Ensemble learning based classifier to predict depression caused due to pandemic

P Vaishali¹ and P L S Kumari²
Department of CSE, Koneru Lakshmaiah Education Foundation, Aziznagar, Hyderabad, Telangana State, India.

Email: vaishupurude@gmail.com

Abstract. Pandemic caused due to Corona Virus Disease 2019 (COVID-19) affected each and every person life throughout the world. First wave of COVID-19 followed by second wave made situation more panic. Government declared Lockdown imposed strict prohibition on social gathering, unnecessary outing, travelling, and education. During home quarantine, people shared opinion, expressed views, feelings on social media. Home isolation and quarantine affected mental health of people which may lead to depression. Hence in this research article depression is predicted by implementing Neural Network based model. At first level this model implements text classification of COVID-19 based Tweets. Neural network model accuracy is 86.85%. In next level, using same tweet dataset as input, Ensemble learning based model is constructed. This model uses one of the boosting techniques known as Adaboost. Model is executed by varying Train-test-validation ratio. It is observed that accuracy of the model is improved. The model showed accuracy of 99.33 % successfully in every execution. Obtained results are compared with previous work in same area.

Keywords: Pandemic; Neural network; Ensemble Learning; Boosting.

1. Introduction
In March 2020, COVID-19 was declared as a pandemic situation throughout the world as the virus has spread in more than 190 countries within a short time span [1]. For reducing the speed of virus spread, nearly in all countries worldwide declared complete lockdown and which caused restrictions on public life [2]. “Social distancing” tried to avoid physical contact. Government enforced closure of educational institutes and non-essential businesses. Lockdown caused bans on travel, public gathering. It encouraged online education and work from home. These new guidelines resulted in a negative impact as loss of hope, frustration and depression. Depression affects physical and mental health [3].

People like labor relocated to native places. All kind of business was paused. Many people lost job due to cost-cutting strategy of management. These all circumstances caused economical burden nearly in each family. This kind of pressure also affected mental state. If mental health is monitored and depression predicted at early stage, then risk of mental disorder and suicide attempts can be controlled.

In this paper depression is predicted using social media posts. Social media platform has become first choice for sharing frustration feelings, to expressing their opinion. Twitter is most popular choice in social media due to number of characters permitted in tweet. Hence Twitter is good source for text classification [4].

So far research is not performed to predict depression caused due to Pandemic using Ensemble Learning based Text Classification. In this paper efforts are put in that direction. Dataset is formed by using Tweets related to corona virus on Twitter. These tweets contains keywords related to virus such
as ‘Lockdown’, ‘schools closed’, ‘pandemic’, ‘no social gatherings’, ‘home isolation’, ‘quarantine’, ‘infection’, ‘pandemic’ etc. Dataset contains nearly 2000 tweets divided into two classes with label 0 for non depressive tweets and Label 1 for depressive. A neural network is constructed to recognize depressive tweets. In this research, experiments are performed on Convolutional Neural Network (CNN) and adaboost classifiers are applied by varying Test-Train-Validation ratio. It is observed that accuracy obtained by adaboost based Ensemble Model is always better in every execution.

2. Previous Work
In text classification research work is already started. Text classification can be implemented using fuzzy neural network [5]. While artificial neural network can also be used for automatic text classification [6]. Research has also carried on comparison of algorithms used for text classification [7]. Some authors carried out literature survey in this area [8]. Text classification is applied for sentiment analysis of COVID-19 disease using tweets by applying machine learning and also conducted Geographic analysis [9]. In [10], the psychological effect of the pandemic in Italy has been evaluated by adopting psychometric self reporting methodology, as restrictions to conduct experimental research by meeting participants. Data was collected with technique CATI and CAWI means Computer assisted Telephonic and Web interview. Researchers’ job was made quite easy by making availability of open source data [11]. As research is being conducted on diagnosis of COVID-19 from X-ray and CT scan images, knowledge based recovery, emotion and sentiment analysis, these research works can be compared and analyzed. Sentiment analysis of social media posts can be implemented using Naïve bayes classifier [12]. Online survey can be conducted using email-link to find burden caused due to COVID-19 [13]. By conducting online survey, it is observed that COVID-19 caused negative impact on mental state [2]. During COVID-19, health care workers are playing major role by sacrificing personal life and devoting time to cure patients. But it increased stress in them, which can be assessed using self reports and artificial intelligence [14]. Social media use was increased during lockdown and social distancing. It is needed for reducing the spread of virus. People are online most of the time, to find latest updates and information. The extent upto which Social media was used and stress symptoms caused due to it can be studied [3]. Study was carried out in Italy and Germany through an online survey. Predictive model can be constructed to find stress, anxiety and depression in students by conducting survey using questionnaire [15-17]. An international online survey in 18 languages was conducted by researchers around the world for people having age greater than 18 to find impact of COVID-19 on mental health [18]. In [19] review of methods used for Sentiment Analysis in Healthcare workers has been performed before COVID-19, using their tweets. Some researchers found negative emotions in tweets from Brazil and USA during COVID-19 [20].

Ensemble learning can be used to improve accuracy of the model [21]. Research was carried on volatile stock exchange prediction using Ensemble learning and found improved accuracy. Using cluster ensemble robustness, scalability and accuracy can be improved [22]. By constructing model based on ensemble classifier, Malaria disease can be classified among individuals [23]. When searched for previous work based on Natural Language processing, found model based on deep learning which answers queries asked in natural language [24]. This is one of the applications of text processing. For extracting information from huge data machine learning tools like classification, regression etc can be used [25,26]. To extract huge information generated from patient’s medical data in healthcare applications, mining tools should be used [27]. If healthcare is based on network then its quality of service can be improved by modifying segment size [28].

From previous study it is clear that till now research has not carried out using Ensemble Learning to improve accuracy for predicting depression caused due to Pandemic. In this paper using Boosting techniques accuracy of prediction is improved.
3. Novel Prediction Methodology

3.1 Introduction

In this paper an effort is made to predict depression caused due to COVID-19, using tweets during pandemic. In first stage of research, model uses Neural Network. In the next level model uses Ensemble Learning.

3.1.1 Neural Network

Neural network is a model that behaves similar to human brain consisting biological neurons and nervous system. It is used to find complex patterns and relationship existing within a labeled data. Neural network is made up of layers called input layer, hidden layer, and output layer as indicated in Figure 1.

![Figure 1. Neural network](image1)

3.1.2 Ensemble Learning

Accuracy obtained by neural network is not satisfactory. Hence to improve Accuracy Ensemble model based on boosting is built. Ensemble model combines multiple machine learning models to show improvement in accuracy. It is implemented either using bagging or boosting. Bagging divides dataset and model is applied on sampled dataset and final prediction is average of each prediction. In boosting whole dataset is inputted to model, output of one model acts as input to next model. And final accuracy is better than original. Ensemble learning model is as shown in Figure 2.

![Figure 2. Ensemble learning](image2)
3.2 Proposed Methodology

3.2.1 Proposed Model
Proposed Model is shown in Figure 3. After collecting COVID-19 tweets dataset, preprocessing should be performed on data as it affects accuracy of model. During preprocessing data is cleaned means removal of emoji symbols, stopwords, punctuation marks etc. After cleaning data, it is labeled based upon negative emotion keywords like hopeless, locked, lonely, jobless, financial problem, no income, suicide, depressed, not interested, and useless. This labeled dataset is split into train-test and validation sets required for Neural Network and Adaboosting technique of Ensemble Learning. Then model is trained using Algorithm and evaluated.

![Proposed Network Model](image)

**Figure 3.** Proposed network model

3.2.2 Algorithm
Input : Tweets from ‘Tweets COVID-19’ dataset (3.3.1)
Steps:
1. Preprocessing: Data is cleaned with removal of emoji symbols, stopwords, punctuation marks etc.
2. Labeling: Tweets are divided into depressive tweets (labeled 1) and non depressive tweets (labeled 0) based on search of negative emotion keywords.
3. Splitting dataset: Labeled Dataset is divided into training set, validation set and test set.
4. Training phase: Using training set model is trained.
5. Validation: Using test and validation sets trained model is validated.
6. Evaluation: Model performance is evaluated based on training and validation accuracy and loss.
Output: Training and Validation Accuracy and Loss.

3.3 Experimental Results
3.3.1 Dataset
Dataset used for experiments is ‘Tweets COVID-19’. It contains the Twitter data on COVID-19 from early 2020 (March-June). It has parameters: Serial No, Date of the tweet(s), Time of the tweet(s) in IST, Source link, User ID, Language, Retweet counts, Hashtags (only related to COVID-19), User mentions (if any), Place of the tweet(s), Tweet body [29].

3.3.2 Results of Neural Network
Loss and Accuracy obtained from Algorithm on Neural Network is 0.389761140346527, 0.8685969114303589 respectively. Output obtained in each step is shown in Figure 4. Epoch in the output indicates that inputted dataset is divided into batches and epoch contains these batches. Each batch is used to train neural network. Output step indicates training and validation loss and accuracy obtained for each batch.

![Figure 4. Stepwise loss and accuracy](image)

Figure 5 and Figure 6 indicates plots of training and validation loss and accuracy.

![Figure 5. Training and validation accuracy](image)

![Figure 6. Training and validation loss](image)

In Figure 5, validation accuracy is more than training. In Figure 6 validation loss is less than training. This indicates that model is trained and validated correctly.

3.3.3 Results of Boosting Technique
In this paper Adaptive boosting known as Adaboost which is one of the boosting algorithms is used.
Adaboost builds strong classifier by combining many weak classifiers. It is operated in iterations. In each iteration by assigning different weights to weak classifiers, it gets predictions with improved accuracy. In this paper by varying train-test-validation ratio when model is executed multiple times, accuracy obtained is minimum 0.959910913 and maximum 0.993333333.

3.4 Result Analysis
Results obtained with application of Adaboost are compared with results of neural network as shown in Table 1.

| Exe No. | Neural Network | Adaboost1 | Adaboost2 | Adaboost3 |
|---------|----------------|-----------|-----------|-----------|
| Exe1    | 0.866369724    | 0.959910913 | 0.986666667 | 0.983333333 |
| Exe2    | 0.866369724    | 0.973273942 | 0.986666667 | 0.983333333 |
| Exe3    | 0.850779533    | 0.968819599 | 0.98 | 0.981666667 |
| Exe4    | 0.859688222    | 0.962138085 | 0.991111111 | 0.993333333 |
| Exe5    | 0.868596911    | 0.964365256 | 0.98 | 0.981666667 |

Table 1. Comparison of results
In Table 1, Exe1 to Exe5 are executions 1 to 5. Out of 2000 tweets Adaboost1 uses 70% dataset for training, 22.5% for validation and 7.5% for testing. Adaboost2 splits dataset as 70% for training and 22.5% for testing and does not use validation set. Adaboost3 splits dataset as 70% for training and 30% for testing. Same output is plotted in Figure 7.

Figure 7. Comparison of results
In Figure 7, it can be observed that in each execution Accuracy of Adaboost classifier is better than Neural Network ranging from 95.99 % to 99.33 %.

Results are compared with previous work in same area [30], as shown in Table 2.
### Table 2. Comparison with previous work

| Model     | Accuracy |
|-----------|----------|
| RF        | 0.86     |
| XGBoost   | 0.85     |
| SVC       | 0.85     |
| ETC       | 0.88     |
| DT        | 0.83     |
| Neural Network | 0.86   |
| Adaboost  | 0.99     |

Same comparison is shown graphically in the Figure 8.

In [30] the performance of machine learning models on COVID-19 tweets has been compared. Techniques used are RF (Random forest), XGBoost (eXtreme Gradient Boosting), SVC, ETC (meta-estimator), DT (Decision Tree). Using these 5 techniques accuracy observed is minimum 83% to maximum 88%. This accuracy when compared with neural network model and Adaboost implemented in this paper, Adaboost gives best accuracy upto 99%. Hence Adaboost improves accuracy of model by boosting its performance.

### 4. Conclusion and Future Work

When World was relaxing from First wave of COVID-19, second wave entered in India with increase in cases every day. Nearly each house was having people affected by virus and death rate was also increased. This created fear in the mind of people. Due to lockdown, restrictions were imposed on movement. Hence people relied on social media to get updated news and to share their feelings. Use of Facebook, Twitter increased. The messages getting circulated in social media impacts the mind of people. Some people may take news in positive way and utilize available time for improvement of health and skills. While some people will take it negatively, get panic and which makes them mentally unstable. In this paper COVID-19 Tweets are divided based on keywords and used to predict depression caused due to pandemic. In model1 Neural Network is built with accuracy 86.85 %. At next level Ensemble model is constructed using Adaboost classifier and accuracy was improved upto
99.33 %. Hence it can be concluded that boosting techniques of Ensemble Learning improves accuracy of prediction.

In future Ada boosting can be applied on other base classifiers.

References

[1] https://www.who.int/india/emergencies/coronavirus-disease-(COVID-19).
[2] S. K. Schäfer, M. R. Sopp, C. G. Schanz, M. Staginnus, A. S. Göritz, and T. Michael, “Impact of COVID-19 on Public Mental Health and the Buffering Effect of a Sense of Coherence,” Psychother. Psychosom., vol. 89, no. 6, pp. 386–392, 2020, doi: 10.1159/000510752.
[3] J. Brailovskiaia, F. Cosci, G. Mansueto, and J. Margraf, “The relationship between social media use, stress symptoms and burden caused by coronavirus (COVID-19) in Germany and Italy: A cross-sectional and longitudinal investigation,” J. Affect. Disord., vol. 3, no. December 2020, p. 100067, 2021, doi: 10.1016/j.jad.2020.100067.
[4] S. Das and A. K. Kolya, “Predicting the pandemic: sentiment evaluation and predictive analysis from large-scale tweets on COVID-19 by deep convolutional neural network,” Evol. Intell., no. 0123456789, 2021, doi: 10.1007/s12065-021-00598-7.
[5] U. S. Krishna, H. Shreek, K. Jayadeep, and P. L. Prasanna, “Text classification using fuzzy neural network,” Int. J. Recent Technol. Eng., vol. 7, no. 6, pp. 193–198, 2019.
[6] P. L. Prasanna and D. R. Rao, “Text classification using artificial neural networks”, International Journal of Engineering and Technology(UAE), vol. 7, pp. 603–606, 2018.
[7] M. Syamala and N. J. Nalini, “Comparative Analysis of Document level Text Classification Algorithms using R”, IOP Conference Series: Materials Science and Engineering, 2017, doi: 10.1088/1757-899X/225/1/012076.
[8] Lakshmi Prasanna, P. Rajeswara Rao, D., “Literature survey on text classification: A review”, Journal of Advanced Research in Dynamical and Control Systems, vol. 9, special issue 12, pp 220-2280, 2017.
[9] M. Hung et al., “Social Network Analysis of COVID-19 Sentiments: Application of Artificial Intelligence,” J. Med. Internet Res., vol. 22, no. 8, p. e22590, Aug. 2020, doi: 10.2196/22590.
[10] M. Delmastro and G. Zamarilia, “Depressive symptoms in response to COVID-19 and lockdown: a cross-sectional study on the Italian population,” Sci. Rep., vol. 10, no. 1, pp. 1–10, 2020, doi: 10.1038/s41598-020-79850-6.
[11] J. Shuja, E. Alanazi, W. Alasmary, and A. Alashaikh, “COVID-19 open source data sets: A comprehensive survey,” Appl. Intell., pp. 1296–1325, 2020, doi: 10.1101/2020.05.19.20107532.
[12] Minduru, S., Eknath, A., Velgapudi, V.S.C., Mohammed, M., “Analysis of text classification of dataset using NB-classification”, International Journal of Innovative Technology and Exploring Engineering, vol. 8, no.5, pp 1175-1179,2019.
[13] J. Brailovskiaia and J. Margraf, “Predicting adaptive and maladaptive responses to the Coronavirus (COVID-19) outbreak: A prospective longitudinal study,” Int. J. Clin. Heal. Psychol., vol. 20, no. 3, pp. 183–191, 2020, doi: 10.1016/j.jchp.2020.06.002.
[14] K. Sosić, S. Popović, M. Šarlija, I. Kesedžić, and T. Jovanovic, “Artificial intelligence in prediction of mental health disorders induced by the COVID-19 pandemic among health care workers,” Croat. Med. J., vol. 61, no. 3, pp. 279–288, 2020, doi: 10.3325/cmj.2020.61.279.
[15] M. C. Zurlo, M. F. Cattaneo Della Volta, and F. Vallone, “COVID-19 Student Stress Questionnaire: Development and Validation of a Questionnaire to Evaluate Students’ Stressors Related to the Coronavirus Pandemic Lockdown,” Front. Psychol., vol. 11, no. March, pp. 1–11, 2020, doi: 10.3389/fpsyg.2020.576758.
[16] A. J. Rodríguez-Hidalgo, Y. Pantaleón, I. Dios, and D. Falla, “Fear of COVID-19, Stress, and Anxiety in University Undergraduate Students: A Predictive Model for Depression,” Front. Psychol., vol. 11, no. November, 2020, doi: 10.3389/fpsyg.2020.591797.
[17] M. Akhterul Islam, S. D. Barma, H. Raihan, M. Nafiul Alam Khan, and M. Tanvir Hossain, “Depression and anxiety among university students during the COVID-19 pandemic in Bangladesh: A web-based cross-sectional survey,” PLoS One, vol. 15, no. 8 August, pp. 1–12, 2020, doi: 10.1371/journal.pone.0238162.
[18] A. T. Gloster et al., “Impact of COVID-19 pandemic on mental health: An international study,” PLoS One, vol. 15, no. 12 December, pp. 1–20, 2020, doi: 10.1371/journal.pone.0244809.
[19] S. Gohil, S. Vuik, and A. Darzi, “Sentiment Analysis of Health Care Tweets: Review of the Methods Used,” JIMIR Public Heal. Surveill., vol. 4, no. 2, p. e43, Apr. 2018, doi: 10.2196/publicheal.5789.
[20] K. Garcia and L. Berton, “Topic detection and sentiment analysis in Twitter content related to COVID-19 from Brazil and the USA,” Appl. Soft Comput., vol. 101, no. January, p. 107057, Mar. 2021, doi: 10.1016/j.asoc.2020.107057.
[21] P. Rajesh, N. Srinivas, K. V. Reddy, G. Vamsipriya, V. D. M, and D. Himaja, “Stock trend prediction using Ensemble learning techniques in python”, International Journal of Innovative Technology and Exploring Engineering, no. 5, pp. 150–155, 2019.
[22] Y. L. S. Rani, V. Sucharita, K. V. V Satyanarayana, and Y. L. S. Rani, “Extensive Analysis on Generation and Consensus Mechanisms of Clustering Ensemble: A Survey”, International Journal of Electrical and Computer Engineering, vol. 8, no. 4, pp. 2351–2357, 2018, doi: 10.1159/ijence.v8i4.pp2351-2357.
[23] T. Sajana and M. R. Narasingarao, “AN ENSEMBLE FRAMEWORK FOR CLASSIFICATION OF MALARIA DISEASE”, ARPN Journal of Engineering and Applied Sciences, vol. 13, no. 9, pp. 3299–3307, 2018.
[24] A. Alava, B. Bai, M. Blushan, “Deep Learning Techniques and Models for improving Machine Reading
Comprehension System”, *IJAST*, vol. 29, no. 04, pp. 9692–9710, Oct. 2020.

[25] K. Verma, S. Bhardwaj, R. Arya, M.S.U. Islam, M. Bhushan, A. Kumar, and P. Samant, “Latest tools for data mining and machine learning”, *International Journal of Innovative Technology and Exploring Engineering (IJITEE)*, vol. 8, no. 9S, pp. 18-23, ISSN: 2278-3075, 2019. Available: https://doi.org/10.35940/ijitee.I1003.0789S19

[26] Resham Arya, Ashok Kumar, and Megha Bhushan, “Affect Recognition using Brain Signals: A Survey”, In: Singh V., Asari V.K., Kumar S., Patel R.B. (eds) *Computational Methods and Data Engineering. Advances in Intelligent Systems and Computing*, vol 1287, 2020. Springer, Singapore. https://doi.org/10.1007/978-981-15-7907-3_40

[27] Sayeedakhanum Pathan, Megha Bhushan, Anita Bai, “A Study on Health Care using Data Mining Techniques”, *Journal of Critical Reviews*, 7 (19), 7877-7890, 2020. doi:10.31838/jcr.07.19.896

[28] Vikram Jeet Singh, Megha Bhushan, Vikram Kumar, Kishori Lal Bansal, “Optimization of Segment Size Assuring Application Perceived QoS in Healthcare”, *Lecture Notes in Engineering and Computer Science*, World Congress on Engineering 2015, London, United Kingdom, *volume 2217, no. 1*, 2015, pp. 274-278.

[29] 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”, *Evol. Intel.* (2021). https://doi.org/10.1007/s12065-021-00598-7.

[30] F. Rustam, M. Khalid, W. Aslam, V. Rupapara, A. Mehmood, and G. S. Choi, “A performance comparison of supervised machine learning models for COVID-19 tweets sentiment analysis,” *PLoS One*, vol. 16, no. 2, p. e0245909, Feb. 2021. doi: 10.1371/journal.pone.0245909.