RESEARCH ARTICLE OPEN ACCESS

Comparative Analysis of Machine Learning and Deep Learning Algorithms for Classification of Social Media data related to COVID-19

Gohar Ali [1], Khawar Iqbal Malik [2], Unsa Maheen [3]
[1], [2], [3] Department of Computer Science, University of Lahore - Pakistan

ABSTRACT

The Coronavirus disease struck Wuhan in late December 2019, and in March it affects 169 countries. It effected the lungs and also mental health, economy loss and behaviors of peoples. To know the quickly impact of Corona on the society we need to analysis the public opinion thorough social media to know the pandemic effect on the society, social media is the platform where people express their feelings and we can quickly examine the effect of pandemic through public opinion on social media. This research examines the different studies of the classification of social media text data posted by individuals in the Pandemic of COVID-19. In our research, we assess the Naïve Bayes, Logistic Regression, SVM, Deep LSTM, and BERT models to classifying the different classes of social media text data & their sentiments. We observe the efficiency of these models with accuracy and compare the performance of the models, which shows that Logistic regression and Deep LSTM models give the better performance with an accuracy of 83.11% and 82.2%, which is greater than as compared to BERT, Naïve Bayes and Support Vector Machine. Our study may be helpful for the researchers to fine-tune the Logistic Regression and LSTM models for quick and accurate classification of text.

Keywords — COVID-19, social, classification, Twitter, Deep LSTM, Logistic Regression,

I. INTRODUCTION

Coronavirus (COVID-19) was first classified in January 2020 by Chinese scientists in Wuhan, Hubei province, China. The disease struck the Wuhan when a high number of individuals were affected with pneumonia of unclear origin in late December 2019. The indications of this deadly virus, which affects the lungs, vary from minor health signs including some cough, difficulty in breathing, scratchy feeling in throat, and temperature to serious pneumonia infection in lungs, the Acute Respiratory Distress Syndrome (ARDS), some shocks, and eventually paralysis. [1, 2]. This dangerous situation, as announced by the World Health Organization (WHO) on March 11, 2020, affects 169 countries, and affects all continents except Antarctica. The numbers of coronavirus affected peoples has crossed the 1.8 million by the ending of year 2020, with 0.1 million patients infected worldwide. [3]. To end the epidemic, human cooperation is necessary. According to the WHO, it is essential to link danger and involve with many groups in order to develop an effective plan to prevent and prepare against coronavirus, as several researches have proven that taking preventative actions at the individual level is helpful in avoiding the increase of spreading the virus. [4]. The WHO also recommends general population to take certain simple precautions, such as using the sanitizer on hands, wearing the face mask, washing the hands with soap, staying in touch with others, and remaining at home, each of them the government of Pakistan has now tightened these precautions. [5].

Social media gives the public the opportunity to express their feelings. Where peoples share their opinion & behaviours. It gives us the opportunity to understand the People, Group & Society. These applications run on real-time data. Twitter, Facebook, Instagram and many more are examples where people can vote for their speech in public.

As Coronavirus Disease starts from China and covered the whole world in it. Pakistan is also affected by it. As it effects the health related issues of human body but it also effects the behaviour of society. In this Covid-19 Epidemic, people around the world are sharing their thoughts and feelings through these platforms. Each country contributes its best to the fight against the corona virus but the situation remains difficult everyone who will manage. There is social tension between the virus and the action taken to prevent. Social Media Emotions refer to thoughts, attitudes, and ideas about any situation. Feelings are hidden behind people's response because social media is a platform that is easily accessible in this situation exposure. We can examine the behaviour of Peoples easily at sitting our home place by doing social media analysis, which also help us to understand the opinion of the peoples of any society.

Since the current focus of most research is focused on the pathogenesis, symptoms, identification and preparation of medication of this harmful virus, the psychological health element of this pandemic is sometimes ignored and a small reduction in research provided to understand the psychological impact and behavioural changes in affected individuals and their families. Seeing that the present scenario has put an impact on several people with mental health problems, and in order to response this issue there is need of an hour, The World Health Organization has issued certain mental health guidelines to be followed at this important time. Avoid viewing and hearing the news on a daily basis, remaining in contact with friends and family via social media, encouraging and helping one another, and look after of own health (workout, good diet, and sleep on time) are just a few of them. [8].

In this study We observe the Naive Bayes, Logistic Regression classifiers, Support Vector Machine BERT and Deep LSTM models for the classification of text. NLP and information extraction are two ways in which emotional analysis can be used to analyses emotions of users.

A. Problem Statement

During this period of epidemiological analysis, it is a helpful tool to help predict the user's mood and behavioural changes in the Country Population in time. It is estimated that about 80 percent of the world's data is not created, so it is organized. Large number of text data is processed daily but it is very difficult to translate. Therefore, social media analysis and classification may be helpful to organizations in understanding the data in this pandemic. For the analysis of public opinion on COVID-19 pandemic and its effect on society we need a tool that classify the information into some classes to correctly know the COVID 19 pandemic situation effect on the society. which will be beneficial for the Government to easy understand the society behaviour and overcome the issues immediately.

B. Study Background

Micro-blogging sites connect people around the world by lending a helping hand to changing ideas, times and opinions on the situation. [9] identifies the effect of the COVID-19 disease on the aspect of social life & on other economic activities and shows that how people affected in their social life due to corona and explain the effect of corona on economy, health and transportation etc.

Note that when talking on social media somehow the way of people's thoughts and feelings are exposed, which helps to separate people from themselves ideas. Twitter is taken into account one among the most important social media platforms within the world. By Using real-time data from Twitter, maintain them in an understandable format and continue analysing these data sets using NLP (Natural Language Processing) and machine learning. Through Sentiment Analysis we will be looking at the polarity of the text that will help us to organize thinking there are three categories namely Positive, Neutral and Negative and classify the text into some other categories like social, health etc. Inspired by the way this is, a lesson we have learned to use emotional testing to reflect the general feelings of people and to reach out to people’s opinions and feelings about this epidemic.

The researcher[10]discussed the harmful effects of coronavirus on families in China, as well as policy intimation of support needed throughout the epidemic related to family violence, and made suggestions for new researchers. The research concentrated on New Zealand's mental and emotional well-being, taking into account the decrease in social interaction period, interests of the public, loss of jobs, and economic inadequacy. The goal of this study was to look at the impact and consequences of lockdown on some kind of particular group of communities. [11].

The main goal of this investigation is to look at pandemic-related discussions, concerns, and sentiments among Twitter users. Machine learning algorithms are utilized to discover common embedding popular themes and trends, and ideas in the gathered tweets. [12]. The study's major goal is to look into the covid-19 debate and see how the people feel about the pandemic, get the topics from tweets and their sentiments. They utilize a machine learning technique to analyses 4 million Twitter posts linked to COVID-19 using hashtags and do the classification of the tweets and emotion analysis on the issue.

The researcher[13] utilizes a variety of ML algorithms to categories emotions of tweets and compare the accuracy, they use the tweets from India during the COVID-19 lockdown period and perform the sentiment analysis on it, and compare the accuracy of the models, which shows that the machine learning models get the highest accuracy.

The researcher[14] use the natural language processing, to investigates the polarity of discussion related to COVID -19 and their sentiments posted on Twitter from January to March 2020. During the specified dates, a total of 29,514 tweets were gathered, with 10% of them manually categorized to train a Multinomial Naive Bayes classification model that obtained 72 percent accuracy. According to the findings, 52 percent of the remaining tweets are favourable, while 48 percent are negative.

Not all parties are able to identify ideas or feelings in the use of non-official language as well this can crave a more robust testing process. Human language is hard to understand by machines, therefore requires the need for Natural Language Processing. The subfield of artificial intelligence that facilitates in the analysis, interpretation, and evaluate comprehensible personal data in machine processing data. The purpose of this study to examine behavioural changes in the Pakistan and their opinions about COVID-19 that how it effects the society. All information will be gather through social media like Twitter, Facebook, Blogs etc.

Now a day the sentiment analysis methods are improving and many deep learning & machine learning approaches are using in this technique like SVM, Naïve Bayes, ANN and linear regression for prediction, training, modelling, and emulating people actions. Sentiment analysis and classification is use to analyses the data and predict the outcome. Data is the main part today for the analysis and almost all data associated to people to identity the way they think is present on social media. This data will help to the researchers and different company’s experts to easily know the behaviour of peoples and their requirements more proficiently. These sentiments of people analysed by researchers and scholars using various approaches such as Naive Bayes method, linear regression, and other deep learning algorithms. The obtained results from sentiment analysis, we can classify the reactions of people on the particular event and categorizing in the form of positive, neutral or negative.

II. LITERATURE REVIEW & RELATED WORK

A. Tweet Analysis of Coronavirus Outbreak Using Machine Learning

Social media is a platform which having huge volume of data related to the opinion of the individuals [15]. Apart from providing entertainment to the community who are searching relevant material about the condition, it is a forum for everyone to express their thoughts, opinions, and experiences. As unexpected as the emergence of harmful virus illness 2019 (coronavirus) was, which had a profound impact on people all across the world, necessitating an examination of public opinion on the pandemic COVID-19. The emotional analysis of this pandemic situation by utilizing tweets data is subject of this study. To do the analysis of this discussion the Machine learning algorithms were used. They used twitter data to investigate the corona virus epidemic, which expanded across multiple nations and became a pandemic. This study contributes to a better understanding of the public's perspective of coronavirus and its consequences. The public's reaction to the epidemic was analysed once the feelings from the time period were retrieved. The dataset was compiled using tweets from Twitter's public API. The data, which includes the IDs of tweets & sentiment ratings of tweets on the coronavirus epidemic, is analysed in five stages. Collection of Tweets and Performing Pre Processing on tweets, Cleaning of Tweets, Clustering, Apply Model and Evaluation. For data analysis, the Naive Bayes classifier was used, and the model's accuracy was approximately 70%. Positive tweets account for 30% of all tweets, whereas negative tweets account for 16%. On the corona virus epidemic, almost 56% of tweets were neutral. People were well-informed on government policies, safety precautions, symptoms, and preventative measures to be performed at this time. They adhered to the social distancing and sterilizing procedures to a tee. Their research aids organizations in gaining a better understanding of public opinion during the Corona Virus epidemic. Because the virus is spreading rapidly, the research should be conducted on a weekly basis to have a better grasp of public attitude.

B. Analysis on COVID-19 Discussion Using SVM

The research conducted by [16] stated that social distance is such a preventative strategy. Individuals share their opinion freely using social websites like Facebook and Twitter, that may be shared among other people. The general public’s feelings on social distance can be discovered by analysing articulated messages from Twitter. The purpose of their work were to decipher and assess public perceptions of social distance as expressed in twitter unstructured data. The SentiStrength programme was used to extract sentiment polarity from tweet texts using Twitter data unique to Canada and words including social distance keywords. After that, to classify the sentiments they uses support vector machine (SVM) method. The Tweets data used in this study came from an accessible, publicly available source. This is due to the Twitter API's restriction on accessing data older than 1 week. The publicly accessible dataset comprised primarily geotagged worldwide tweets. The twitter tweets were then filtered utilizing terms relevant to COVID-19 and hydrated by using DocNow hydrator program, a Twitter hydrator program developed as a desktop application that allows to the collection of tweets in CSV and JSON format. This study used a hybrid method to emotional analysis, utilising the SentiStrength v2.3 applications, a lexicon-based technique, to evaluate and recover sentiment score, and afterwards the SVM algorithm, a machine learning technique, for the categorizing and analysis of the tweets. Different performance evaluation parameters like F1, recall and precision were used to evaluate effectiveness of the model. Only 795 tweets out of hundreds of millions had social distancing words and phrases on twitter post texts and Canada as the location of the user, resulting in the compilation of the 629 tweets, with 40 percent of tweets expressing neutral feelings, 35 percent expressing negative feelings, and only 25 percent expressing positive feelings towards social distancing. By dividing the sample dataset into eight percent training and twenty percent testing data, the SVM technique is applied. The accuracy of the test score was 71 percent.

C. Analysis of Public Reactions on COVID-19

According to [17], Microblogging services, particularly Twitter, have become crucial communication tools in recent years, particularly for political and professional leaders, including health experts and the general public, to communicate with one another as well as their intra-domain communication. It has become a popular platform in countries ranging from the third world to the industrialized world. This platform is actively disseminating public health information and employing crowdsourcing approaches to gather real-time health data. They create a supervised machine learning approach to categories public emotions into health and economic concerns in their research. Such knowledge would open up large-scale possibilities for predicting public concerns about health and economic in future catastrophic disasters. For six months, they gathered public tweets from the four nations. Many keywords were utilized in the keyword-based search to retrieve the needed tweets using the Twitter API. Using the given criteria, they were able to retrieve 28, 930 public tweets for the four nations using Twitter's regular search API. They used Japanese social media tweets as a case study to build a classification system based on logistic regression to categories public emotions as "health" and "economic." The count-vectorizer was used to build a numeric feature matrix for categorization after pre-processing. Their suggested logistic regression-based technique correctly categorized Japanese public tweets into three categories: "health" "economic" and "other" class with accuracy of 83.11%. When comparing tweets about economic concerns to those about health concerns, the results indicated a larger number of tweets about economic concerns.

D. Cross Cultural Polarity and Emotion Detection Using Deep Learning

On the Coronavirus Pandemic, the researchers[18] investigated different cultural reaction, responses and sentiment analysis using deep learning. They examine how different cultures respond and behave in the pandemic, as well as how this impacts the norms of society and governmental will to address the issue. The analysis is based on consumers' tweets about the Corona incident and is done using quantitative research techniques. From beginning of 2nd month to through the end of fourth month of the year 2020, they gathered tweets. Tweepy Twitter API was queried with Python scripts to obtain users' tweets and extract feature sets for cataloguing, while NLTK was used to cleaning the collected text. In their work, deep learning methods for sentiment classification were used. A collection of different hidden layers with numerous nodes make up a deep neural network (DNN). A DNN's training method contains 2 steps, the first one is “pre training” and the second one is “fine-tuning”. Pre-training step involves weight initialization on the inputs in unsupervised means. By maximizing predicted log probability, a contrastive divergence method was utilized to estimate the trainable parameters. The LSTM, a type of RNN, were used in their suggested sentiment evaluation algorithm (RNN). LSTMs aid in the preservation of mistake that can be transmitted backwards in time and layers. On the 16,784 good and negative tweets, they evaluated the Model based on LSTM and FastText technology which learned on Sentiment. Total 16 thousand tweets were used for the testing to determine the accuracy of model. The classifier utilised the emoticons as the categories to classify the classes of the twitter text. This model has correctly classify the text with accuracy rate of 82 percent & F1 score of 78 percent. This demonstrates that their model is relatively in accordance with the users' emoticon-based feelings. Their research also shows that mood and emotion identification based on natural language processing can help with more than only identifying cross-cultural patterns. However, it is also possible to make a strong relationship between actual occurrences and users' feelings expressed on social media sites. And that, in the face of a global catastrophe, such as the coronavirus epidemic, there is a strong correlation of feelings voiced despite socioeconomic and cultural disparities.

E. Analysis of COVID-19 Tweets using BERT

In the paper [19] the author described that The Coronavirus Infection of 2019 (COVID-19) poses a significant threat to the planet. While on an epidemic, individual opinion assessment offers useful method for evaluating effective public health actions. Posts with negative feelings on Weibo, a major Chinese social networking site, are useful in evaluating social issues. From first of January to eighteen of February 2020, total 999,978 posts related to COVID-19 were randomly chosen COVID-19-related Weibo postings were examined. To categorise emotion classes (+ve, -ve and neutral), the unsupervised BERT algorithm is utilised, and the TF-IDF technique were used to summarise the themes of text. To detect negative feelings features, trend analysis and theme analysis are used. In overall, the good modified BERT is very accurate when it comes to sentiment categorization. Furthermore, TF-IDF themes accurately transmit COVID-19-related post features. Weibo posts that have been properly categorised are split into two portion, one for the training the model and second for testing the model at randomly in 5:5 ratios. They well modified the BERT-BASE architecture with twelve layers in research study to correctly classify the coronavirus related discussion into their topics and extract the sentiments into their topics accordingly. They chose 4 iterations, a training amount of 2e-5 with having 32 batch size based on the recommendation for hyper parameter selection. Using the training dataset, they utilised the neural network layer named as softmax to train multi sentiment model as neutral, positive and negative. The sentiment categorization model obtains a 75.65 percent accurate result on the test data after learning. The majority of the 999,978 postings 56.2 percent are neutral, with 27.4 percent positive sentiment and 16.5 percent negative feelings. As a consequence, they determined that individuals are concerned about 4 aspects of COVID-19: the origin of pandemic symptoms having lungs infection and some kind of fever, production activity like office job, work, education classes and community health handle like patients and lockdown in the country. The findings from Weibo postings offer helpful advice on global health responses, demonstrating that open effective communication and scientific counsel may help ease community worries.

III. RESULTS AND DISCUSSION

Among all the studies they collected the dataset and performed some pre-processing on the data. The classification of all above model shows that the datasets on which model was performed are in categorized in multi classes and with their sentiments also, which is also discussed above in the study.

A. Comparison of Models.

We observe the classification 5 different models, Logistic regression, Naïve Bayes, SVM, Deep LSTM and BERT, which were used to for training and evaluation of the dataset. The dataset was differently distributed for different models and their detail is mentioned below in the table I. The comparison of different models of observation with detail is also shown in the table I.

TABLE I
COMPARISON OF MODELS

Model	Dataset	Classes	Accuracy
Naïve Bayes	5 days' tweets	Multi	70%
SVM	629 tweets	Multi	71%
Logistic Regression	28,930 tweets	Multi	83.11%
Deep LSTM	27,357 tweets	Multi	82.2%
BERT	999,978 posts	Multi	75.65%

From above we can see that out of Naïve Bayes, SVM and BERT the Logistic regression and Deep LSTM performs better for multiclass classification for the Corona pandemic related dataset with the accuracy of 83.11% and 82.2% respectively.

B. Discussion

In this research we observe the efficiency and accuracy of different machine learning and deep learning models as mentioned as Logistic Regression, Naïve Bayes, SVM, Deep LSTM and BERT for the classification of the social media data expressed by the peoples in COVID-19 Pandemic and examined the effect of COVID-19 on society. This majority of the research on these model is to find out the sentiment analysis of COVID-19 for limited time period and data. In this study we observe the multi classification of the data and we asses that Logistic Regression and Deep LSTM models perform better than other Machine Learning and deep learning models.

The limitation of our research is that we observe some deep learning and machine learning models on different datasets, which can be limited. We observe the performance of different models on some classes which are related to text only.

C. Future Work

In future the more models can be observed, as we observe that LSTM and Logistic Regression perform better. So in future we train enhanced LSTM model with large dataset of Twitter with classification of text into some social norms effected by the COVID-19 Pandemic like economy, education etc. We will also compare the result of different deep learning and machine learning models with LSTM on the balanced and same data set for all the models. Which can be more helpful to know the deep effect of pandemic on the society and more accurate results can be achieved in short frame of time.

IV. CONCLUSIONS

In this research we observe some deep learning & machine learning approaches related to the multi-class classification of the social media data posted by individuals in the COVID-19 epidemic to observe the pandemic effect on the society. We observe some machine learning and deep learning models on different datasets & check the accuracy results of different models. We observe that Logistic regression and Deep LSTM models gives the better performance with accuracy of 83.11% and 82.2% respectively, which is greater than as compared to BERT, Naïve Bayes and Support Vector Machine. This study can be helpful for the researchers to fine tune the Logistic Regression and LSTM and more improve the accuracy of these models which can be helpful for the society also for overcoming the issues of pandemic in time.

ACKNOWLEDGMENT

We say thanks to ALLAH Almighty for everything throughout our research work. He is always with us for help us out of challenging moments of our life. We are grateful to ALLAH Almighty who made us able to complete our research work. It is due to HIS unending mercy that this work moved towards success. ALLAH Almighty bestowed us the opportunity and spirit to make the material assessments in already existing ocean of knowledge of the subject. All respect for his "HOLY PROPHET" (peace be upon him) who enabled us to recognize the purpose of creation.

REFERENCES

[1] N. Chen, M. Zhou, X. Dong, J. Qu, F. Gong, Y. Han, Y. Qiu, J. Wang, Y. Liu and Y. Wei, "Epidemiological and clinical characteristics of 99 cases of 2019 novel coronavirus pneumonia in Wuhan, China: a descriptive study," *The Lancet,* pp. 507-513, 2020.

[2] H. A. Rothan and S. N. Byrareddy, "The epidemiology and pathogenesis of coronavirus disease (COVID-19) outbreak," *Journal of autoimmunity,* p. 102433, 2020.

[3] "Coronavirus disease (COVID-19) pandemic," April 2020. [Online]. Available: https://www.who.int/emergencies/diseases/novel-coronavirus-2019.

[4] W. M. Jang, S. Cho, D. H. Jang, U.-N. Kim, H. Jung, J. Y. Lee and S. J. Eun, "Preventive behavioral responses to the 2015 middle east respiratory syndrome coronavirus outbreak in Korea," *International journal of environmental research and public health,* p. 2161, 2019.

[5] "Coronavirus disease (COVID-19) advice for the public," March 2020. [Online]. Available: https://www.who.int/emergencies/diseases/novel-coronavirus-2019/advice-for-public.

[6] Y.-R. Guo, Q.-D. Cao, Z.-S. Hong, Y.-Y. Tan, S.-D. Chen, H.-J. Jin, K.-S. Tan, D.-Y. Wang and Y. Yan, "The origin, transmission and clinical therapies on coronavirus disease 2019 (COVID-19) outbreak--an update on the status," *Military Medical Research,* pp. 1-10, 2020.

[7] N. Gupta and J. Nusbaum, "Points \& Pearls: Novel 2019 Coronavirus SARS-CoV-2 (COVID-19) An Overview for Emergency Clinicians," *Emergency medicine practice,* 2020.

[8] "Mental Health," March 2020. [Online]. Available:

https://www.who.int/docs/default-source/coronaviruse/mental-health-considerations.pdf.

[9] B. Kundu and D. Bhowmik, "Societal impact of novel corona virus (COVID– 19 pandemic) in India," *kundu2020societal,*, 2020.

[10] H. Zhang, "The influence of the ongoing COVID-19 pandemic on family violence in China," *Journal of family violence,* pp. 1-11, 2020.

[11] S. Every-Palmer, M. Jenkins, P. Gendall, J. Hoek, B. Beaglehole, C. Bell, J. Williman, C. Rapsey and J. Stanley, "Psychological distress, anxiety, family violence, suicidality, and wellbeing in New Zealand during the COVID-19 lockdown: A cross-sectional study," *PLoS one,* 2020.

[12] J. Xue, J. Chen, R. Hu, C. Chen, C. Zheng, Y. Su and T. Zhu, "Twitter Discussions and Emotions About the COVID-19 Pandemic: Machine Learning Approach," *Journal of medical Internet research,* p. e20550, 2020.

[13] N. Chintalapudi, G. Battineni and F. Amenta, "Sentimental Analysis of COVID-19 Tweets Using Deep Learning Models," *Infectious Disease Reports,* pp. 329--339, 2021.

[14] J. P. D. Delizo, M. B. Abisado and L. T. M. I. P. De, "Philippine Twitter Sentiments during Covid-19 Pandemic using Multinomial Na{\"\i}ve-Bayes," *International Journal,* 2020.

[15] D. K. Iyer and D. S. Kumaresh, "Twitter Sentiment Analysis On Coronavirus Outbreak Using Machine Learning Algorithms," *European Journal of Molecular \& Clinical Medicine,* pp. 2663--2676, 2020.

[16] C. Shofiya and S. Abidi, "Sentiment Analysis on COVID-19-Related Social Distancing in Canada Using Twitter," *International Journal of Environmental Research and Public Health,* 2021.

[17] M. K. Bashar, "Event-driven timeseries analysis and the comparison of public reactions on COVID-19," *arXiv preprint,* 2021.

[18] Imran, S. Ali, Doudpota, M. Sher, Kastrati, Zenun, Bhatra and Rakhi, "Cross-Cultural Polarity and Emotion Detection Using Sentiment Analysis and Deep Learning--a Case Study on COVID-19," *arXiv preprint arXiv:2008.10031,* 2020.

[19] T. Wang, K. Lu, K. P. Chow and Q. Zhu, "COVID-19 sensing: negative sentiment analysis on social media in China via BERT model," *Ieee Access,* pp. 138162--138169, 2020.