Sentiment Analysis on Noisy Bangla Texts using Machine Learning Techniques

Md. Omar Faruqe[1], Md. Mehedy[2], Sanjoy Kumar Chakravarty[3], A. F. M. Mahbubur Rahman[4]
[1,2,3,4](Department of Computer Science and Engineering, University of Rajshahi, Bangladesh)
faruqe@ru.ac.bd[1], mdmehedy302@gmail.com[2], sanjoy.cse@ru.ac.bd[3], mmr@ru.ac.bd[3]

Abstract:
In this modern era of Internet, we can express our thoughts through different social media platforms i.e. Facebook, Twitter, Instagram etc. The ongoing novel coronavirus disease (COVID-19) pandemic has created devastating consequences on our overall health condition. Some of the survivors are experiencing serious mental trauma and other mental health issues. To determine the aftereffect of COVID-19 on mental health condition, we performed sentiment analysis on their social media posts. This paper primarily focuses on sentiment analysis of twitter data written in Bengla language and intends to determine whether the user is mentally depressed or not. The expressions of those users are either negative or positive, sometimes neutral. Using different contemporary machine learning (ML) algorithms like Naive Bayes, Support Vector Machines (SVM), Random forest and Multi-layer perceptron (MLP), we determined the mental health condition of individual user. Empirical results demonstrated that MLP outperformed other techniques in terms of overall performance and the frequency of negative tweets has dramatically increased after COVID-19.

***Key Word**: Sentiment Analysis, Machine Learning (ML), COVID-19, Bangla Tweet.*

Date of Submission: 02-04-2022 Date of Acceptance: 15-04-2022

I. Introduction

On December 31, 2019, the first COVID-19 case was identified in Wuhan, China [1]. Within 3 months after the first detection report, the novel coronavirus (SARS-COV-2) spread out all over the world. On 8th March 2020 the first three cases were identified in Bangladesh by Institute of Epidemiology, Disease Control and Research (IEDCR) [2]. The pandemic caused an enormous deprivation of our economy and its inhabitants. As the internet is available in both urban and rural areas and now people can express their emotions in social media i.e., Facebook, Twitter, Instagram etc. Our main target is to identify the mental illness of Bangladeshi people. Therefore, we would like to trace the mental health level of the social media users using their posts and Artificial Intelligence. Sentiment analysis is a powerful tool to determine the mental health of a user. We chose twitter platform as our social media to collect data. At the beginning, we collect data from twitter using twitter API, and then preprocess the data for our classifier. We have classified the data into three classes, Positive, Negative and Neutral (Ternary Classification) [3][4]. The accuracy of ternary classification is lower than binary classification (with only positive and negative polarity) [6] but trained on ternary-labeled data instead of binary-labeled, utilizing sentiment embedding from data sets made with different distant supervision methods [7]. We trained our model using those data and measured the performance of the model by different parameters i.e., Accuracy, Precision, Recall and F1 score. After that, we choose the best model based on different parameters. And finally, we have chosen three distinct tweeter accounts and test their posts using the model. We have performed comparison checking of tweets tweeted before and after the pandemic.

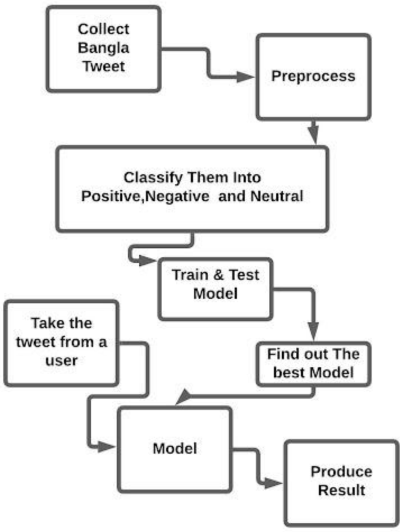


Figure 1: Process of basic sentiment analyzer.

II. Material And Methods

Data Collection

In this analysis, we have used two main data sets, one of them is prepared by us and another data set was taken from Kaggle open dataset (SentNoB)[8]. We prepare the twitter data set by collecting tweets from 7-2-2021 to 28-4-2021 and the data is filtered with the language 'Bangla'. We have collected nearly 2300 tweets and 1968 of them are suitable to use. Those data are not enough to train our model as Bangla is an extremely hard language as well as it has too many dialects [9]. So, we have considered another data set called SentNoB from Kaggle, where they have 12193 data collected online. We have merged two data sets and finally got a large enough data set where we have 14161 online data to train our model. After that we divide our data set into training data set and test set. We take 80% of our data as training data set and 20% of our data as test data set [10].

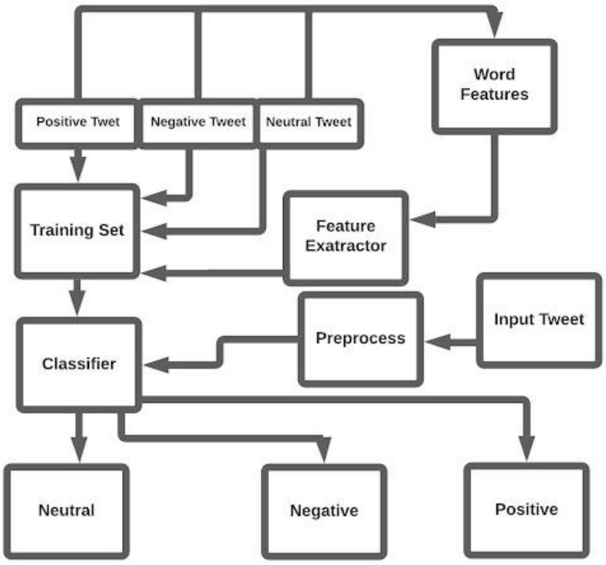


Figure 2: Sentiment analysis architecture

Positive	Negative	Neutral
5787	5119	3255

Table 1: Data Description

Pre-processing of the datasets

The following steps are done to preprocess the tweets.

1. Remove all URLs (example. www.abcd.com), hash tags (example #topic), targets (@username)
2. Remove mentions, reserved words (RT, FAV), emojis, smileys.
3. Remove all punctuation symbols, numbers , Stop Words.
4. Expand Acronyms

Model Training

There are two types of learning methods in machine learning. Supervised and unsupervised learning. We have used supervised learning to train our model. Supervised learning is divided into two sub-parts when data mining classification and regression. We have used classification. It is more convenient and powerful [11].

Evaluation of Sentiment Classification

We have some simple equations to calculate the accuracy, precision, recall, and F1 score.

i. Accuracy = (TP+TN)/(TP+TN+FP+FN) [12]

ii. Precision = TP/(TP+FP) [12]

iii. Recall = TP/(TP+FN) [12]

iv. F1 = (2×Precision×Recall) / (Precision + Recall) [12]

Where, TP stands for True Positive, FN for False Negative, FP for False Positive and TN is True Negative.

	Predicted Positive	Predicted Negative
Actual Positive	TP	FN
Actual Negative	FP	TN

Table 2: Confusion Matrix

III. Result

Naive Bayes

It is a simple technique to build a classifier. It is a probabilistic classifier. It is based on the Bayes' Theorem where it finds the probability of an event occurring given the probability of another event that has already occurred. It examines a set of documents that have been categorized. It consists of a group of algorithms where they share a common feature. We apply this method to train and test our model. The result of our analysis is

Table 3: Confusion Matrix for Multinomial Naive Bayes

	Neutral	Positive	Negative
Neutral	62	304	282
Positive	85	725	354
Negative	61	293	667

Table 3: Confusion Matrix for Multinomial Naïve Bayes

	precision	recall	F1-score	support
Neutral	0.30	0.10	0.14	648
Positive	0.55	0.62	0.58	1164
Negative	0.51	0.65	0.57	1021
Accuracy			0.51	2833
Macro avg	0.45	0.46	0.43	2833
Weighted avg	0.48	0.51	0.48	2833

Table 4: Classification Report for Multinomial Naïve Bayes

Support Vector Machine

Support vector machine is a supervised learning algorithm which is used in both regression and or classification purposes. The input data is nothing but two vectors with size m. Every data which is represented as a vector is assumed as a class. Secondly, we find a margin between the two classes which is far from any instance. The distance represents the margin of the classifier [5]. It uses a special technique called kernel. It simply converts the input dimension. It transforms the lower order dimension to higher order dimension. We apply this method to train and test our model. The result of our analysis is

	Neutral	Positive	Negative
Neutral	8	451	189
Positive	11	883	270
Negative	18	449	554

Table 5: Confusion Matrix for SVM

	precision	recall	F1-score	support
Neutral	0.22	0.01	0.02	648
Positive	0.50	0.76	0.60	1164
Negative	0.55	0.54	0.54	1021
Accuracy			0.51	2833
Macro avg	0.42	0.44	0.39	2833
Weighted avg	0.45	0.51	0.45	2833

Table 6: Classification Report for SVM

Random Forest

Random forest consists of a vast number of individual decision trees that operate as an ensemble. Each individual tree in the random forest expectorate out a class prediction. The class with the highest value is chosen as the model prediction. We apply this method to train and test our model. The result of our analysis is

	Neutral	Positive	Negative
Neutral	198	274	176
Positive	97	872	195
Negative	77	266	678

Table 7: Confusion Matrix for Random Forest

	precision	recall	F1-score	support
Neutral	0.53	0.31	0.39	648
Positive	0.62	0.75	0.68	1164
Negative	0.65	0.66	0.66	1021
Accuracy			0.62	2833
Macro avg	0.60	0.57	0.57	2833
Weighted avg	0.61	0.62	0.60	2833

Table 8: Classification Report for Random Forest

Multi-layer perceptron (MLP)

MLP is one kind of feedforward artificial neural network. It has at least three layers, an input layer, an output layer, and a hidden layer. All the nodes of the network are fully connected and work as a neuron and they use a nonlinear activation function. MLP promotes a supervised learning technique called backpropagation for training [13][14]. We train the MLP network with our train data set and test the network with the test data set. We set epoch (1 epoch is equal to 1 forward and 1 backward pass) =100 and batch size=32. We got a test accuracy as 87.31 % and a train accuracy as 88.79%.

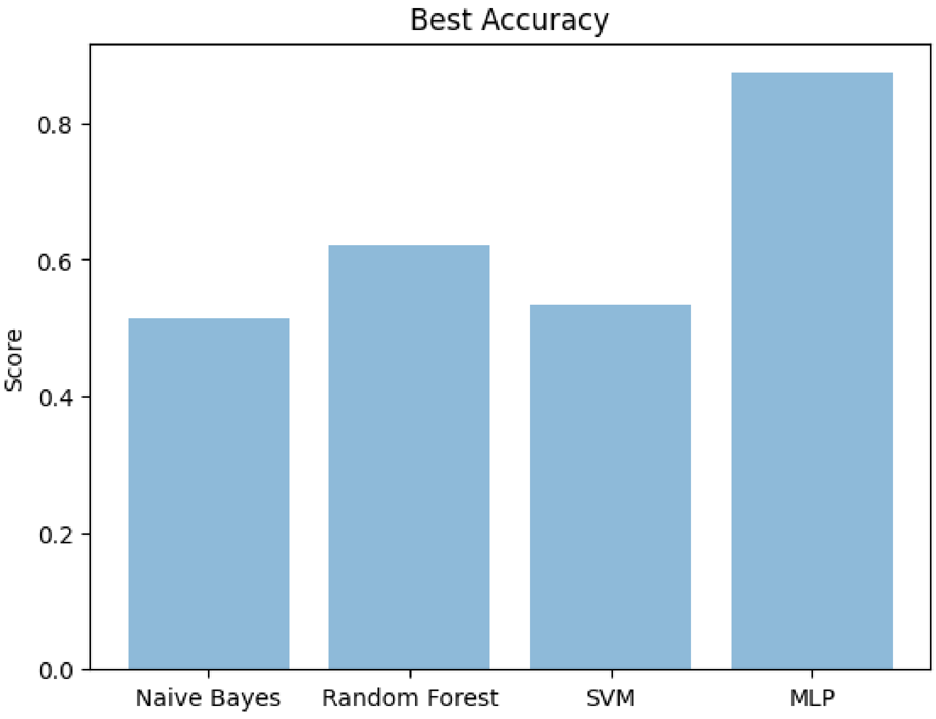


Figure 3: Model performance graph

Model Performance Summary

Here we use three classifier algorithms, the Naive Bayes, Random Forest, and Support Vector Machine and one artificial neural network the MLP. We apply those algorithms on the same data set. And get the outcome which are shown above. We can see from the figure Fig:3 that the MLP provide the best accuracy and it is 87.31% It does not have overfitting error. Random forest provides 62% and both SVM and Naive Bayes provide 51%. The main purpose of our work is to find out the mental illness of social media user, that is why we take ten tweeter account randomly and check their tweet written in Bangla. We find out the percentage of Negative tweets of a user before 5 months and after 5 months of the Covid-19 pandemic. We chose MLP to build our proposed model as it ensures the best accuracy. The result is shown in figure Fig:4.

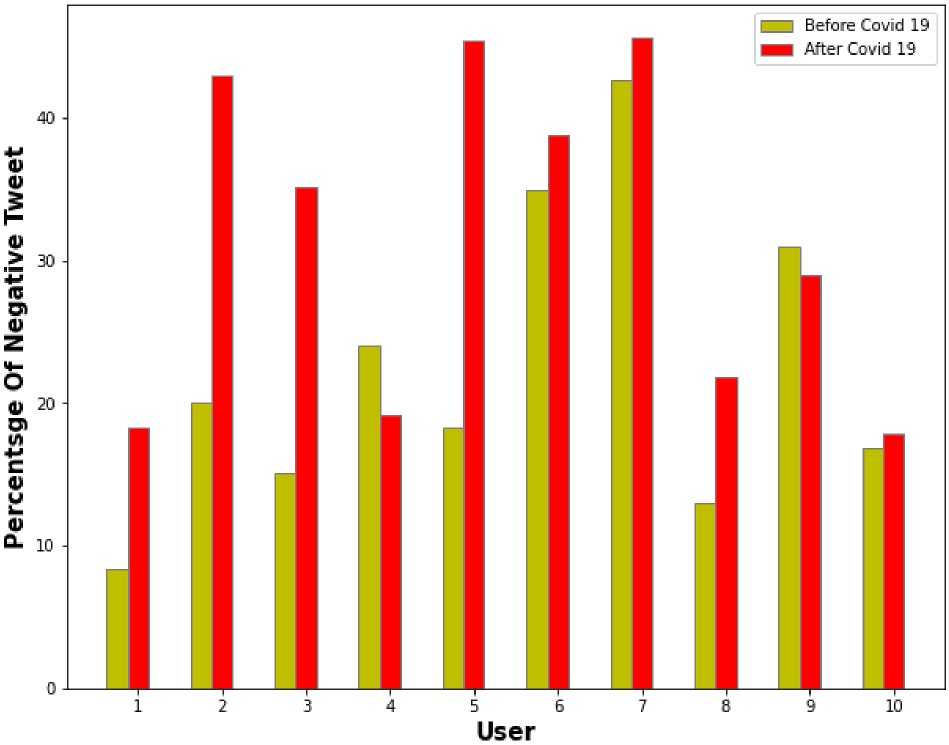
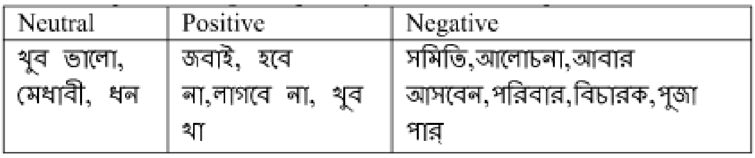


Figure 4: Percentage of Negative tweets of user 1 to 10

IV. Discussion & Conclusion

In this paper we present the mental condition of social media user. We chose twitter account as it is easy to collect the twitter data. Here we have used SentNotB dataset, and another dataset prepared by us. Our merged dataset contains nearly 15000 instances which are positive, negative, or neutral in label. We applied different machine learning approach to train our model using those data. We found that the MLP provide the best accuracy. We used this technique to train our model and choose randomly ten twitter account to test their tweet.We found that the frequency of negative tweets after the pandemic is more than before the pandemic. From figure 4 and 5 we can comment that most of the user’s post has negative polarity than before the pandemic.



Neutral	Positive	Negative
খুব ভালো, মেধাবী, ধন	জবাই, হবে না,লাগবে না, খুব খা	সমিতি,আলোচনা,আবার আসবেন,পরিবার,বিচারক,পূজা পার্

Table 9: Examples of some of the strongest words from each class

References

[1]. Novel Coronavirus (2019-nCoV) SITUATION REPORT - 1, 21 JANUARY 2020.

[2]. COVID-19 Situation Report No. 10 04 May 2020

[3]. Chen B, Cheng L, Chen R, Huang Q, Phoebe Chen Y-P. Deep neural networks for multiclass sentiment classification. In: IEEE 20th International Conference on high performance computing and communications, IEEE 16th International Conference on Smart City, IEEE 4th International Conference on Data Science and Systems 2018; pp. 854–59.

[4]. Sethi M, Pande S, Trar P, Soni P. Sentiment identification in COVID-19 specific tweets. In: International Conference on electronics and sustainable communication systems (ICESC 2020), pp. 509–16, https://doi.org/10.1109/ICESC48915.2020.9155674.

[5]. Cristianini, N., & Shawe-Taylor, J. An introduction to support vector machines and other kernel-based learning methods (Cambridge University Press, 2000).

[6]. Shakir, Anjume; Arora, Jyoti.ACCURACY IN BINARY, TERNARY AND MULTI-CLASS CLASSIFICATION SENTIMENTAL ANALYSIS-A SURVEY International Journal of Advanced Research in Computer Science; Udaipur Vol. 9, Iss. 2, (Mar 2018)

[7]. Duyu Tang, Furu Wei, Bing Qin, Nan Yang, Ting Liu, and Ming Zhou. 2016. Sentiment embeddings with applications to sentiment analysis. IEEE Transactions on Knowledge and Data Engineering, 28(2):496–509.

[8]. Khondokar Ittehadul Islam, Md Saiful Islam, Sudpta Kar, Muhammad Ruhul Amin.2021. A Dataset for Analysing Sentiment on Noisy Bangla Texts.

[9]. Muhammad Azizul Hoque. Chittagonian Variety: Dialect, Language, or Semi-Language? IIUC STUDIES ISSN 1813-7733 Vol.-12 December 2015 (P. 41-62)

[10]. The Pareto Principal H. Benjamin Harvey, MD, JD Susan T. Sotardi, MD Published:April 26, 2018

[11]. R. Quinlan, Learning efficient classification procedures and their applications to chess end games, in: R.S. Michalski, J. Carbonell, T. Mitchell (Eds.), Machine Learning: An Artificial Intelligence Approach 1, Tioga Publishing, Palo Alto, CA, 1983, pp. 463 – 482

[12]. Dalianis H. (2018) Evaluation Metrics and Evaluation. In: Clinical Text Mining. Springer, Cham.https://doi.org/10.1007/978-3-319-78503-5_6

[13]. Rosenblatt, Frank. x. Principles of Neurodynamics: Perceptron's and the Theory of Brain Mechanisms. Spartan Books, Washington DC, 1961

[14]. Rumelhart, David E., Geoffrey E. Hinton, and R. J. Williams. ”Learning Internal Representations by Error Propagation”. David E. Rumelhart, James L. McClelland, and the PDP research group. (editors), Parallel distributed processing: Explorations in the microstructure of cognition, Volume 1: Foundation. MIT Press, 1986.