श्रमम् बिना न किमपि साध्यम्

Sentiment Analysis on Large-Scale Covid-19 Tweets using Hybrid Convolutional LSTM Based on Naïve Bayes Sentiment Modeling

Arunava Kumar Chakraborty

Under Guidance of

Dr. Anup Kumar Kolya

Assistant Professor, Dept. of Computer Science and Engineering,

RCC Institute of Information Technology, Beleghata, Kolkata - 700015, India

Project – Part 2 (Dissertation II + Defence of Project – II) [PGCSE 491]

Presented By:

Arunava Kumar Chakraborty

Course – MTech (CSE); **Year –** 2nd; **Semester –** 4th; **Roll No. –** MCS2019/001

University Roll No. – 11711219006; **University Reg. No.** – 026712

RCC Institute of Information Technology

Introduction

- Sentiment Analysis is a type of Data Mining that measures the inclination of people's opinions through Natural Language Processing (NLP), computational linguistics and text analysis [2].
- By using sentiment analysis on social media, we can get incredible insights on a particular topic or incident what is happening around the world.
- This is often used to identify and extract human sensations within a given text across blogs, reviews, social media, forums, news etc.
- Using the sentiment analysis, we can analyze and extract subjective information from the social media to determine whether the data is *positive*, *negative* or *neutral* [3, 10].

Motivation & Objectives

- The Novel Coronavirus (COVID-19) was identified in 31st Dec, 2019 in Wuhan, Hubei Province, China [1].
- WHO announced Covid-19 outbreak as pandemic on 11th March, 2020.
- Several countries implemented strict lockdowns to prevent spread of disease by stopping chains of transmission of COVID-19 [2].
- Lot of people shared their expression on Twitter on this pandemic during this lockdown period [3].
- The motivation behind this work is to analyze the irrational behaviors of people throughout the world.
- We used Sentiment Analysis to measure the trend of public opinions throughout the world [9].

Related Work I

“Analysis of Twitter Data Using Evolutionary Clustering during the COVID-19 Pandemic” ***By*** *Ibrahim Arpaci, Shadi Alshehabi, Mostafa Al-Emran, Mahmoud Khasawneh, Ibrahim Mahariq, Thabet Abdeljawad and Aboul Ella Hassanien*, **2020**. [6]

- **Proposed Methodology:** Evolutionary K-means Clustering Model.
- **Purpose:** Analyse the public attention during the epidemic.
- N-gram Model for identification of tweet patterns.
- **Result:** The difference between the occurrences of n-gram.

Related Work II

“Prediction and analysis of COVID-19 positive cases using deep learning models: A descriptive case study of India” ***By*** *Parul Arora, Himanshu Kumar and Bijaya Ketan Panigrahi,* ***2020****.* [5]

- **Proposed Methodology:** Covid-19 trend prediction model.
- **Purpose:** Predicting the COVID-19 trend for positive cases in different states of India.
- **Training:** Different LSTM variants such as stacked, convolutional and bi-directional LSTM used for training on historical data.
- **Result:** The proposed Bi-LSTM model gives more accurate results over other LSTM models.

Related Work III

"Datastories at semeval-2017 task 4: Deep LSTM with Attention for Message-level and Topic-based Sentiment Analysis" ***By*** *Baziotis Christos, Nikos Pelekis and Christos Doulkeridis*, **2017**. [8]

- **Proposed Methodology:** Message-level and Topic-based sentiment analysis.
- **Purpose:** The proposed deep learning system used for -
  - ✓ Short-text sentiment analysis using an attention mechanism.
  - ✓ Topic-based sentiment analysis, with a context-aware attention mechanism utilizing the topic information.
- **Training:** The LSTM network augmented with two kinds of attention mechanisms.

Related Work IV

“Long Short Term Memory Hyperparameter Optimization for a Neural Network Based Emotion Recognition Framework” ***By*** *Bahareh Nakisa, Mohammad Naim Rastgoo, Andry Rakotonirainy, Frederic Maire and Vinod Chandran*, **2018**. [9]

- **Proposed Methodology:** The Deep Neural Network-based Emotion Recognition framework.
- **Purpose:** To find an optimized LSTM classier with high performance in the context of emotion classification.
- **Training:** LSTM Deep Neural Network.
- **Result:** Optimized LSTM classifier achieved 77.68% accuracy by using the Differential Evolution algorithm.

Related Work V

“COVID-19 Public Sentiment Insights and Machine Learning for Tweets Classification” ***By*** *Jim Samuel, G. G. Md. Nawaz Ali, Md. Mokhlesur Rahman, Ek Esawi and Yana Samuel*, **2020**. [11]

- **Purpose:** Comparative analysis of four different machine learning classifiers on Coronavirus Tweets data.
- **Machine Learning Classifiers:** Naïve Bayes classifier, Logistic Regression.
- **Result:** Naïve Bayes classifier achieved 91% classification accuracy for short Tweets whether using the Logistic Regression classifier provides classification accuracy of 74% on the same number of tweets.

Related Work VI

“Deep Sentiment Classification and Topic Discovery on Novel Coronavirus or COVID-19 Online Discussions: NLP Using LSTM Recurrent Neural Network Approach” ***By*** *Hamed Jelodar, Yongli Wang, Rita Orji and Shucheng Huang*, **2020**. [12]

- **Proposed Methodology:** Sentiment Evaluation using deep learning model.
- **Purpose:** To detect the topics on Coronavirus related issues from online healthcare forums.
- **Training:** LSTM Recurrent Neural Network.
- **Result:** The proposed model achieved 81.15% of classification accuracy on COVID-19 comments for classifying them into positive, negative and neutral classes.

Novelty of Our Work

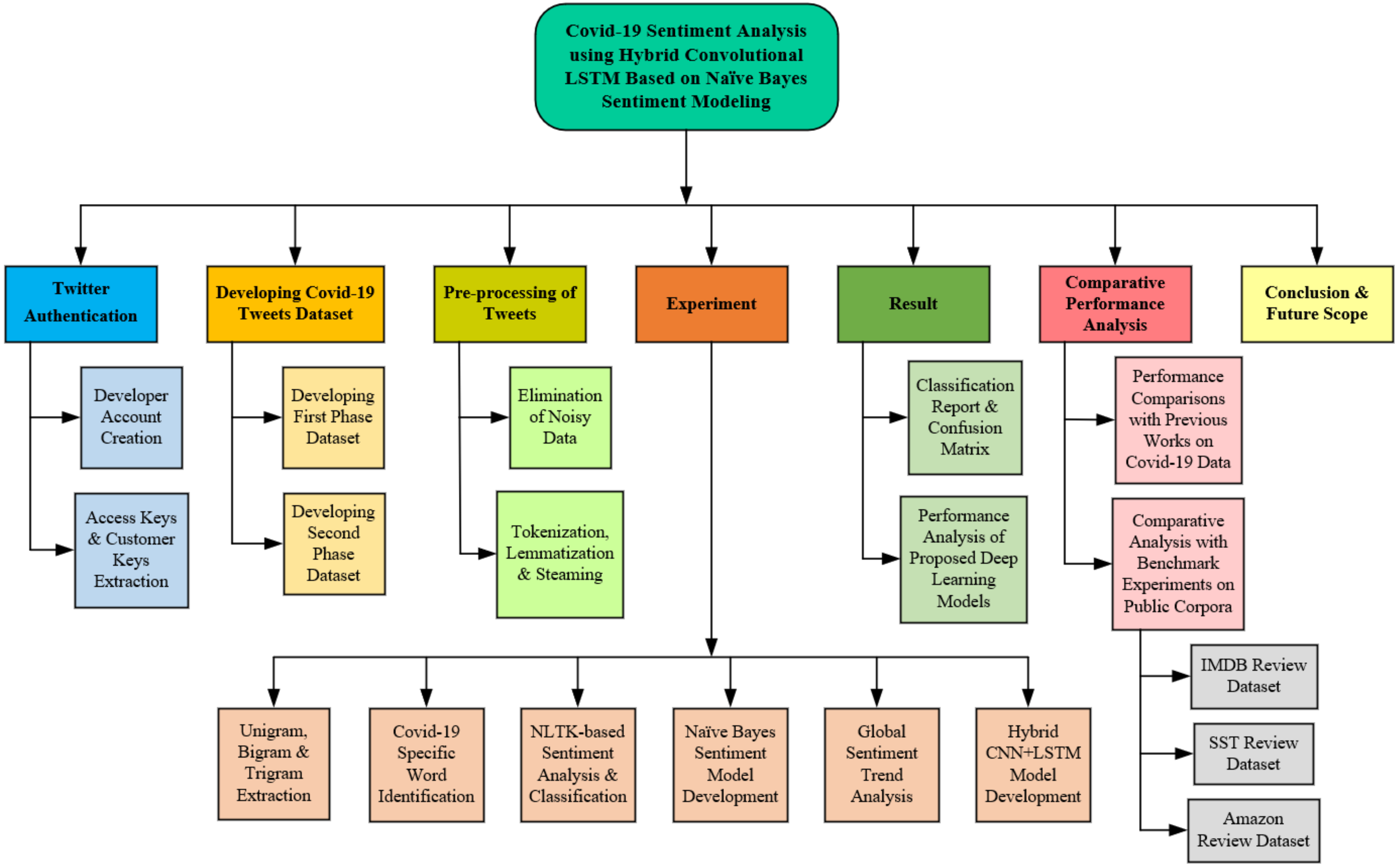
- **Originality :** Till now there are only few numbers of significant research works demonstrating the public sentiments on Covid-19 pandemic all over the world.
- **Preparing real time dataset :** We have started streaming tweets from the twitter after WHO declared Covid-19 as pandemic. We have collected almost 235k and 320k worldwide English tweets related to Covid-19 during April – June, 2020 and August – October, 2020.

Contribution

- N-gram Model
- Covid-19 Specified Word Identification
- NLTK-based Sentiment Analysis
- Naïve Bayes Sentiment Model
- Global Sentiment Trend Analysis
- Sentiment Modeling using Hybrid Convolutional LSTM

Design Methodology

- **Work Breakdown Structure**



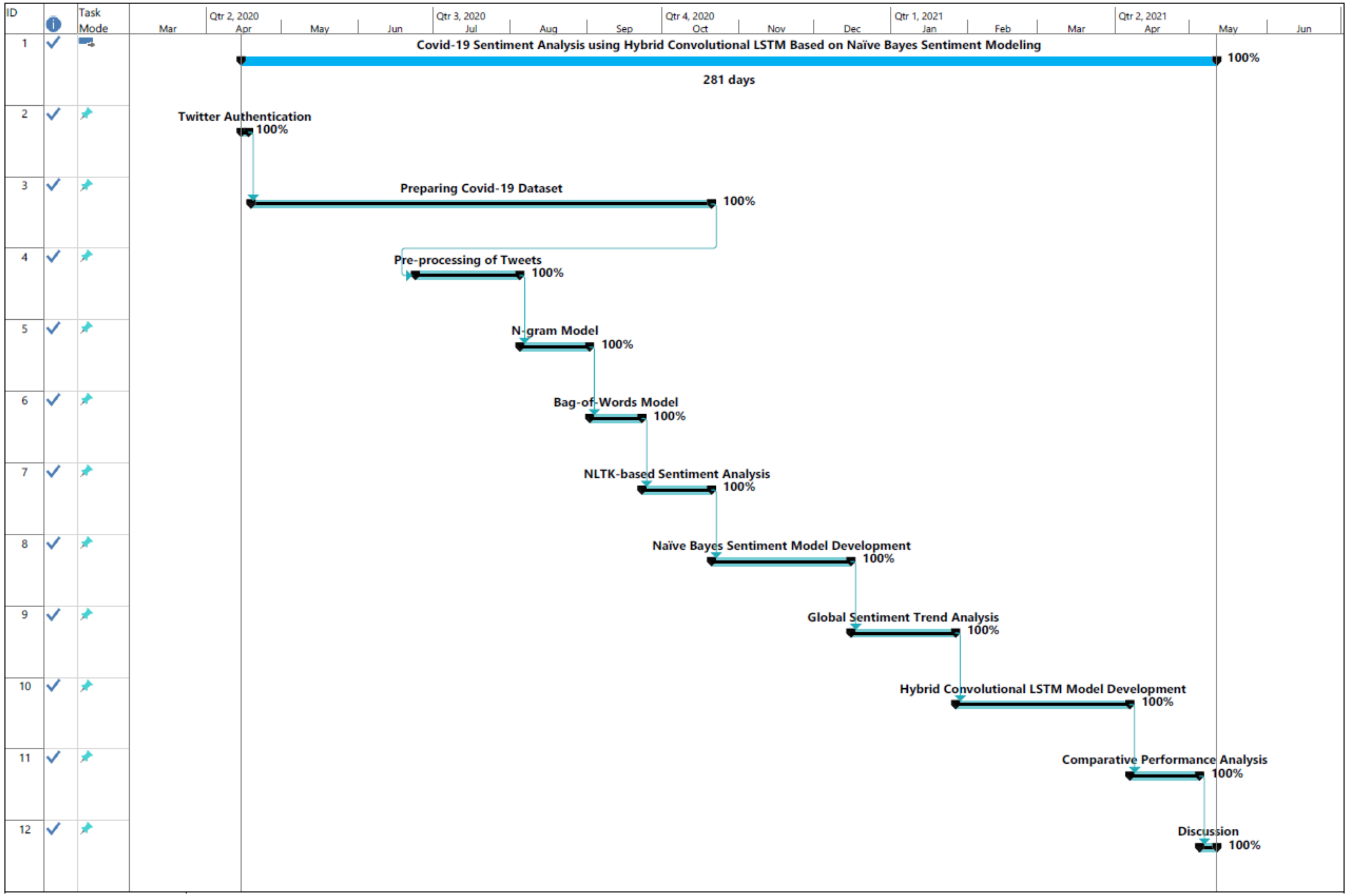
Design Methodology

- **Timeline Task Sheet**

ID		Task Mode	Task Name	Duration	Start	Finish	Predecessors
1			**Covid-19 Sentiment Analysis using Hybrid Convolutional LSTM Based on Naïve Bayes Sentiment Modeling**	**281 days**	**Wed 15-04-20**	**Tue 11-05-21**	
2			**Twitter Authentication**	**3 days**	**Wed 15-04-20**	**Fri 17-04-20**	
3			Developer Account Creation	2 days	Wed 15-04-20	Thu 16-04-20	
4			Access Keys & Customer Keys Extraction	1 day	Fri 17-04-20	Fri 17-04-20	3
5			**Preparing Covid-19 Dataset**	**133 days**	**Sun 19-04-20**	**Tue 20-10-20**	**2**
6			Construction of First Phase Dataset	46 days	Sun 19-04-20	Sat 20-06-20	4
7			Construction of Second Phase Dataset	44 days	Thu 20-08-20	Tue 20-10-20	6
8			**Pre-processing of Tweets**	**30 days**	**Wed 24-06-20**	**Tue 04-08-20**	**6**
9			Elimination of Noisy Data	20 days	Wed 24-06-20	Tue 21-07-20	6
10			Tokenization, Lemmatization & Steaming	10 days	Wed 22-07-20	Tue 04-08-20	9
11			**N-gram Model**	**20 days**	**Wed 05-08-20**	**Tue 01-09-20**	**8**
12			Unigram Extraction & Visualization	8 days	Wed 05-08-20	Fri 14-08-20	10
13			Bigram Extraction & Visualization	6 days	Sat 15-08-20	Mon 24-08-20	12
14			Trigram Extraction & Visualization	6 days	Tue 25-08-20	Tue 01-09-20	13
15			**Bag-of-Words Model**	**15 days**	**Wed 02-09-20**	**Tue 22-09-20**	**11**
16			Covid-19 Specific Word Identification	10 days	Wed 02-09-20	Tue 15-09-20	14
17			Word Popularity & Probability, Visualization	5 days	Wed 16-09-20	Tue 22-09-20	16
18			**NLTK-based Sentiment Analysis**	**20 days**	**Wed 23-09-20**	**Tue 20-10-20**	**15**
19			Sentiment Classification & Distribution	15 days	Wed 23-09-20	Tue 13-10-20	17
20			Visualization	5 days	Wed 14-10-20	Tue 20-10-20	19
21			**Naïve Bayes Sentiment Model Development**	**40 days**	**Wed 21-10-20**	**Tue 15-12-20**	**18**
22			Naïve Bayes Sentiment Classifier Implementation	20 days	Wed 21-10-20	Tue 17-11-20	20
23			Fine-grained Sentiment Classification	6 days	Wed 18-11-20	Wed 25-11-20	22
24			Peperation of Final Data	8 days	Thu 26-11-20	Mon 07-12-20	23
25			Confusion Matrix & Classification Report	6 days	Tue 08-12-20	Tue 15-12-20	24
26			**Global Sentiment Trend Analysis**	**30 days**	**Wed 16-12-20**	**Tue 26-01-21**	**21**
27			Overall Average Sentiment Trend	10 days	Wed 16-12-20	Tue 29-12-20	25
28			Average Sentiment Trend Shift Detection	15 days	Wed 30-12-20	Tue 19-01-21	27
29			Visualization	5 days	Wed 20-01-21	Tue 26-01-21	28
30			**Hybrid Convolutional LSTM Model Development**	**50 days**	**Wed 27-01-21**	**Tue 06-04-21**	**26**
31			Model Creation & Data Fitting	35 days	Wed 27-01-21	Tue 16-03-21	28
32			Sentiment Prediction, Error Analysis & Visualization	15 days	Wed 17-03-21	Tue 06-04-21	31
33			**Comparative Performance Analysis**	**20 days**	**Wed 07-04-21**	**Tue 04-05-21**	**30**
34			Previous Works on Covid-19 Data	5 days	Wed 07-04-21	Tue 13-04-21	32
35			IMDB Movie Review Dataset	5 days	Wed 14-04-21	Tue 20-04-21	34
36			SST Review Dataset	5 days	Wed 21-04-21	Tue 27-04-21	35
37			Amazon Customer Review Dataset	5 days	Wed 28-04-21	Tue 04-05-21	36
38			**Discussion**	**5 days**	**Wed 05-05-21**	**Tue 11-05-21**	**33**
39			Final Result & Conclusion	3 days	Wed 05-05-21	Fri 07-05-21	37
40			Future Scope	2 days	Sat 08-05-21	Tue 11-05-21	39

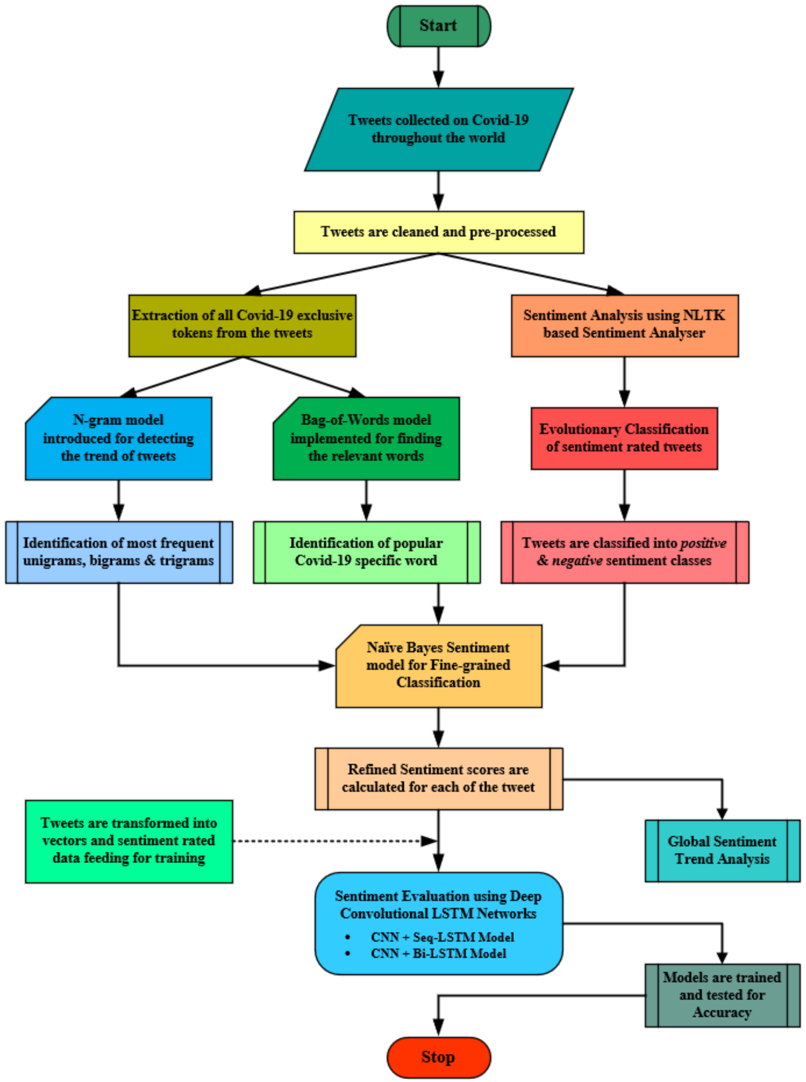
Design Methodology

- **Gantt Chart**



Design Methodology

- **Proposed Methodology**



Developing Covid-19 Dataset

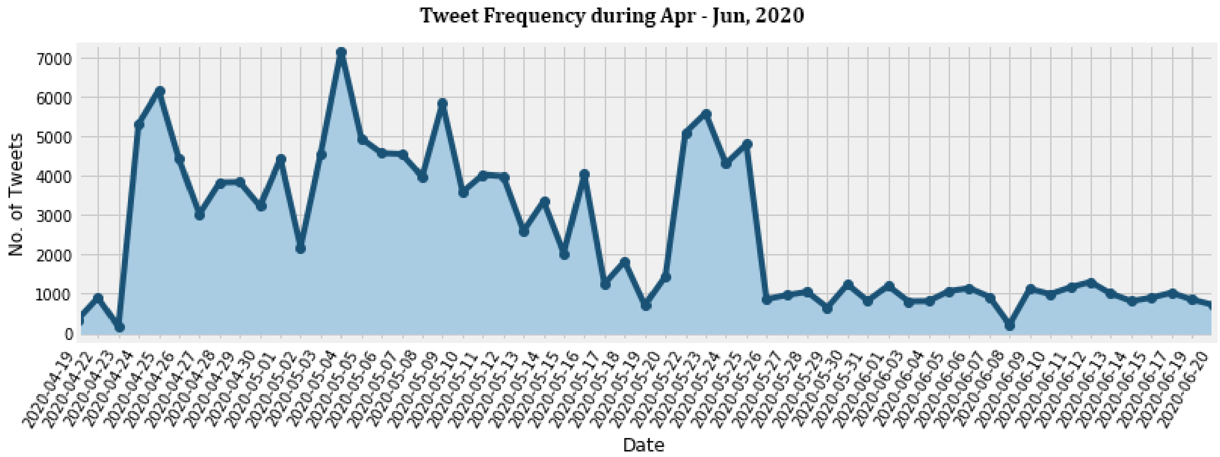
- In 2020, out of 320 million active Twitter users from all over the world, India has 18.9 million active Twitter users, i.e. currently the third highest in the world [4].
- Twitter API mostly allows the users to live stream only 1-2% of the total tweets on a particular keyword [3].
- However, we have collected 2,35,240 tweets for first phase dataset and 3,20,316 tweets for second phase dataset containing the hash-tagged keywords like – ***#covid-19***, ***#coronavirus***, ***#covid***, ***#covaccine***, ***#lockdown***, ***#homequarantine***, ***#quarantinecenter***, ***#socialdistancing***, ***#stayhome***, ***#staysafe*** etc.

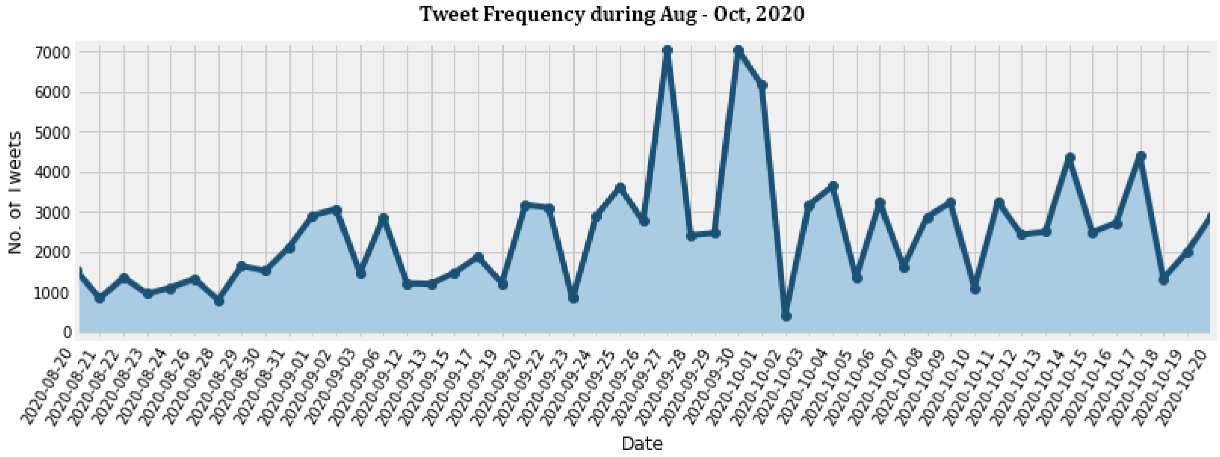
id	created_at	source	original_text	lang	favorite_count	retweet_count	original_author	hashtags	user_mentions	place
125193476721131	Sun Apr 19	<a href="http:	RT @Ash_The	en	0	705	EmpoweringGo	Jehanaba	Ash_TheLoneW	Panjim Goa India
125193473331715	Sun Apr 19	<a href="http:	RT @NGvision	en	0	2646	Ibilola_Amao	lockdown	NGvision2020	London, England
125193466618305	Sun Apr 19	<a href="http:	RT @Barnes_	en	0	5593	cliff_skidmore		Barnes_Law	Texas, USA
125193460755922	Sun Apr 19	<a href="http:	RT @Joelpatri	en	0	108	GMA4Trump_	Covid_19	Joelpatrick1776	Choctaw, OK
125193458229277	Sun Apr 19	<a href="http:	RT @GlblCtzn	en	0	4	Macgirl730	Together,	GlblCtzn	Menomonee Falls, WI

- [4] www.statista.com

Data Pre-Processing

- Initially, we removed all duplicate tweets from the datasets and found 1,43,903 and 1,20,509 tweets from first phase and second phase Covid-19 datasets.



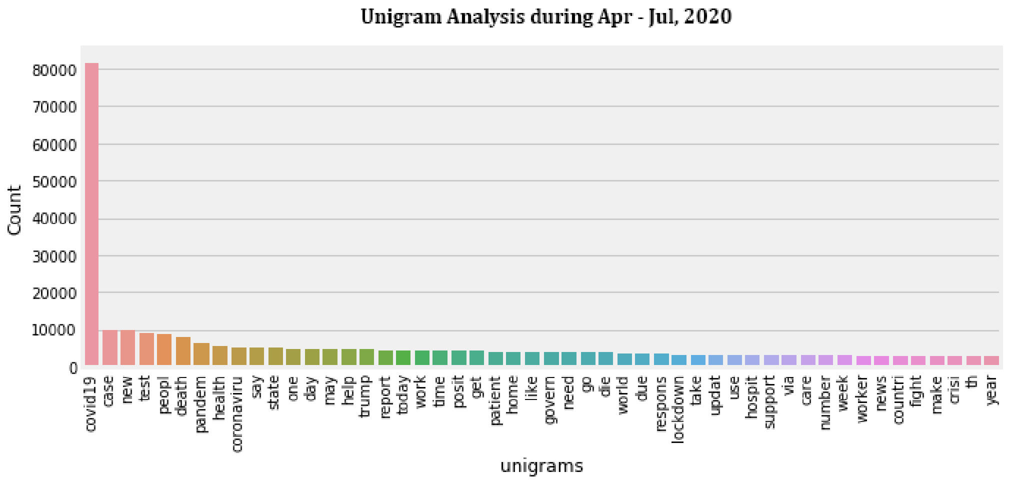


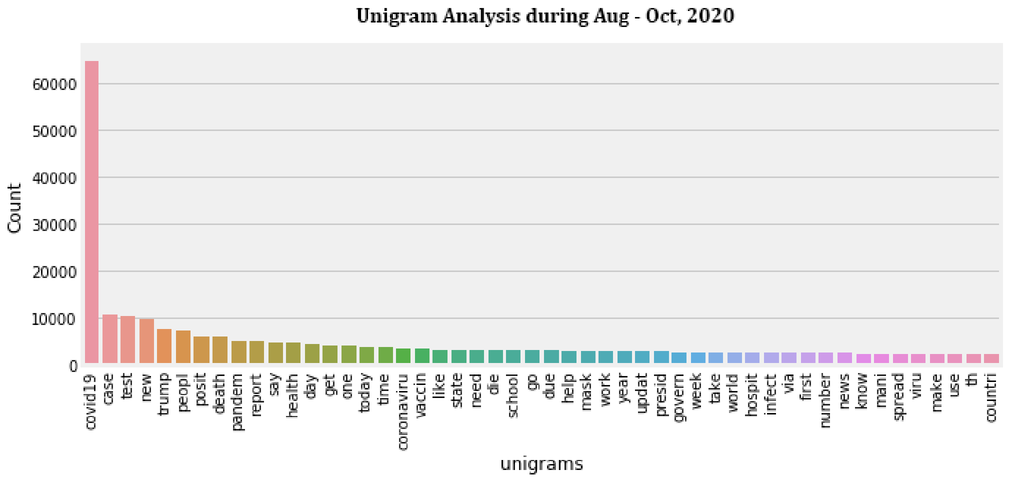
Data Pre-Processing

- The pre-processing function follows the following steps to clean raw tweets-
  1. *Convert the whole tweet into lower case.*
  2. *Remove all URLs from the tweet.*
  3. *Remove all punctuations from the tweet.*
  4. *Remove all stop words from the tweet.*
  5. *Tokenize the tweet to split each of the sentence into smaller parts of word.*
  6. *Stemming used to reduce inflected words to their word stem.*

Feature I : Word Trend Detection using N-gram

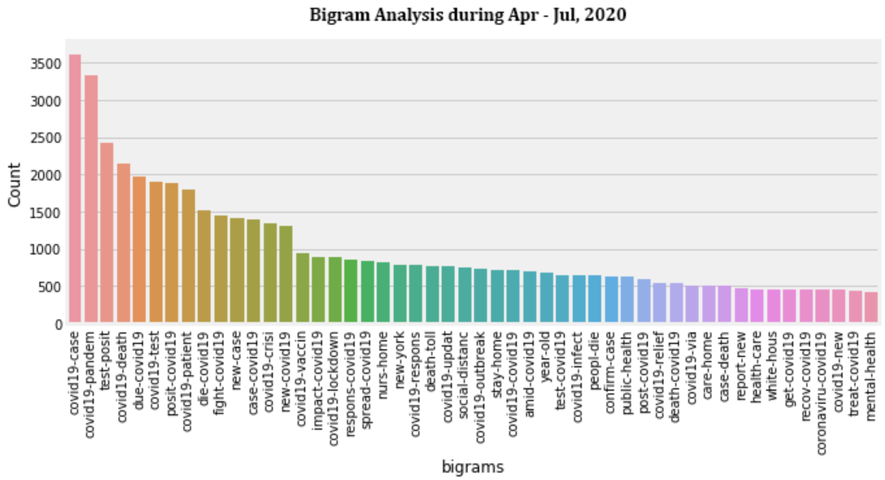
- **lexical n-gram model** used for Identification of the most popular unigrams, bigrams and trigrams within our corpus.
- **Unigram Analysis**

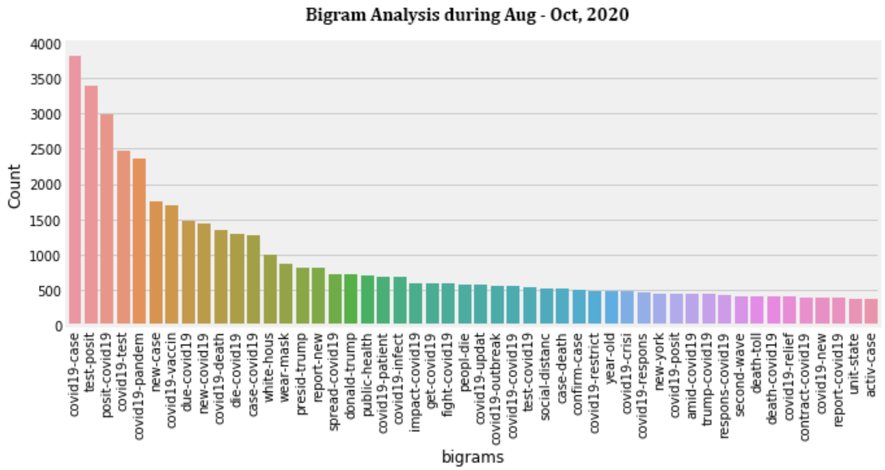




Feature I : Word Trend Detection using N-gram

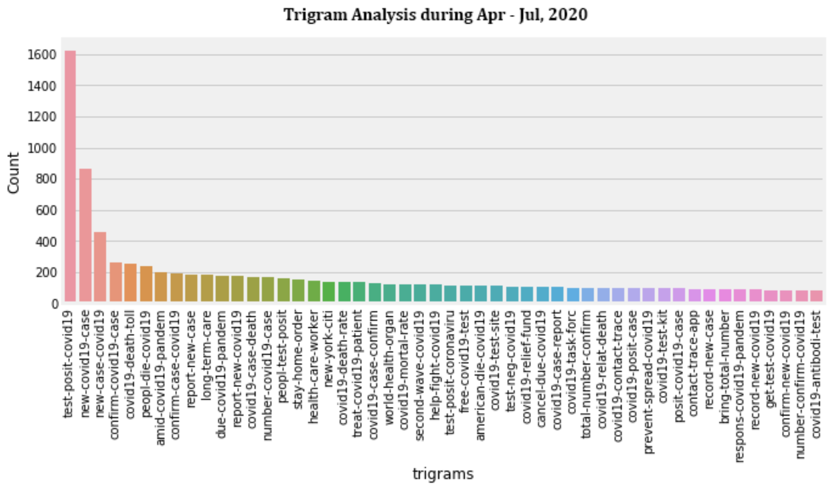
- **Bigram Analysis**

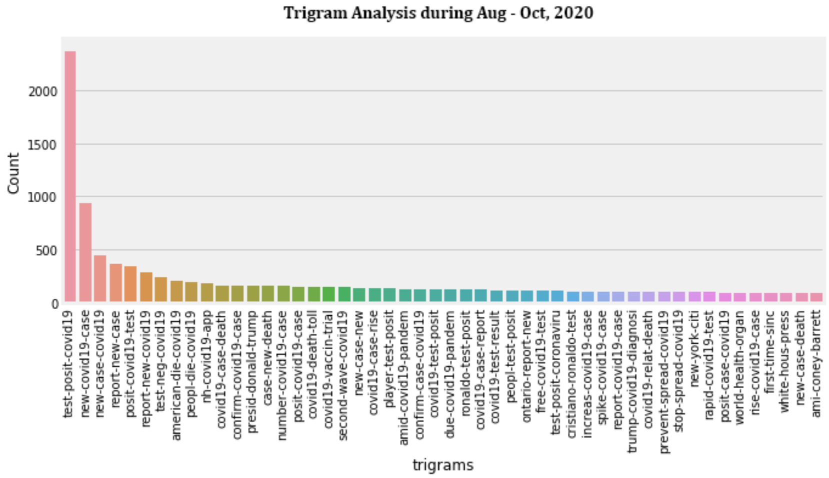




Feature I : Word Trend Detection using N-gram

- **Trigram Analysis**



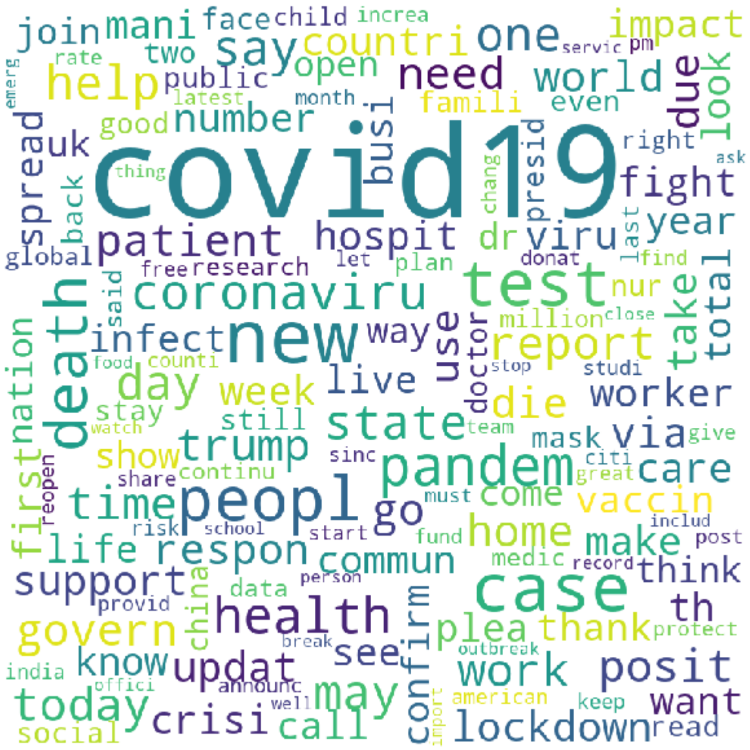


- According to the n-gram model, the popularity of trigrams is lesser than that of bigrams and the unigrams popularity is the highest.

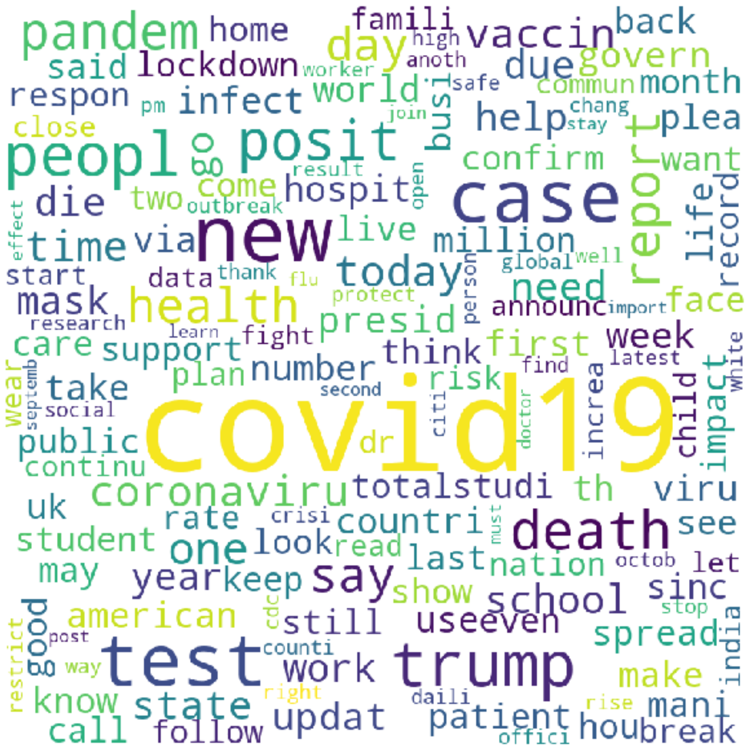
Feature II : Covid-19 Specified Word Identification

- The **Bag-of-Words** (BOW) model used to identify frequently occurred grams from the word lexicon.
- Finally, we obtained a list of most frequent Covid-19 exclusive words.

Popular words found in Tweets during Apr - Jun, 2020



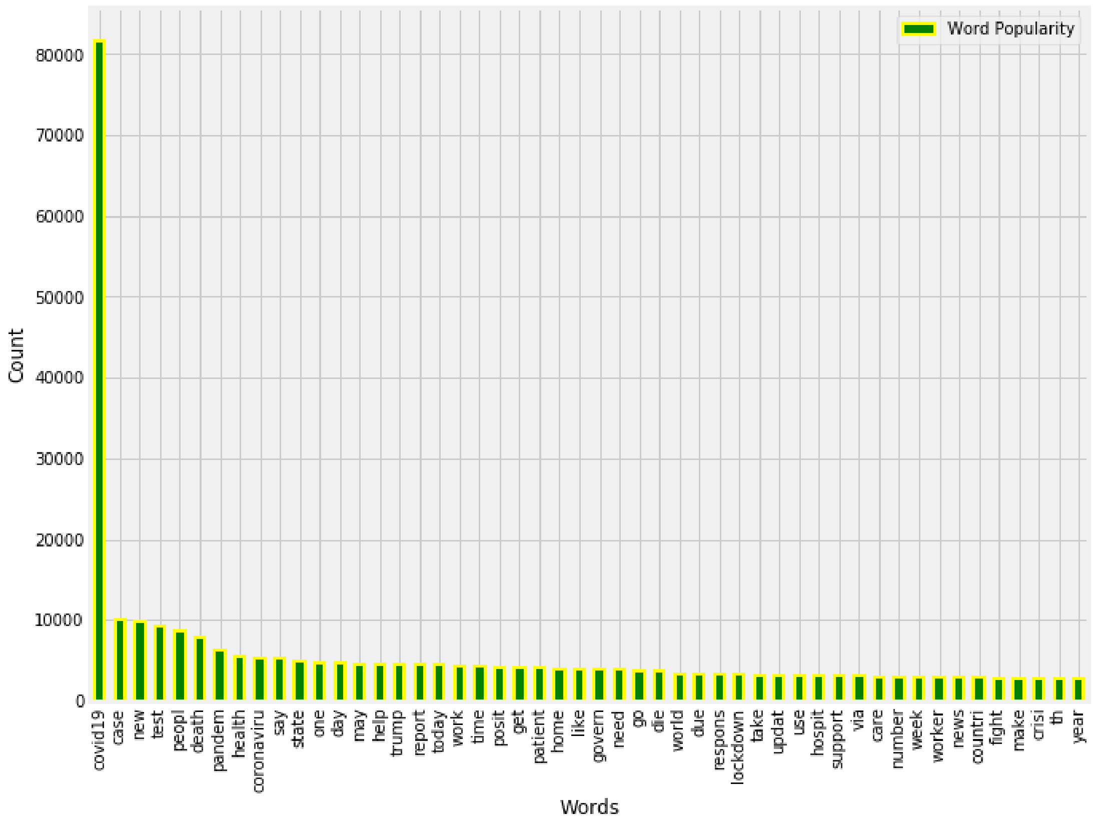
Popular words found in Tweets during Aug - Oct, 2020



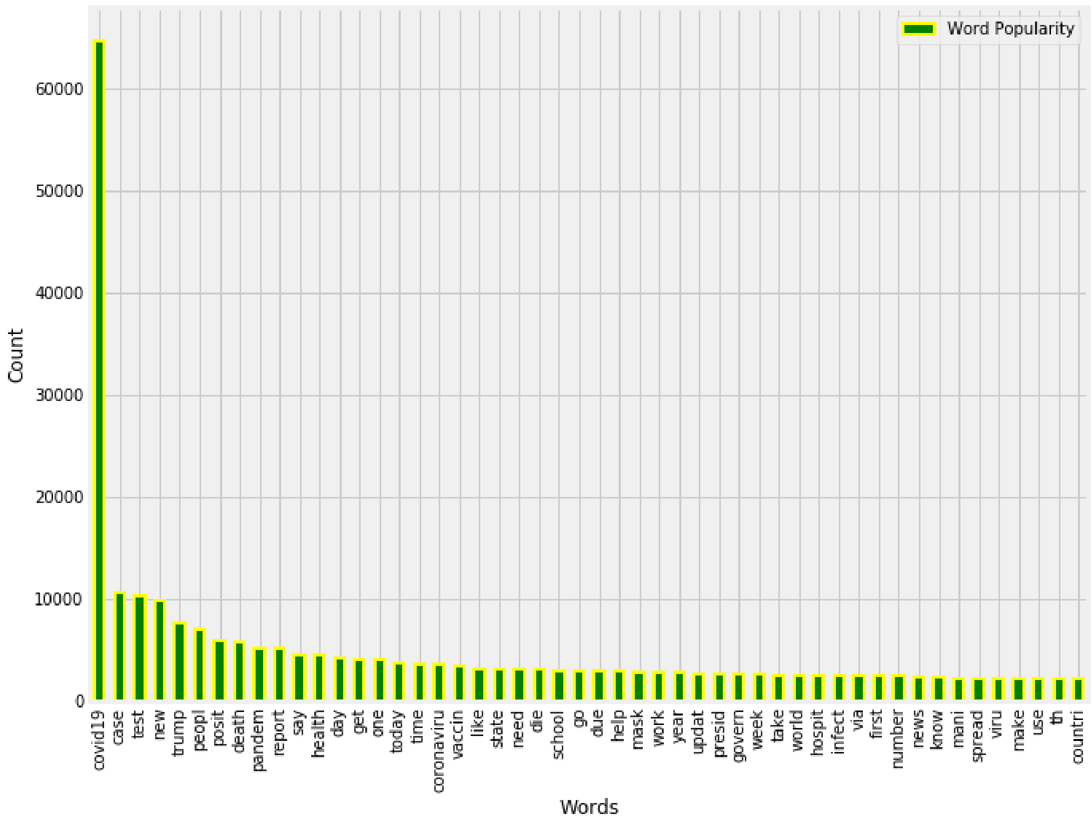
Feature II : Covid-19 Specified Word Identification

- From the obtained list of most popular Covid-19 exclusive words we have found the frequency for the recurrence of each words.

Frequency of Covid-19 specific words during Apr - Jun, 2020



Frequency of Covid-19 specific words during Aug - Oct, 2020



Feature II : Covid-19 Specified Word Identification

- Probability of repetition for each word on the basis of total 13,79,835 and 11,33,188 words from both of the generated corpora.

$$P(W_i) = \frac{count(W_i)}{\sum_{i=0}^{n} count(W_{i=0}^{n})} \qquad (1)$$

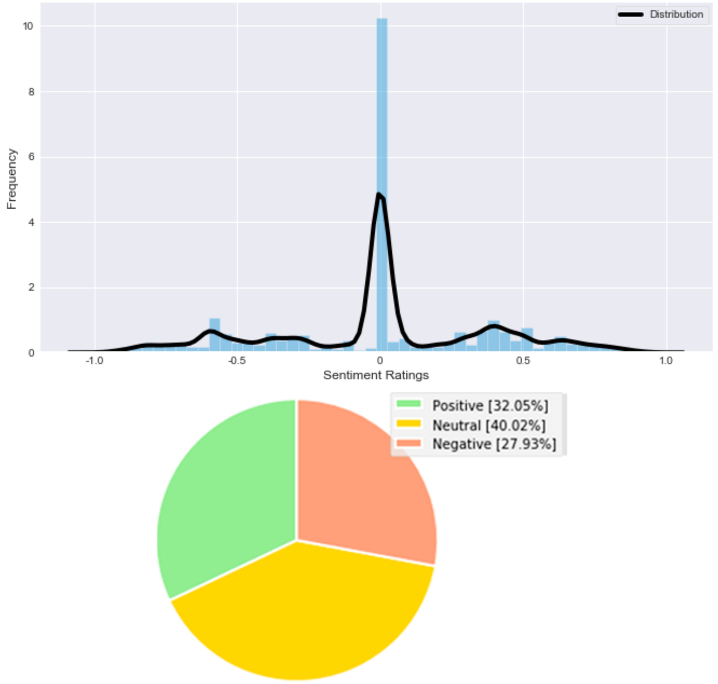
- We represented the popularity and probability of recurrence for top five most frequent Covid-19 related words.

	Covid-19 Dataset I (Apr – Jun, 2020)				**Covid-19 Dataset II (Aug – Oct, 2020)**		
	Words	**Popularity**	**Probability**		**Words**	**Popularity**	**Probability**
1	covid19	81696	0.059207	**1**	covid19	64672	0.057071
2	case	10001	0.007248	**2**	case	10656	0.009404
3	new	9875	0.007157	**3**	test	10323	0.009110
4	test	9204	0.006670	**4**	new	9797	0.008646
5	peopl	8624	0.006250	**5**	trump	7605	0.006711

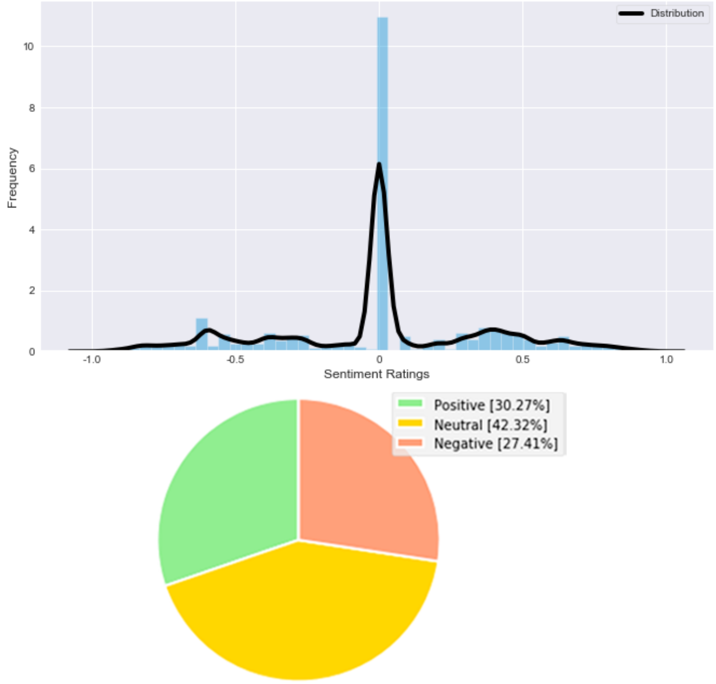
NLTK-Based Sentiment Analysis & Classification

- Evaluation of Sentiment polarity for each preprocessed tweets.
- Classification of tweets into ***positive***, ***negative*** and ***neutral*** classes based on their sentiment score.
- Sentiment distribution of three class polarity along with the percentage of Covid-19 tweets occurred from each class.

Sentiment Distribution of tweets during Apr - Jun, 2020



Sentiment Distribution of tweets during Aug - Oct, 2020



Naïve Bayes Sentiment Analysis

- Evaluation of refined Sentiment ratings for each classified tweets on the basis of extracted features.
- Classification of tweets into ***most positive (1.0)***, ***positive (0.5)***, ***neutral (0.0)***, ***negative (-0.5)*** and ***most negative (-1.0)*** classes based on their sentiment score.

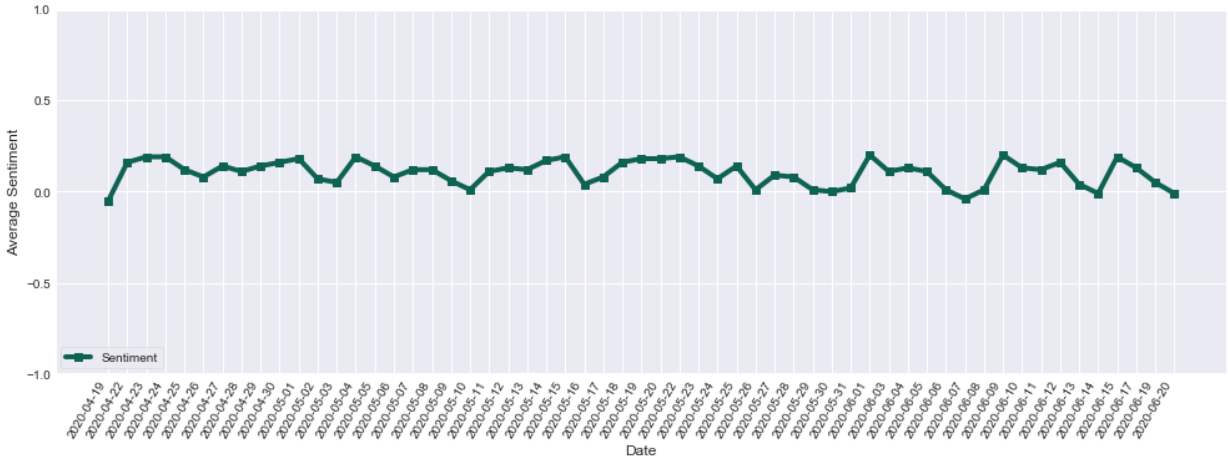
Algorithm Fine-Grained Sentiment Classification (sentiment, refined_sentiment):

```
for each i in range (0, len(tweet.index)):
    if tweet_i[sentiment] >= 0.0 and tweet_i[sentiment] < 0.25:
        tweet_i[refined_sentiment] ← -1.0      # assigned -1.0 for Most Negative Tweets
    elif tweet_i[sentiment] >= 0.25 and tweet_i[sentiment] < 0.5:
        tweet_i[refined_sentiment] ← -0.5      # assigned -0.5 for Negative Tweets
    elif tweet_i[sentiment] == 0.5:
        tweet_i[refined_sentiment] ← 0.0       # assigned 0.0 for Neutral Tweets
    elif tweet_i[sentiment] > 0.5 and tweet_i[sentiment] <= 0.75:
        tweet_i[refined_sentiment] ← 0.5       # assigned 0.5 for Positive Tweets
    else:
        tweet_i[refined_sentiment] ← 1.0       # assigned 1.0 for Most Positive Tweets
end
```

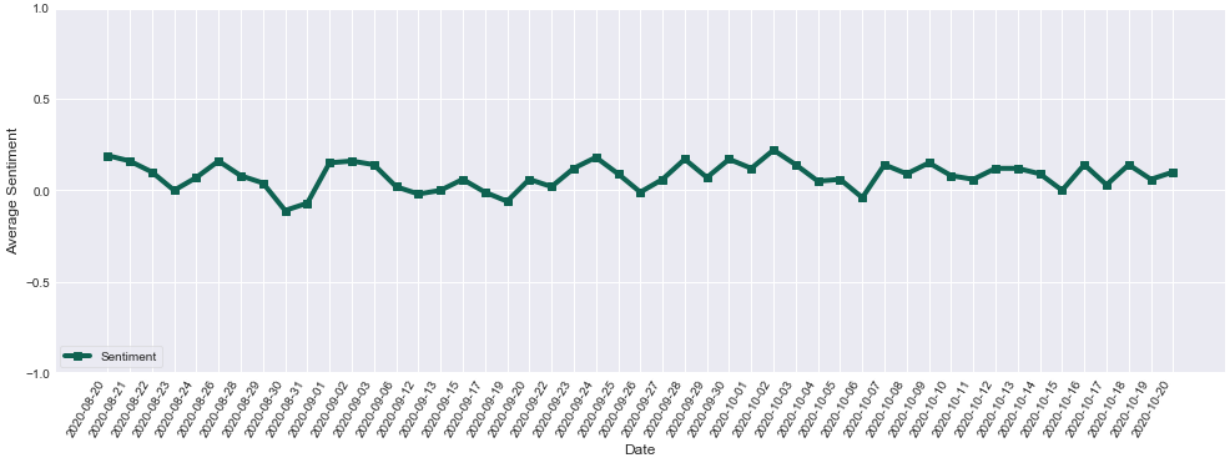
Global Sentiment Trend Analysis

- Overall Average Sentiment Trend throughout the world.

Average Sentiment Trend of Covid-19 during Apr - Jun, 2020



Average Sentiment Trend of Covid-19 during Aug - Oct, 2020



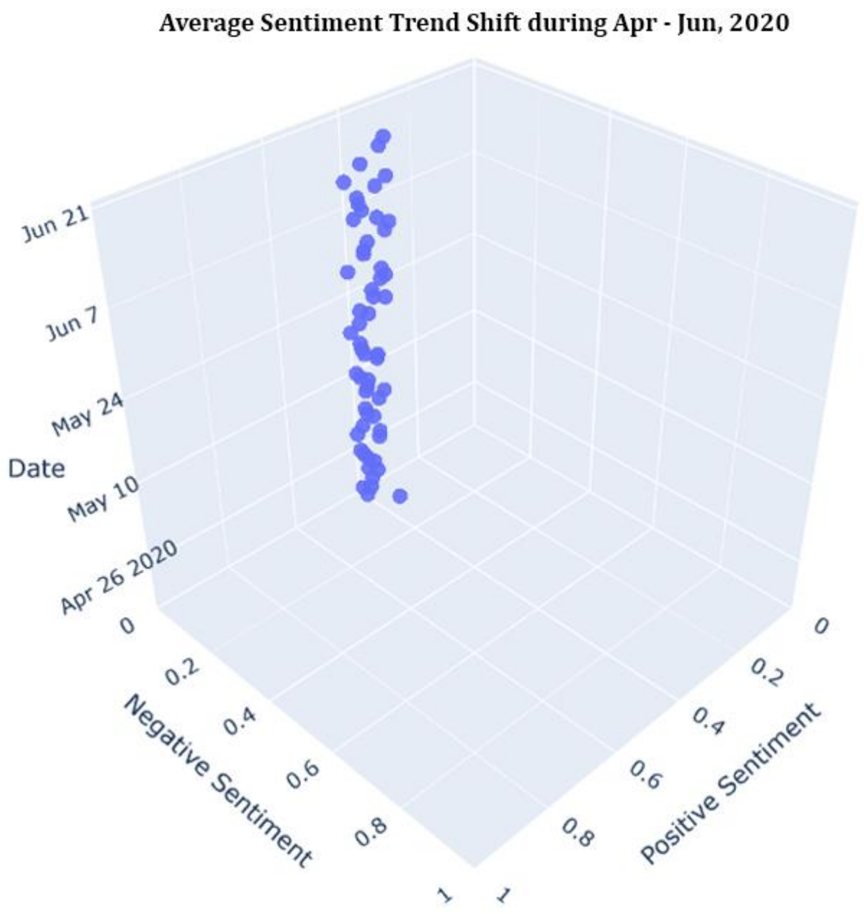
Global Sentiment Trend Analysis

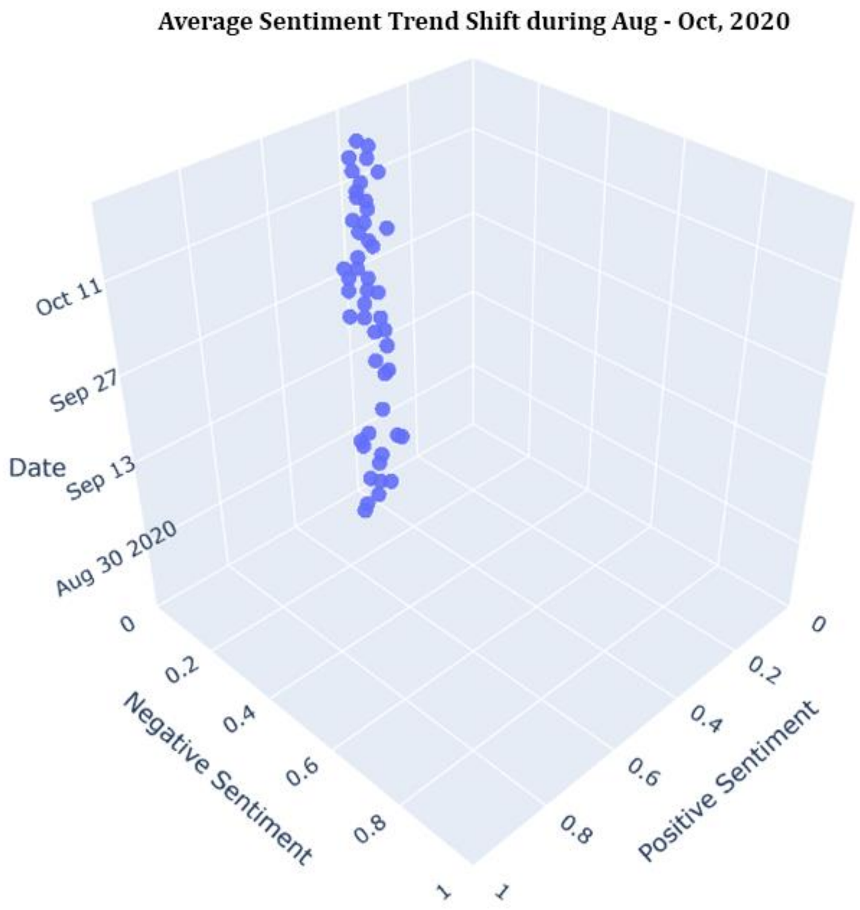
- Some Covid-19 tweets from different sentiment classes.

Date	Original Tweet	Naïve Bayes Sentiment Rating	Refined Sentiment Rating	Sentiment Class
05-05-2020	**Barbara Walton #FBPE @LesTroisChenes** More Than 60 Doctors in Italy Have Died in COVID-19 Pandemic. Why? Lack of PPE and information. UK just a fraction behind them - far ahead of other countries. We threw our health worker under the #COVID bus and then clapped for them.	0.24	-1	Most Negative
27-04-2020	**Elizabeth Stein @lizstein** A tale of two cities. “More than fifteen thousand people in New York are believed to have died from covid-19. Last week in Washington State, the estimate was fewer than seven hundred people.”	0.26	-0.5	Negative
10-05-2020	**Volker Stollorz @Stollovo** “Anyway, I remain a born optimist. And now that I have faced death, my tolerance levels for nonsense and bullshit have gone down even more than before. So, I continue calmly and enthusiastically, although more selectively than before my illness” #COVID19	0.50	0.0	Neutral
26-04-2020	**Muh’d @smj_esq** People are recovering from COVID-19 and it’s a good development. I think Government should tell the public the drugs administered on the recovered patients. Just my own thoughts though!	0.73	0.5	Positive
07-05-2020	**CNN Philippines @cnnphilippines** There are now more than 400 health personnel who have recovered from the coronavirus disease, the Department of Health says.	0.10	1.0	Most Positive

Global Sentiment Trend Analysis

- Worldwide Average Sentiment Trend Shift Detection.



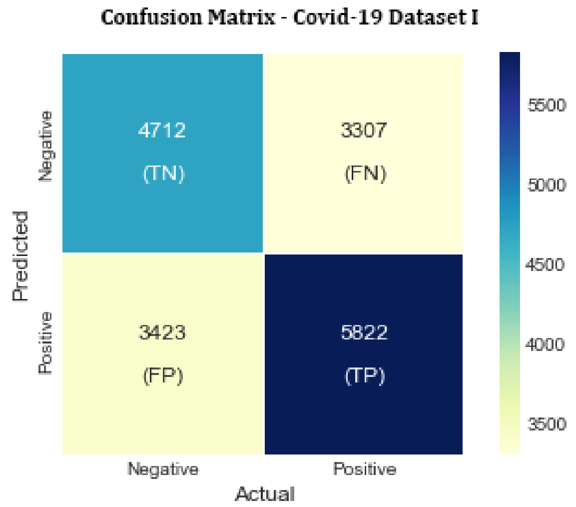


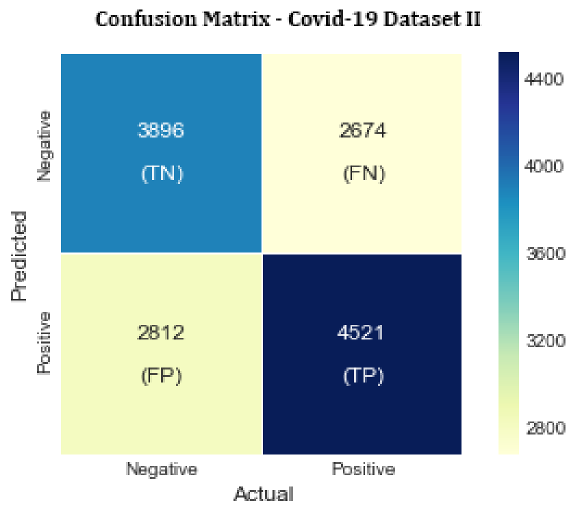
Classification Report & Confusion Matrix

- **Classification Report**

		Precision	Recall	F1-Score	Support
Covid-19 Dataset I (Apr – Jun, 2020)	*Positive (1.0)*	0.64	0.63	0.63	9245
	Negative (0.0)	0.58	0.59	0.58	8019
	Avg. / Total	**0.61**	**0.61**	**0.61**	**17264**
Covid-19 Dataset II (Aug – Oct, 2020)	*Positive (1.0)*	0.63	0.62	0.62	7333
	Negative (0.0)	0.58	0.59	0.59	6570
	Avg. / Total	**0.61**	**0.61**	**0.61**	**13903**

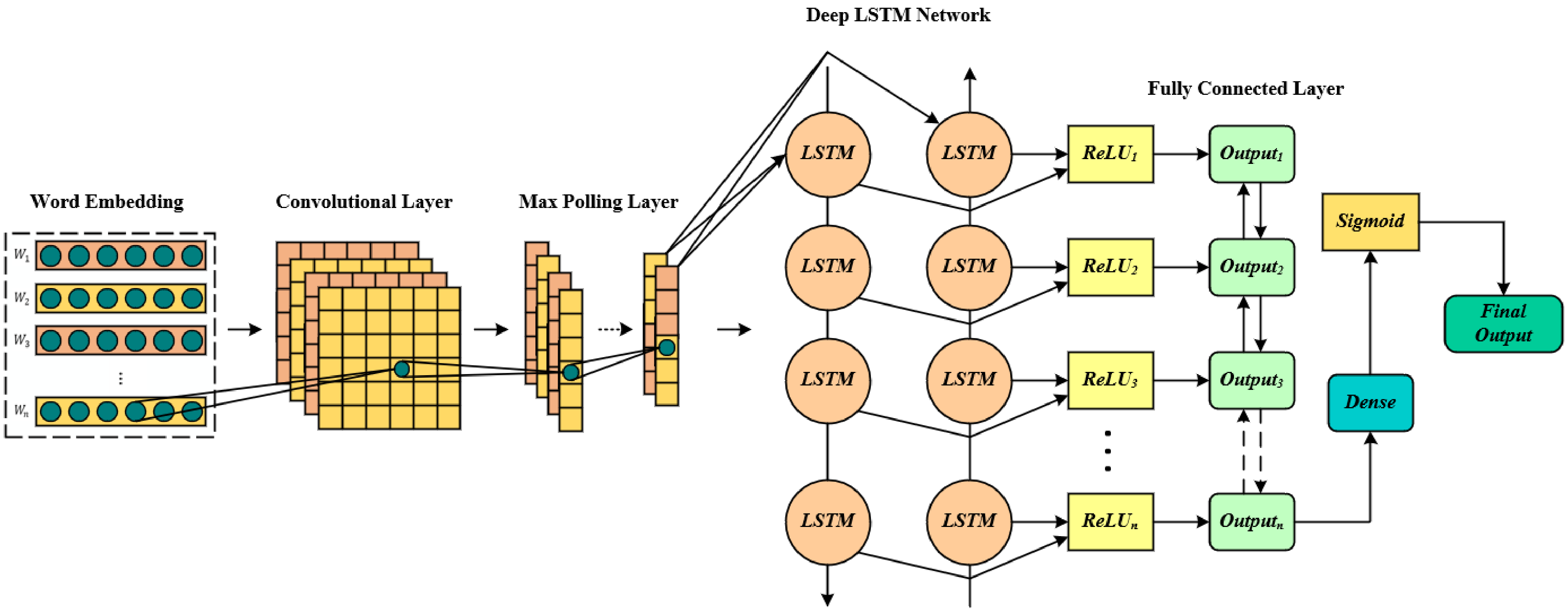
- **Confusion Matrix**





Sentiment Modeling using Hybrid Convolutional LSTM

- Proposed System Architecture of **Hybrid Convolutional LSTM model** for sentiment prediction of Covid-19 tweets.



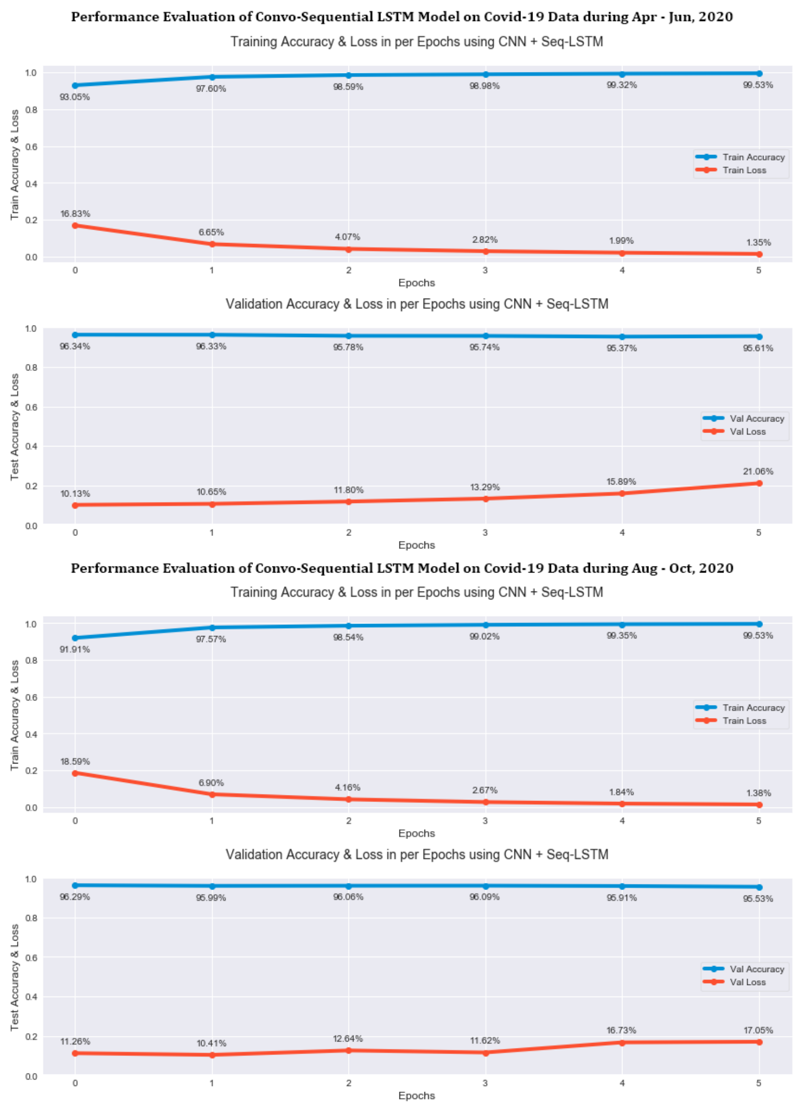
Sentiment Modeling using Hybrid Convolutional LSTM

- The new dataset contains all preprocessed tweets along with their corresponding ***positive*** (1.0) and ***negative*** (0.0) sentiments.
- We split the dataset into **80:20** ratio. i.e., **80%** for **training** and **20%** for **validation** purposes respectively.
- Word vectors are calculated for each Covid-19 exclusive words.
- ***TensorFlow*** framework and ***keras*** library are used to add ***Sequential*** and ***Bidirectional LSTM*** models with ***Embedding***, ***Convolutional***, ***Max Pooling*** and ***Dense*** layers.
- The **Convo-Sequential LSTM** and **Convo-Bidirectional LSTM** models trained for **6 epochs** on certain parameters**.**
- We have finally achieved **95.61%** and **95.81%** of **validation accuracy** for **first phase dataset** whereas on the **second phase dataset** these models obtained the validation accuracy as **95.53%** and **95.75%** respectively.

Sentiment Modeling using Hybrid CNN + Seq-LSTM

- Training accuracy vs. loss, validation accuracy vs. loss using **CNN + Seq-LSTM** network.

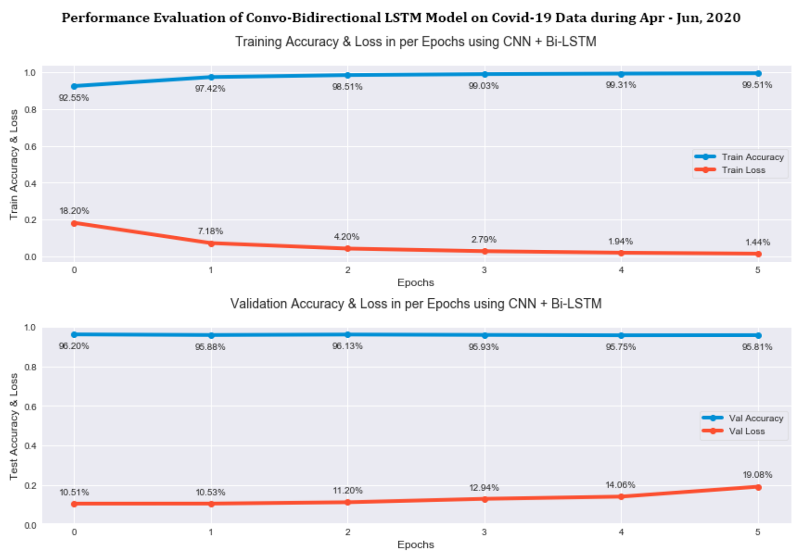
	Epochs	Train Loss	Train Accuracy	Val Loss	Val Accuracy
Covid-19 Dataset I (Apr – Jun, 2020)	*Initially*	16.83%	93.05%	10.13%	96.34%
	2nd	06.65%	97.60%	10.65%	96.33%
	3rd	04.07%	98.59%	11.80%	95.78%
	4th	02.82%	98.98%	13.29%	95.74%
	5th	01.99%	99.32%	15.89%	95.37%
	6th	01.35%	**99.53%**	21.06%	**95.61%**
Covid-19 Dataset II (Aug – Oct, 2020)	*Initially*	18.59%	91.91%	11.26%	96.29%
	2nd	06.90%	97.57%	10.41%	95.99%
	3rd	04.16%	98.54%	12.64%	96.06%
	4th	02.67%	99.02%	11.62%	96.09%
	5th	01.84%	99.35%	16.73%	95.91%
	6th	01.38%	**99.53%**	17.05%	**95.53%**

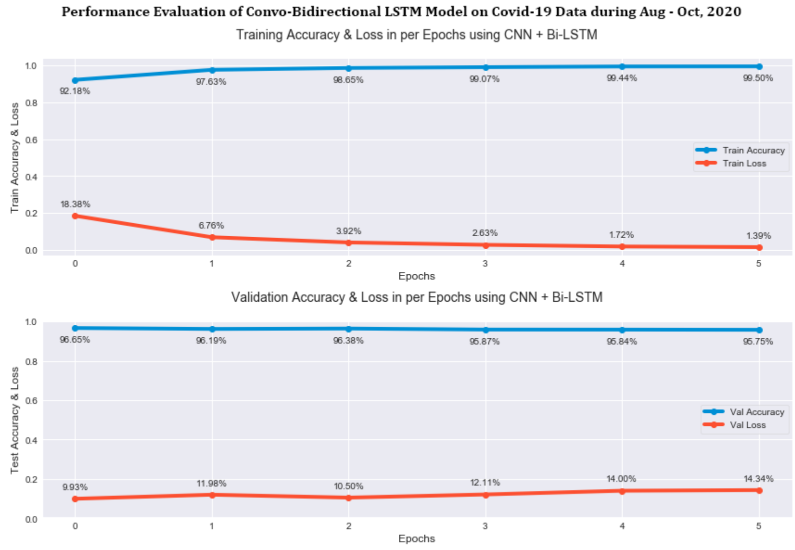


Sentiment Modeling using Hybrid CNN + Bi-LSTM

- Training accuracy vs. loss, validation accuracy vs. loss using **CNN + Bi-LSTM** network.

	Epochs	Train Loss	Train Accuracy	Val Loss	Val Accuracy
Covid-19 Dataset I (Apr – Jun, 2020)	*Initially*	18.20%	92.55%	10.51%	96.20%
	2^{nd}	07.18%	97.42%	10.53%	95.88%
	3^{rd}	04.20%	98.51%	11.20%	96.13%
	4^{th}	02.79%	99.03%	12.94%	95.93%
	5^{th}	01.94%	99.31%	14.06%	95.75%
	6^{th}	01.44%	**99.51%**	19.08%	**95.81%**
Covid-19 Dataset II (Aug – Oct, 2020)	*Initially*	18.38%	92.18%	09.93%	96.65%
	2^{nd}	06.76%	97.63%	11.98%	96.19%
	3^{rd}	03.92%	98.65%	10.50%	96.38%
	4^{th}	02.63%	99.07%	12.11%	95.87%
	5^{th}	01.72%	99.44%	14.00%	95.84%
	6^{th}	01.39%	**99.50%**	14.34%	**95.75%**





Comparative Performance Analysis I

- Benchmark Comparison of NLP based Experiments on **Covid-19**.

Machine Learning or Deep Neural Network Models	Prediction Accuracy
CNN + Bi-LSTM w. Naïve Bayes Sentiment Model (Ours)	**95.81%**
Multilingual BERT on single-level classification [19]	95.62%
CNN + Seq-LSTM w. Naïve Bayes Sentiment Model (Ours)	**95.61%**
CNN w. *GloVe* Embeddings [13]	90.67%
Bayesian Regression w. Tf-Idf [17]	89.40%
ERNIE on Chinese Weibo message [18]	88.00%
Bert on Chinese Weibo message [18]	83.00%
Random Forest [17]	82.20%
LDA + Deep LSTM [14]	81.15%
Logistic Regression w. trigrams + Tf-Idf [15]	81.00%
SVM w. Gaussian Membership based Fuzzy logic [15]	79.00%
LSTM on Chinese Weibo message [18]	78.00%
Deep RNN Model [16]	76.71%

Comparative Performance Analysis II

- Comparative Analysis of State-of-Art Experiments on **IMDB Dataset**.

Deep Neural Network Models	Prediction Accuracy
CNN + Bi-LSTM w. Naïve Bayes Sentiment Model (Ours)	**90.44%**
CNN + Seq-LSTM w. Naïve Bayes Sentiment Model (Ours)	**90.26%**
Ensemble LSTM + CNN [25]	90.00%
CNN + LSTM w. Combined Kernels [21]	89.50%
CNN [25]	89.30%
CNN - LSTM [23]	89.20%
LSTM [25]	89.00%
CNN + LSTM w. Vanilla or Multiword Pre-processing [24]	88.90%
CNN w. Multiword Pre-processing [24]	87.90%
CNN [23]	87.70%
MLP [23]	86.74%
Vanilla Neural Network [22]	86.67%
LSTM [23]	86.64%
LSTM w. Tuning and Dropout [26]	86.50%
SA-LSTM w. Joint Training [26]	85.30%
Recursive RNN [20]	83.88%

Comparative Performance Analysis III

- Comparisons with Benchmark Experiments on **Other Public Corpora**.

Public open-source corpora	**Previous Benchmark results**	**Our results**	
		CNN + Seq-LSTM	**CNN + Bi-LSTM**
Amazon customer review dataset	90.00% [27]	**99.91%**	**99.92%**
Stanford Sentiment Treebank (SST) dataset	86.99% [28]	**90.07%**	**90.25%**

Recent Achievement

- Regarding the progress of this research work so far I like to inform you that my paper has been accepted in the **First International Conference on Research and Applications in Artificial Intelligence (RAAI 2020)** organized by *Dept. of Information Technology, RCC Institute of Information Technology, A Unit of an Autonomous Society of Department of Higher Education. Govt. of West Bengal, India.*
- The paper is entitled as : **Chakraborty, A.K.**, Das, S. and Kolya, A.K., 2021. Sentiment Analysis of Covid-19 Tweets Using Evolutionary Classification-Based LSTM Model. In *Proceedings of Research and Applications in Artificial Intelligence* (pp. 75-86). Springer, Singapore. https://doi.org/10.1007/978-981-16-1543-6_7
- This paper has received **Best Paper Award** for the respective track in the **International Conference on Research and Applications in Artificial Intelligence (RAAI 2020).**

Conclusion and Future Scope

- For the future work, we will extract the tweets from the datasets for which the Convo-Bidirectional LSTM model performs better than Convo-Sequential LSTM model.

- Further we will use fine-grained classified tweets to train the hybrid deep learning models to ensure the versatility of my proposed system.

Acknowledgment

I would like to express humbly my absolute amenity and heartfelt gratitude to my respected guide ***Dr. Anup Kumar Kolya, Assistant Professor, Dept. of Computer Science and Engineering, RCCIIT*** *whose kind and valuable suggestions and excellent guidance enlightened to give me the best opportunity in preparing this Project Paper.*

I am also grateful to ***Mr. Rajib Saha, Assistant Professor and Head of the Department of Computer Science and Engineering, RCCIIT*** *for providing inspiration and comprehensible instructions in connection with this Paper.*

In accordance with that, I would like to express humbly my absolute amenity and reverence to respected ***Mr. Sourav Das, Assistant Professor, Dept. of Computer Science & Engineering, Future Institute of Technology*** *for guiding and mentoring me on and off throughout my thesis work.*

- Thank You

References

1. Tuli, S., Tuli, S., Tuli, R. and Gill, S.S., 2020. Predicting the Growth and Trend of COVID-19 Pandemic using Machine Learning and Cloud Computing. *Internet of Things*, p.100222.

2. Dubey, A.D., 2020. Twitter Sentiment Analysis during COVID19 Outbreak. *Available at SSRN 3572023*.

3. Das, S., Das, D. and Kolya, A.K., 2020. Sentiment classification with GST tweet data on LSTM based on polarity-popularity model. *Sadhana*, *45*(1).

5. Arora, P., Kumar, H. and Panigrahi, B.K., 2020. Prediction and analysis of COVID-19 positive cases using deep learning models: A descriptive case study of India. *Chaos, Solitons & Fractals*, *139*, p.110017.

6. Arpaci, I., Alshehabi, S., Al-Emran, M., Khasawneh, M., Mahariq, I., Abdeljawad, T. and Hassanien, A.E., 2020. Analysis of Twitter Data Using Evolutionary Clustering during the COVID-19 Pandemic. *CMC-COMPUTERS MATERIALS & CONTINUA*, *65*(1), pp.193-203.

7. Jurafsky, D., 2000. *Speech & language processing*. Pearson Education India.

8. Baziotis, C., Pelekis, N. and Doulkeridis, C., 2017, August. Datastories at semeval-2017 task 4: Deep lstm with attention for message-level and topic-based sentiment analysis. In *Proceedings of the 11th international workshop on semantic evaluation (SemEval-2017)* (pp. 747-754).

9. Nakisa, B., Rastgoo, M.N., Rakotonirainy, A., Maire, F. and Chandran, V., 2018. Long short term memory hyperparameter optimization for a neural network based emotion recognition framework. *IEEE Access*, *6*, pp.49325-49338.

10. Wang, J.H., Liu, T.W., Luo, X. and Wang, L., 2018, October. An LSTM approach to short text sentiment classification with word embeddings. In *Proceedings of the 30th conference on computational linguistics and speech processing (ROCLING 2018)* (pp. 214-223).

11. Samuel, J., Ali, G.G., Rahman, M., Esawi, E. and Samuel, Y., 2020. Covid-19 public sentiment insights and machine learning for tweets classification. *Information*, *11*(6), p.314.

12. Jelodar, H., Wang, Y., Orji, R. and Huang, S., 2020. Deep sentiment classification and topic discovery on novel coronavirus or covid-19 online discussions: Nlp using lstm recurrent neural network approach. *IEEE Journal of Biomedical and Health Informatics*, *24*(10), pp.2733-2742.

13. Das, S., Kolya, A.K.: Predicting the pandemic: sentiment evaluation and predictive analysis from large-scale tweets on Covid-19 by deep convolutional neural network. *Evolutionary Intelligence*. 1–22 (2021).

References

14. Jelodar, H., Wang, Y., Orji, R. and Huang, S., 2020. Deep sentiment classification and topic discovery on novel coronavirus or covid-19 online discussions: Nlp using lstm recurrent neural network approach. *IEEE Journal of Biomedical and Health Informatics*, *24*(10), pp.2733-2742.

15. Chakraborty, K., Bhatia, S., Bhattacharyya, S., Platos, J., Bag, R. and Hassanien, A.E., 2020. 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*, *97*, p.106754.

16. Al-Shaher, M.A., 2020. A hybrid deep learning and NLP based system to predict the spread of Covid-19 and unexpected side effects on people. *Periodicals of Engineering and Natural Sciences (PEN)*, *8*(4), pp.2232-2241.

17. Serrano, J.C.M., Papakyriakopoulos, O. and Hegelich, S., 2020, July. NLP-based feature extraction for the detection of COVID-19 misinformation videos on Youtube. In *Proceedings of the 1st Workshop on NLP for COVID-19 at ACL 2020*.

18. Yang, Q., Alamro, H., Albaradei, S., Salhi, A., Lv, X., Ma, C., Alshehri, M., Jaber, I., Tifratene, F., Wang, W. and Gojobori, T., 2020. Senwave: Monitoring the global sentiments under the covid-19 pandemic. *arXiv preprint arXiv:2006.10842*.

19. Li, I., Li, Y., Li, T., Alvarez-Napagao, S., Garcia-Gasulla, D. and Suzumura, T., 2020, December. What Are We Depressed About When We Talk About COVID-19: Mental Health Analysis on Tweets Using Natural Language Processing. In *International Conference on Innovative Techniques and Applications of Artificial Intelligence* (pp. 358-370). Springer, Cham.

20. Timmaraju, A. and Khanna, V., 2015. Sentiment analysis on movie reviews using recursive and recurrent neural network architectures. *Semantic Scholar*, pp.1-5.

21. Yenter, A. and Verma, A., 2017, October. Deep CNN-LSTM with combined kernels from multiple branches for IMDb review sentiment analysis. In *2017 IEEE 8th Annual Ubiquitous Computing, Electronics and Mobile Communication Conference (UEMCON)* (pp. 540-546). IEEE.

22. Shaukat, Z., Zulfiqar, A.A., Xiao, C., Azeem, M. and Mahmood, T., 2020. Sentiment analysis on IMDB using lexicon and neural networks. *SN Applied Sciences*, *2*(2), pp.1-10.

23. Ali, N.M., Abd El Hamid, M.M. and Youssif, A., 2019. Sentiment analysis for movies reviews dataset using deep learning models. *International Journal of Data Mining & Knowledge Management Process (IJDKP) Vol*, *9*.

24. Camacho-Collados, J. and Pilehvar, M.T., 2017. On the role of text preprocessing in neural network architectures: An evaluation study on text categorization and sentiment analysis. *arXiv preprint arXiv:1707.01780*.

References

25. Minaee, S., Azimi, E. and Abdolrashidi, A., 2019. Deep-sentiment: Sentiment analysis using ensemble of cnn and bi-lstm models. *arXiv preprint arXiv:1904.04206*.

26. Dai, A.M. and Le, Q.V., 2015. Semi-supervised sequence learning. *arXiv preprint arXiv:1511.01432*.

27. Guner, L., Coyne, E. and Smit, J., 2019. Sentiment analysis for Amazon. com reviews.

28. Dong, X.L. and De Melo, G., 2018, July. A helping hand: Transfer learning for deep sentiment analysis. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 2524-2534).