Public Sentiment Analysis on COVID-19 Vaccination from Social Media Comments in Bangladesh.

Noor Nafeur Rahman	**170204034**
Rafsan Rahman	**170204036**
Tamanna Nazmin	**170204052**

Project Report

Course ID: CSE 4214

Course Name: Pattern Recognition Lab

Semester: Spring 2021

Department of Computer Science and Engineering

Ahsanullah University of Science and Technology

Dhaka, Bangladesh

March 2022

Public Sentiment Analysis on COVID-19 Vaccination from Social Media Comments in Bangladesh.

Submitted by

Noor Nafeur Rahman	**170204034**
Rafsan Rahman	**170204036**
Tamanna Nazmin	**170204052**

Submitted To

Faisal Muhammad Shah, Associate Professor

Md. Tanvir Rouf Shawon, Lecturer

Department of Computer Science and Engineering

Ahsanullah University of Science and Technology

Department of Computer Science and Engineering

Ahsanullah University of Science and Technology

Dhaka, Bangladesh

March 2022

ABSTRACT

One year during the pandemic of COVID 19, numerous viable possibilities have been created in worldwide efforts to create and disseminate a viable vaccine. The rapid development of numerous vaccinations is remarkable; generally, the procedure takes 8 to 15 years. The vaccination of a critical proportion of the global population, which is vital for containing the pandemic, is now facing a new set of hurdles, including hazardous new strains of the virus, worldwide competition over a shortage of doses, as well as public suspicion about the vaccinations. A safe and efficacious vaccine COVID-19 is borne fruit globally. There are presently more than a dozen vaccinations worldwide authorized; many more continue to be developed. This paper used COVID-19 vaccine-related public reactions to present an overview of the public's reactions to the current vaccination situation in Bangladesh. The authors have collected vaccine related-comments and posts from social media and applied traditional machine learning and ensemble classification algorithms and drew a comparison between the performances of each model. Results have shown that the Support Vector Machine(SVM) algorithm outperformed the other classifiers with 72.8% accuracy with unigram and bigram feature extracted.

Keywords: COVID 19, Vaccination, SVM, Feature extraction

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Problem Statement

Covid-19 has already been recognized as one of the deadliest pandemics worldwide. Strong Immunity is the only solution to stop such epidemic and a vaccine is the fastest means to reach there. To take control over the sparsity of this threat, the government of Bangladesh has taken enormous approaches to vaccinate the population as well as educate and make people aware of the necessity of taking the vaccine eradicating the misinformation about it. There are different opinions and judgments present among the general public. In our project, we will try to analyze the sentiment towards Covid vaccination in the context of Bangladeshi demographics using social media posts and comments. In this era of machine learning, it is way easier to analyze sentiment and use it for several noble purposes.

1.2 Motivation

The Covid-19 pandemic has shaken the entire world with its devastation. Millions of people died and suffered from it. Different vaccines have been discovered and experimented till now and the majority of them helped the immune system of the human body to fight against the virus. But we have noticed a mixed reaction to the vaccination process which is filled up with uncertainty, agitation and doubtful assumptions. We have felt a necessity to bring this to light and reveal the actual facts about the after-maths of covid vaccination so that we can decide what is better for us.

1.3 Uniqueness

For our project purpose, we have constructed a unique dataset that is human annotated. For that purpose, we have searched on popular social media platforms in Bangladesh. We have found that there are approximately 88% of Facebook users and 3% of Twitter users in 2021. These two platforms are optimal sources of data to achieve our target. A balanced dataset containing 3,807 data has been constructed for this research purpose.

Chapter 2

Literature Reviews

Recent studies have shown the versatility of the public sentiment on covid vaccine-related data on social media. Several approaches have been developed to classify, extract, and summarize crisis-relevant comments and information from social media. Puja et al. [1] collected 2,313 real-time data from Bangladeshi people and labeled them into three categories: Positive, Negative and Neutral. Further, they applied six popular classification algorithms which are Naive Bayes, Random Forest, SVM, Decision Tree, K-Nearest Neighbors(KNN), Logistic Regression and two deep learning algorithms, which are LSTM and CNN. Among them, CNN performed the best with 65.41% accuracy obtained. Charlyn et al. [2] implemented Naive Bayes classification model to classify English and Filipino language tweets into positive, neutral and negative polarities through the RapidMiner data science software. The result yielded 81.7% accuracy. Their model was trained on 993 tweets which were fetched among 11,974 tweets. Only TF-IDF feature was extracted within the data and 10-fold cross-validation were measured in order to get the performance. Md Tarique and Naseen [3] gathered 820,000 tweets all across the world to analyze the public sentient towards covid vaccination in different regions. They experimented the variety in positive, neutral and negative sentiments towards the covid-19 vaccines in their research. Ebelechukwu et al. [4] utilized machine learning models such as logistic regression, Support Vector Machines and Naive Bayes to develop baseline models while extracting TF-IDF. Furthermore, they developed Transformer-based models that provide classification of sentiments. In addition, the use of deep learning finetuned BERT language models were employed in providing classification of the sentiments. The results from our analysis showed that finetuning our dataset on a Covid-BERT v2 model performed better than the baseline models. A total of 14,725 tweets were extracted initially. In order to handle the imbalance dataset, random undersampling was executed and 9,702 data were chosen finally. The researchers split the dataset into 80-20 ratio as training and test sets.

Chapter 3

Data Collection & Processing

3.1 Data Collection

We have constructed a new dataset for our project purpose. Data for our study were retrospectively collected from two eminently popular social media platforms: Facebook and Twitter. Assorted Facebook pages and tweets that may contain urgent posts were our main focus. These include popular social organization page, renowned newspapers, news channels and education. The dataset was collected manually by professional data collectors and completely preprocessed and annotated by the authors.

3.2 Annotation Process

We have classified the collected data into two categories:

- **Positive:** The aspect of the comments refers to positive review. This category is annotated as ‘1’. e.g. করোনার সাথে আপোষ করার কিছু নাই। আমি এমনিতেই ভালো থাকবো। টিকা লাগবো

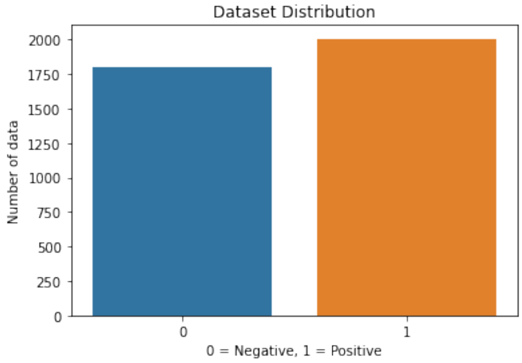

Figure 3.1: Distribution of Total Labeled Data

না(There is nothing to compromise with Corona. I'll be fine anyway. I don't need a vaccine).

- **Negative:** The comments conveys negative reviews . This category is annotated as ‘0’. e.g. আমি টিকা নিয়েছি, আমার এখন পর্যন্ত কোনো অসুবিধা হয়নি।(I got vaccinated, I haven't had any problems so far).

The dataset contains 3862 Bengali posts that include 53 duplicate values and 2 blanks. The number of data after removing those rows is 3,807.

Chapter 4

Methodology

In this section, we will be discussing our methodology of sentiment analysis on Covid-19 vaccine broadly. Fig. 4.1 represents our proposed approach. We have distributed our proposed approach into five phases. At the onset, our first phase contains dataset preparation. The second phase involves dataset preprocessing by utilizing a few natural language preprocessing techniques. Subsequently, the third phase consolidates the feature selection process. This process includes Term Frequency-Inverse Document Frequency(TF-IDF) and N-grams for both traditional machine learning models and boosting algorithms. The fourth phase is the preeminent phase of our proposed method that is the appliance of multiple machine learning models and boosting techniques to classify text as positive or negative. At last, we have examined the performance of each algorithm and presented the best-performed one in the last phase. In the following subsections, we have discussed all the phases briefly.

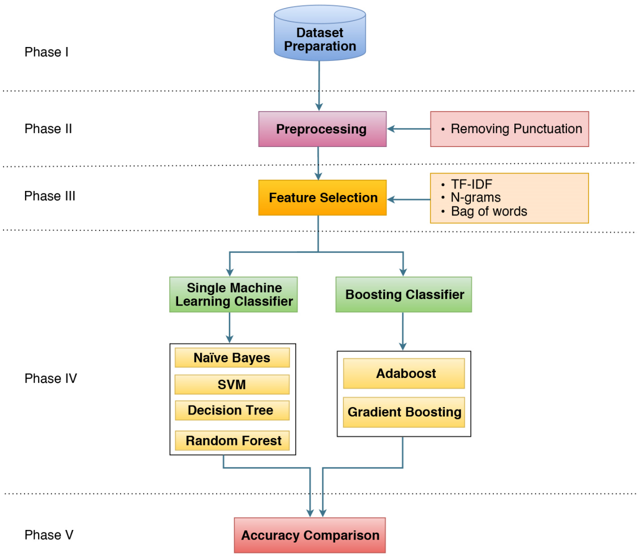

Figure 4.1: Proposed Model For Sentiment Analysis on Covid-19 vaccine

4.1 Phase I- Dataset Preparation

We have gathered 3,034 data from Facebook and Twitter written in Bengali. The overall dataset construction process is described in Chapter 3.

4.2 Phase II- Preprocessing

In the field of NLP, data preprocessing is the first step of building a model. Depending on how well the data has been preprocessed, the results are seen. It transforms text into a more digestible form so that machine learning algorithms can perform better. The foremost problems are the fact that the posts and tweets from Facebook and Twitter comprise abounding unnecessary information, symbols and errors. In order to clean the dataset, we have gone through the following steps of preprocessing:

4.2.1 Removing and altering unnecessary elements

There are countless errors in the public posts of social media. Facebook posts are very noisy and mostly contain errors. On the other hand, Twitter posts are very short and abbreviated that it is difficult to understand the meaning of a post sometimes. For the sake of a decent dataset we applied the following steps: There are countless errors in the public posts of social media. Facebook posts are very noisy and mostly contain errors. On the other hand, Twitter posts are very short and abbreviated that it is difficult to understand the meaning of a post sometimes. For the sake of a decent dataset we applied the following steps:

- Altering unintentional spelling mistakes: Spelling mistakes are fully corrected by human annotators manually. Four annotators were engaged to check the spelling of the complete dataset.
- Removing duplicate data, links and URLs, emoticons, hashtags and underscores and unnecessary signs, convert them if necessary.

It is not rational if all the steps stated for preprocessing are depleted with coding. The result will still enclose unnecessary elements for some of the steps. For a precisely preprocessed dataset, we have accomplished some of the steps manually by human annotators.

4.3 Phase III- Feature Selection

Feature selection is the third phase of our proposed model. We have used two feature extraction approaches: N-grams and TF-IDF. For traditional machine learning classifiers, we have used TF-IDF and N-grams methods.

4.3.1 N-gram

An N-gram is an N-character slice of a longer string. [5]

N-gram creates a document-term matrix where columns represent all columns of adjacent words of length n and cells represent count. It helps machines to get a better understanding of the meaning of the words in a particular context.

4.3.2 TF-IDF

TF-IDF (Term Frequency-Inverse Document Frequency) is a technique that measures how important a word is to a document in a collection of documents. Term Frequency(TF) measures how frequently a word occurs in a document. Suppose we have a number of N documents. Define fij to be the frequency of word i in document j. Then, the term frequency TFij can be defined:

$$TFij = \frac{f_{ij}}{max_k f_{kj}}$$

Inverse Document Frequency(IDF) measures how significant a word is. Suppose word i appears in ni of the N documents in the collection. Then, the Inverse Document Frequency IDFi can be defined: [6] [7]

$$IDF_i = log_2(N/n_i)$$

Each word or term has its own TF and IDF score. The TF and IDF product scores of a term is referred to the TF*IDF weight of that term. Simply we can state that the higher the TF*IDF score (weight) the rarer the term and vice versa.

4.3.3 Bag of Words

A very common feature extraction procedures for sentences and documents is the bag-of-words approach (BOW). In this approach, we look at the histogram of the words within the text, i.e.

considering each word count as a feature.The intuition is that documents are similar if they have similar content. Further, that from the content alone we can learn something about the meaning of the document.

4.4 Phase IV- Appliance of different algorithms

4.4.1 Traditional Machine Learning Classifiers

We have implemented multiple machine learning classifiers to choose the most suitable classifier for our model. The classifiers are Naive Bayes (NB), Decision Tree (DT) and Support Vector Machine (SVM), Random forest. Furthermore, we have applied Unigram, Bigram and also extracted the features using TF-IDF individually. We have split the dataset 80:20 training set-test set ratios to examine the performances. Additionally, we have run the model under 10-fold cross-validation also.

4.4.2 Boosting algorithms

Boosting is an ensemble learning strategy for minimizing training errors by combining a group of weak learners into a strong learner. A chunk of data is chosen, fitted with a model, and then trained progressively in boosting—that is, each model attempts to compensate for the shortcomings of its predecessor and generate a strong prediction rule. In this work our main goal is the appliance of Gradient boosting and Adaptive boosting(AdaBoost) techniques.

The concept of transforming weak learners into strong learners with improved accuracy was initially proposed by the authors of [8]. The authors of [9] and [10] then followed suit, making the boosting popular. They used it as a functional logistic regression approximation. ‘Gradient boosting’ is the name given to this method of issue solving. Gradient boosting is a robust ensemble method which can deal well with the sparsity of dataset. It can optimize different loss function and provides several hyper parameter options that make the function fit with flexibility.This is why gradient boosting performs so well on binary classification models. We have also used AdaBoost algorithm [11] based on Decision Tree classifier for its performance with weak learners.

In our research, both the algorithms were trained and validated with the same train-test splits and validation sets after the feature selection process was conducted. The boosting algorithms used here generally have several hyperparameters. We used Grid Search to find the best hyperparameter set. The simulation procedure is the same as the traditional classification approach keeping the training data and feature selection techniques same for comparison purposes.

Chapter 5

Experiments and Results

In this section, we displayed the performances of all the algorithms with the extraction of features individually. We conduct our observation by splitting the dataset into 80:20 as well as implementing 10-fold cross-validation for both machine learning classifiers and boosting algorithms. K-fold cross-validation is used to estimate how precise the accuracy of a model on a limited data sample presents.

Table 5.1: Comparison of Different Algorithms Using Different Features Performed with 10-fold

		Feature Sets		
Classifiers	Evaluation Metrics	TF-IDF	Unigram + Bigram	Bag of words
Naïve Bayes	Accuracy	0.530	0.627	0.510
	Precision	0.758	0.726	0.748
	Recall	0.235	0.535	0.188
	F1 score	0.358	0.616	0.300
SVM	Accuracy	0.669	0.722	0.664
	Precision	0.674	0.769	0.695
	Recall	0.727	0.677	0.650
	F1 score	0.699	0.720	0.672
Random Forest	Accuracy	0.657	0.656	0.661
	Precision	0.711	0.718	0.716
	Recall	0.593	0.576	0.596
	F1 score	0.647	0.639	0.650
Decision Tree	Accuracy	0.630	0.635	0.643
	Precision	0.687	0.696	0.705
	Recall	0.622	0.617	0.622
	F1 score	0.653	0.654	0.661
Adaboost	Accuracy	0.613	0.627	0.617
	Precision	0.582	0.663	0.609
	Recall	0.909	0.607	0.666
	F1 score	0.709	0.634	0.636
Gradient Boosting	Accuracy	0.678	0.694	0.655
	Precision	0.711	0.715	0.643
	Recall	0.685	0.716	0.718
	F1 score	0.698	0.716	0.678

We have evaluated the performance in both approaches of splitting and cross-validating. Addi-

Table 5.2: Comparison of Different Algorithms Using Different Features Performed with 10-fold

		Feature Sets		
Classifiers	Evaluation Metrics	TF-IDF	Unigram + Bigram	Bag of words
Naïve Bayes	Accuracy	0.540	0.672	0.528
	Precision	0.755	0.750	0.747
	Recall	0.188	0.565	0.157
	F1 score	0.300	0.644	0.258
SVM	Accuracy	0.687	0.728	0.686
	Precision	0.692	0.769	0.719
	Recall	0.732	0.694	0.665
	F1 score	0.711	0.729	0.690
Random Forest	Accuracy	0.689	0.684	0.686
	Precision	0.728	0.747	0.737
	Recall	0.655	0.609	0.629
	F1 score	0.689	0.669	0.678
Decision Tree	Accuracy	0.646	0.660	0.654
	Precision	0.675	0.700	0.686
	Recall	0.635	0.622	0.635
	F1 score	0.654	0.659	0.659
Adaboost	Accuracy	0.639	0.644	0.645
	Precision	0.641	0.664	0.664
	Recall	0.741	0.658	0.661
	F1 score	0.683	0.661	0.662
Gradient Boosting	Accuracy	0.684	0.686	0.681
	Precision	0.696	0.702	0.706
	Recall	0.718	0.690	0.679
	F1 score	0.706	0.698	0.688

tionally, we have examined the effects of Unigram and Bigram, TF-IDF and Bag of Words(BoW). The simulation procedure is the same in both traditional classifiers and boosting algorithms approaches. We kept the training data and feature selection techniques same for the sake of ideal comparison. In Table 3 and Table 4 we have shown the accuracy, precision, recall and F1 scores for each of the algorithms with individual features. Table 1 specifies the performance record for 80:20 split. Table 2 specifies the performance record of each algorithms for 10-fold cross-validation.. From Table 2, it is clearly noticeable that SVM performed really well on 10-fold cross-validation while being trained with Unigram and Bigram. It shows 72.8% accuracy.

Chapter 6

Future Work and Conclusion

Sentiment Analysis of COVID-19 Vaccination Process in Bangladesh Using Machine Learning Algorithm from Bengali Text Dataset is the subject of our research. In our work and model, we have some limitations and deficiencies. The data set we used was not particularly large; in order to achieve good accuracy, it would have been preferable to use a larger and more diverse data set. We were unable to collect data from the people of various occupations, counties, and social classes due to some limitations. And we will also use deep learning models for better performance.

References

[1] P. Sarker and D. K. Sarkar, "Sentiment analysis of general peoples reaction about covid19 vaccination in bangladesh using machine learning algorithm from bengali text dataset," 2021.

[2] C. Villavicencio, J. J. Macrohon, X. A. Inbaraj, J.-H. Jeng, and J.-G. Hsieh, "Twitter sentiment analysis towards covid-19 vaccines in the philippines using naïve bayes," *Information*, vol. 12, no. 5, p. 204, 2021.

[3] M. T. J. Ansari and N. A. Khan, "Worldwide covid-19 vaccines sentiment analysis through twitter content.," *Electronic Journal of General Medicine*, vol. 18, no. 6, 2021.

[4] E. Nwafor, R. Vaughan, and C. Kolimago, "Covid vaccine sentiment analysis by geographic region," in *2021 IEEE International Conference on Big Data (Big Data)*, pp. 4401–4404, IEEE, 2021.

[5] W. B. Cavnar, J. M. Trenkle, *et al.*, "N-gram-based text categorization," in *Proceedings of SDAIR-94, 3rd annual symposium on document analysis and information retrieval*, vol. 161175, Citeseer, 1994.

[6] A. Rajaraman and J. D. Ullman, *Mining of massive datasets*. Cambridge University Press, 2011.

[7] S. Sarker, "Bnlp: Natural language processing toolkit for bengali language," *arXiv preprint arXiv:2102.00405*, 2021.

[8] R. E. Schapire, "The strength of weak learnability," *Machine learning*, vol. 5, no. 2, pp. 197–227, 1990.

[9] J. Friedman, T. Hastie, and R. Tibshirani, "Additive logistic regression: a statistical view of boosting (with discussion and a rejoinder by the authors)," *The annals of statistics*, vol. 28, no. 2, pp. 337–407, 2000.

[10] J. H. Friedman, "Greedy function approximation: a gradient boosting machine," *Annals of statistics*, pp. 1189–1232, 2001.

[11] Y. Freund and R. E. Schapire, “A decision-theoretic generalization of on-line learning and an application to boosting,” *Journal of computer and system sciences*, vol. 55, no. 1, pp. 119–139, 1997.

Generated using Undegraduate Thesis LaTeX Template, Version 1.4. Department of Computer Science and Engineering, Ahsanullah University of Science and Technology, Dhaka, Bangladesh.

This project report was generated on Wednesday 9th March, 2022 at 7:42am.