecessary Characters**

   As the data are scraped from online, the dataset contains many unnecessary characters like ‘@’, ‘=’, ‘&’, and so on. These characters have zero impact and likely to mislead the learning of our models. To improve the model performance, we have normalized our dataset by removing these characters.

4.3 Feature Extraction

Once the data is processed, we can move toward our next part that is feature extraction As the text data cannot be fed directly to neural network, this is important to extract the numeric feature from the dataset. The process of extracting features are as follows:

1. **Tokenization**

   Tokenization is the process of creating a vocabulary from the data set. We have tokenized each sentence of the corpus to make a vocabulary of the whole data set. We have found 96691 unique words in the dataset. But only 10,000 important words have been taken to carry out the experiment.

2. **Text to Sequence**

   After tokenization, a sequence of the word has been made for each sentence. That converts the sentences of the dataset to numerical form which is understandable for a DNN model.

3. **Padding**

   After making sequences of text, the numeric sentences were not of in the same length. We have padded the sentences to normalized the length. We added ‘0’ to the last of each numeric until the length reaches 40.

4.4 The Architecture of DNN Model

As the deep neural network can derive high-level function from input information; designing the input layer, hidden layer, output layer, choosing the activation function, and to optimize the performance, choosing the best-suited optimizer are the main important key works for designing the DNN.

Input, Output, and Hidden Layers

In the experiment, 256 input layers have been implemented to maintain equality to the embedding dimension. 2 hidden layers have been implemented with 512 neurons each. As the aim is to analyze and classify the sentiments into five classes; thus we have taken 5 output layers. The architecture of our deep neural network model is shown in figure 4.2

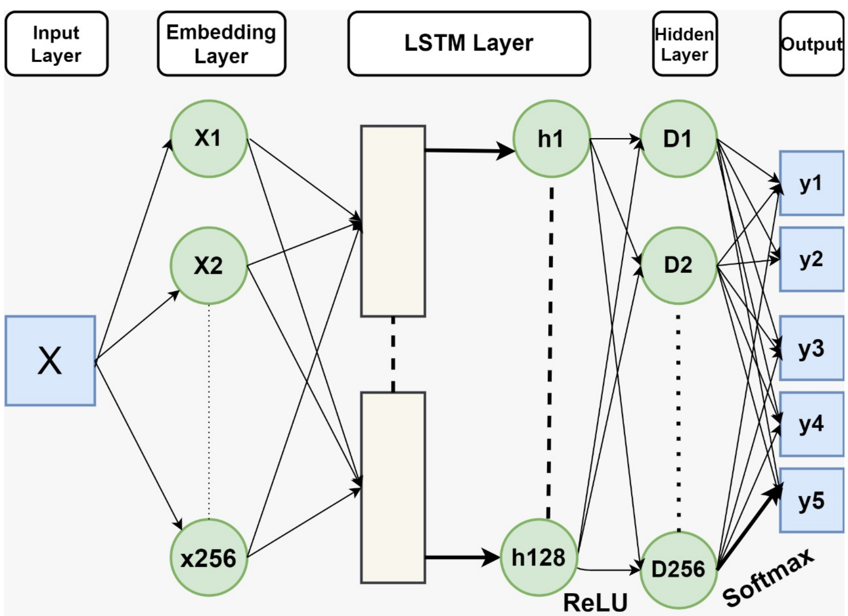


Figure 4.2: Deep Neural Network

LSTM Layer

Long Short Term Memory is a special kind of Recurrent Neural Network. It is capable of learning long-term dependencies. We have applied an LSTM layer on our model with 128 neurons. Our applied LSTM layer is represented in figure 4.3

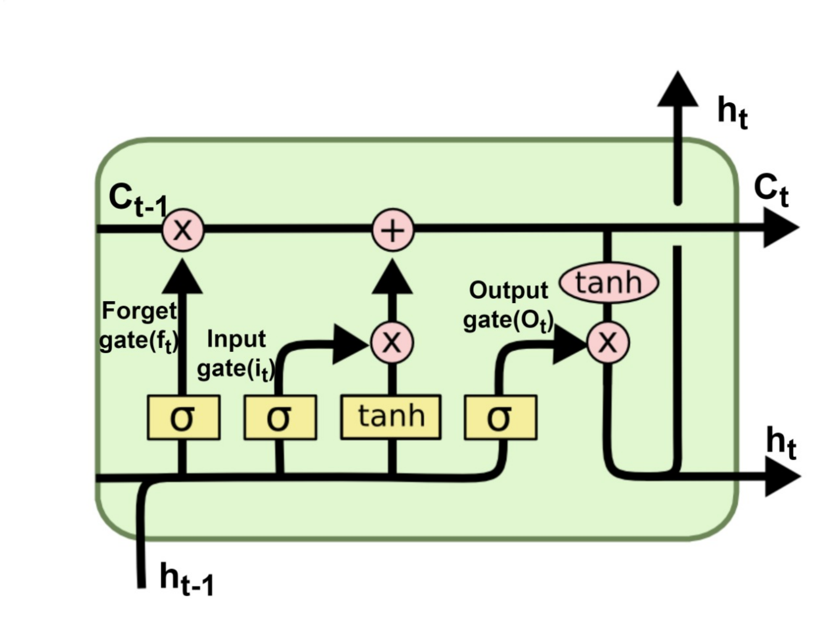


Figure 4.3: LSTM Layer

Activation Functions

Activation functions perform the functional mappings between the inputs and response variables. It fundamentally changes over an input signal of a node into an output signal. That output signal is at that point utilized as an input within the next layer within the stack. Most of the actuation capacities are to make the systems non-linear. In this study, we have utilized the following two activation functions.

- **ReLU:** Researchers are implementing ReLU largely in their respective research work as it is one of the largely used activation function in machine learning. We extracted features from text and their range is between zero to infinity. We used ReLu as an activation function as we have a positive value as a feature. The mathematical expression of ReLu is:

$$f(x) = tanh(x) = \frac{2}{1+e^{-2x}} - 1 \tag{4.1}$$

- **Softmax:** The softmax function is also known as softargmax or normalized exponential function. It is a general form of the logistic function to multiple dimensions. As our goal was to classify the sentiments into five classes, softmax is the suitable activation function for our neural network model. The standard softmax function is defined as follows:

$$S(y_i) = \frac{e^{y_i}}{\sum_{j=1}^{i} e^{y_j}} \tag{4.2}$$

Optimization Algorithm

In neural networks, optimization calculation plays the foremost imperative part in achieving a promising result from the model. It shapes the model into its most exact conceivable frame by altering the weights. In our proposed model, we have applied 'Adam optimizer'. The memory requirement for Adam is relatively low and works with a little number of tuning parameters. We have applied 0.001 learning rate on our model. The Adam optimizer equation is noted as follows:

$$v_t = \beta_1 \ddot{\mathrm{O}} v_{t-1} - (1-\beta_1) \ddot{\mathrm{O}} g_t, \tag{4.3}$$

$$s_t = \beta_2 \ddot{\mathrm{O}} s_{t-1} - (1-\beta_2) \ddot{\mathrm{O}} g_t, \tag{4.4}$$

$$\Delta w_t = -\eta \frac{v_t}{\sqrt{s_t + \epsilon}} \ddot{\mathrm{O}} g_t, \tag{4.5}$$

$$w_{t+1} = w_t + \Delta w_t. \tag{4.6}$$

where

η = Initial learning rate

g_t = Gradient at time t along w_j

v_t = Exponential average of gradients along w_j

s_t = Exponential average of squares of gradients along w_j

$\beta_1; \beta_2$= Hyperparameters

4.5 Evaluation Metrics

For validating our model performance, we have used different evaluation matrices specifically Accuracy, Precision, Recall, and F1 score. The accuracy provided us the ratio of the number of correct predictions to the total number of given data; precision provided the ratio of a true positive result to the predicted positive result; recall provided the ratio of a correct positive result to all the relevant data. F1 score helped to measure the accuracy of the test so that we could determine the performance of our model based on this score. We have generated the confusion matrix from the test dataset and evaluated the classification report by using the following equations:

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \tag{4.7}$$

$$Precision = \frac{TP}{TP+FP} \tag{4.8}$$

$$Recall = \frac{TP}{TP+FN} \tag{4.9}$$

$$F1-Score = 2 * \frac{Recall * Precision}{Recall + Precision} \tag{4.10}$$

In all equation of Accuracy,Precision,Recall & F-1 Score,

TP= True Positive

TN= True Negative

FP= False Positive

FN= False Negative

4.6 Result Analysis

We divided the process of analyzing our results into two parts. First, we created a graph plotting epochs and an accuracy level to determine whether our proposed model was overfitting. And then, From the score of precession, recall, and f-1 score, we analyzed the performance of prediction of our model.

4.7 Training-Validation Curve

For generating the training curve and validation curve, we trained our model up to 60 epochs. The model reached up to the accuracy of 0.7 for our training data. Validation curved reached up to 0.67. The graphical analsis is given in figure

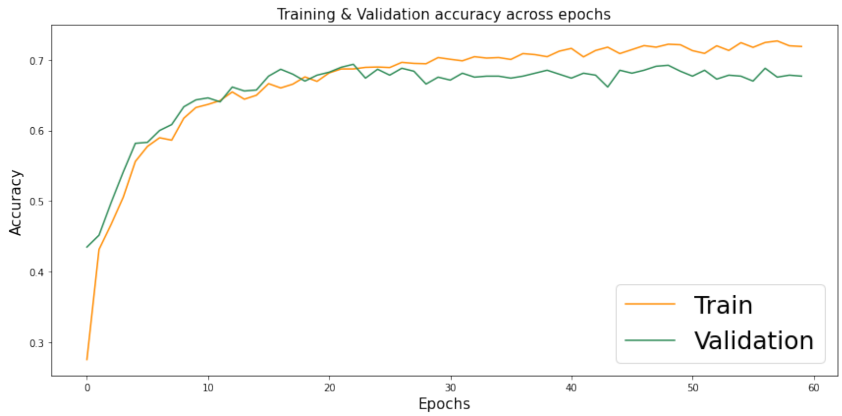


Figure 4.4: Training & Validation Accuracy across Epochs

From this observation from figure ,we reached the conclusion that there were no overfitting issues in our model.

4.8 Confusion Matrix

We created an evaluation matrices mentioned in our methodology. From the confusion matrix, we got a visualization of how accurately the model was predicting compared to the actual data. We presented a section of our confusion matrix in the figure 4.5

In the presented section, our model predicted extremely negative data 1028 times, whereas the actual number was 1229; predicted extremely positive 697 times out of actual 590 data. Samely, predicted negative data 893 times out of 1060 actual data, the amount of predicted neutral data was 596 out of 519 actual data predicted 494 positive data out of actual 307 positive data.

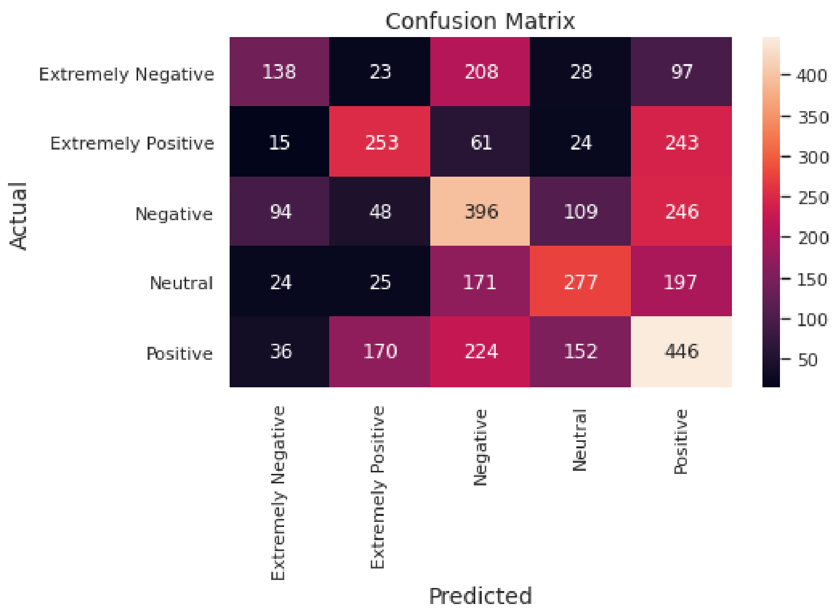


Figure 4.5: Confusion Matrix

4.9 Classification Report

We generated a classification report from the confusion matrix which is shown in Table .The classification report shows that the f1-score for 'Extremely Positive' class is maximum which is 0.65 that means our model can predicted the 'extremely positive' class so well. but f1-score for 'Extremely negative' class is poor that is why our model have predicted this class less efficiently. And We can see the accuracy of our model on test data-set is 0.67 which is better than baseline performance. Tha cleassification report is presented in table 4.2

Table 4.2: Classification Report

Classes	**Precision**	**Recall**	**F1 Score**
Extremely Negative	0.65	0.48	0.55
Extremely Positive	0.69	0.62	**0.65**
Negative	0.57	0.64	0.60
Neutral	0.67	0.60	0.63
Positive	0.56	0.63	0.59
Accuracy	0.67		

From these analyses, we settled with the outcome that our proposed model was able to perform with better accuracy in terms of classifying the sentiments in the real world.

Chapter 5

Conclusions

This final chapter concludes this thesis, summarizes the key results and suggests a direction for future works.

5.1 Conclusions

In this research work, we have represented a natural language processing approach on sentiment analysis to detect negativity from social media and to classify public sentiments regarding COVID-19. To detect negativity, we applied logistic regression, SVM, SGD, Naïve Bayes, and designed DNN to classify non-negative and negative data. Over six thousand of data have been collected manually from Facebook, Instagram, Twitter, and different news portals were classified into non-negative and negative labeling. From the result we saw, DNN has showed outstanding performance out while performing with precision, recall f-1 score, and accuracy. It has achieved the highest f-1 score of 0.92 and the accuracy was 0.93 which was far more promising to perform in the real world. Also from the curve, we found out there was no overfitting issue in our model. To classify covid-19 related sentiments, We have used LSTM to design the Deep Neural Network (DNN), model. Overall50000 tweets, collected from Kaggle, were classified into five categories: positive, negative, neutral, extremely positive, extremely negative. From the experiment result, We have got a 0.67 accuracy score which is better than the baseline performance. we have also observed that the train-validation vs epoch curve is converging so there were no overfitting issues. that our model is able to classify ‘extremely positive’ sentiment best with an f1 score of 0.65 whereas ,it provides a little poor performance while performing with extremely negative sentiments. Overall, our model can work up to an accuracy of 0.67 with test data. In both cases, we have performed some data pre-processing, to clean unwanted data from the dataset. All of these were fed into DNN after feature extraction.

All the necessary dataset, excel graph generation, and codes can be found in https://github.com/Towrutislam/Thesis1504023

5.2 Directions of Future Research

In future research, to classify sentiments, advanced DNN models like CNN or Transformation models like BERT can be applied which can provide better performance and accuracy score.

References

[1] S. Singh, "Relu as an activation function in neural networks." [Online]. Available: https://deeplearninguniversity.com/relu-as-an-activation-function-in-neural-networks/

[2] B. J. van der Zwaag, "Using domain-specific basic functions for the analysis of supervised artificial neural networks," Ph.D. dissertation, 12 2003.

[3] "Multi-layer networks." [Online]. Available: https://cnl.salk.edu/~schraudo/teach/NNcourse/multilayer.html

[4] Chris, "How does the softmax activation function work?" January 2020. [Online]. Available: HowdoestheSoftmaxactivationfunctionwork?

[5] J. Clement, "Number of global social network users 2017-2025," Nov 2020. [Online]. Available: https://www.statista.com/statistics/278414/number-of-worldwide-social-network-users/#:~:text=Social%20media%20usage%20is%20one,almost%204.41%20billion%20in%202025.

[6] W. H. Organisation, "Coronavirus disease 2019 (covid-19) situation report – 94," April 2020. [Online]. Available: https://www.who.int/docs/default-source/coronaviruse/situation-reports/20200423-sitrep-94-covid-19.pdf

[7] "Covid-19 pandemic." [Online]. Available: https://en.wikipedia.org/wiki/COVID-19_pandemic

[8] "Coronavirus disease 2019 (covid-19) situation report-94," April 2020. [Online]. Available: https://www.who.int/docs/default-source/coronaviruse/situation-reports/20200423-sitrep-94-covid-19.pdf

[9] "Coronavirus: The world in lockdown in maps and charts." [Online]. Available: https://www.bbc.com/news/world-52103747

[10] "Coronavirus disease (covid-19) advice for the public." [Online]. Available: https://www.who.int/emergencies/diseases/novel-coronavirus-2019/advice-for-public?gclid=Cj0KCQiAzZL-BRDnARIsAPCJs70zeoTXhqpCEkEGmTsshCS-pUptOa7FFjvOTKYKlRcqtPyBZAvJnx4a wcB

[11] "'rolling shock' as job losses mount even with reopenings." [Online]. Available: https://www.nytimes.com/2020/05/14/business/economy/coronavirus-unemployment-claims.html

[12] "Social media and mental health." [Online]. Available: https://www.helpguide.org/articles/mental-health/social-media-and-mental-health

[13] A. Samet, "How the coronavirus is changing us social media usage," July 2020. [Online]. Available: https://www.emarketer.com/content/how-coronavirus-changing-us-social-media-usage

[14] C. Wang, R. Pan, X. Wan, Y. Tan, L. Xu, C. S. Ho, and R. C. Ho, "Immediate psychological responses and associated factors during the initial stage of the 2019 coronavirus disease (covid-19) epidemic among the general population in china," *International journal of environmental research and public health*, vol. 17, no. 5, p. 1729, 2020.

[15] D. K. Ahorsu, C.-Y. Lin, V. Imani, M. Saffari, M. D. Griffiths, and A. H. Pakpour, "The fear of covid-19 scale: development and initial validation," *International journal of mental health and addiction*, 2020.

[16] I. W. C. Mak, C. M. Chu, P. C. Pan, M. G. C. Yiu, and V. L. Chan, "Long-term psychiatric morbidities among sars survivors," *General hospital psychiatry*, vol. 31, no. 4, pp. 318–326, 2009.

[17] C.-Y. J. Peng, K. L. Lee, and G. M. Ingersoll, "An introduction to logistic regression analysis and reporting," *The journal of educational research*, vol. 96, no. 1, pp. 3–14, 2002.

[18] M. Ahmad, S. Aftab, and I. Ali, "Sentiment analysis of tweets using svm," *Int. J. Comput. Appl*, vol. 177, no. 5, pp. 25–29, 2017.

[19] J. Wang and G. Joshi, "Cooperative sgd: A unified framework for the design and analysis of communication-efficient sgd algorithms," *arXiv preprint arXiv:1808.07576*, 2018.

[20] M. M. Saritas and A. Yasar, "Performance analysis of ann and naive bayes classification algorithm for data classification," *International Journal of Intelligent Systems and Applications in Engineering*, vol. 7, no. 2, pp. 88–91, 2019.

[21] J. Schmidt-Hieber *et al.*, "Nonparametric regression using deep neural networks with relu activation function," *Annals of Statistics*, vol. 48, no. 4, pp. 1875–1897, 2020.

[22] M. R. Zadeh, S. Amin, D. Khalili, and V. P. Singh, "Daily outflow prediction by multi layer perceptron with logistic sigmoid and tangent sigmoid activation functions," *Water resources management*, vol. 24, no. 11, pp. 2673–2688, 2010.

[23] R. A. Dunne and N. A. Campbell, "On the pairing of the softmax activation and cross-entropy penalty functions and the derivation of the softmax activation function," in *Proc. 8th Aust. Conf. on the Neural Networks, Melbourne*, vol. 181. Citeseer, 1997, p. 185.

[24] F. Haque, M. M. H. Manik, and M. Hashem, "Opinion mining from bangla and phonetic bangla reviews using vectorization methods," in *2019 4th International Conference on Electrical Information and Communication Technology (EICT)*. IEEE, 2019, pp. 1–6.

[25] O. Sharif, M. M. Hoque, and E. Hossain, "Sentiment analysis of bengali texts on online restaurant reviews using multinomial naïve bayes," in *2019 1st International Conference on Advances in Science, Engineering and Robotics Technology (ICASERT)*. IEEE, 2019, pp. 1–6.

[26] M. Rahman, S. Haque, and Z. R. Saurav, "Identifying and categorizing opinions expressed in bangla sentences using deep learning technique," *International Journal of Computer Applications*, vol. 975, p. 8887.

[27] M. H. Alam, M.-M. Rahoman, and M. A. K. Azad, "Sentiment analysis for bangla sentences using convolutional neural network," in *2017 20th International Conference of Computer and Information Technology (ICCIT)*. IEEE, 2017, pp. 1–6.

[28] M. Anand and R. Eswari, "Classification of abusive comments in social media using deep learning," in *2019 3rd International Conference on Computing Methodologies and Communication (ICCMC)*. IEEE, 2019, pp. 974–977.

[29] A. Hassan, M. R. Amin, A. K. Al Azad, and N. Mohammed, "Sentiment analysis on bangla and romanized bangla text using deep recurrent models," in *2016 International Workshop on Computational Intelligence (IWCI)*. IEEE, 2016, pp. 51–56.

[30] M. M. J. Soumik, S. S. M. Farhavi, F. Eva, T. Sinha, and M. S. Alam, "Employing machine learning techniques on sentiment analysis of google play store bangla reviews," in *2019 22nd International Conference on Computer and Information Technology (ICCIT)*. IEEE, 2019, pp. 1–5.

[31] M. G. Sarowar, M. Rahman, M. N. Y. Ali, and O. F. Rakib, "An automated machine learning approach for sentiment classification of bengali e-commerce sites," in *2019 IEEE 5th International Conference for Convergence in Technology (I2CT)*. IEEE, 2019, pp. 1–5.

[32] T. Rabeya, N. R. Chakraborty, S. Ferdous, M. Dash, and A. Al Marouf, "Sentiment analysis of bangla song review-a lexicon based backtracking approach," in *2019 IEEE International Conference on Electrical, Computer and Communication Technologies (ICECCCT)*. IEEE, 2019, pp. 1–7.

[33] H. Jelodar, Y. Wang, R. Orji, and H. Huang, "Deep sentiment classification and topic discovery on novel coronavirus or covid-19 online discussions: Nlp using lstm recurrent neural network approach," *arXiv preprint arXiv:2004.11695*, 2020.

[34] G. Barkur and G. B. K. Vibha, "Sentiment analysis of nationwide lockdown due to covid 19 outbreak: Evidence from india," *Asian journal of psychiatry*, 2020.