Project Report

on

CORONA VIRUS TWEETS ANALYSIS

(A dissertation submitted in partial fulfillment of the requirements of Bachelor of Technology in Computer Science and Engineering of theMaulana Abul Kalam Azad University of Technology, West Bengal)

Submitted by

1. Sk Mustak Ahammed
2. Shuvadip Roy
3. Ankan Deria
4. Shatanik Mahanty
5. Sayan Datta

Under the guidance of
Shri/Dr. Debayan
Ganguly

Professor/Asst. Prof./Lecturer,
Dept. of Computer Science and Engineering

Government College of Engineering and Leather Technology

(Affiliated to MAKAUT, West Bengal)

Kolkata - 700106, WB

2022-2023

Certificate of Approval

This is to certify that the project report on “Measuring Heart Rate Using Video” isa record of bonafide work, carried out by Shri Ankan Deria, Sk Mustak Ahammed, Shri Shuvadip Roy, Shri Shatanik Mahanty and Sayan Datta under myguidance and supervision

In my opinion, the report in its present form is in conformity as specified by Government College of Engineering and Leather Technology and as per regulations ofthe Maulana Abul Kalam Azad University of Technology, West Bengal. To the best of my knowledge the results presented here are original in nature and worthy of incorporation in project report for the B.Tech. Program in Computer Science and Engineering.

Signature of
Supervisor/ Guide

Signature of
Head, Dept. of CSE

ACKNOWLEDGEMENT

With great pleasure, I would like to express my profound gratitude and indebtedness to Shri Debayan Ganguly, Computer Science Department, GovernmentCollege of Engineering and Leather Technology, W.B. for his continuous guidance, valuable advice and constant encouragement throughout the project work. His valuable and constructive suggestions at many difficult situations are immensely acknowledged. I am in short of words to express his contribution to this thesis through criticism, suggestions and discussions.

I would like to express my gratitude to Mr. ..... .and Dr... . for their valuable suggestions and help.

1. SKMustak Ahammed-11200120026
2. Shuvadip Roy - 11200120027
3. Ankan Deria – 11200120028
4. Shatanik Mahanty – 11200120030
5. Sayan Dutta - 11200120031

ABSTRACT

Nowadays, the whole world is confronting an infectious disease called the coronavirus. No country remained untouched during this pandemic situation. Due to no exact treatment available, the disease has become a matter of seriousness for both the government and the public. As social distance is considered the most effective way to stay away from this disease. Therefore, to address the people eagerness about the Corona pandemic and to express their views, the trend of people has moved very fast towards social media. Twitter has emerged as one of the most popular platforms among those social media platforms. By studying the same eagerness and opinions of people to understand their mental state, we have done sentiment analysis using the BERT model on tweets. In this paper, we perform a sentiment analysis on two data sets; one data set is collected by tweets made by people from all over the world, and the other data set contains the tweets made by people of India. We have validated the accuracy of the emotion classification from the GitHub repository.

CONTENTS

INTRODUCTION

Nowadays, the Internet is becoming worldwide popular, and it is serving as a cost-effective platform for information carrier by the rapid enlargement of social media. Several social media platforms like blogs, reviews, posts, tweets are being processed for extracting the people's opinions about a particular product, organization, or situation. The attitude and feelings comprise an essential part in evaluating the behavior of an individual that is known as sentiments. These sentiments can further be analyzed towards an entity, known as sentiment analysis or opinion mining. By using sentimental analysis, we can interpret the sentiments or emotions of others and classify them into different categories that help an organization to know people's emotions and act accordingly. This analysis depends on its expected outcomes, e.g., *analyzing the text depending on its polarity and emotions, feedback about a particular feature, and analyzing the text in different languages require detection of the respective language*. It requires a large amount of data that may not be properly structured. Therefore, some preprocessing techniques are used to construct final data set from the extracted data. Moreover, the real-time analysis helps us to look into the current scenario and make decisions to get better results. The COVID-19 or Corona Virus has a major outbreak around the various parts of the world, and people are affected on a very large scale. This leads to a major loss in human life even though the people have different views on the outbreak of the Corona Virus.

In this paper, the authors analyses people's opinions around the world to understand the favorable or unfavorable situations for them. Therefore, our main focus is to do the sentiment analysis on COVID-19 to draw some conclusions on people's opinion. Recently, it has been observed that the number of people actively participated in social media like Facebook, twitter, etc. However, this work uses the tweeter, a social media platform, to collect the public opinions in the form of reviews, comments, post on COVID-19. In this proposed model, we scrape data from the twitter using the existing twitter APIs, and prepare two data sets. After that the sentiment analysis is performed using different matrices like Average Likes and Re-tweets a period, Intensity Analysis, Polarity & Subjectivity, and Wordcloud. Along with this, the LSTM(Long Short Term Memory) is also use for NLP(Natural Language Processing)

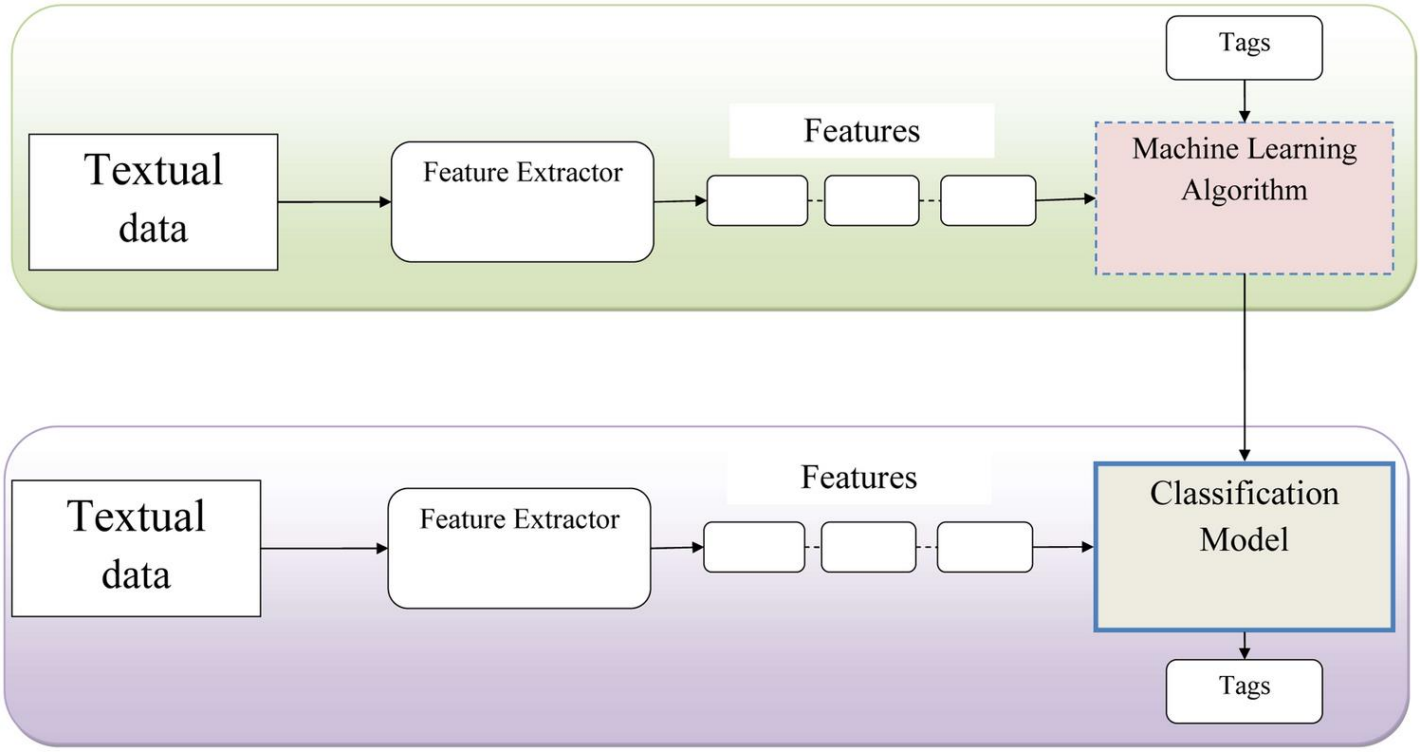
1.1 SOFTWARE USED :-

- Jupyter Notebook
- Pandas, Numpy, Matplotlib, Seaborn, Sklearn, Plotly, Tensorflow, wordcloud, nltk, keras

TECHNICAL APPROACH

Natural Language Processing provides a way to analyse textual data using different approaches that work on different parameters ranging from manual work to an automatic process using in-built libraries. The rule-based approach is one form where the rules are manually defined to perform stemming and covering the text data into tokens, and then classifying the tokens in positive, neutral, and negative categories according to their essence.

In the training phase, the tagged data are provided as the input to a machine learning algorithm to build a classifier model. In the prediction phase, untagged textual data are categorized using the built classifier.



The sentiment analysis process requires two phases:

1. Data set preparation phase and
2. Sentiment analysis phase

The *data set preparation phase* requires the following steps: *scraping data from twitter, cleaning the data, and selecting the relevant features*. We scrape tweets from the twitter using the scraper and filter the scraped data according to our requirements to perform sentiment analysis.

Data set preparation phase

The data set preparation phase comprises of the following steps: *Data scraping and cleaning,* and *selection of the relevant features*.

Data scraping and cleaning

We select Twitter, a social media platform, for extracting tweets on COVID-19, and use the Twitter scraper and the tweepy APIs for data scraping. Table 1 exhibits the total number of tweets extracted from the twitter to prepare the data sets. Thereafter, we clean the scraped data set using regression where tweet text is mapped with an equation and filters out links, images, and emotions from the text. A time stamp feature is a composite entity of date and time in that we require only date of a tweet. Twitter scrapper API is used to extract data with hashtags (#COVID2019 OR #COVID19 OR corona&virus) from 20 Jan 2020 to 25 April 2020 concerning the beginning and end date of each statement to get the required attributes. We construct two data sets, first contains the tweets from the entire world, and another contains the tweets from India only. We use keywords like 'India' and 'Modi' to filter out tweets from India. The Indian tweet 1 data set is created using the keyword 'India' while the Indian tweet 2 data set is created by using the keyword 'Modi'. To get the resultant data set from Indian tweets, we merge both the Indian tweets 1 and the Indian tweets 2 data sets. We do not have an appropriate method to use location filter, therefore, we use the specific keywords to create the second data set. English language is the only medium of expression of emotions in our data set. The APIs yield a large number of features from derived tweets and among them some features may be irrelevant. Therefore, a feature selection technique is used to extract only the relevant features

Data Balancing

We need to balancing the dataset for classification. Here we apply a new technique for oversampling the dataset. After we preprocess the data , we sorting the dataset while keeping the sentence with most no of important words at the first. For example there are 'n' sentences, and we want to increase it to maximum 'm' no of sentences, for that we have to arrange the sentences in descending order of their length. High ranked sentences in the order will be duplicated and hence finally increase the no of sentences.

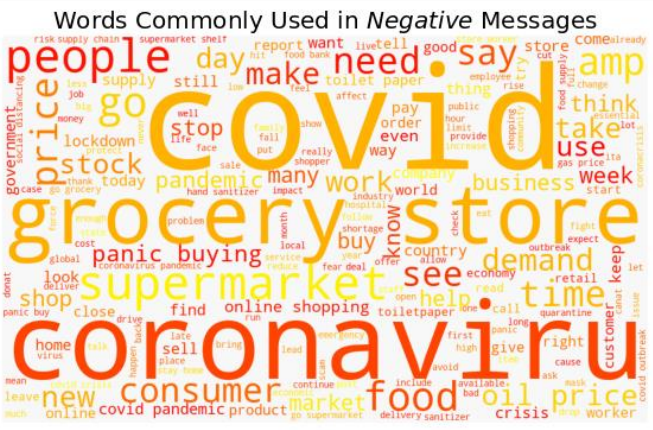
Sentiment analysis phase

After preparing the resultant data sets, we analyse both data sets to calculate various measures using the in-built libraries and functions provided by Numpy and Pandas.

We select five metrics, *Average Likes over the period, Average Re-tweets over the period, Intensity Analysis, Polarity & Subjectivity*, and *Wordcloud*, for analysis, and use the LSTM model for classification. The details of each metric are given in the following subsections.

Wordcloud

Each sentence consists of several words having different intensity and behaviour. In the previous step, we have calculated the type of polarity i.e., positive, negative, and neutral. Each of them has a wordcloud showing different words that come in that category. Wordcloud shows all the words and frequencies that indicate the size of respective words. The bigger word represents a high frequency of occurring in the text. Stop words also have a high frequency than any other words, but these words do not make any sense and do not show the emotion. Therefore, we remove the stop words before making the wordcloud.



RESULTS AND APPROACH

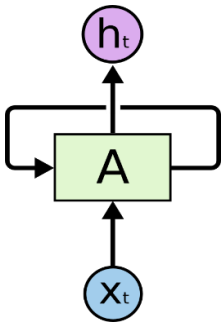
- **Emotion classification using LSTM and RNN**

In this phase, Long Short Term Memory (LSTM) model is used for emotion classification. The meaning of a word in a given sentence depends on the other words surrounding it.

Humans don't start their thinking from scratch every second. As you read this essay, you understand each word based on your understanding of previous words. You don't throw everything away and start thinking from scratch again. Your thoughts have persistence.

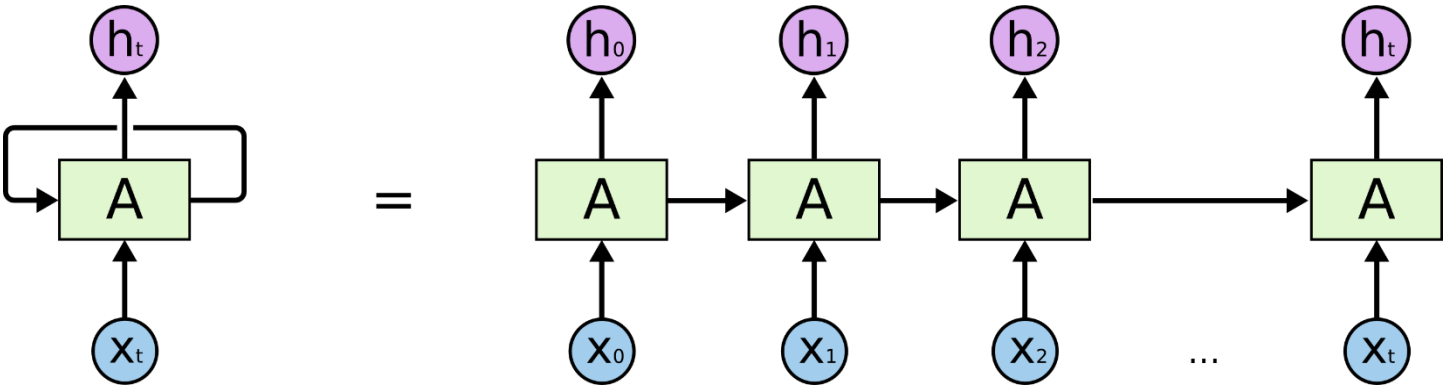
Traditional neural networks can't do this, and it seems like a major shortcoming. For example, imagine you want to classify what kind of event is happening at every point in a movie. It's unclear how a traditional neural network could use its reasoning about previous events in the film to inform later ones.

Recurrent neural networks address this issue. They are networks with loops in them, allowing information to persist.



Recurrent Neural Networks have loops.

In the above diagram, a chunk of neural network, A, looks at some input x_t and outputs a value h_t. A loop allows information to be passed from one step of the network to the next.

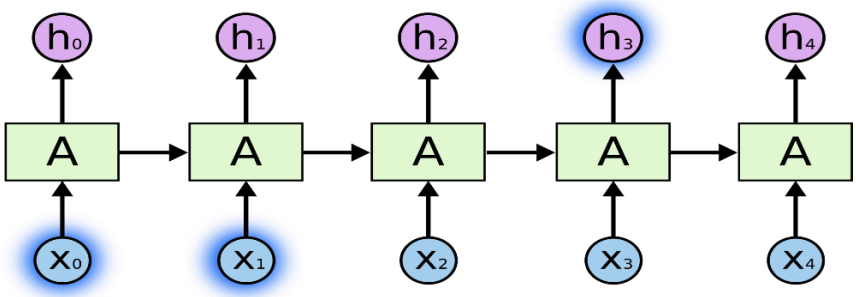


An unrolled recurrent neural network.

- ## The Problem of Long-Term Dependencies

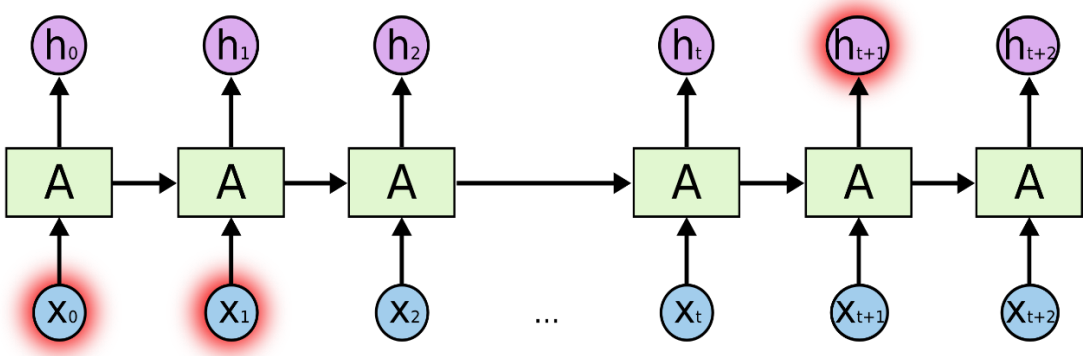
One of the appeals of RNNs is the idea that they might be able to connect previous information to the present task, such as using previous video frames might inform the understanding of the present frame. If RNNs could do this, they'd be extremely useful. But can they? It depends.

Sometimes, we only need to look at recent information to perform the present task. For example, consider a language model trying to predict the next word based on the previous ones. If we are trying to predict the last word in "the clouds are in the *sky*," we don't need any further context – it's pretty obvious the next word is going to be sky. In such cases, where the gap between the relevant information and the place that it's needed is small, RNNs can learn to use the past information.



But there are also cases where we need more context. Consider trying to predict the last word in the text "I grew up in France… I speak fluent *French*." Recent information suggests that the next word is probably the name of a language, but if we want to narrow down which language, we need the context of France, from further back. It's entirely possible for the gap between the relevant information and the point where it is needed to become very large.

Unfortunately, as that gap grows, RNNs become unable to learn to connect the information.

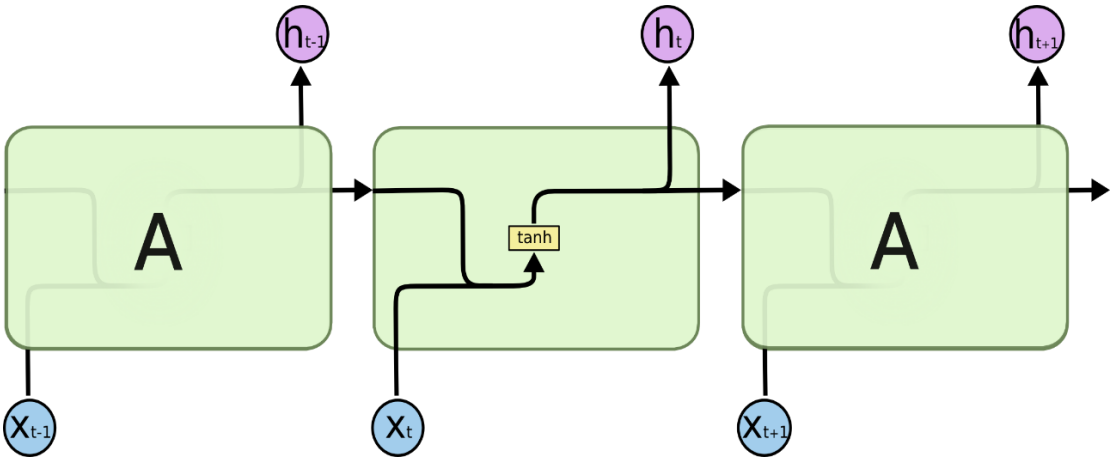


- **LSTM Networks**

Long Short Term Memory networks – usually just called "LSTMs" – are a special kind of RNN, capable of learning long-term dependencies. They were introduced by Hochreiter & Schmidhuber (1997), and were refined and popularized by many people in following work. They work tremendously well on a large variety of problems, and are now widely used.

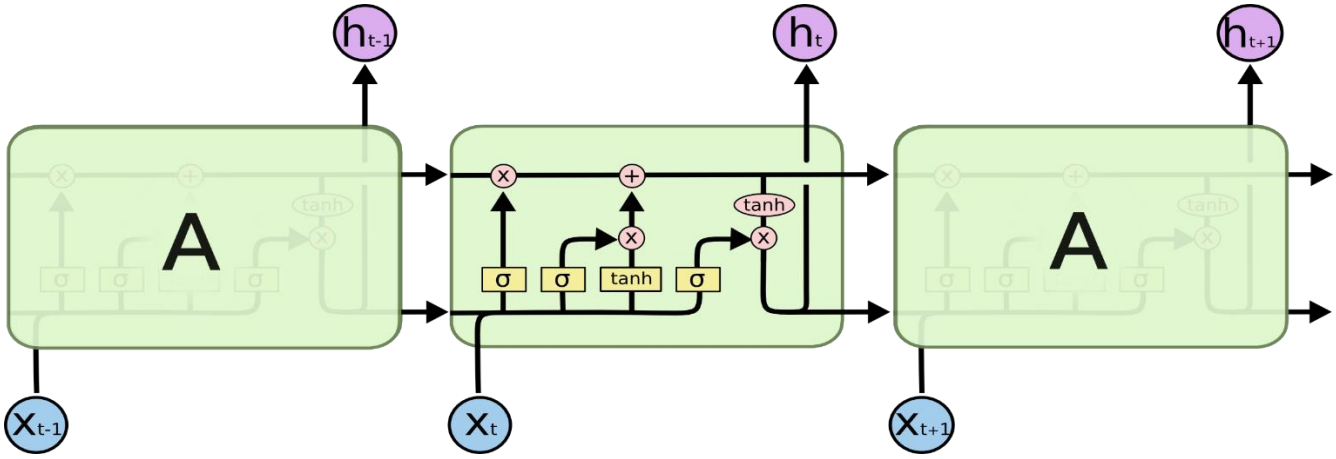
LSTMs are explicitly designed to avoid the long-term dependency problem. Remembering information for long periods of time is practically their default behavior, not something they struggle to learn!

All recurrent neural networks have the form of a chain of repeating modules of neural network. In standard RNNs, this repeating module will have a very simple structure, such as a single tanh layer.



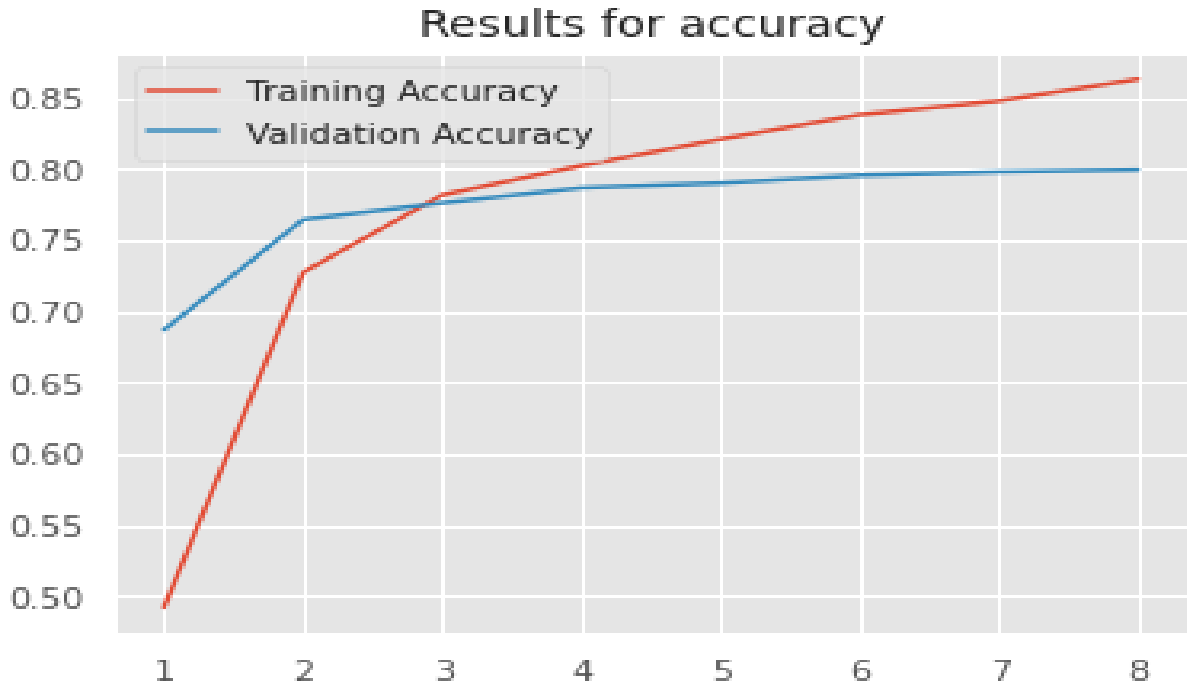
The repeating module in a standard RNN contains a single layer.

LSTMs also have this chain like structure, but the repeating module has a different structure. Instead of having a single neural network layer, there are four, interacting in a very special way.



Accuracy

After applying all these process, we got the validation accuracy is ≈≈ 80% and test accuracy ≈≈ 70%



As we can see from the graph that our accuracy is still increasing, so if we train it with more epochs, there is a possibility that our accuracy may increase.

CONCLUSION

In this work, the sentiment analysis is performed with the help of the LSTM model on the twitter data sets. The data set is categorized on the basis of location of tweets made by people of India and rest of the world. The collected tweets are taken at the time when there was lack of negitiveness about the coronavirus around the world that impact their personal and professional lives. Simultaneously, it has been observed that people from India have relatively more positive communication on the twitter and less

tendency towards spreading negativity. The emotion analysis indicates the success or failure of the measures adopted by the government of a country in various circumstances. Further, it can be observed that the efficacy of taken measures for the people of a country that can support the government in taking more significant decisions to tackle novel coronavirus. The overall performance of the proposed model in terms of the validation accuracy on the collected data sets is approximately 70%.

You can find our code here : https://colab.research.google.com/drive/1j8N9SXEoh-2Wn5US4hSfDE2jHuRSfEJ0?usp=sharing

References

[1] https://link.springer.com/article/10.1007/s13278-021-00737-z#Sec14

[2] Hands-On Machine Learning with Scikit-Learn Keras & TensorFlow – Aurelien Georn

[3] https://www.kaggle.com/code/tracyporter/udemy-nlp-covid19-tweets

[4] Alhajji M, Al Khalifah A, Aljubran M, Alkhalifah M (2020) Sentiment analysis of tweets in saudi arabia regarding governmental preventive measures to contain covid-19