



Sentiment Analysis of Covid-19 Vaccination in Under-Resourced Bangla Mixed-Text from Social Media

Authors

Mahamudur Rahaman Khan
Student ID: 170042017

Md Fuadul Islam
Student ID: 170042069

SM Nawsad Rahmatullah
Student ID: 170042084

Supervisor

Dr. Md. Azam Hossain
Assistant Professor, Department of CSE, IUT

A thesis submitted to the Department of CSE in partial fulfillment of the requirements for the degree of B.Sc. in Software Engineering

**Department of Computer Science and Engineering (CSE)
Islamic University of Technology (IUT)
A Subsidiary organ of the Organization of Islamic Cooperation
(OIC)**

Academic Year: 2020-2021

May 2022

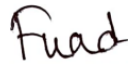
Declaration of Authorship

This is hereby declared that the work presented in this literature is the result of scrutinized experiments carried out by the candidates under the supervision of Dr. Md. Azam Hossain in the Department of Computer Science and Engineering, Islamic University of Technology, Gazipur, Dhaka, Bangladesh. In addition, neither this thesis nor any part of this thesis has been included in any degree, diploma, or other certifications to this or any other institution. The guidelines of conduct have been acknowledged and respected by the authors, along with existing literature from their respective authors, which are mentioned at the closing chapter.

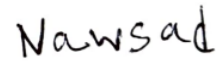
Authors:



Mahamudur Rahaman Khan
Student ID: 170042017
Academic Year: 2020-21

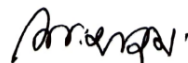


Md. Fuadul Islam
Student ID: 170042069
Academic Year: 2020-21



SM Nawsad Rahmatullah
Student ID: 170042084
Academic Year: 2020-21

Supervisor:



Dr. Md Azam Hossain
Assistant Professor
Department of Computer Science and Engineering
Islamic University of Technology (IUT)

Abstract

Vaccine reluctance is one of the top ten global health concerns to confront. In this day and age, social media plays a critical role in disseminating vaccination information, even material that is incorrect or misleading. Monitoring the emotion expressed in vaccine-related social media interactions can assist the health authority in introducing the public safety procedure and guiding the government in developing appropriate policies. Newly developed vaccines for COVID-19 are causing widespread reactions all around the globe. Trust is an essential factor to success in vaccine inoculation and sentiment analysis may help assess public opinion. Social media is prevalent in Bangladesh where more than 80 million Internet users express their opinions in Bangla, English, and a mixture of Bangla and English text which are commonly referred to as codemixed language. Since sentiment analysis on Bangla has not progressed significantly compared to other prominent languages like English, this proved to be a major undertaking on our part. In this paper, we propose a method for determining vaccination-related sentiment from public comments on Facebook written in Bangla, English, or a combination of both texts. The proposed model is constructed on the basis of the multilingual BERT model. It achieves a validation accuracy of around 97.3% and a training accuracy of approximately 98.8%.

Contents

1	Introduction	1
1.1	Problem Statement	4
1.2	Research Challenges	4
1.3	Thesis Objective	4
1.4	Key Contribution	5
1.5	Thesis Organization	5
2	Background	6
2.1	What is vaccine hesitancy?	6
2.2	Factors that influence vaccine hesitancy	6
2.3	Consequences of vaccine hesitancy	7
2.4	Codemix	7
2.5	Proposed Solution	8
3	Literature Review	10
3.1	LETS: A Label-Efficient Training Scheme for Aspect-Based Sentiment Analysis by Using a Pre-Trained Language Model	11
3.2	Sentiment analysis with NLP on Twitter Data	13
3.3	Sentiment Analysis of Comment Texts Based on BiLSTM	14
3.4	Detecting Multilabel Sentiment and Emotions from Bangla YouTube Comments	15
3.5	Important Factors	16
4	Dataset	19
4.1	Dataset Labelling	19
4.2	Data Augmentation	20
4.3	Dataset Undersampling	21
4.4	Dataset Pre-Processing	21
4.5	Proposed Dataset:	21

5	Methodology and Experiment	23
5.1	Foundational Methodology	23
5.2	Training Arrangement	28
5.2.1	Activation Function	28
5.2.2	Optimizer	29
5.2.3	Momentum	29
5.2.4	Root Mean Square Propagation(RMSP)	30
5.2.5	Loss Function	31
5.3	Training Experiments	32
5.3.1	Bangla-BERT With Early-Stopping	32
5.3.2	Multilingual BERT Without Early-Stopping	32
5.3.3	Multilingual BERT with Early-Stopping	33
6	Result and Discussion	34
6.1	Result metrics	34
6.1.1	Precision	34
6.1.2	Recall	34
6.1.3	F1-Score	35
6.1.4	Confusion Matrix	35
6.1.5	Total polarity score	35
6.2	Model Evaluation	35
6.2.1	Bangla-BERT With Early-Stopping	36
6.2.2	Multilingual BERT Without Early-Stopping	36
6.2.3	Multilingual BERT with Early-Stopping	37
6.2.4	End to End Inference	38
6.3	Discussion	39
7	Conclusions	41
7.1	Reflection	41
7.2	Future work	42
7.3	Conclusion	43

List of Figures

1.1	Social media usage of the last one year [42]	3
1.2	Bangladesh Coronavirus Vaccination Rate: Any Dosage[17]	3
2.1	Example of a Bengali codemixed word with English and Bengali	8
4.1	CoVaxBD word cloud from Bangla text samples.	22
4.2	CoVaxBD word cloud from English text samples.	22
5.1	General architecture of the BERT	24
5.2	Attention mask for both X and Y set of tokens.	25
5.3	An instance of input for BERT embedding layer.	26
5.4	Model architecture.	27
5.5	Learning curve for Bangla-BERT with early-stopping.	32
5.6	Learning curve for Multilingual BERT without early-stopping.	33
5.7	Learning curve for Multilingual BERT with early-stopping.	33
6.1	Confusion matrix for Bangla-BERT with early-stopping.	36
6.2	Confusion Matrix for Multilingual Bert without early-stopping	37
6.3	Confusion matrix for Multilingual-BERT with early-stopping.	37
6.4	End-to-end inference example.	38

Chapter 1

Introduction

The coronavirus COVID-19 pandemic has changed the life of every person in the world. Ever since the pandemic, the whole world has struggled against it. After almost two years, it has affected 222 countries and territories by 30th December, 2021[47]. Governments all around the world are trying to fight against it by taking various measures like imposing lockdowns, making people maintain social-distancing, wearing masks etc. The latest weapon in this list is the vaccine. With the rolling out of the first vaccine, a lot of questions have been raised in social media. People are sharing their thoughts about the pandemic on platforms like facebook, twitter, youtube, reddit etc. Social media usage is increased a lot due to the lockdowns. Like other countries, the Government of the People's Republic of Bangladesh is trying to vaccinate its people. Runu Veronica Costa, a senior staff nurse at Kurmitola General Hospital, Dhaka received the first ever shot of Covid-19 vaccine in Bangladesh on January 27, 2021[38] and the vaccination campaign for general mass started on February 7, 2021[20]. But, vaccination topic produced fear, rumors, misinformation among the general public. As a result, only 26.9%[8] of people have been vaccinated as of January 16,2022 and 52.6%[8] of people at least received one dose of the vaccines. The government is very concerned about the issue as it is hampering the vaccination campaigns.

Coarse-grained sentiment analysis technique classifies emotions into three polarities which are positive, negative and neutral. While there are more fine grained systems that classify sentiments into even more categories such as sad, annoyed, official and joking, they are typically not that more useful in our sentiment analysis as the main goal is to find out the overall mood and take steps according to it. Most vaccine sentiment analysis therefore lean

into the category of coarse-grained sentiment. So our work also includes these three sentiments which can capture the public sentiment regarding covid-19 vaccine [29].

One of the most challenging aspect of the task was to find good quality dataset that has a reasonable probability of capturing various opinions on the covid-19 vaccination scenario to make it less prone to over-fitting on a small range of opinion. Secondly cleaning and annotating properly was another key aspect that needed to be done correctly as a well cleaned and annotated dataset provides good quality training result for the model. Finally there was an issue with training time as the large dataset with a sophisticated model can require a lot of computation power. So, we had to go with vertical scaling of our hardware which required expensive CPU and GPU capability.

Table 1.1: Social media usage of Bangladesh

Social Media	Usage
Facebook	92.65%
Youtube	4.81%
LinkedIn	0.95%
Pinterest	0.49%
Instagram	0.42%
Twitter	0.4%

We choose facebook over all kind of social media to scrap our data from because facebook alone contributes to 92.65% [Table 1.1] of total social media usage in Bangladesh [42]. We can say from Figure- 1.1 that most Bangladeshis are not present in other social media platforms as much as in facebook. Facebook is a mirror to Bangladeshi society.

We also observe that other social media sites(except for twitter for a brief period of time) do not pick up steam even during delta variant, meaning that no other social media trend will represent the broad user base of Bangladesh like facebook on a consistent basis.

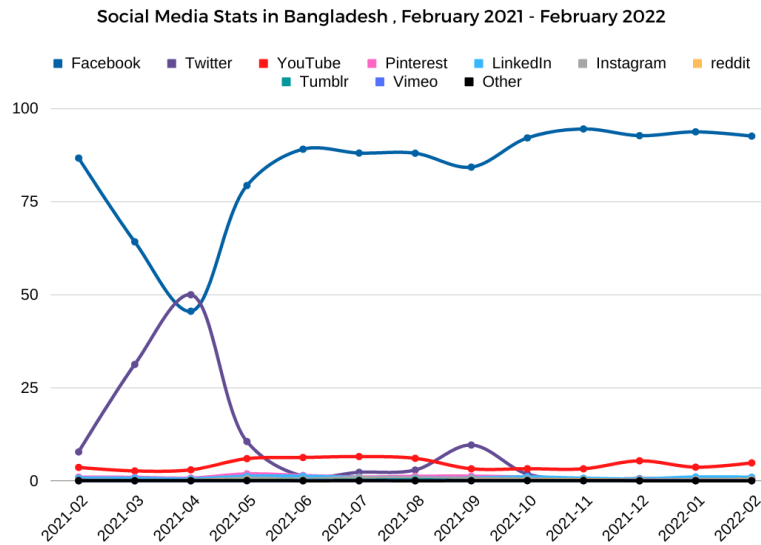


Figure 1.1: Social media usage of the last one year [42]

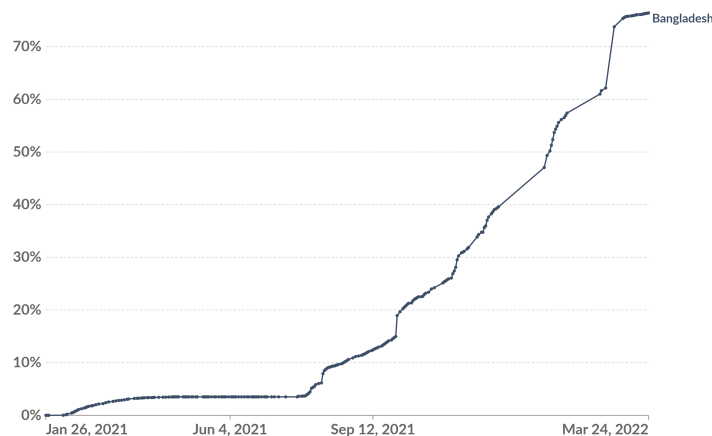


Figure 1.2: Bangladesh Coronavirus Vaccination Rate: Any Dosage[17]

As we can see from the graph above that the COVID-19 vaccination rate in Bangladesh was going pretty slowly. From our background study we concluded that vaccine hesitancy may play a part in it since sentiment towards vaccine can be quite elastic from the public. Specially if that vaccine is being talked about in the social media a lot. From January 26th September 12, the vaccination rate barely hovered around 10%. This encouraged us to carry out sentiment analysis on the Bangladeshi public on social media regarding the COVID-19 vaccine.

1.1 Problem Statement

Given a set of Bangla datasets that are often mixed with English,

- Find the sentiment score.
- Predict the overall sentiment of a person's expression in the social media network.
- Provide good accuracy compared to similar models.

1.2 Research Challenges

There were many research challenges we had to face along the way. They are mentioned below:

- Finding good dataset that accurately conforms with the vaccination topic. Since there are various social media platforms, finding the right platform that allows for data collection through api was challenging as many platforms do not allow for simple api data collection.
- Secondly the dataset we collected did not come in pure form and had mixtures of Bangla and English words. In some cases we had to translate the English into Bangla or exclude problematic expressions altogether.
- The The third problem we faced in our implementation is to make a model that is not too overfitting and is able to classify unique vaccine related expressions outside of our training and test dataset. We applied regularization with dropout to get around this problem.
- Finally training a huge dataset with our BERT model was computationally expensive and required a lot of time.

1.3 Thesis Objective

The main objectives of this thesis can be described below.

- Creating a comprehensive Bangla dataset with accurate labels.

- A multilingual sentiment analysis model that is capable of predicting sentiment of both codemixed Bangla and English as well as individual languages.
- Improvement in terms of accuracy over the comparable models.

1.4 Key Contribution

Here are the key contributions of our work.

- Our primary contribution to this work was creating a robust dataset that correctly reflects the varying opinion of Bangladeshi people regarding COVID-19. Experts opinion were considered to remove any sort of bias and subjectivity though this is an area that can still be improved.
- Taking into account the context of surrounding text as many of the comparable models did not take into account the surrounding context of words which made it difficult for accurate sentiment analysis. Context of each of the word that surrounds it was taken into account to assign a more accurate weight, instead of assigning an arbitrary weight or using TF-IDF to assign weight.
- The BERT model's embedded layer was retrained with the training dataset. Instead of using the pre-assigned weights in the BERT model's embedded layer, the whole layer was retrained with the training dataset. As a result it produced better F1-score and training accuracy.

1.5 Thesis Organization

The remaining segments of the thesis has been included in the following configuration:

Chapter 2 consists of the background study to better understand this literature. The following chapters 3, 4, 5, and 6 provide a comprehensive view of the literature review, dataset, methodologies, and results in order. Finally, chapter 7 draws a conclusion to this literature.

Chapter 2

Background

2.1 What is vaccine hesitancy?

Vaccine hesitancy is delaying in acceptance or refusing a vaccine despite being available [4]. Historically vaccine hesitancy emerged when vaccines were first introduced way back in the 18th century.[11],[27]. Individuals that show this sort of trait of refusing vaccines is commonly known as "anti-vax". Some may refuse to take a certain vaccine while refusing others [11].

2.2 Factors that influence vaccine hesitancy

There are a few factors that influence vaccine hesitancy in public. Among the most common are:

- Health Concerns
- Vaccine Administration
- Social-Attributes of Individuals

Among the vaccine-related factors, mistrust in the safety and effectiveness of the vaccine causes vaccine hesitancy among the skeptical public.[35]. But a very significant factor that has led to high vaccine hesitancy and skepticism is disinformation in social media.[28]. A study in 1998 associated measles, mumps and rubella vaccine with autism which resulted in public mistrust for the vaccine [25]. The second factor that influences vaccine hesitancy is public health related system. If a health system that has a history of mistreatment and inefficiency, then the public hesitation for vaccine from that health system grows. The Third

major system that can increase vaccine hesitancy is different religious beliefs of an individual or a lack of knowledge about vaccines and public safety [22], [50]. Low education rate also points to increased vaccine hesitancy [16].

2.3 Consequences of vaccine hesitancy

Vaccine hesitancy can lead to increased likelihood of epidemic diseases. For example, unwillingness to take vaccine has resulted in uptick of measles outbreak in Europe and unvaccinated individuals make up the majority of the cases [36],[46]. That's why identifying current sentiment about covid-19 vaccination is key in tackling misinformation and to understand the amount of actions that are needed to be taken by the government to get people vaccinated. Specially in a developing country where access to trusted and reliable information can be challenging due to technological barrier.

2.4 Codemix

Codemix refers to the use of multiple languages together [13]. Codemixing mixes two or more languages to express meanings or sentiments. Codemixed sentences usually have two or more languages, often in singular grammar or using different grammars for unique languages found in a sentence [13]. Pure form usually indicates non code mixed expressions where there are no more than one languages involved, either contextually or in written form. This results in a challenging sentiment analysis problem as different grammars from different languages can convey different meaning based on the context.

One of our main motivation to undertake sentiment analysis is to tackle codemixing commonly found in social media. Codemixing is a very common phenomenon among Bangla social media users as writing in pure Bangla can be cumbersome for a lot of people due to the lack of a clear keyboard layout and the difficulty of writing Bangla with many different letter combinations.

Don't take this vaccine মরার শখ নাই ভাই|

Figure 2.1: Example of a Bengali codemixed word with English and Bengali

The above picture shows a common codemixed language expression found among Bangla social media users. The most common form of codemixed language among Bangla users can be found with the mixture of Bengali and English letters. Often times both Bangla and English are used to communicate expression with the audience in the social media where both proper Bangla and English is used. Other times we can observe the use of pure English letters to express Bangla words, commonly known as Bangla written in English(Banglish).

2.5 Proposed Solution

To solve the issue described so far, the present paper analyzes the general sentiment of People's Republic of Bangladesh on the topic of vaccination. Human languages often use structural and grammatical expressions that can be quite difficult for the computer to understand. Therefore, making the computer understand natural language accordingly can be a daunting task. Natural Language Processing (NLP) aims to process the language used by humans understandable for the computer to understand. Sentiment analysis has become a big phenomenon in NLP. As many sentences and opinion reflect different sentiment on various topics, sentiment analysis can be quite helpful in determining the public opinion on something useful. Sentiments may even provide useful information for various people groups to make decision based upon it. Ever since the COVID-19 pandemic approached, the opinion among Bangladeshi people in social media has exploded and has led to a lot of different sentiment on the vaccines being developed.

In this study, we present model that can correctly analyze the sentiment of the general public regarding COVID-19 vaccine. Since the early days of natural language processing, sentiment analysis has been a popular tool for analyzing human emotions. Sentiment Analysis is a crucial to determine the sentiment towards something without the aid from a human.Sentiment

analysis is used to identify the specific segment on a variety of topics. For example, the experience of a customer in a restaurant and whether the food served was good or not. This helps determine the approach a restaurant owner might take. In modern times there is a requirement for handling large volume of data quickly and respond to the sentiment expressed in a short period of time. This is where sentiment analysis comes into play. But developing an accurate model that is always accurate is the tricky part which can provide quick solutions at the same time.

Finding suitable datasets to train an NLP model for classification, specifically for Bangla and mixture of English and Bangla, is one of the key challenging aspects of the process. Due to a lack of annotated corpora, named dictionaries, and morphological analyzers, little research has been done on Bangla text, especially mixed of Bangla and English text for social media sentiment analysis. The following are the key contributions made by this work in order to address the difficulties raised above:

- We have collected datasets from social media users that are more accurate representations of real-world events and sentiments. Our collection contains text in both Bangla and English, as well as Bangla with a variety of local accents.
- This work presents a model for determining the sentiment regarding the vaccination program of the mass audience, which is commonly expressed in Bangla or English, or a mixture of both English and Bangla in text. The proposed model is built on the basis of the multilingual BERT model.
- A text classifier based on BERT which does not arbitrarily assigned weights to the word vectors or uses traditional TF-IDF but rather uses the weights.

Chapter 3

Literature Review

In relation to the current process being implemented, there are quite a few related works that caught our attention in this field that has been shown in the related works table. Sentiment analysis using TF-IDF on twitter data is a related work that uses traditional NLP methods. However it is only limited to two classes and does not use state of the art models like *BERT*[15].

Another interesting one is sentiment analysis using aspect based NLP methods where data augmentation is used to auto label based on aspects [41]. But this process uses survey data so aspects are much more limited whereas our process includes social media comments that are much more, numerous and varied in aspects. Plus it uses more iterations than necessary which reduces performance. In addition, the task specific pre-training method is not clearly stated and it's not shown whether this methodology is the one that can improve performance. It also only supports a single language. The sentiment analysis using BiLSTM uses a binary classification with a single language [48]. But it has a very lengthy training time and only classifies English language. There are also lexicon based sentiment analysis found in Paper 4 that crawls through reddit platform. However it does not state it's accuracy and only limited to a single language.

Most of the models that we have found are single language models focused on only one language. We did find one research work that did hotel research work with BERT model featuring deep neural networks, but it only contained two classes and only had a single language. Though in another paper which measures sentiment by geographic region, we saw that simply using BERT encoder is not enough as it still had a relatively low accuracy for 3 classes. It

does show that the BERT encoder model improves on its accuracy over traditional baseline models such as Logistic Regression, SVM and Naive Bayes, however. Traditional Non Neural Network methods with SVM and KNN also works but we have found that they tend hover just under 90 in terms of accuracy. They also tend to not catch the surrounding context around a sentence unlike BERT model, so it becomes rather difficult for them to properly classify Neutral class.

The main takeaway here is that most of them classify single language and lacking in Bengali language classification specifically on the topic of covid. We propose to improve on that by using BERT model to capture the surrounding aspects that predict classification of Positive, Negative and Neutral classes accurately. The next sections will discuss the individual literature in more detail.

3.1 LETS: A Label-Efficient Training Scheme for Aspect-Based Sentiment Analysis by Using a Pre-Trained Language Model

Label-Efficient-Training-Scheme is a scheme to improve the labelling of datasets in NLP using automated labelling instead of manual labelling. Label-Efficient Training Scheme consists of three phases such as, task-specific pre-training to label unlabeled data, label augmentation to maximize labelled data and active learning to label data strategically. Pre-trained language models are often quite good at learning contextualized language, however they still require labelling[9]. Since our goal is to find out the sentiment of various written expression, labelled dataset through automated labeling is a good pathway for our intended work[41]. As we are trying to analyze sentiment on facebook comments, this analysis shows us an efficient way we to label data instead of doing it manually. Aspect Based Sentiment Analysis is a fairly new phenomenon compared to Non Aspect Based which can aid us in getting greater accuracy in our own experiment.

The literature proposes a workflow that collected datasets using crowd sourcing platform like twitter. A total of 1,000 participants with 12,000 answers were used and among them a

random subset of data was used to manually apply them before applying the LETS model. The dataset was divided to use active learning algorithms and label augmentation technique. Task-specific pre-training is used to exploit the unlabelled task-specific corpus data (or linguistic data) and Label them. A novel pre-training strategy which uses mask language modelling (MLM) from the well known BERT model. During MLM the input is formulated with a sequence of tokens that are randomly masked out with a special token at a certain percentage. Based on the data findings, Aspect based sentiment analysis model using BERT encoder was used to determine the accuracy of Labelled data. Two more important was used to evaluate the model.

As per the ACD metric in micro average polling, LETS significantly outperforms other labelling schemes and reduces the manual labelling Cost by 2-3 times. However there are some drawbacks to the procedure they used. The dataset is semi-realistic, meaning they are not real feedbacks but rather surveys taken from people which could create bias as they may not be properly randomized. Moreover the aspects that were defined had Majority and Minority classes based on importance. But those classes were based on frequency rather than empirical evidence. Plus after a certain iteration, the LETS model tends to return diminishing results and uses more iterations than necessary which unnecessarily speeds up the computation process.

The model takes into account the various sentiment of a sentence expressed in different words by making a sentence pair classification of input data where the sentence is paired with the aspect category and it's sentiment and tries to make an overall prediction on whether its positive or negative. It uses label augmentation and active learning for faster pre-processing. It claims to reduce the cost of labeling by at least 2-3 times and performs better than random sampling. It also suffers from limitations such as semi realistic dataset and more than necessary iterative steps leading to diminishing returns.

3.2 Sentiment analysis with NLP on Twitter Data

The next paper emphasizes the use of social networks like twitter and how it can reflect public sentiment regarding smartphone brand. Using NLP in TF-IDF(Term- Frequency Inverse domain Frequency), it claims to have achieved close to 86% accuracy in Sentiment Analysis. Since our project also involves gathering data from a social media site like facebook, it goes without saying that a model that involves in making prediction based on social media is a good model we could work on. Although this particular research is based on Sentiment of two major smartphone brands, it explores a TF-IDF model that has shown promise in making accurate prediction in Sentiment analysis. A lot sentiment analysis is based on Statistical model uses TF-IDF.

The paper briefly discusses related works that occurred in similar sentiment analysis. It specifically mentions *Parts of Speech Polarity* and a tree kernel for classification and also traditional non neural networking methods like *K-Nearest-Neighbour*. The proposed workflow included fetching data from twitter API, then using an NLP toolkit for preprocessing. This pre-processing includes Tokenizations, Stemming, Lemmatization, Parts of Speech tagging and various other methods. The NLP framework then used two algorithms which incorporated *Bag of Words* and *TF-IDF* to filter tweets. It was then fed into the classifier model which analyzed the sentiments.

While it shows good results with an accuracy up to 86 percent, it falls short of much higher accepted accuracy. Also it does not provide alternative comparison to unsupervised neural networking method which are much better at learning sentiments from given data. It only compares between other Machine Learning methods like *SVM*, *Naive Bayes* etc. To summarize, twitter sentiments were classified using TF-IDF and Bag of Words that analyzed the general sentiment of two big smartphone brands, Samsung and iPhone. Based on the related works with tree-kernel and other traditional machine learning methods like K-Nearest-Neighbour, it built a model to classify whether the sentiment was positive or negative for the two brands. It shows impressive results at up to 86 percent, though fails to compare itself with other neural

network models.

3.3 Sentiment Analysis of Comment Texts Based on BiLSTM

In this paper, an improved representation method of word is used integrating the contribution of sentiment information into the traditional TF-IDF algorithm and generating weighted word vectors. It gathered 15000 comments from an website. The comment vectors are input into bidirectional long short term memory (BiLSTM) to capture the context of the information. In this paper, they compared the result with the methods of RNN, CNN, LSTM, NB etc. The paper shows that their analysis method has higher precision, recall and F1 score.

A Bidirectional LSTM, or biLSTM, is a sequence processing model that consists of two LSTMs: one taking the input in a forward direction, and the other in a backwards direction. This paper is based on Sentiment of customers on hotels, it explores a TF-IDF model and BiLSTM model that has shown promise in making accurate prediction in sentiment analysis. It scraped dataset from online social media sites. It used *Word2Vec*, *Bi-LSTM*, *LSTM*, *RNN* and *CNN* to extract sentiments of Chinese hotel review comments with an F1-score of 92%. Their result was overall very promising. To overcome the shortcomings of current methods in sentiment analysis, a sentiment analysis method of comments based on BiLSTM is proposed in this paper. BiLSTM model combines bidirectional recurrent neural network (BiRNN) models and LSTM units to capture the context information.

Despite all of the advantages mentioned above, the model's training time is very lengthy which is a major shortcoming. It also suffered from insufficient coverage of sentiment words and lack of domain words. It is also a *binary classification* as it only classifies the comment on two classes - positive and negative where there is a broad spectrum of sentiment and we want to classify at-least three of them. In short, this paper achieves a respectable accuracy of 91.54 in terms of precision and 92.82 in terms of recall. It also managed to achieve an F1-score of 92.18. Since BiLSTM model fully considers the context information and can better obtain the

text representation of the comments, it achieved higher than average precision, recall and F1 score. However it does have some major shortcomings due to its lengthy training time and not having dynamically assigned weights to each individual word vector based on context.

3.4 Detecting Multilabel Sentiment and Emotions from Bangla YouTube Comments

Unlike the other papers discussed above, this one features multilabel sentiment analysis on Bangla. This paper's findings and results highlight the need of more sentiment research for Bangla language as the results clearly prove that Bangla is an under resourced language. This paper features an LSTM and CNN model with a softmax activation function at the final layer for classification.

The proposed methodology first goes through a series of pre-processing steps like removing stopwords through tokenization, removing unnecessary punctuation etc. Then the input word is tokenized to get a vector representation of each sentence to capture the syntactic meaning of the word. The input vector is then fed into the CNN and LSTM model. The model outputs 3 class sentiment of positive, neutral and negative as well as a 5 class sentiment with positive and negative further divided into 2 sentiments that results in 5 classes of very positive, positive, neutral, negative and very negative. Finally it also uses baseline method as a comparison baseline to show its improvement.

The model improves the baseline estimation of by almost 10%. Despite this improvement, the paper also has its drawback as has low accuracy on both the LSTM and CNN models. The model has a word limit of 50 in each feature vector representing each sentence, which may fail to capture valuable sentiment information due to limit. It also drops emojis and typos that may contribute to the overall sentiment.

3.5 Important Factors

The studies that are relevant to this literature have been accumulated and their impressions are summarized in this section. Table 3.1 is comprised of work focused on the under-resourced languages. Tables 3.2 and 3.3 extend on with work associated with well-resourced languages. From this section, it can be communicated that the Bangla language, being under-resourced, lacks promising results in this area compared to the English language.

Works Associated with Under-Resourced Languages

The following table shows the key findings from some works which contribute to this area for under-resourced languages:

Table 3.1: Works Associated with Under-Resourced Languages

Study	Year published	Type of dataset	Models used	No. of sentiments	Result	Bangla text
Tripto and Ali [44]	2018	YouTube	LSTM with embedding layer and CNN as its core layers	3	Accuracy: 65.97% and F1 Score: 0.64	Yes
Hande et al.[14]	2020	YouTube	SVM, Multinomial Naive Bayes, KNN, Decision tree and Random forest	5	Multi-class precision: 0.55 and weighted average: 0.58 with Random forest	No
AlturayEIF and Luqman [5]	2021	Twitter	CNN model with AraBERT and MARBERT	11	Multiclass accuracy 93% for MarBERT model and AraBERT model with baseline scoring about 88%.	No
Our work	2022	Facebook	BERT embedding layer with DNN layers	3	Accuracy: 97%, F1 Score: 0.93 and Recall rate: 0.93	Yes

Works Associated with Well-Resourced Languages

The following tables show the key findings from some works which contribute to this area for well-resourced languages:

Table 3.2: Works Associated with Well-Resourced Languages I

Study	Year published	Type of dataset	Models used	No. of sentiments	Result	Bangla text
Gutti Gowri et al. [18]	2021	Social media	Logistic regression and TF-IDF	3	Logistic regression average accuracy: 91.925% along with TF-IDF class: 92%	No
Kazi Nabiul et al. [23]	2021	Social media	A RNN oriented architecture, including LSTM and BLSTM	3	LSTM Model accuracy: 90.59% and Bi-LSTM: 90.83%.	No
Gloria J et al. [21]	2021	Social media	Semantic network analysis	5	the approach shows 65.97% and 54.24% accuracy in three and five labels sentiment	No
Adamu et al. [3]	2021	Social media	SVM and KNN	3	SVM accuracy: 88% and with KNN, 78% accuracy.	No

Table 3.3: Works Associated with Well-Resourced Languages II

Study	Year published	Type of dataset	Models used	No. of sentiments	Result	Bangla text
Ebele chukwu et al. [34]	2021	Social media	logistic regression, SVM, and Naive Bayes as Base-line Model and Transformer-based model, Covid-BERT v2	3	Covid-BERT V2 yeilds best results. No validation accuracy mentioned.	No
Rui et al. [30]	2021	Hotel website reviews	BERT and the deep CNN	2	Accuracy: 90%, Recall rate: 0.90 and F1 score: 0.90	No
Xu et al. [49]	2019	Booking website comments	BiLSTM is used; method is compared with the sentiment analysis methods of RNN, CNN, LSTM, and NB.	2	BiLSTM Model - Precision: 91.54, Recall: 92.82, F1 Score: 92.18	No
Piedrahita-Valdés et al. [37]	2021	Social media	Hybrid Model-a combination of lexicon and machine learning approaches	3	Statistics: (69.36%) neutral, (21.78%) positive and (8.86%) negative tweets. No validation accuracy mentioned.	No
Chad A. et al. [32]	2021	Social media	A lexical-based sentiment analysis	3	Statistics: 56.68% positive, 27.69% negative, and 15.63% neutral tweets. Polarity: 0.0520, variance: 0.0415. No validation accuracy mentioned.	No

Chapter 4

Dataset

Empirical studies have shown that one of the major challenges of training a model for classification is finding good dataset [6]. This has been a major drawback in the area of NLP classification in Bengali. So we took a different approach. Rather than taking Bangla Datasets from available resources, we took datasets from social media users that better reflect the real world events and sentiments. This dataset were suitable for our model as they contained accented Bengali which is used regularly compared to traditional running language. This helps the BERT embedding pickup various sentence structures not conforming to strict grammatical Bengali. We manually collected public comments from popular Bangladeshi news pages on facebook. Our dataset is comprised of Both Bangla and English text as well as Bangla which had various local accents.

4.1 Dataset Labelling

The entries were hand labeled as *Negative*, *Neutral* and *Positive* sentiment classes by three annotators with their expertise as:

- NLP researcher
- Native Bangla speaker
- Bilingual speaker with English proficiency

There is a serious concern of getting personal emotions in the way of annotation. The task of annotation manually is challenging since the lack of specification leaves the annotators in a very uncomfortable situation over the doubt of how to label certain texts[33]. Our strategy

was collecting as much information as possible when scrapping data because a text may not reveal the true intention to the annotators, but the following replies to the post/comment can very easily give the annotators a upper hand. So, whenever the annotators can't agree on a sentiment undoubtedly, extensive research on the following replies to that text is done to fill the gap.

4.2 Data Augmentation

Data Augmentation is an effective way to improve the accuracy of the predictions by increasing the variation within the existing dataset. This also reduces the chances of over-fitting. Cropping and Translation are common techniques used in data augmentation of language datasets [12].

We preferred to preserve the unique mixture of Bangla and English characters for BERT multilingual encoder, auto augmentation technique like translation was not feasible. Plus manual cropping tends to produce sentences without substantial meaning as most of them are not in a single language. This makes it difficult for the transformer model to make clear and accurate predictions. Instead, we did the augmentation of our dataset on our own and considered the following:

- Bangla text and their romanized form
- Narration
- Context
- Entities

Since the dataset had unbalanced class distribution to begin with, we maximized on our effort to attain impartiality.

4.3 Dataset Undersampling

Initially the dataset had good variety in terms of *Negative*, *Neutral* and *Positive*, eventually there was too much bias towards negative samples. We attributed it to having too much oversampling of datasets since we were manually hand-collecting data from facebook and there was no automated randomized crawling to balance out the classes. So we had to deliberately undersample our dataset to match the ratio of classes we needed. In some cases even undersampling yielded an unsatisfactory ratio for training, so we had to resample again.

4.4 Dataset Pre-Processing

During this phase, we applied customary pre-processing techniques. This was to ensure that our dataset does not contain the following,

- Unnecessary keywords
- Duplicate words
- Emojis and Special Characters
- URLs and Spams

4.5 Proposed Dataset:

Our compiled dataset was finally named *CoVaxBD* where we carefully selected **1113** samples. Each classes have the following configuration as Table 4.1 exhibits:

Table 4.1: Summary of CoVaxBD Samples

Text	Sentiment
ইন্সালিঙ্কাহ! ভ্যাক্সিন নেওয়ার দুইদিন পর-ই মারা গেলেন তিনি।	NEGATIVE
Astrazeneca vaccine ভারতের delta variant এর জন্য বেশি কার্যকর.	POSITIVE
Breast feeding mother ki corona tika dite parbe?	NEUTRAL
আমার ফুল ফ্যামিলি Moderna Vaccine দুই ডোজই নিয়েছি।	POSITIVE
এই দেশে vaccine এর জন্য মানুষ মরে। বিচার নাই	NEGATIVE
Pfizer vaccine গুলো school students দেব দেওয়া হোক.	NEUTRAL

Chapter 5

Methodology and Experiment

5.1 Foundational Methodology

We developed our classifier model for sentiment analysis of COVID-19 vaccination in Bangladesh using our *CoVaxBD* dataset. We followed transfer learning approach [43] with a compatible BERT(Bidirectional Encoder Representational from Transformers) deep learning model where every output element is connected to every input element with the exception that the weighting between them is dynamically calculated based on the connection created between each of the nodes. These are type of neural network that are specialized for analyzing complex tasks like naturally processing languages. The main benefit of transformer come from positional encoding, attention and self-attention which makes it suitable for language specific tasks. Transformers are a significant improvement over the traditional Recurrent Neural Network or architecture that was used to process languages. They also required sequential data at large volumes for training[9]. Bert model uses the same underlying benefits of transformer models. One of the advantages of the BERT model is that it can read text input sequentially from left to right and from right to left at the same time. This allows for greater bidirectional capability and weight adjustment. BERT model is an extension of the Transformer based model. This type of encoders can be divided up-to two blocks, the attention block and the Feed-Forward-Network block.[51] The general architecture of the BERT model is described at figure: 5.1

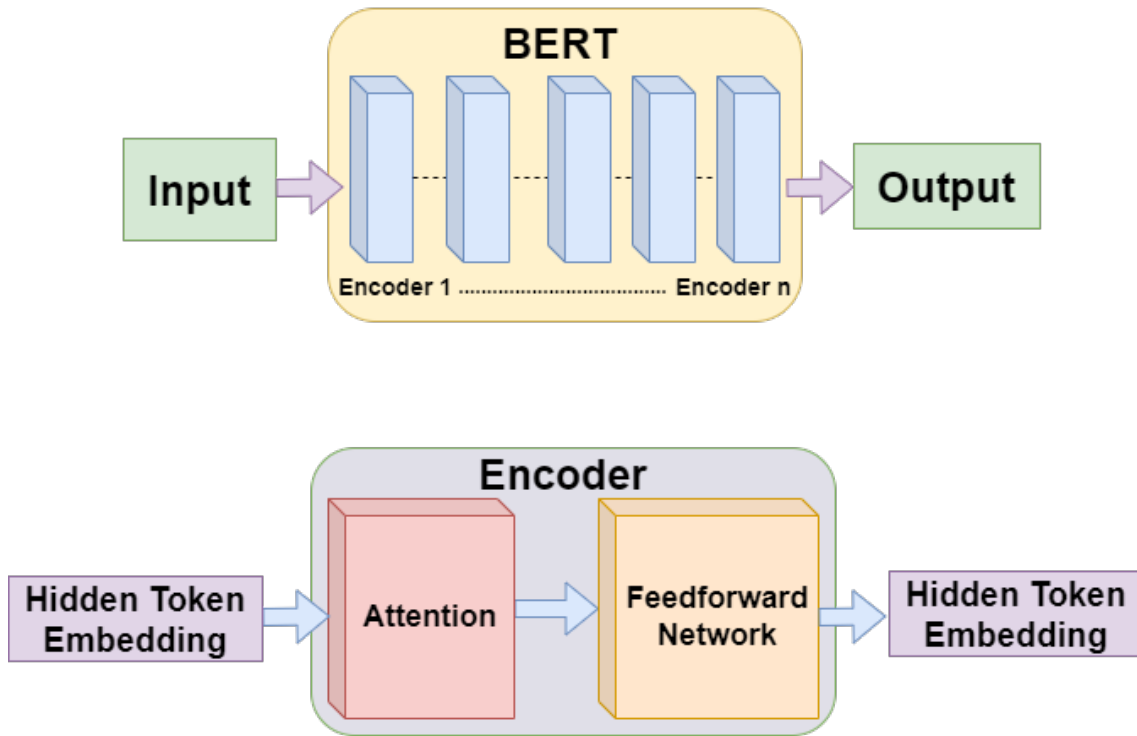


Figure 5.1: General architecture of the BERT

The BERT encoders take sequence of inputs, which are considered tokens. BERT model has two steps in it's framework: 1. Pre-training, and 2. Fine-tuning [9]. BERT is pre-trained on two tasks, Masked Language Modeling and Next Sentence Prediction [9]. Masked language modeling masks some input tokens at random and then tries to predict those masked tokens. During the Next Sentence Prediction phase, it trains the model by having pairs of sentences to make the model understand the relationship between sentences. This allows for understanding of context and ambiguity surrounding sentences[9]. Say, the model is trying to predict the answer to questions, then it pairs the question with the actual sentence to pre-train the model. Similarly in sentence prediction which predicts the next sentence, it pairs the actual sentences that are besides each other during the pre-train phase. In the final stage of the framework which is fine tuning, we simply put the task specific inputs and outputs and fine tune all the parameters from end to end.[9]. The self attention mechanism of the transformer allows for modeling of many downstream work, so fine tuning is simpler and easier.[9]. The transformer part allows for BERT to understand the surrounding context of a word or a sentence and

the overall ambiguity of a sentence. It processes words with respect to the other words in the sentence, rather than processing them one at a time.[9]. So a word could have different meaning based on the context surrounding it and the sentence it belongs to. It also has a self attention mechanism, meaning a particular word will always be aware of it's surrounding words and adjust itself dynamically to the structure of the sentence, rather than getting fixated at a particular meaning.[9]. The attention mask for the bidirectional encoder models attends to all tokens from a sequence as illustrated on figure 5.2

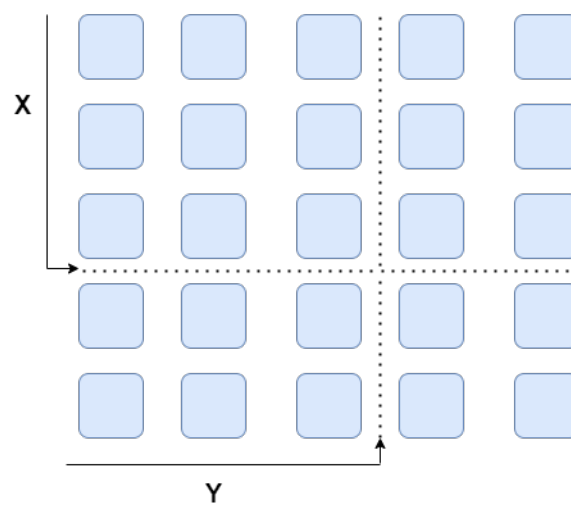


Figure 5.2: Attention mask for both X and Y set of tokens.

The bidirectional nature of the model makes it possible for the words to update themselves. We used the BERT multilingual base cased model [9] which was trained on 104 language using a masked language modeling that gives BERT a sentence and optimizing the weights inside BERT to output the same sentence on the other side [7]. Input texts are tokenized as a sequence of input ids and fed into the input layer of our model [Figure: 5.3].

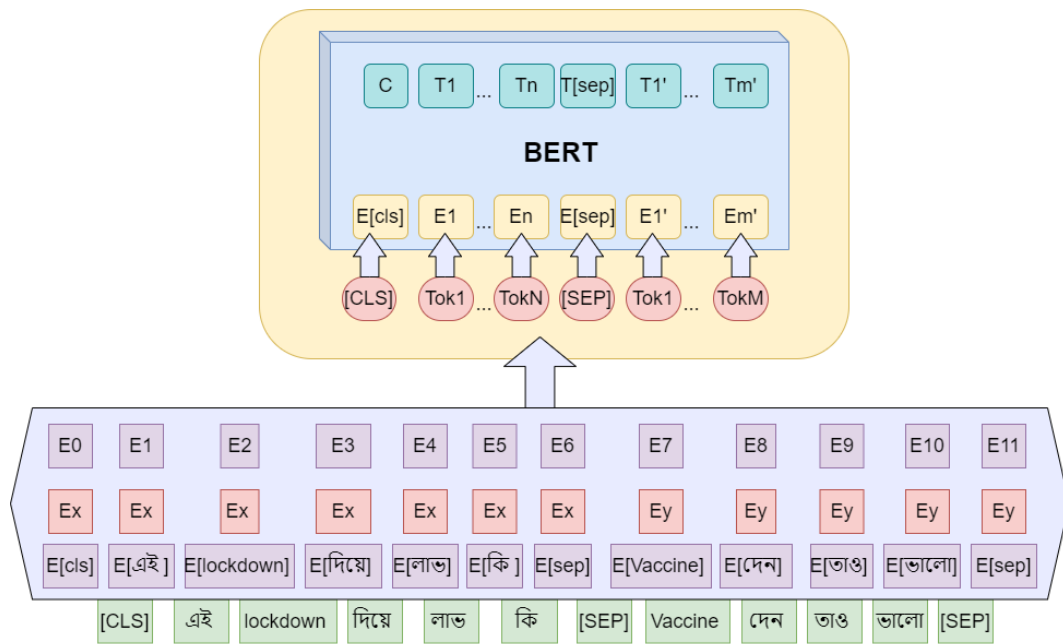


Figure 5.3: An instance of input for BERT embedding layer.

The model has the maximum input sequence length of 512 where the attention mask also has the same dimension. Shorter sequences are padded while sequences exceeding the length of 512 are truncated. [Figure: 5.4] Masked language modeling gives BERT a sentence and optimizes the weights inside BERT to output the same sentence on the other side of the feed-forward neural network. It processes words with respect to the other words in the sentence.

Similarly, for inference, unlabeled Bangla and English texts are fed into the input layer of our neural network and labels are predicted which has the maximum output value. Figure 5.4 depicts the architecture of our proposed model along with an instance of input text and the output from the model.

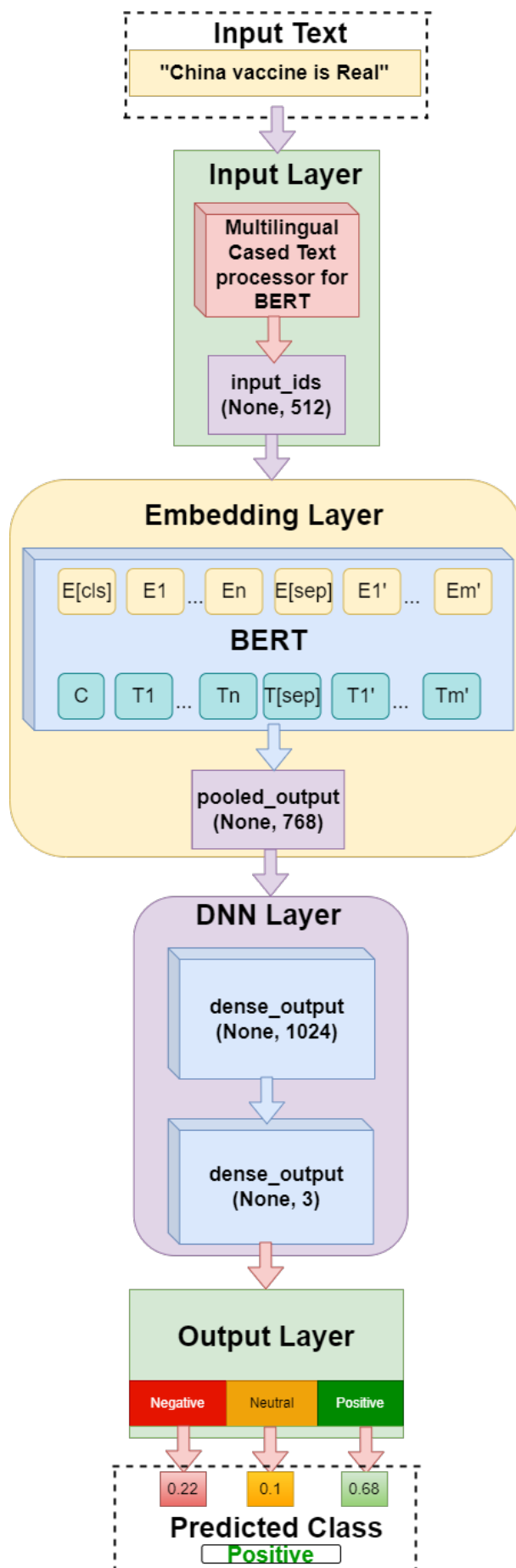


Figure 5.4: Model architecture.

5.2 Training Arrangement

Along with supportive Dense Neural Network layers [40] with softmax activation, the text classifier model was compiled with the ADAM optimizer [24] with learning rate of 10^{-5} and Categorical Cross-entropy [31] as loss function.

Our implementation leveraged the TensorFlow [1] library for Python programming language [45]. The training was lengthy and required extensive computational capabilities. The code has been included in chapter 7.3.

- Due to hardware limitations, we went for a batch size of 10 to prevent us from running out of memory while training.
- Throughout the training pipeline, we have utilized the architecture depicted in Figure 5.4 for our proposed model.

5.2.1 Activation Function

We know that neurons in a neural network work in accordance with their weight, bias, and activation function. We would change the weights and biases of the neurons in a neural network based on the output error. Back-propagation is the term for this procedure. Back-propagation is possible with activation functions since the gradients are supplied together with the error to update the weights and biases.

Softmax activation [26] is a generalization of the logistic function to multiple dimensions. This is used as an activation function of a neural network to normalize the output of a network to a probability distribution meaning adding non-linearity to the function. The softmax equation is given below:

$$\sigma(\vec{z}_i) = \frac{e^{z_i}}{\sum_{j=1}^K e^{z_j}} \quad (5.1)$$

Where,

- $\sigma \implies$ softmax
- $\vec{z}_i \implies$ input vector
- $e^{z_i} \implies$ standard exponential function for input vector
- $K \implies$ number of classes in the multi-class classifier
- $e^{z_j} \implies$ standard exponential function for output vector

5.2.2 Optimizer

The adaptive moment estimation algorithm (ADAM) calculates an exponential weighted moving average of the gradient and then squares the calculated gradient. Being a derived from the gradient descent with moments estimation and the Root Mean Square Prop (RMSP) algorithms, ADAM has two decay parameters that control the decay rates of these calculated moving averages. This algorithm has been efficient for the huge number of parameters we have been working with excellent memory space management. The adam optimizer involves combining two gradient descent methodologies.

5.2.3 Momentum

The momentum algorithm is used to accelerate the gradient descent process which takes into consideration the exponentially weighted average of the gradients. This makes the algorithm converge towards the global minima at a faster pace which initiates the idea of emerging as an integral part of the optimizer. The equations for momentum is given below:

$$w_{t+1} = w_t - \alpha m_t \tag{5.2}$$

$$m_t = \beta m_{t-1} + (1 - \beta) \left[\frac{\delta L}{\delta w_t} \right] \tag{5.3}$$

Where,

- $m_t \implies$ aggregate of gradients at time t [current] (initially, $m_t = 0$)
- $m_{t-1} \implies$ aggregate of gradients at time $t - 1$ [previous]
- $w_t \implies$ weights at time t
- $\alpha_t \implies$ learning rate at time t
- $\delta L \implies$ derivative of Loss Function
- $\delta w_t \implies$ derivative of weights at time t
- $\beta \implies$ Moving average parameter

5.2.4 Root Mean Square Propagation (RMSP)

The RMSP is an adaptive learning algorithm that tries to improve adaptive gradient descent by taking the *exponential moving average* instead of taking the cumulative sum of square gradients like adaptive gradient descent. The equation is given below:

$$w_{t+1} = w_t - \frac{\alpha_t}{(v_t + \epsilon)^{1/2}} * \left[\frac{\delta L}{\delta w_t} \right] \quad (5.4)$$

$$v_t = \beta v_{t-1} + (1 - \beta) * \left[\frac{\delta L}{\delta w_t} \right]^2 \quad (5.5)$$

Where,

- $w_t \implies$ weights at time t
- $w_{t+1} \implies$ weights at time $t + 1$
- $\alpha_t \implies$ learning rate at time t
- $\delta L \implies$ derivative of Loss Function
- $\delta w_t \implies$ derivative of weights at time t
- $v_t \implies$ sum of square of past gradients. (initially, $v_t = 0$)
- $\beta \implies$ Moving average parameter

- $\epsilon \implies$ A small positive constant

Combining the equations of momentum algorithm and RMSProp gives the following final equation:

$$m_t = \beta_1 m_{t-1} + (1 - \beta) \left[\frac{\delta L}{\delta w_t} \right] \quad (5.6)$$

$$v_t = \beta_2 v_{t-1} + (1 - \beta_2) \left[\frac{\delta L}{\delta w_t} \right]^2 \quad (5.7)$$

The optimizer then fixes the bias of both β_1 and β_2 which tends to hover towards 1 since m_t and v_t have both been initialized as 0. The adam optimizer fixes that by computing bias corrected m_t and v_t . It is done to prevent oscillation when the weights reach global maxima. The below formula is used:

$$\hat{m}_t = \frac{m_t}{1 - \beta_1^t} \hat{v}_t \quad (5.8)$$

$$= \frac{v_t}{1 - \beta_2^t} \quad (5.9)$$

Now, instead of using normal weight parameters m_t and v_t we take the bias corrected weights. Since the gradient descent process is being adapted after every iteration, it is called the adam optimizer. The final equation is given below :

$$w_{t+1} = w_t - \hat{m}_t \left(\frac{\alpha}{\sqrt{\hat{v}_t} + \epsilon} \right) \quad (5.10)$$

5.2.5 Loss Function

The categorical cross-entropy is a loss function that is designed to quantify the difference between probability distributions making it suitable for multi-class classification tasks.

Considering \hat{y}_i as the i -th scalar value and the output size as n ,

$$loss = - \sum_{i=1}^n y_i \times \log \hat{y}_i \quad (5.11)$$

This function ensures that the loss gets smaller as the distributions converge, making our effectively distinguish discrete probability measures.

5.3 Training Experiments

The following pre-trained models were used to train our proposed model:

- Bangla-BERT [39]
- Multilingual-BERT [10]

5.3.1 Bangla-BERT With Early-Stopping

Training graph [figure 5.5] of Bangla-BERT with early-stopping to avoid overfitting:

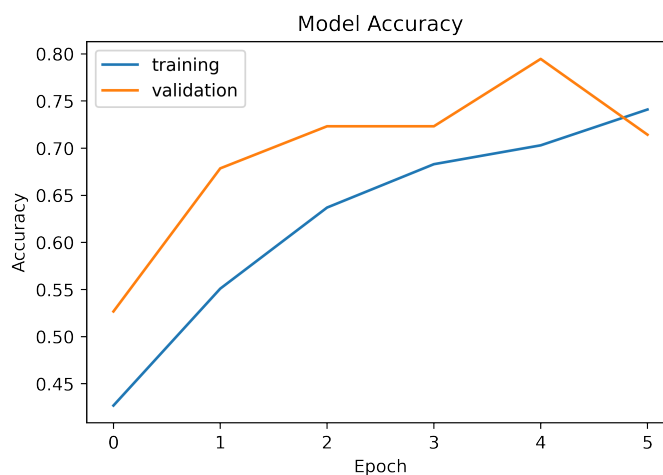


Figure 5.5: Learning curve for Bangla-BERT with early-stopping.

The validation accuracy starts to drop right before the 5th epoch. So, the training is stopped after that point.

5.3.2 Multilingual BERT Without Early-Stopping

Training graph [figure 5.6] of Multilingual BERT for 8 epochs:

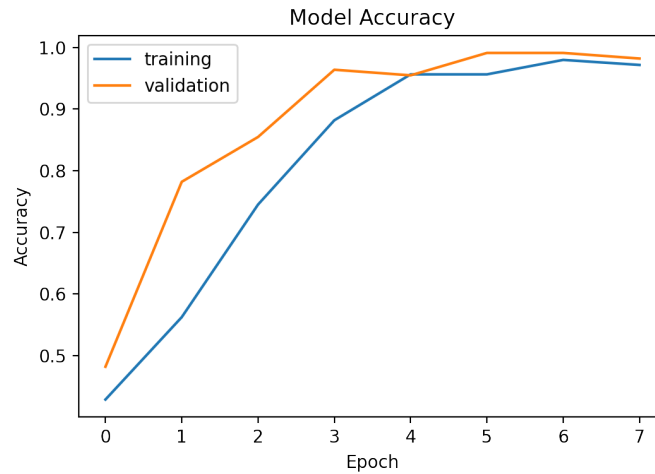


Figure 5.6: Learning curve for Multilingual BERT without early-stopping.

The validation accuracy starts to drop after the 6th epoch. Continuing up to 8 epoch, the accuracy measures indicate overfitting.

5.3.3 Multilingual BERT with Early-Stopping

Training graph [figure 5.7] of Multilingual BERT with early-stopping to avoid overfitting:

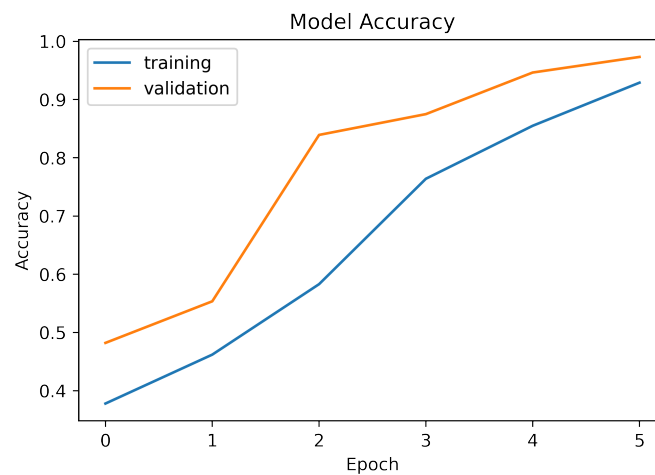


Figure 5.7: Learning curve for Multilingual BERT with early-stopping.

This arrangement retains higher validation accuracy which indicate lower chances overfitting.

Chapter 6

Result and Discussion

6.1 Result metrics

The results gathered from the model we constructed came out much better than anticipated [Figure: 6.4]. To give an overview of the results, we decided to use machine learning metrics such as *Precision*, *Recall* and *F1 – score* [19].

6.1.1 Precision

This is the metric that is used to find out the ratio of correctly predicted observation relative to the total no of observations. To calculate precision, the following equation was used:

$$Precision = \frac{TP}{TP + FP} \quad (6.1)$$

where,

- $TP = TruePositives$
- $FP = FalsePositives$

6.1.2 Recall

This is the metric that is used to find out the ratio of correctly predicted observations to the observations in the actual class. The following equation was used to calculate recall:

$$Recall = \frac{TP}{TP + FN} \quad (6.2)$$

where,

- $TP = TruePositives$
- $FN = FalseNegatives$

6.1.3 F1-Score

F1 score is the metric we used to find the balance between precision and recall. The equation is as follows:

$$F1 = 2 * \frac{Precision * Recall}{Precision + Recall} \quad (6.3)$$

6.1.4 Confusion Matrix

A confusion matrix is a matrix that visualizes the performance of a supervised machine learning model by putting instances in the actual class in the row and putting instances of the predicted class in the column. The goal here is to compare between actual result and predicted result to see how confused the model is. This can contribute to identifying biases among the class distributions.

6.1.5 Total polarity score

Total polarity score refers to the combined polarity of all classified sentiments on a scale from 0-1. The model evaluates the sentences on many different criteria that allows it do numerically aggregate whether a sentence ends up being positive, negative or neutral. Figure 6.4 shows the models polarity using end to end inference, where the model evaluates the percentage in each of the 3 classes, *positive*, *neutral* and *negative*. We can see that negative scored 97% percent which is by far the highest compared to only 2% for positive and none for neutral. So the class with the highest percentage gets predicted.

6.2 Model Evaluation

We evaluate our experimented models with the metrics mentioned in this section. For confusion matrices, the class ID for the respective sentiment categories is represented in table 6.1.

Table 6.1: Class ID for the respective sentiment categories

Class	ID
Negative	0
Neutral	1
Positive	2

6.2.1 Bangla-BERT With Early-Stopping

Table 6.2: Results for Bangla-BERT With Early-Stopping

Class	Precision	Recall	F1-Score
Negative	0.87	0.71	0.78
Neutral	0.72	0.87	0.79
Positive	0.76	0.74	0.75

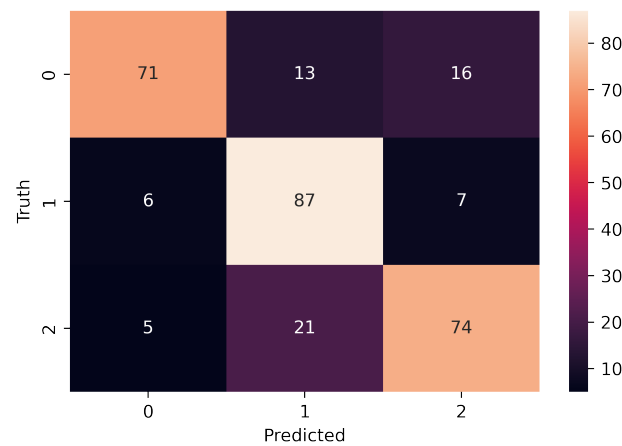


Figure 6.1: Confusion matrix for Bangla-BERT with early-stopping.

6.2.2 Multilingual BERT Without Early-Stopping

Table 6.3: Results for Multilingual BERT Without Early-Stopping

Class	Precision	Recall	F1-Score
Negative	0.97	0.93	0.95
Neutral	0.88	0.98	0.92
Positive	0.97	0.89	0.93

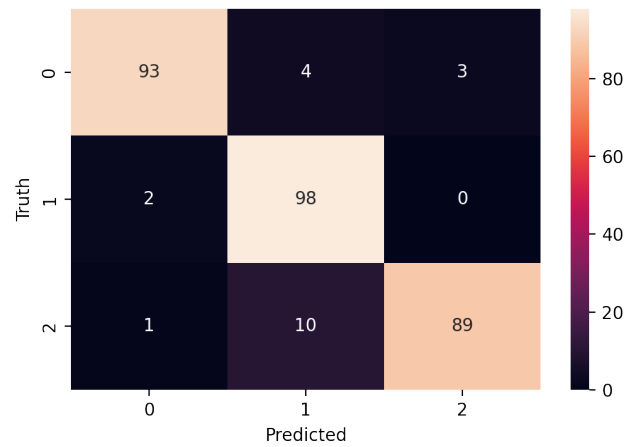


Figure 6.2: Confusion Matrix for Multilingual Bert without early-stopping

6.2.3 Multilingual BERT with Early-Stopping

Table 6.4: Results for Multilingual BERT With Early-Stopping

Class	Precision	Recall	F1-Score
Negative	0.95	0.99	0.97
Neutral	0.99	0.98	0.98
Positive	0.98	0.95	0.96

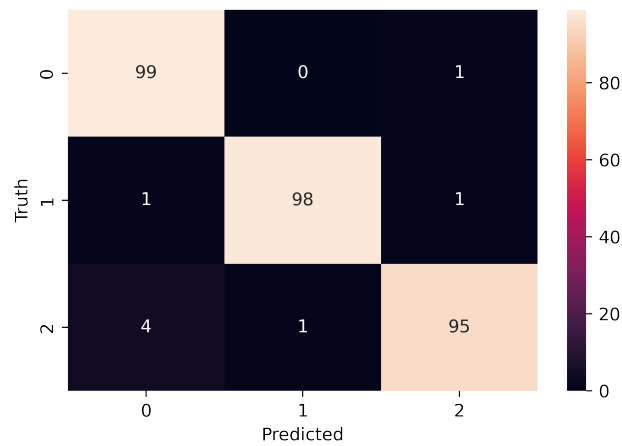


Figure 6.3: Confusion matrix for Multilingual-BERT with early-stopping.

6.2.4 End to End Inference

Table 6.5: Sample of inference

Text	Prediction	Truth
Instead of arguing here check On Vaccine Tracker for vaccine % & Also check Worldometer where also shows number of infected people across the world. Whatever DS mentioned here all are copied from there	POSITIVE	POSITIVE
Ai vaccine chai na . Govt otirikto Suru korse	NEGATIVE	NEGATIVE
এগুলিরে ভ্যাক্সিন দিয়া মাইরা ফালানো দরকার	NEGATIVE	NEGATIVE
এক সময় শোনা যাবে যে টিকার গায়ে লেখা Covid-19 vaccine কিন্তু ভিতরে লবণ পানি ভইরা রাখছেন।	NEUTRAL	NEGATIVE
জাপানে বাতিল কৃত ভ্যাসিন বাংলাদেশের সেরা ভ্যাক্সিন হিসাবে গন্য	NEGATIVE	NEGATIVE
China vaccine is Real	POSTIVE	POSITIVE
kon ta hoise Indian variant nki	POSITIVE	NEUTRAL
কারোনায় আক্রান্ত হলে কতদিন পর ভ্যাক্সিন নেয়া উচিৎ?	NEUTRAL	NEUTRAL

The end-to-end sentiment analysis interface depicted in figure 6.4 is a web application based on gradio [2]. The polarity of predictions can also be analyzed from this interface.

INPUT TEXT

don't take this vaccine! মরার শখ নাই ভাই

Clear Submit

OUTPUT 2.1s

NEGATIVE

NEGATIVE 97%

POSITIVE 2%

NEUTRAL 0%

Figure 6.4: End-to-end inference example.

6.3 Discussion

For the overall view of our model we can observe the learning curve graph from 5.6 without early stopping. As we can see from figure 5.6, the final accuracy we gathered over the course of 8 epoch is around 95.5% in terms of validation and around 98.8% in terms of training accuracy. The model quickly gathers accuracy from 0 to 4 epochs but drops down in gain after that. The steep upward climb flattens after epoch no 4 and rises insignificantly up to epoch 6. Finally, the graph flattens afterwards pointing to over-fitting. We can observe the effect of overfitting in this graph so we apply early stopping on multilingual BERT to mitigate the effect of overfitting as after a certain no of epochs the model barely improves. The accuracy graph of multilingual BERT with early stopping can be seen in figure 5.7. Similarly the effect on accuracy through early stopping can be seen on the Bangla BERT model as well in figure 5.5. Further discussion on early stopping and its effect are discussed later on in this section.

False positive rate appears to be low since the precision value is 0.97 for both *Negative* and *Neutral* classes respectively. Although the *Positive* class suffers from having the lowest value of 0.78, the recall value for that class is 0.95. Which makes it better than any other in terms of predicting true positives. Considering both of the metrics, this *F1 – score* hints a good performance for such minimal samples of data.

The confusion matrix portrayed in figure 6.2 has support for 100 classes each. The classes 0, 1, 2 are *Negative*, *Neutral* and *Positive* respectively. It can be observed that 10% of the positive inputs were falsely predicted as neutral. A marginal bias towards the neutral class can be inferred.

Let's now discuss the effect of using early stopping on both Bangla-BERT and multilingual BERT. Early stopping was the technique that was used to prevent the side-effect of overfitting. As we could see from the learning curve in figure 5.6, more epochs lead to the levelling of accuracy due to overfitting of the model as the model can't improve on accuracy on the same data. If the no of epoch is too high, it leads to the model learning more features of the dataset

thus increasing accuracy but it also leads to overfitting. Moreover after a certain iteration the level of increment in accuracy flattens and the model does not improve any further. Early stopping is used to mitigate the effect of overfitting as stopping at the correct no of epoch will lead to close to maximum accuracy without wasted epoch while also reducing the chance of overfitting. We can observe the effect of early stopping for Bangla-BERT in figure 6.1. The epoch is being stopped at 5 rather than at 7, hence leading to early stopping. The same phenomenon can be observed in figure 6.2 and figure 6.3 for multilingual BERT without early stopping and early stopping respectively. In the figure 6.2, the learning curve stagnates after about 6th epoch, leading to static or even validation accuracy. So the epochs after the 6th ones become stagnant where the training and validation accuracy hardly improves or even start to decrease while also overfitting with the training data. So using early stopping we can see in figure 6.3 that the model's training and testing accuracy always climbs up within the no of epochs that was selected for early stopping. Since the validation accuracy drops from 6th epoch, the early stopping is enacted after the 5th epoch as there are no more improvements to be found in terms of accuracy. This leads to less overfitting and in turn makes the model predict unique sentiments more accurately.

Chapter 7

Conclusions

7.1 Reflection

NLP is vast and broad and learning to apply different methods and techniques has been absolutely thrilling from a research point of view as the learning experience has been immense.

The knowledge that we were able to develop while working on our research are:

- Natural language processing with Neural Networks
- Learning different language pre-processing methods such as tokenization, stemming, lematization, parts of speech tagging etc
- Different type of Neural Network models like RNN, BiLSTM, BERT etc

The main challenges that we faced in this undertaking can be categorized below:

- Explainability of the model.
- Ambiguity in Neutral classification with respect to Positive and Negative classification likely due to the subjectivity of the Neutral class.

While initially after reviewing the literature of the project, we set out to develop a solution based on the BERT model and while it did provide tremendous success, there are a few approaches we could take if similar problems in the future. These are some of the different steps we could have taken compared to what we took for the initial problem based on our learning experience:

- Have expert opinion much earlier in our research to clearly identify sentiment classes compared to our late approach in bringing expert opinion which left us with ambiguity in Neutral class.

- Focus on collecting a larger dataset compared to what we got since it would have reduced our chances of overfitting our model.
- Analyze some of the aspects of each sentences and classify based on the sentiment towards that aspect rather than considering the whole aspect. While considering the whole sentence might theoretically make the model more accurate, in our case it also increased training time and made the model more difficult to explain. Plus in a lot of cases only analyzing the aspects rather than the whole sentence gives close to perfect accuracy while improving performance.

7.2 Future work

The work that we plan to carry out in future are:

- While our proposed solution managed to achieve admirable accuracy compared to many other model in this field, there were a few limitations that we could not overcome such as clear distinction between positive and neutral. Since there is very little margin for error on neutral classification, sometimes the sentence that may appear slightly positive could be classified as neutral. Secondly the perspective for neutral can be very subjective as a lot of the sentences which may seem neutral may seem negative or positive from different viewers perspective. Our takeaway is that this could be considered as a limitation of human resource rather than model. So, we plan to get more expert opinion on the matter of sentiment classification that would be less confusing for the model.
- The second step that we would like to take is to add explainability to the model that better describes the behavior of our model and how it processed the text input and outputs the classified sentiment. Since we learned that hard models are tough to replicate and doing future work on models that are un-explainable without proper documentation is difficult.
- The final improvement that we want to do in the future is to generalize the model towards different model. One of our main objective in our proposed model is to improve upon sentiment analysis of topic in under-resourced Bangla dataset which are often

mixed with English. Even though we improved on the general Bengali sentiment, our current model only focuses on a single topic and that is regarding COVID-19 vaccine sentiment in Bangladesh. In the future we would like to expand the domain of our work and generalize our model towards different Bangla topic on different datasets to classify sentiment.

7.3 Conclusion

Bangla Sentiment analysis is an under-resourced field and since knowing the sentiment of the general public in Bangladesh during COVID-19 was crucial in determining public mood regarding vaccination. So, undertaking the challenge to improve sentiment analysis in Bengali language was our primary goal. In our findings, we saw that using classifier model based on BERT which constitutes of multilingual cased text processor yielded a significant improvement over a lot of other traditional model. The transformer based BERT which constitutes of self attention heads allows for contextual embedding of word. Overall, our contribution to good dataset collection with proper augmentation, assigning weights based on context for each word rather than TF-IDF and training the whole embedded layer with our own training dataset instead of pre-assigned weights in regular BERT model significantly improved accuracy of the sentiment prediction.

Appendix

Code for utility and pre-processing the data

```
1 import pandas as pd
2 import re
3 from tqdm import tqdm
4 import tensorflow as tf
5 import numpy as np
6 import matplotlib.pyplot as plt
7 from transformers import BertTokenizer, TFAutoModel
8
9 def plotTraining(history):
10     acc = history.history['accuracy']
11     val_acc = history.history['val_accuracy']
12     loss = history.history['loss']
13     plt.plot(acc)
14     plt.plot(val_acc)
15     plt.title('Model Accuracy')
16     plt.ylabel('Accuracy')
17     plt.xlabel('Epoch')
18     plt.legend(['training', 'validation'], loc='upper left')
19     plt.savefig("history.jpg", dpi=200)
20     plt.show()
21
22
23 def sentiment_to_id(row):
24     if row['Sentiment'] == "NEGATIVE" :
25         return int(0)
26     if row['Sentiment'] == "NEUTRAL" :
27         return int(1)
28     if row['Sentiment'] == "POSITIVE" :
29         return int(2)
30     return -1
31
32 def id_to_sentiment(row):
33     if row['Class'] == 1 :
34         return "NEUTRAL"
35     if row['Class'] == 2 :
36         return "POSITIVE"
37     if row['Class'] == 0 :
38         return "NEGATIVE"
39     return "UNKNOWN"
40
41
42 def get_prediction_labels(id):
43     if id == 0 :
44         return "NEGATIVE"
45     if id == 1:
46         return "NEUTRAL"
47     if id ==2:
48         return "POSITIVE"
49     return "NONE"
50
51
52 def remove_emoji(string):
53     emoji_pattern = re.compile("[
54         u"\U0001F600-\U0001F64F" # emoticons
55         u"\U0001F300-\U0001F5FF" # symbols & pictographs
56         u"\U0001F680-\U0001F6FF" # transport & map symbols
57         u"\U0001F1E0-\U0001F1FF" # flags (iOS)
58         u"\U00002500-\U00002BEF" # chinese char
59         u"\U00002702-\U000027B0"
60         u"\U00002702-\U000027B0"
61         u"\U000024C2-\U0001F251"
62         u"\U0001f926-\U0001f937"
63         u"\U00010000-\U0010ffff"
```

```

64     u"\u2640-\u2642"
65     u"\u2600-\u2B55"
66     u"\u200d"
67     u"\u23cf"
68     u"\u23e9"
69     u"\u231a"
70     u"\ufe0f" # dingbats
71     u"\u3030"
72     "]"+"", flags=re.UNICODE)
73     return emoji_pattern.sub(r'', string)
74
75 def clean_en(value:str)->str:
76     ''' Get clean English-only Text '''
77     clean_value = re.sub("[^A-Za-z0-9\\/\- ]+", '', value)
78     return clean_value
79
80
81
82 def remove_links(text):
83     text = re.sub('http[s]?://(?:[a-zA-Z]|[0-9]|[$-_@.&+]|[*\(\)]|(?:%[0-9a
84     -fA-F][0-9a-fA-F]))', '', text, flags=re.MULTILINE)
85     return text

```

Code for TensorFlow Model

```

1 def map_data(input_ids, masks, labels):
2     return {'input_ids': input_ids, 'attention_mask': masks}, labels
3
4 def tokenize_text(text, tokenizer, seq_len=512):
5     tokens = tokenizer.encode_plus(text, max_length=seq_len, truncation=True,
6     padding='max_length', add_special_tokens=True, return_token_type_ids=
7     False, return_tensors='tf')
8     return {
9         'input_ids': tf.cast(tokens['input_ids'], tf.float64),
10        'attention_mask': tf.cast(tokens['attention_mask'], tf.float64)
11    }
12
13 def tf_dataset_from_df(dataset, tokenizer, textColumn='TEXT', classColumn='
14 Class', seq_len=512):
15     num_samples = len(dataset)
16     Xids = np.zeros((num_samples, seq_len), dtype=int)
17     Xmask = np.zeros((num_samples, seq_len), dtype=int)
18
19     for i, TEXT in enumerate(dataset[textColumn]):
20         tokens = tokenizer.encode_plus(TEXT, max_length=seq_len, truncation=
21         True, padding='max_length', add_special_tokens=True, return_tensors='tf'
22         )
23         Xids[i,:] = tokens['input_ids']
24         Xmask[i,:] = tokens['attention_mask']
25
26     classes = dataset[classColumn].values
27     classes = np.array(classes).astype(int)
28     labels = np.zeros((num_samples, classes.max()+1))
29     labels[np.arange(num_samples), classes] = 1
30     dataset_tf = tf.data.Dataset.from_tensor_slices((Xids, Xmask, labels))
31     dataset_tf.take(1)
32     dataset_tf = dataset_tf.map(map_data)
33     dataset_tf.take(1)
34     batch_size = 2
35     dataset_tf = dataset_tf.shuffle(3000).batch(batch_size, drop_remainder=
36     True)
37     dataset_tf.take(1)
38     split = 0.9
39     size = int((num_samples/batch_size)* split)
40     train_data = dataset_tf.take(size)
41     val_data = dataset_tf.skip(size)
42     del dataset_tf
43     return train_data, val_data, classes
44
45 def infer_text_sentiment(text, inferModel, tokenizer):
46     prediction = inferModel.predict(tokenize_text(text, tokenizer))
47     output = np.argmax(prediction[0])

```

```
45     return output, get_prediction_labels(output), prediction
46
47 def getModel(bert, seq_len = 512):
48     input_ids = tf.keras.layers.Input(shape=(seq_len,), name='input_ids',
49     dtype='int32')
50     mask = tf.keras.layers.Input(shape=(seq_len,), name='attention_mask',
51     dtype='int32')
52     embeddings = bert.bert(input_ids, attention_mask=mask)[1]
53     x = tf.keras.layers.Dense(1024, activation='relu')(embeddings)
54     y = tf.keras.layers.Dense(classes.max()+1, activation='softmax', name='
55     outputs')(x)
56     model = tf.keras.Model(inputs=[input_ids, mask], outputs=y)
57     model.layers[2].trainable = True
58     optimizer = tf.keras.optimizers.Adam(learning_rate=1e-5, decay=1e-6)
59     loss = tf.keras.losses.CategoricalCrossentropy()
60     acc = tf.keras.metrics.CategoricalAccuracy('accuracy')
61     model.compile(optimizer=optimizer, loss=loss, metrics=[acc])
62     return model
```

Bibliography

- [1] Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S. Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Ian Goodfellow, Andrew Harp, Geoffrey Irving, Michael Isard, Yangqing Jia, Rafal Jozefowicz, Lukasz Kaiser, Manjunath Kudlur, Josh Levenberg, Dandelion Mané, Rajat Monga, Sherry Moore, Derek Murray, Chris Olah, Mike Schuster, Jonathon Shlens, Benoit Steiner, Ilya Sutskever, Kunal Talwar, Paul Tucker, Vincent Vanhoucke, Vijay Vasudevan, Fernanda Viégas, Oriol Vinyals, Pete Warden, Martin Wattenberg, Martin Wicke, Yuan Yu, and Xiaoqiang Zheng. TensorFlow: Large-scale machine learning on heterogeneous systems, 2015. Software available from tensorflow.org.
- [2] Abubakar Abid, Ali Abdalla, Ali Abid, Dawood Khan, Abdulrahman Alfozan, and James Zou. Gradio: Hassle-free sharing and testing of ml models in the wild. *arXiv preprint arXiv:1906.02569*, 2019.
- [3] Hassan Adamu, Mat Jasri Bin Mat Jiran, Keng Hoon Gan, and Nur-Hana Samsudin. Text analytics on twitter text-based public sentiment for covid-19 vaccine: A machine learning approach. In *2021 IEEE International Conference on Artificial Intelligence in Engineering and Technology (IICAJET)*, pages 1–6, 2021.
- [4] AH Alamoodi, BB Zaidan, Maimonah Al-Masawa, Sahar M Taresh, Sarah Noman, Ibrahim YY Ahmaro, Salem Garfan, Juliana Chen, MA Ahmed, AA Zaidan, et al. Multi-perspectives systematic review on the applications of sentiment analysis for vaccine hesitancy. *Computers in Biology and Medicine*, 139:104957, 2021.
- [5] Nora Alturayef and Hamzah Luqman. Fine-grained sentiment analysis of arabic covid-19 tweets using bert-based transformers and dynamically weighted loss function. *Applied*

Sciences, 11(22), 2021.

- [6] Mithun Biswas, Rafiqul Islam, Gautam Kumar Shom, Md Shopon, Nabeel Mohammed, Sifat Momen, and Anowarul Abedin. Banglalekha-isolated: A multi-purpose comprehensive dataset of handwritten bangla isolated characters. *Data in brief*, 12:103–107, 2017.
- [7] James Briggs. Masked-language modeling with bert, 2021. [Accessed December 2, 2021].
- [8] covidvax.live. Vaccination in bangladesh, (Dec. 31, 2021). [Accessed Dec. 31, 2021][Online].
- [9] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805:1–16, 2018.
- [10] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805, 2018.
- [11] Eve Dubé, Maryline Vivion, and Noni E MacDonald. Vaccine hesitancy, vaccine refusal and the anti-vaccine movement: influence, impact and implications. *Expert review of vaccines*, 14(1):99–117, 2015.
- [12] Steven Y Feng, Varun Gangal, Jason Wei, Sarath Chandar, Soroush Vosoughi, Teruko Mitamura, and Eduard Hovy. A survey of data augmentation approaches for nlp. *arXiv preprint arXiv:2105.03075*, 2021.
- [13] Adeep Hande, Ruba Priyadharshini, and Bharathi Raja Chakravarthi. Kancmd: Kannada codemixed dataset for sentiment analysis and offensive language detection. In *Proceedings of the Third Workshop on Computational Modeling of People’s Opinions, Personality, and Emotion’s in Social Media*, pages 54–63, 2020.
- [14] Adeep Hande, Ruba Priyadharshini, and Bharathi Raja Chakravarthi. KanCMD: Kannada CodeMixed dataset for sentiment analysis and offensive language detection. In

- Proceedings of the Third Workshop on Computational Modeling of People's Opinions, Personality, and Emotion's in Social Media*, pages 54–63, Barcelona, Spain (Online), December 2020. Association for Computational Linguistics.
- [15] Md. Rakibul Hasan, Maisha Maliha, and M. Arifuzzaman. Sentiment analysis with nlp on twitter data. In *2019 International Conference on Computer, Communication, Chemical, Materials and Electronic Engineering (IC4ME2)*, pages 1–4, Rajshahi University, Bangladesh, 2019. IEEE Bangladesh Section.
- [16] Hatice İkişik, Mehmet Akif Sezerol, Yusuf Taşçı, and Işıl Maral. Covid-19 vaccine hesitancy: A community-based research in turkey. *International Journal of Clinical Practice*, 75(8):e14336, 2021.
- [17] Our World in Data. Bangladesh coronavirus vaccination rate: Any dosage, (Jan 2021-Mar 2022).
- [18] Gutti Gowri Jayasurya, Sanjay Kumar, Binod Kumar Singh, and Vinay Kumar. Analysis of public sentiment on covid-19 vaccination using twitter. *IEEE Transactions on Computational Social Systems*, pages 1–11, 2021.
- [19] Renuka Joshi. Accuracy, precision, recall and f1 score: Interpretation of performance measures, 2016. [Accessed November 23, 2021].
- [20] Md. Kamruzzaman. Bangladesh starts nationwide covid vaccination drive, (Feb. 07, 2021). [Accessed Dec. 31, 2021][Online].
- [21] Gloria J. Kang, Sinclair R. Ewing-Nelson, Lauren Mackey, James T. Schlitt, Achla Marathe, Kaja M. Abbas, and Samarth Swarup. Semantic network analysis of vaccine sentiment in online social media. *Vaccine*, 35(29):3621–3638, 2017.
- [22] Emilie Karafillakis, Heidi J Larson, et al. The benefit of the doubt or doubts over benefits? a systematic literature review of perceived risks of vaccines in european populations. *Vaccine*, 35(37):4840–4850, 2017.

- [23] Mohammad Khan. Deep learning-based sentiment analysis of covid-19 vaccination responses from twitter data. *Computational and Mathematical Methods in Medicine*, 2021, 12 2021.
- [24] Diederik Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *International Conference on Learning Representations*, 12 2014.
- [25] Alison Knopf. Time to remember: Vaccines don't cause autism. *The Brown University Child and Adolescent Behavior Letter*, 37(7):9–10, 2021.
- [26] Kiprono Elijah Koech. Softmax activation function — how it actually works [online], 2020. [Accessed January 23, 2022].
- [27] Mary Koslap-Petraco. Vaccine hesitancy: Not a new phenomenon, but a new threat. *Journal of the American Association of Nurse Practitioners*, 31(11):624–626, 2019.
- [28] The Lancet. Looking beyond the decade of vaccines, 2018.
- [29] Dehong Ma, Sujian Li, Xiaodong Zhang, and Houfeng Wang. Interactive attention networks for aspect-level sentiment classification. *arXiv preprint arXiv:1709.00893*, 2017.
- [30] Rui Man and Ke Lin. Sentiment analysis algorithm based on bert and convolutional neural network. In *2021 IEEE Asia-Pacific Conference on Image Processing, Electronics and Computers (IPEC)*, pages 769–772, 2021.
- [31] Shie Mannor, Dori Peleg, and Reuven Rubinfeld. The cross entropy method for classification. In *Proceedings of the 22nd International Conference on Machine Learning, ICML '05*, page 561–568, New York, NY, USA, 2005. Association for Computing Machinery.
- [32] Chad A. Melton, Olufunto A. Olusanya, Nariman Ammar, and Arash Shaban-Nejad. Public sentiment analysis and topic modeling regarding covid-19 vaccines on the reddit social media platform: A call to action for strengthening vaccine confidence. *Journal of Infection and Public Health*, 14(10):1505–1512, 2021. Special Issue on COVID-19 – Vaccine, Variants and New Waves.

- [33] Saif Mohammad. A practical guide to sentiment annotation: Challenges and solutions. In *Proceedings of the 7th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 174–179, San Diego, California, June 2016. Association for Computational Linguistics.
- [34] Ebelechukwu Nwafor, Ryan Vaughan, and Christopher Kolimago. Covid vaccine sentiment analysis by geographic region. In *2021 IEEE International Conference on Big Data (Big Data)*, pages 4401–4404, 2021.
- [35] Douglas J Opel, Bernard Lo, and Monica E Peek. Addressing mistrust about covid-19 vaccines among patients of color, 2021.
- [36] Varun K Phadke, Robert A Bednarczyk, and Saad B Omer. Vaccine refusal and measles outbreaks in the us. *Jama*, 324(13):1344–1345, 2020.
- [37] Hilary Piedrahita-Valdés, Diego Castillo, Javier Bermejo, Patricia Guillem-Saiz, Juan-Ramón Higuera, Javier Guillem-Saiz, Juan Antonio Montalvo, and Francisco Machio. Vaccine hesitancy on social media: Sentiment analysis from june 2011 to april 2019. *Vaccines*, 9:28, 01 2021.
- [38] Star Online Report. Nurse runu veronica first to receive covid-19 vaccine in bangladesh, (Jan. 27, 2021). [Accessed Dec. 31, 2021][Online].
- [39] Sagor Sarker. Banglabert: Bengali mask language model for bengali language understanding, 2020.
- [40] Piotr Semberecki and Henryk Maciejewski. Deep learning methods for subject text classification of articles. In *2017 Federated Conference on Computer Science and Information Systems (FedCSIS)*, pages 357–360, 2017.
- [41] Heereen Shim, Dietwig Lowet, Stijn Luca, and Bart Vanrumste. Lets: A label-efficient training scheme for aspect-based sentiment analysis by using a pre-trained language model. *IEEE Access*, 9:115563–115578, 2021.
- [42] StatCounter. Social media usage of bangladesh, (Feb 2021-Feb 2022).

- [43] Lisa Torrey and Jude Shavlik. Transfer learning. In *Handbook of research on machine learning applications and trends: algorithms, methods, and techniques*, pages 242–264. IGI global, 2010.
- [44] Nafis Tripto and Mohammed Eunus Ali. Detecting multilabel sentiment and emotions from bangla youtube comments. In *2018 International Conference on Bangla Speech and Language Processing (ICBSLP)*, pages 1–6, 09 2018.
- [45] Guido Van Rossum and Fred L Drake Jr. *Python reference manual*. Centrum voor Wiskunde en Informatica Amsterdam, 1995.
- [46] Annika B Wilder-Smith and Kaveri Qureshi. Resurgence of measles in europe: a systematic review on parental attitudes and beliefs of measles vaccine. *Journal of epidemiology and global health*, 10(1):46, 2020.
- [47] Worldometer. Coronavirus update (live), (Dec. 31, 2021). [Accessed Dec. 31, 2021][Online].
- [48] Guixian Xu, Yueting Meng, Xiaoyu Qiu, Ziheng Yu, and Xu Wu. Sentiment analysis of comment texts based on bilstm. *IEEE Access*, 7:51522–51532, 2019.
- [49] Guixian Xu, Yueting Meng, Xiaoyu Qiu, Ziheng Yu, and Xu Wu. Sentiment analysis of comment texts based on bilstm. *IEEE Access*, 7:51522–51532, 2019.
- [50] Ohid Yaqub, Sophie Castle-Clarke, Nick Sevdalis, and Joanna Chataway. Attitudes to vaccination: a critical review. *Social science & medicine*, 112:1–11, 2014.
- [51] Sumu Zhao, Damian Pascual, Gino Brunner, and Roger Wattenhofer. Of non-linearity and commutativity in bert. In *2021 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE, 2021.