IJRAR.ORG E-ISSN: 2348-1269, P-ISSN: 2349-5138

INTERNATIONAL JOURNAL OF RESEARCH AND ANALYTICAL REVIEWS (IJRAR) | IJRAR.ORG

An International Open Access, Peer-reviewed, Refereed Journal

An Optimized Artificial Intelligence-enabled Sentiment Analysis on the COVID-19 Pandemic

Shyam Prakash Mishra
AIT- CSE
Chandigarh University,Punjab, India

Mohit Choudhary
AIT- CSE
Chandigarh University,Punjab, India

Bharat Yadav
AIT- CSE
Chandigarh University,Punjab, India

Siddharth Kumar
Supervisor
Chandigarh, University

Ayush Jha
AIT- CSE
Chandigarh University,Punjab, India

Abstract- COVID-19 is an infectious disease and its first recorded cases was identified in March of 2020 it was declared as a pandemic. The outbreak of this disease has led to a tremendous increase in posts and comments from social media users. This paper discusses the of sentiment analysis of COVID-19 on the basis of tweets from social media, in this paper we have focused on the classification of users' sentiment from posts related to COVID-19 that are posted on Twitter. The time period taken to examined is from June to August of 2020, when pandemic mostly affected the whole world. The data used and processed are deeply analyzed with the help of several natural language processing techniques. In this Sentiment analysis we implemented many different deep learning models based on LSTM neural networks, and also made comparison with simple machine learning mode. The models are trained in order to distinguish the tweets in three main classes, namely negative, neutral and positive.

Keywords- COVID-19, Sentiment analysis, Twitter, Machine Learning, NLP

I. Introduction

Internet growth is tremendously increasing and affecting every action of our lives. This growth is continuously increasing day by day due to the massive growth in volume of data and information. Most of these data and information is generate through human interaction at social media platforms like Twitter, Facebook and LinkedIn. Among this most popular social media platform is Twitter, that provides all types of information and allows its users to create and post messages and media called "tweets."

COVID-19 pandemic was started in December 2019. The virus was firstly found in the Wuhan region of China and affected 221 countries and territories around the world. This virus mainly affects the respiratory system of our body, although other organ systems are also affected. Symptoms of covid-19 is that lower respiratory tract infections, fever, dry cough and shortness of breath.

In this study we present a number of deep learning models whose work is to categorize the sentiment of people written in the posts of Twitter. The words that models are utilizing to distinguish whether the sentiment is negative, neutral or positive, and the meaning of these tweets analyze the COVID-19 pandemic. The dataset we have used in the paper is the period of time between July and August 2020, at that time disease had already affected the

many parts of world and the cases were being constantly increasing day by day. The number of tweets was 41, 157 and that was continuously increasing every second of time and it was categorized on the basis of people sentiment. For knowing the sentiment of the tweets, we used different deep learning models comparing with simple machine learning models.

In this paper we have introduced different deep learning and machine learning model that used information from social media platforms to understand the public sentiment during the COVID-19 pandemic. The trained model we have used to compares the different types of sentiments expressed by the people through word tokenizer during

Fig 1: Covid-19 Testing

the lock downs. We have used the LSTM and machine learning models to compare the results from different types of classification algorithms. These models are mainly focused on multi-label sentiment classification, consisting of three main different classes and 2 sub classes namely negative, extremely negative, neutral, positive and extremely positive.

Fig 2: Covid-19 Detecting Machine

Two pre-trained BERT models in Italian and 7 multi-language pre-trained models were used, to make the deep learning model capable for sentiment analysis on COVID-19 pandemic. In all models, deep learning algorithms are used, which is directly connected to the end of BERT and also connected to a Softmax activation function. We have also added linear layer which has 421 nodes and an output of two nodes, first one for the classification of category and second one for the text processing training for sentiment analysis.

The Comparison between the performance of the BERT models in Italian with machine learning model such as SVM, Decision Trees, Logistic Regression, and Naive Bayes models with same data set and parameters on COVID-19 tweets sentiment analysis. In the proposed work, we have used deep learning-based models, Bidirectional long/short-term

memory (Bi-LSTM) and Valence Aware Dictionary for Sentiment Reasoning (VEDAR), to recoganise the public sentiments related to COVID-19. Our model focuses on analysing the public emotions based on tweets posted by people from India. Our study analysed the sentiment of people whether it is positive, neutral, or negative value called polarity. It also provides insights of emotions such as happiness, anger, neutrality, etc. The proposed methodology described in this paper can be helpful for Indian government to take proper decisions by knowing the public's emotions about COVID-19 pandemic and vaccination.

This analysis can help government to analyse the sentiments of people and help them through following ways:

(i.) Understand the issues and demands asked by people related to COVID-19 pandemic and vaccination;

(ii.) To Ensure sufficient facility are managed;

(iii.) Understand the actual vaccination count and problem faced during vaccination;

(iv.) Take best suitable initiatives to create awareness about the present situation.

In this study we have used a large dataset of around 89,000 tweets from India related to COVID-19 and vaccination was used.

The unique contributions of this study are provided below :

1.The methodology described in this paper can be helpful for Indian government to take proper decisions by knowing the public's emotions about COVID-19 pandemic and vaccination.

2. The main purpose of this research is to prove the significance of sentiment analysis

using deep learning and machine learning methods.

3. The size of tweets datasets used by our team is large and accurate compared with other team studies.

4. The accuracy of our model is higher than the previous work proposed.

II. Literature Review

From the time COVID-19 was declared as a pandemic, many researchers and students have investigated and analysed the sentiment analysis of COVID-19 from people posts extracted from social media. Most of the perspective of sentiment analysis include:

1. Trend analysis of different time intervals using COVID-19 datasets and pre drained models.

2. Word modelling.

3. Sentiment analysis on social awareness, vaccination

4. Disease surveillance.

Twitter is a one of the most used social media platform for conveying anyone's opinions. The recent approaches used for Twitter sentiment analysis based on deep learning and machine learning such as:

machine learning techniques, deep learning, and hybrid methods. In the past few years there have been many research studies done to analyze sentiments of the tweets with the above approaches. Time series analysis was proposed out on these data. The authors evaluated sentiment scores for vaccine, lockdown and mask. The score were compared with the sentiment scores of the tweets posted from the people from India. This analysis has limitations in interpreting the meaning of neutral sentiments.

LSTM neural networks has been widely used for analysing COVID-19 infection in multiple countries especially in the period of lockdown. The posts were classified into 3 categories (negative, neutral and positive) with the help of a fine-tuned unsupervised BERT model and a Tf-Idf model for topic post verification. Sentiment classification was, also done by the authors of above research. Mainly, negative and positive sentiment classification was done with the help of SVM, Naive Bayes and logistic regression machine learning techniques. In this research paper we have compared the results of these classifiers on the basis of given datasets that consisted of tweets of shorter and longer length, the LSTM technique being the most accurate in both. Another approach that deals with sentiment classification is supervised machine learning technique. The authors examined how the covid-19 in August of 2020 affected people of India based on a dataset consisting of 87,000 tweets, which were extracted from twitter using prominent hashtags. The results showed that less negative sentiment that appeared but positive tweets were in the dominant ones.

Furthermore, the we used deep neural networks for better result, where we utilized various models and identified the best implementation among several models. Mainly, the classification task was improved as the deep learning method reduced the execution time by values ranging from 45 to 63%. Hence, the accuracy and speed of LSTM neural network and its valuable contribution for specific tasks was demonstrated.

III. Methodology

We have used a dataset that contains tweets related to COVID-19. We preprocessed the data using Natural Language Processing (NLP) techniques, such as tokenization, stemming, and stop words removal. We then used various machine learning algorithms, including Naive Bayes, Random Forest, and Logistic Regression, to classify the sentiment of the tweets as positive, negative, or neutral.

Text Preprocessing

The text preprocessing phase has of many steps as we aim to minimize the complexity of our proposed model. Mainly, all characters are converted to lowercase, and the hyperlinks are removed as they do not add any useful linguistic information. Hence we used, Part-of-speech (POS) tagging in our model to extract the useful features and enhance the performance of deep learning model. In this, each token captured a special tag, called the POS Tag. This tag shows the part of speech (noun, verb, adjective, etc.) to which the word belongs and it also contains additional information about its grammatical part.

Tokenization and lemmatization were also used in the process of dividing the text into smaller parts called “tokens”. Each term of every tweet is stored within a token list, and the text’s tokens appear based on their base order. This lemmatization is implemented by converting bigger words into a root word that exists in the vocabulary.

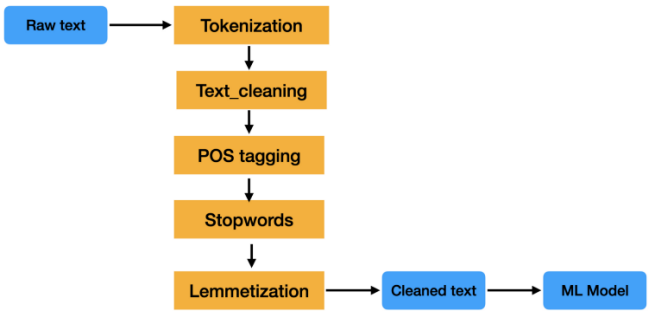


Fig 3: Post tagging

In this research different deep learning models were trained and tested consisting of long short-term memory (LSTM/BiLSTM) neural networks with the help of the Tensorflow and Keras libraries.

1.Simple LSTM is the first layer of the classifier is the Keras embedding layer, which is mostly used for text data. In this the input data must be encoded to integers so that each word can be represented by a single integer. After this, the embedding layer is assigned with random weights, so that during the training, a vector representation of each word from dataset is created. This process is done to prevent the problem of overfitting.

2.GloVe LSTM is a pre-trained embeddings in which 3 LSTM and 3 Dropout layers are used. The first layer, as in the previous part, it is the embedding layer. However the difference is that an already pre-trained vector representation of the words is developed, as a result of which present knowledge is transferred to the proposed model. After applying this, there is no need of embedding layer training with this number of learning parameters is dramatically reduced.

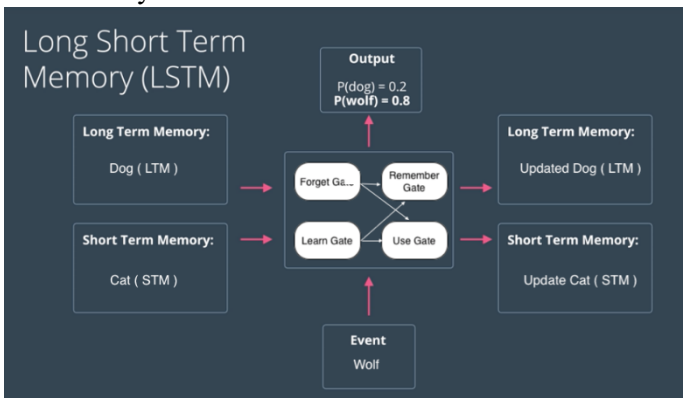


Fig 4: Long Short Term Memory (LSTM)

3. Bert Tokenizer LSTM is the tokens that the embedding layer get as input which are based on the Bert Tokenizer, called WordPiece Tokenizer. This Tokenizer was chosen separately to test due to the innovative way in which it splits words into tokens and we aim to evaluating its behavior throughout our whole data.

4.Text & Numerical Data LSTM is the model which receives as input, in addition to the text data with a set of numerical data. The numerical data is related to the certain features. The text data is transferred to an embedding layer, whose output is inserted into BiLSTM model and then in the Dropout layer. The numerical data, after getting normalized make the input of a dense layer, whose output is get added with the output of the Dropout layer of the text data. The output of the added layer forms the input for the final dense layer.

Parameter	Value
Batch size	16–64 (depending on the classifier)
Embedding size	200
Learning rate	0.00001–0.001 (depending on the classifier)
Optimizer	Adam
Loss function	Categorical cross-entropy
Evaluation metric	Accuracy

Every deep learning model receives a input which is preprocessed textual data and makes use of an embedding layer that changes each word into a vector representation of size equal to 210. In the present work, different kinds of embedding layers and models are tested. All models, except for Bert LSTM, contain a SpatialDropout1D layer which is used for avoiding overfitting during the training. In this layer, the output of the embedding layer is taken as input and the output of SpatialDropout1D is put into an LSTM or BiLSTM, depending on the model and situations. After every LSTM or BiLSTM layer, a Batch Normalization or a Dropout layer is used where these layers are used for avoiding overfitting.

1.Tf-Idf & Multinomial Naive Bayes: The operation of this classifier is depend on the Tf-Idf technique and it calculate how relevant a word is in a document in terms of a document collection. The Tf-idf Vectorizer function of the sklearn library is used to implement the Tf-Idf technique. By generating the matrix that uses the corresponding Tf-Idf value to each word in every tweet, and then the data were used into a Multinomial Naive Bayes model. This algorithm was used due to the fact that it is mostly used in NLP problems.

2.Tf-Idf & Decision Tree : In this model, the same Tf-Idf technique was used, where it produced matrix that was inserted into a Decision Tree classifier .

3.BoW & Multinomial Naive Bayes: This classifier take the conversion of words as a numbers with the bag of words (BoW) technique. It is a widely used NLP algorithm, which is depend on the frequency of words in the sentence. The difference between Tf-Idf is that it just generate a set of vectors for calculating the count of word occurrences in the document, while the Tf-Idf model store the information whether it is more or less important words. At last, the data were inserted into the Multinomial Naive Bayes classifier.

4.BoW & Random Forest: The last classifier which is implemented on the one basis of Random Forest algorithm in combination with the BoW technique.

Evaluation and Results:

The evaluation of the uesd models has been done with the help of the Kaggle dataset, and the models were computed on the basis of their ability to classify tweets sentiment effectively. The performance and efficiency of the models was measured in terms of accuracy, which is one of the most widely used metrics for the measuring of a system performance. Moreover, precision, recall and F1_score were, also used to present the following equations. Due to the fact that the problem we are facing is related to multi-class classification, for the determining the values of the evaluation indices such as accuracy, precision, recall and F1_score the one vs. all approach was taken.

$$\text{accuracy} \stackrel{\text{def}}{=} \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}}.$$

$$\text{precision} \stackrel{\text{def}}{=} \frac{\text{TP}}{\text{TP} + \text{FP}}.$$

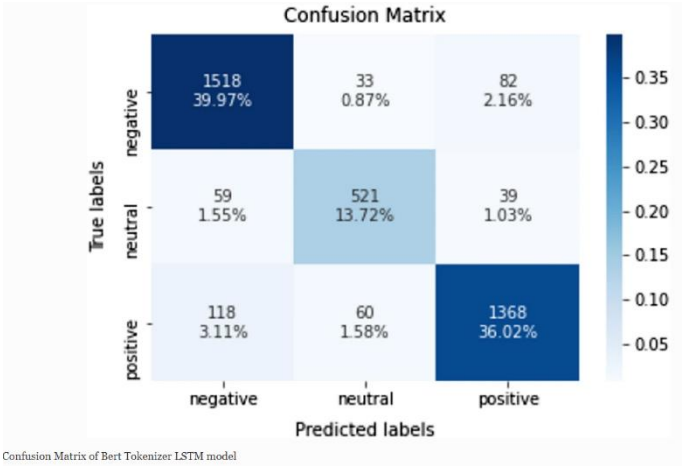
$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

$$F1 = 2 * \frac{\text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}} = \frac{\text{TP}}{\text{TP} + \frac{1}{2}(\text{FP} + \text{FN})}$$

The accuracy, precision, recall and F1 score of the trained models are provided above. In our research Bert Tokenizer LSTM achieved the best performance based on four metrics, accompanied by Text & Numerical Data LSTM, BiLSTM and Bert LSTM. The basic and normal machine learning approaches, e.g., Tf-Idf and BoW performed as we expected, performed worse than LSTM deep learning techniques. The highest value of accuracy is 91%, whereas the lowest is 61%

and the same applied for the other remaining metrics as well for Bert Tokenizer LSTM and Tf-Idf & Decision Tree, respectively.

The confusion matrix of the Bert Tokenizer LSTM model is explained below: The diagonal elements of matrix represent the number of tweets for which the predicted label is same as true label, while off-diagonal elements are those numbers that are mislabeled by the models. The higher the diagonal values of the confusion matrix ,better the performance and accuracy which indicating many correct predictions, which is the correctly classified tweets. We can see from below results that for the majority of techniques, positive tweets achieve highest percentage to be correctly classified while negative and neutral are low. Bert Tokenizer LSTM attain the highest values for all remaining metrics in order to performance, whereas Tf-Idf and bag of words (BoW) models have the lowest values.



Confusion Matrix of Bert Tokenizer LSTM model

Fig 5: Confusion Matrix of Bert Tokenizer LSTM Model

IV. Conclusions

In conclusion, our research present the high pervasiveness of keywords and related words among Indian tweets during COVID-19. Based on the data, the tweets are classified into three main different classes and 2 sub classes namely negative, extremely negative, neutral, positive and extremely positive. Some words like “death", “kill”, “hospitalized”, “die” offend people have created unnecessary fear and words such as “God”, “well”, “good” create a positive sentiment among healthcare authorities and people. These results encourage local and central governments to apply fact-checkers on social media to overcome false news and propaganda. Previous research only focuses on the effects of social media and impact of fake news on people but does not discuss the validation, counts, accuracy and classification of tweets. Hence, we applied a best deep-learning model called BERT to achieve high classification accuracy compared to conventional ML models. Our research and results provided enough proof that the BERT model achieved 91% accuracy, which is more than other models like LR, SVM, and LSTM. In this manner, our work clarifies public opinion on pandemics and guides medical authorities, the public, and private workers to overcome from these types of pandemics. From above analysis we can see that the BERT model has performed better than the other machine learning models in the classification of COVID-19 tweets. It produced 91% accuracy, which is much better than other models. Hence, it proved that the BERT model is the best solution for understanding the fake tweets and its sentiment while compared with other models.

References

1. S.; Mittal, Chawl Chawla, M.; Goyal, L. Corona Virus-SARS-CoV-2: EAI Endorsed Trans. Pervasive Health Technol. 2020, An Insight to Another way of Natural Disaster.

2.S.; Salemink, E.; Mertens, G.; Gerritsen, L.; Duijndam, Engelhard, I.M. Fear of the coronavirus (COVID-19): Anxiety Disord. 2020, 74, 101258.Predictors in an online study conducted in March 2020. J.

3.Socio-Economic Impact of COVID-19|UNDP. Available online:https://www.undp.org/content/undp/en/home/coronavirus/socio-economic-impact-of-covid-19.html (accessed on 15 October 2020).

4.I.Staszkiewicz, P.; Chomiak-Orsa, Dynamics of the COVID-19 Contagion and Mortality: Country Factors, Social Media, and Market Response Evidence from a Global Panel Analysis. IEEE Access 2020, 8, 106019–106422.

5. A. Effects of COVID-19 on business and research. J. Bus. Staszkiewicz, P.; Chomiak-Orsa, Res. 2020, 117, 284–249.

6.D.-Y.; Yan, Guo, Y.-R.; Cao, Q.-D.; Hong, Z.-S.; Tan, Y.-Y.; Chen, S.-D.; Jin, H.-J.; Tan, K.-S.; Wang, Y. The origin, transmission and clinical therapies on coronavirus pandemic 2019 (COVID-19) tremendous increase—An update on the status. Mil. Med. Res. 2020, 7, 1–9.

7.N.; Amenta, Mittal, M.; Battineni, G.; Goyal, L.M.; Chhetri, B.; Oberoi, S.V.; Chintalapudi, F. Cloud-based infrastructure to mitigate the impact of COVID-19 on seafarers' mental health. Int. Marit. Health 2020, 71, 213–234.

8.T.A.; Akindele, A.T.; Arulogun, Akande, O.N.; Badmus, O.T. Dataset to support the adoption of social media and emerging technologies for students which is useful as well as harmful ' continuous engagement. Data Brief 2020, 31, 105926.

9. Duarte, E Garcia, L.P.;. Infodemic: Excess quantity to the decrease of quality of information about COVID-19. Epidemiol. Serv. Health 2020, 29.

10.Park, J.; Dang, P.; Hung, M.; Lauren, E.; Hon, E.S.; Birmingham, W.C.; Xu, J.; Su, S.; Hon, S.D.; Lipsky, M.S. Social media platforms Analysis of COVID-19 Sentiments: Application of Artificial Intelligence. J. Med. Internet Res. 2020, 22, e22890.

11.A.; Nunan, Di Domenico, G.; Sit, J.; Ishizaka D. Fake news, marketing and social media: A systematic review. J. Bus. Res. 2021, 124, 329–343.

12.Omar, B. Apuke, O.D.; Fake news and COVID-19: Modelling the predictors of fake news sharing among social media users. Telemat. Inform. 2021, 56, 10045.

13.Zaman, A. COVID-19-Related Social Media spreading Fake News in India. J. Media 2021