Covid-19 Sentiment Analysis using Bidirectional Encoder Representations from Transformers

Praveen [#1], Basavesha D[*2], Dr. Piyush Kumar Pareek[*3]

[#]*PG Student, CSE Department , Sridevi institute of technology and management, Visveswaraya Technological University*

** Assistant Professor, CSE Department, Sridevi institute of technology and management, Visveswaraya Technological University*

[*]*Professor, Department of Computer Science & Engineering, East West Institute of Technology, Bangalore*

praveenkatti16@gmail.com, basavesha@gmail.com, Piyushkumarpareek88@gmail.com

Abstract - Corona coronavirus (COVID-19) is a progressive pandemic that is being recognized worldwide. However, spreading false news on social media platforms such as Twitter creates unnecessary concern about the disease. The motto of this study analyzes tweets by Indian netizens during the closure of COVID-19. The data included tweets collected between 23 March 2020 and 15 July 2020 and the text was written as fear, sadness, anger and happiness. Data analysis was performed by the Bidirectional Encoder Representations from Transformers (BERT) model, which is a new in-depth study model for text analysis and performance and was compared with three other models such as logistic regression (LR), vector support (SVM). The accuracy of all the words was calculated separately. The BERT model produced 86% accuracy. Our findings point to a significant increase in keywords and related names among Indian tweets during the COVID-19 era. In addition, this work clarifies public opinion on epidemics and leads public health authorities to a better society.

Keywords — Covid-19, Sentiment Analysis, BERT, Deep Learning.

I. INTRODUCTION

The COVID-19 can be a deadly illness currently days and conjointly the folks are plagued by this and much of people last their lives World Health Organization declares it as common illness. The primary case of COVID-19 in Asian nation, that derive against China, move elaborate on thirty January 2020 were made public among the state of Kerala. To manage this such an outsized quantity of measures are taken by government like Lockdowns are declared among the country on twenty five March. A assist flap starting in March 2021 move abundant larger than the first, with varied issues are Janus-faced by Medical hospitals like scarcity about dose, cot, oxygen gas barrel and alternative medicines in elements of the nation. Asian nation began its vaccination program on sixteen January 2021[1].

India has licensed people Oxford– AstraZeneca vaccinium (Covishield), the Indian BBV152 (Covaxin) vaccinium, thus the Russian satellite V vaccinium for emergency use. As World Health Organization Director-Generic Tedros Adhanom Ghebreyesus announced found in the urban center Security Conference on fifteen February 2020, "We're not simply determined AN endemic; we're determined AN information emic ". It's even been claimed that the unfold of COVID-19 is supported by information. However, information does not solely contribute to the spread: information would possibly bolster concern, drive social disagree, and can even lead to direct injury[2].

This paper introduced BERT model to classify the COVID tweets. We considered the massive dataset of coronavirus tweets by Kaggle.com. With the intention to totally make the foremost this information inside the facts. It far necessary to recommend a possible and affordable category technique for the textual content of the people's livelihood hotline.

II. RELATED WORK

This sentimental analysis process by using the technique called NLP to predict the people opinion,emption it is also called as opinion mining. By consodering the recent history ,the researches analyse that type of data and classify the emotions into varient. In literature, multiple ways to measure on the market to resist sentiment analysis that involves extracting varbal sentiments with the records (Kim and Hovy 2006). However, options square measure related with the sentiment victimization metal and trigrams. As a result of the emotions square measure currently a typical because of express feelings, so emojis square measure usually used for negative, positive, and neutral thoughts. This exhibits the among operating of system by using victimization anybody of the prevailing ways to carryout Sentiment Analysis. This above image that exhibits the strategy to categorize the text or wordings among completely different sentiment teams like positive, negative, and neutral[3][4].

These days, the increase of Coronavirus is modified social and individual lives all over the universe [5]. That's why, several experimenters square measure operating to notice the emotions towards novel coronavirus from a specific point of view, and depiction their result or observations in various ways that victimization on the market tools and techniques. The sequence of this work, have developed the outburst of COVID-19 to look at the emotions by victimization varied Machine Learning techniques(Muthusamy et al. 2020Pastor et. al.)have raked the tweets from the Philippines region and study the output of the symptom's of the covid on community quarantine. The centre purpose to work is that the covid and its implications by exploring the text of social media. Therefore, the authors have given the sentiment analysis on the twitter topic and specify the case in Asian nation as among the rest of the world.

III. BERT Model

BERT by Google AI is one in each of the foremost standard language illustration models. many organizations, together with Facebook more as domain, are researching NLP victimization this electrical device model. that is that the foremost recent deep-learning model for the analysis of matter knowledge.it is meant to pre-train deep two-way representations from the untagged text by together acquisition on each left and right context all told layers. This makes use of electrical device, Associate in Nursing attention mechanism that learns discourse relations between words (or sub-words) throughout a text. electrical device includes 2 separate mechanisms .

BERT's aim is to come up with a model, solely the encoder mechanism is significant. As opposition directional models, that scan the input consecutive, the electrical device encoder reads the complete sequence of words right away. Therefore, it's thought of two-way, though it'd be additional correct to mention that it's non-directional. The supply content is additionally single text sentence or a combine of sentence Associate in Nursing need to be picturized in an extremely linear sequence and for every word inside the input text, the input illustration is performed by 3 components of embedding summed up. The illustration of embedding is shown in Figure1.

BERT's key specialized advancement is applying the bidirectional preparing of Transformer, a mainstream consideration model, to language demonstrating. This is an advanced form, past advancements that took a look at text sequencing either from left to right or right to left and combined the left to the right model.

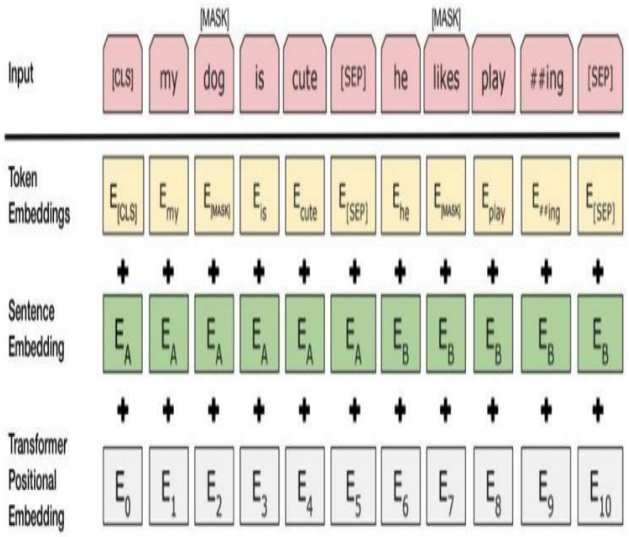

Fig 1 BERT input model

- Token embedding is the word vector, the primary text is in the CLS flag, These area unit the embeddings learned for the precise token from the Word Piece token.
- Phase embeddings is employed to tell apart between 2 sentences, as a result of pre-training does classification jobs with it takes multiple sentence as input.
- Position embedding that tells us about position information, that is get through by model learning

Bert Classifier:

The steps for classifying texts exploitation the Bert model are :

Alter the codes: First of all, code must adjusted consistent with the info set. The changes are: Alter the processor technique and also the processor wordbook.

fine-tuning: We are going to set the acceptable parameters, making ready the text information, we are be able to run the Bert model to finetune the pre-trained model.

Prediction: Tuning the corresponding parameters and exploitation the take a look at information and also the model when finetuning to forecast the classes of the texts.

Accuracy calculation: The shrewd a general classification of the accuracy rate. From the analysis during this particular paper is focus toward multiple classification, it is going to necessary to live the classification of result of the complete text data. That's why the quantitative relation of the amount of genuinely classified texts to the full variety of texts is employed because the analysis guide of the general classification result.

IV. METHODOLOGY

The BERT model is instructed on a collection of 160 Million tweets related to the COVID-19 gathered through before training, the first corpus was about to retweet tags. Each tweet was pseudonymized by restoring all Twitter user names with a unique text Token. an analogous procedure was carryout on all Uniform Resource Locaters(URL) to the sites. We also replaced all Unicode emoticons with textual ASCII representations (e.g.: smile: for,) using the Python emoji library 3. In the last, all re tweets, duplicates and shut duplicates are away from the given dataset, leading to a final outcome corpus of twenty-two.5Million tweets that contain a complete of 0.6B words. The domain specific pretraining dataset consists of 1/7th the scale of what's used for training the most base model .

Tweets were treated as single files and separated into sentences using the spacy library thereby increasing the training batch sizes to 1024 examples. We use a dupe factor of 10 on the dataset, leading to 285M training examples and a couple of.5M validation examples. a relentless learning rate of 2e-5, as recommended on the official BERT GitHub4 when doing domain-specific pretraining. Loss and precision was evaluated through the pretraining procedure.. Distributed training was performed using TensorFlow.

To evaluate the efficiency of the BERT model, another three conventional models, such as LR, SVM, and LSTM, were further considered. LR is a kind of binary classification ML algorithm. It uses the weighted combination of the input text features and is authorized by the sigmoid function. This function changes any real numerical input that ranges from 0 to 1. Instead of calculating the total number of words in a text document, we used Tf -Idf (term frequency–inverse document frequency) for normalizing the words into a number. It estimates normalized count, such as the count of each word that is divided by a total number of documents, where the same word appears. We applied a similar ratio of dataset distribution (85:15) as followed in BERT modelling. SVM is another ML classification algorithm based on the identification of hyperplanes defined by data classes. It operates on large data sizes and calculates the separation of data margin rather than matching key features.

The similar approach of 85% of tweets is used for SVM model training. Alternatively, we used the LSTM model, which is a type of recurrent neural network (RNN). LSTM network models are a sort of intermittent neural organization that can learn and recollect large information groups. They are expected for use with information that is contained in the long sequence of input data, up to 200 to 400-time steps. That is why we considered LSTM as one of the accepted models for text analysis. The model can uphold various parallel information sequences and figures out how to remove features from data sequences. Like BERT, this model also trained by AdamW optimizer

Classification Methods:

Further research has used a variety of texting techniques to test communication sentiments. These dividers are grouped into several categories according to their similarities. The next section discusses details about the four key categories we have reviewed, including line order and the closest neighbor to K, and focuses on the two divisions we have chosen to compare, namely the Naïve Bayes and the regression of assets, their main concepts, strengths and weaknesses. The focus of this study is to present a machine-based perspective on the performance of the widely used Naïve Bayes and asset disposal methods.

Linear Regression Model

Although linear regression is used to predict relationships between continuous variables, line spacing can be used to separate text and text. A standard measurement method using line division is a square algorithm that reduces objective function (e.g., double the difference between predicted outcomes and factual categories). The square algorithm at least resembles the maximum probability of the result when the variability of the result is influenced by the Gaussian sound [8]. The Linear ridge regression classifier improves the installation work by adding a penaltyizer to it. The Ridge separator converts binary results into −1, 1 and treats the problem as deferred (multi-class deferment for multi-stage problem.

Naïve Bayes Divide

The Naïve Bayes (NBC) is a proven, simple and effective method of text separation [5]. It has been widely used to classify texts since the 1950s [6]. This phase of doctrinal division is based on the concept of the Bayes theorem [7]. A discussion on NBC's statistical structure from the point of view of textual analytics is provided under the methods section. NBC uses a limited posteriori rating to find a category (i.e., features are assigned to a category based on the highest conditional possibilities). There are only two types of NBC: Multinomial Naïve Bayes [9] (i.e., binary representation of elements) and Bernoulli Naïve Bayes (i.e., features are represented by frequency) [10]. Many studies have used NBCs to find text, documents and product classification.

Logistic Regression

Logistics (LR) editing is one of the most popular and original methods of editing. LR was first developed by David Cox in 1958 [11]. In the LR model, the possibilities for describing the possible results of a single trial are measured using a systematic task [12]. Using a structured function, the probability of results is converted into binary values (0 and 1). Limitations of great opportunities methods are commonly used to minimize error in the model. A comparative study

classifying product reviews reported that logistic regression multi-class classification method has the highest accuracy compared to Naïve Bayes, Random Forest, Decision Tree, and Support Vector Machines classification methods.[13][14]

V. RESULTS

In this section we have described the results obtained fro our experiment. Fig 2 shows the Word Cloud of Frequently Tweeted Words and Fig 3 descibes the unigram representation of tweets and Fig 4 shows the BERT model accuracy of 86.83 %.

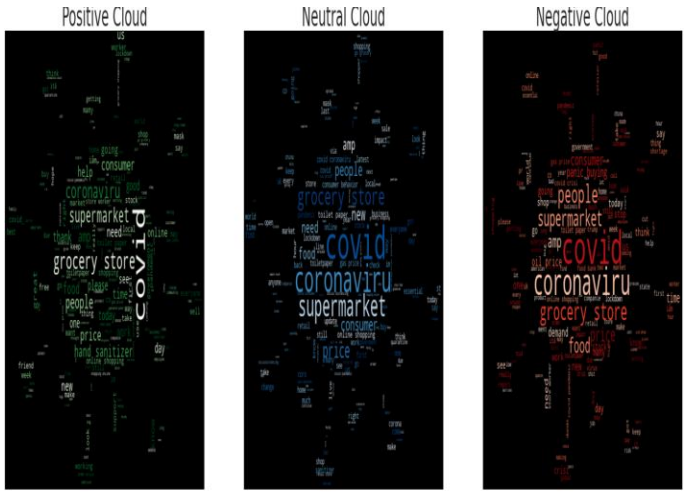

Fig 2: Word Cloud of Frequently Tweeted Words

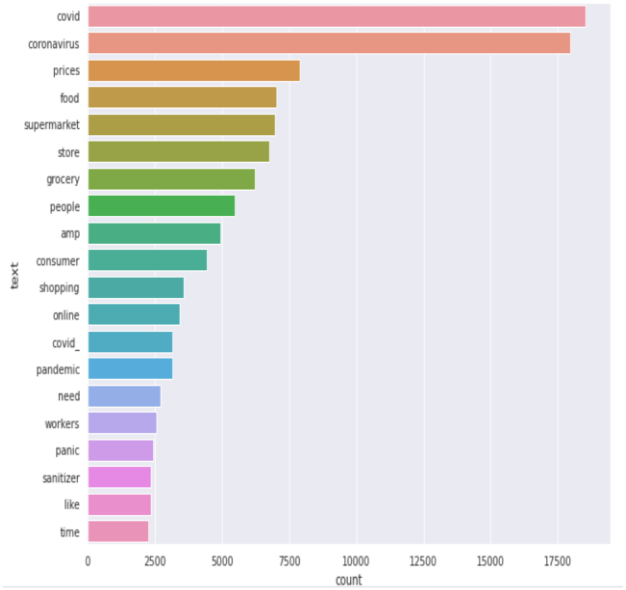

Fig3: Unigram of Covid Tweets

```
Accuracy Score =  0.8683187560738581
```

```
print(classification_report(true_category, predicted_category))
```

	precision	recall	f1-score	support
-1	0.86	0.89	0.87	3084
0	0.89	0.75	0.81	1553
1	0.87	0.90	0.89	3595
accuracy			0.87	8232
macro avg	0.87	0.85	0.86	8232
weighted avg	0.87	0.87	0.87	8232

Fig4: BERT Model Accuracy

VI. CONCLUSIONS

In conclusion, our findings point to a significant increase in keywords and related names among Indian tweets during the COVID-19 era. In terms of content, tweets are divided into four emotions such as fear, sadness, anger and happiness. Some words like “trump”, “kill”, “death”, “die” make people unnecessarily fearful and words like “thank”, “well”, “good” create a good setting for health authorities. These findings encourage local governments to force truth inspectors on social media to overcome false propaganda. Earlier posts focus only on social media outcomes and their implications for non-news broadcasts but do not discuss the validity and segregation of tweets. Therefore, we used an in-depth learning model called BERT to achieve high differentiated accuracy unlike conventional ML models. The results provided sufficient evidence that the BERT model achieved 86% accuracy, beating other models such as LR, SVM, and LSTM. In this way, the project clarifies public opinion on epidemics and guides medical authorities, the public, and the private sector to overcome unnecessary anxiety among epidemics.

REFERENCES

[1] Rashmi, T. V . "Load Balancing As A Service In Openstack-Liberty." International Journal of Scientific & Technology Research 4.8 (2015): 70-73

[2] Khanum, Salma,. "Meta Heuristic Approach for Task Scheduling In Cloud Datacenter for Optimum Performance." International Journal of Advanced Research in Computer Engineering & Technology (IJARCET) Volume 4

[3] Mertens, G.; Gerritsen, L.; Duijndam, S.; Salemink, E.; Engelhard, I.M. Fear of the coronavirus (COVID-19): Predictors in an online study conducted in March 2020. J. Anxiety Disord. 2020, 74, 102258.

[4] Mittal, M.; Battineni, G.; Goyal, L.M.; Chhetri, B.; Oberoi, S.V.; Chintalapudi, N.; Amenta, F. Cloud-based framework to mitigate the impact of COVID-19 on seafarers' mental health. Int. Marit. Health 2020, 71, 213–214.

[5] Pranav T P, Charan S, Darshan M R. (2021). Devops Methods for Automation of Server Management using Ansible . International Journal of Advanced Scientific Innovation, 1(2), 7-13. https://doi.org/10.5281/zenodo.4782271.

[6] Prajwal, S., M. Siddhartha, and S. Charan. "DDos Detection and Mitigation SDN using support vector machine." International Journal of Advanced Scientific Innovation 1.2 (2021): 26-31.

[7] Sahana, D. S., and L. Girish. "Automatic drug reaction detection using sentimental analysis." *International Journal of Advanced Research in Computer Engineering & Technology (IJARCET) Volume* 4 (2015).

[8] Hung, M.; Lauren, E.; Hon, E.S.; Birmingham, W.C.; Xu, J.; Su, S.; Hon, S.D.; Park, J.; Dang, P.; Lipsky, M.S. Social Network Analysis of COVID-19 Sentiments: Application of Artificial Intelligence. J. Med. Internet Res. 2020, 22, e22590

[9] Devlin, J.; Chang, M.W.; Lee, K.; Toutanova, K. BERT: Pre-training of deep bidirectional transformers for language understanding. arXiv 2018, arXiv:1810.04805.

[10] Chang, Y.W.; Hsieh, C.J.; Chang, K.W.; Ringgaard, M.; Lin, C.J. Training and testing low-degree polynomial data mappings via linear SVM. J. Mach. Learn. Res. 2010, 11, 1471–1490.

[11] Samuel, J.; Ali, G.; Rahman, M.; Esawi, E.; Samuel, Y. COVID-19 Public Sentiment Insights and Machine Learning for Tweets Classification. Information 2020, 11, 314

[12] Abd-Alrazaq, A.; Alhuwail, D.; Househ, M.; Hamdi, M.; Shah, Z. Top Concerns of Tweeters During the COVID-19 Pandemic: Infoveillance Study. J. Med. Internet Res. 2020, 22, e19016.

[13] D, BASAVESHA and NIJAGUNARYA, Y S, Detecting Duplicate Questions in Community Based Websites Using Machine Learning (April 27, 2021). Proceedings of the International Conference on Innovative Computing & Communication (ICICC) 2021, Available at SSRN: https://ssrn.com/abstract=3835083 or http://dx.doi.org/10.2139/ssrn.3835083

[14] Basavesha D, Dr. Y S Nijagunarya. *"Soft Computing based Duplicate Text Identification in Online Community Websites"* International Journal of Engineering Trends and Technology 68.7(2020):1-7.