Sentiment Analysis on Corona Virus Tweets

Submitted by

Fhamid Mottaki Aurnob	**170204058**
Mohammad Najrul Islam	**170204061**
Rafsan Habib Rasan	**170204069**

Project Report

Course ID: CSE 4214

Course Name: Pattern Recognition Lab

Semester: Spring 2021

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 14, 2022

ABSTRACT

As the Covid-19 outbreaks rapidly all over the world day by day and also affects the lives of million, a number of countries declared complete lockdown to check its intensity. During this lockdown period, social media platforms have played an important role to spread information about this pandemic across the world, as people used to express their feelings through the social networks. Considering this catastrophic situation, we developed an experimental approach to analyze the reactions of people on Twitter taking into account the popular words either directly or indirectly based on this pandemic. This paper represents the sentiment analysis on collected large number of tweets on Coronavirus or Covid-19. We have applied Lemmatization Stemming for data preprocessing. For feature extraction, we have used techniques such as TF-IDF Bag of Words. Then the classification is done using Machine Learning algorithms such as Support Vector Machine (SVM), Logistic Regression (LR), Random Forest(RF), Extreme Gradient Boosting (XGBoost), and Naive Bayes(NB). We have compared five machine learning algorithms. Finally, utilizing the Bag of Words model and the Logistic Regression Algorithm, we attained an accuracy of 56.53 percent on the test dataset.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

On 31st December, 2019 the Covid-19 outbreak was first reported in the Wuhan, Hubei Province, China and it started spreading rapidly all over the world. Finally, WHO announced Covid-19 outbreak as pandemic on 11th March, 2020, when the virus continues to spread. Starting from China, this virus infected and killed thousands of people from Italy, Spain, USA, UK, Brazil, Russia, and other many more countries as well. On 21st August 2020, more than 22.5 million cases of Covid-19 were reported in more than 188 countries and territories, yielding more than 7,92,000 deaths; although 14.4 million people have reported to be recovered. While this pandemic has continued to affect the lives of millions, many countries had enforced a strict lockdown for different periods to break the chain of this pandemic. Since the Covid-19 vaccines are still yet to be discovered, therefore maintaining social distancing is the one and only one solution to check the spreading rate of this virus. During the lockdown period a lot of people have chosen the Twitter to share their expression about this disease so we have been inspired to measure the human sensations about this epidemic by analyzing this huge Twitter data.

Chapter 2

Motivation

The purpose of the paper is to raise awareness in society that the widespread use of social media around the world needs to be limited, as it is becoming increasingly important in propagating meaningless information, much as the plague has spread throughout humanity. As a contribution to society, this study has illustrated with statistics how astonishing it is that, despite the fact that people are originally providing positive and neutral material, users are re-tweeting bad tweets. COVID-19 is no longer merely an infectious condition spread through touch and minute droplets formed when people cough, sneeze, or talk; it is now becoming a source of despair, worry, and anxiety as a result of false information spread on social media. Although it is expected that social media will assist people in obtaining accurate and reliable information about Corona cases, an examination of the posts reveals that the majority of them have deceived individuals by presenting incorrect statistics and figures.

Chapter 3

Literature Reviews

We dived into some of the related work with Sentiment Analysis and their classification prediction. Some of the related works are discussed in this section.

3.0.1 A sentiment analysis approach to predict an individual's awareness of the precautionary procedures to prevent COVID-19 outbreaks in Saudi Arabia

Author

Aljameel S.S., Alabbad D.A., Alzahrani N.A., Alqarni S.M., Alamoudi F.A., Babili L.M., Aljaafary S.K., Alshamrani F.M.

Approach

Alike recent relevant works of sentiment analysis, some recent studies have been attempted to scrutinize COVID-19 tweets in bulk for public health research purposes, although it is likely that they have been mined for more commercial purposes. Aljameel et al. [1] gathered 2,42,525 tweets from five regions in Saudi Arabia to analyze their sentiments using support vector machine (SVM), k-nearest neighbor (KNN) and Naïve Bayes (NB).The results show that bigram TF-IDF with the SVM classifier produced the highest accuracy of 85which outperformed KNN and Naïve Bayes.

3.0.2 Sentimental analysis on social media data using R programming

Author

Bhargava, Mandava Geetha and Rao, Duvvada Rajeswara

Approach

Here [2] the authors conducted experiments on twitter data in which they simply extracted the tweets using learning methods. For this they collected data regarding cryptocurrency and applied algorithms like naïve bayes and SVM (Support Vector Machine) on it.

3.0.3 Lies Kill, Facts Save: Detecting COVID-19 Misinformation in Twitter

Author

Mabrook S. Al-Rakhami, Atif M. Al-Amri

Approach

In the paper [3], they gathered 4,00,000 tweets and implemented entropy and correlation based feature selection and ensemble methods using NB, Bayes Net, KNN, C4.5, random forest (RF) and SVM. They chose weak learners through experiments and they were combined to form stacked models.

3.0.4 Predicting Coronavirus Pandemic in Real-Time Using Machine Learning and Big Data Streaming System

Author

Xiongwei Zhang, Hager Saleh , Eman M. G. Younis, Radhya Sahal , and Abdelmgeid A. Ali

Approach

Here [4] the authors gathered tweets by employing N-Gram model and TF–IDF as well as explored sentiments using DT, LR, KNN, RF and SVM respectively. The experimental results show that the RF model using the unigram feature extraction method has achieved the best performance on real time data.

3.0.5 Twitter Sentiment Analysis

Author

Aliza Sarlan, Chayanit Nadam, Shuib Basri

Approach

This paper [5] discusses about the ways and procedure we can analyse data from Twitter. It mentions that Support Vector Machine is to detect the sentiments of tweets, and it stated SVM is able to extract and analyze to obtain upto70%-81.3% of accuracy on the test set.

Chapter 4

Data Collection & Processing

The dataset was collected [6] from Kaggle.The tweets have been pulled from Twitter and manual tagging has been done then. It consists of six columns, giving us information such as Location of the Tweet, time it was Tweeted At, the Original Tweet, Label of the sentiment. It initially bears around 45 thousand data. The labels given were: Extremely Positive, Positive, Neutral, Negative and Extremely Negative.
We only used the tweets and the label column, and dropped the unnecessary column from it. We later performed multi-class classification on the dataset taking all of the target values. We first combined the train and test dataset to preprocess. We had 44523 samples after the operation. We removed null rows from the dataset. There were now, 35174 samples after the preprocessing.

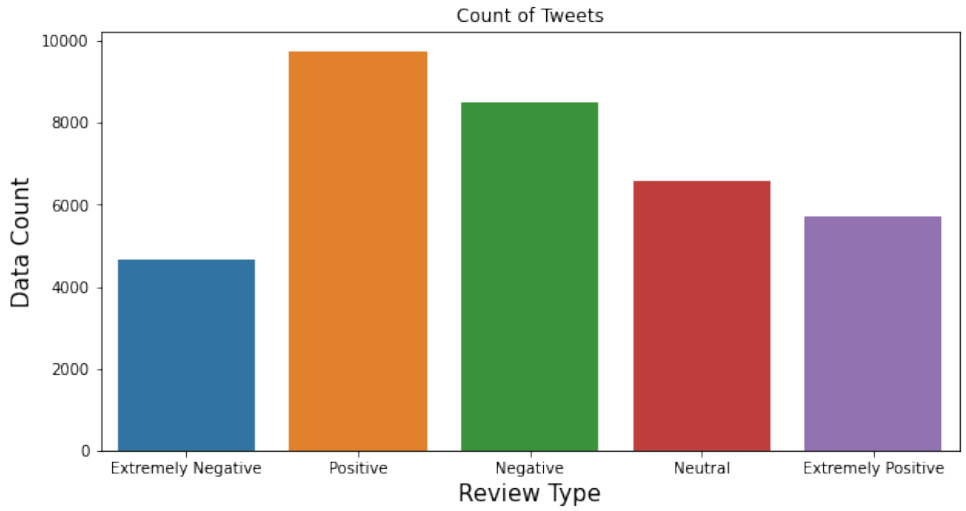

Figure 4.1: Data Visualization

The tweets were all taken into lowercase, stop-words and non-alphabetic characters were removed from them. We performed Text Normalization using Lemmatization.

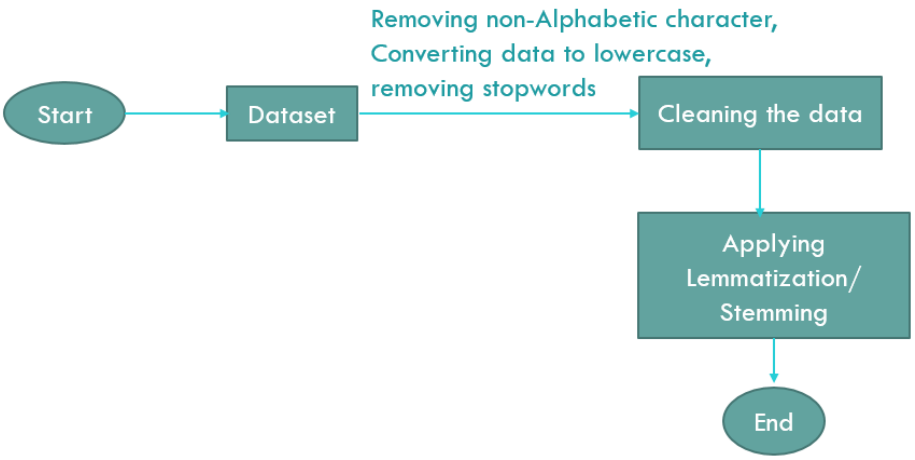

Figure 4.2: Work Process for Preprocessing

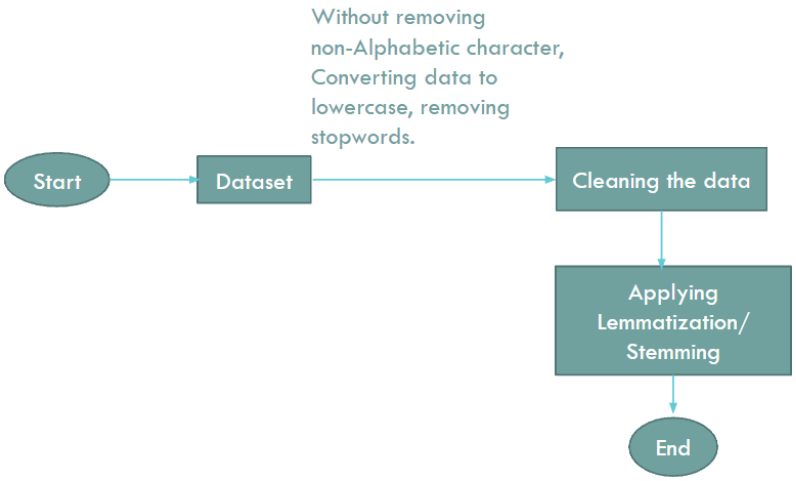

Figure 4.3: Work Process for Preprocessing

For all the model we used Term Frequency — Inverse Document Frequency (tf-idf) and Bag of Words (BoW) as our word embedding technique separately.

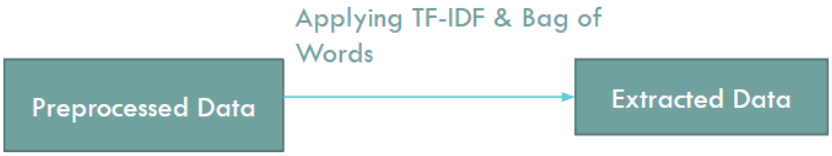

Figure 4.4: Work Process for Feature Extraction

Chapter 5

Methodology

For our project we used five machine learning classifying algorithms like Support Vector Machine, Naive Bayes, Logistic Regression, Random Forest, XGBoost.

For this we preprocessed as described in chapter 4, the data and split the data for training and testing the model. For training we used 80% of the data and rest of the 20% for testing purpose as that is ideal. After training the models, we evaluated the models with the test data based on Accuracy, Precision, Recall and F1 scores. Then compared the results.
We also applied BoW and TF-IDF on the dataset without lemmatization. Split it for training and testing. Then trained our models using all the algorithms and evaluated the models. And compared the results.

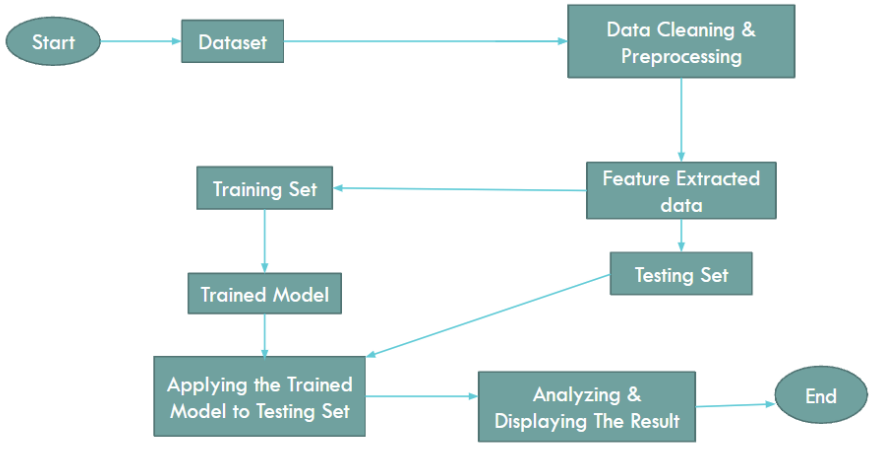

Figure 5.1: Methodology

Chapter 6

Experiments and Results

6.1 Experiment

For this project, we experimented with various machine learning models and fed them data by Lemmatization with and without removing Non Alphabetic character while using TF-IDF as well as Bag Of Words, keeping the other factors constant designing four models.

6.1.1 Model 1

6.1.1.1 Exp 1: without removing non-alphabetic character(TF-IDF)

In model 1 Exp 1 we worked with TF-IDF model and Lemmatization. Then the processed data is sent to the machine learning models. The table shows the comparision of the ML models on the lemmatized data and figure illustrates the results.

Table 6.1: Performance with Lemmatization

No.	Metrics	SVM	LR	NB	RF	XGB
1.	Accuracy	0.475195	0.475480	0.297086	0.349680	0.466525
2.	Precision	0.489163	0.504500	0.351986	0.370531	0.498065
3.	Recall	0.489843	0.471671	0.217117	0.319963	0.465258
4.	F1 Score	0.488485	0.481385	0.126589	0.324036	0.475048

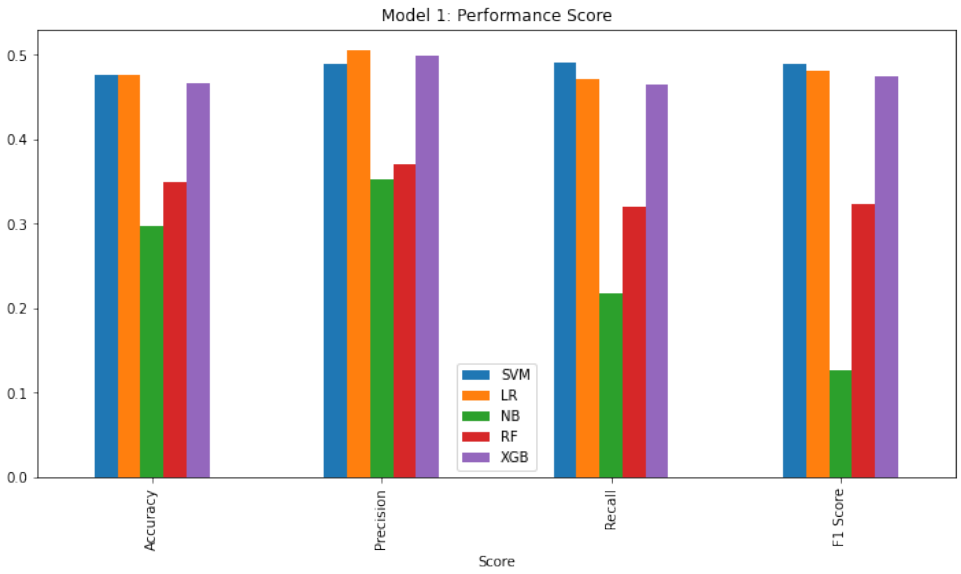

Figure 6.1: Comparison Between the Algorithms

6.1.1.2 Exp 2: without removing non-alphabetic character(BOW)

In model 1 Exp 2 we worked with Bag Of Words model and Lemmatization. Then the processed data is sent to the machine learning models. The table shows the comparision of the ML models on the lemmatized data and figure illustrates the results.

Table 6.2: Performance with Lemmatization

No.	Metrics	SVM	LR	NB	RF	XGB
1.	Accuracy	0.556361	0.552950	0.480171	0.527647	0.488131
2.	Precision	0.561695	0.564772	0.482921	0.554616	0.532390
3.	Recall	0.583045	0.564864	0.505128	0.513995	0.481809
4.	F1 Score	0.568560	0.563846	0.490536	0.522552	0.494584

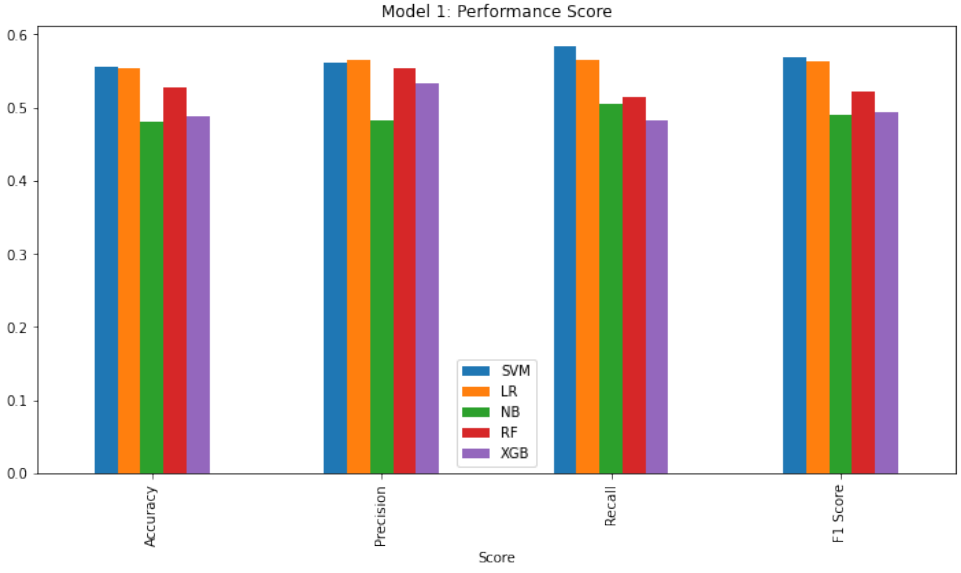

Figure 6.2: Comparison Between the Algorithms

6.1.2 Model 2

6.1.2.1 Exp 1: with removing non-alphabetic character(TF-IDF)

In model 2 Exp 1 we worked with TF-IDF model and Lemmatization. Then the processed data is sent to the machine learning models. The table shows the comparision of the ML models on the stemmed data and figure illustrates the results.

Table 6.3: Performance with Lemmatization

No.	Metrics	SVM	LR	NB	RF	XGB
1.	Accuracy	0.475338	0.478749	0.339303	0.359915	0.464250
2.	Precision	0.486575	0.509689	0.387430	0.393394	0.495868
3.	Recall	0.489704	0.475621	0.261469	0.334058	0.461387
4.	F1 Score	0.487471	0.486649	0.205654	0.342632	0.471959

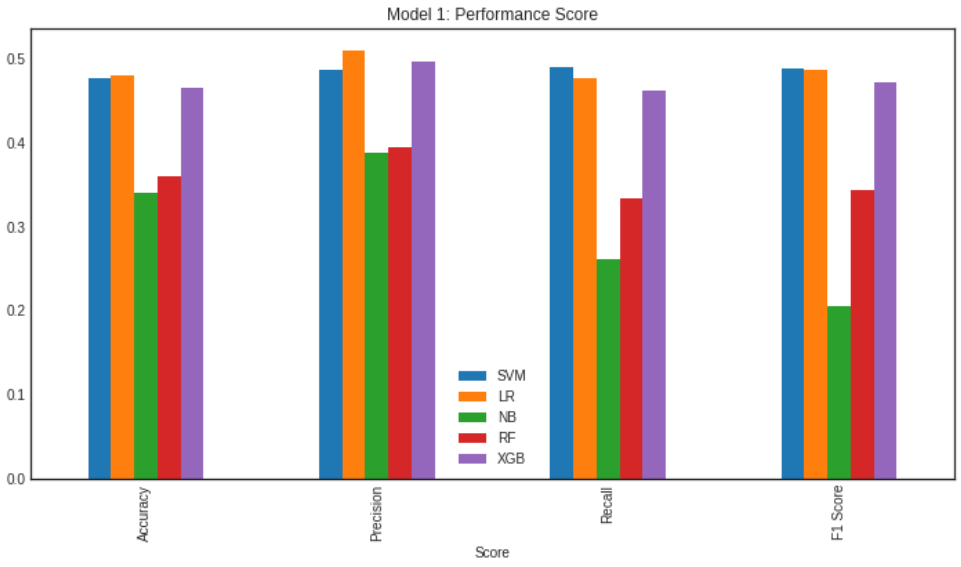

Figure 6.3: Comparison Between the Algorithms

6.1.2.2 Exp 2: with removing non-alphabetic character(BOW)

In model 2 Exp 1 we worked with TF-IDF model and Lemmatization. Then the processed data is sent to the machine learning models. The table shows the comparision of the ML models on the stemmed data and figure illustrates the results.

Table 6.4: Performance with Lemmatization

No.	Metrics	SVM	LR	NB	RF	XGB
1.	Accuracy	0.560910	0.565316	0.484435	0.524947	0.490832
2.	Precision	0.566892	0.574879	0.487434	0.548457	0.533923
3.	Recall	0.586102	0.582482	0.506875	0.508781	0.485388
4.	F1 Score	0.573113	0.577766	0.494458	0.517113	0.498386

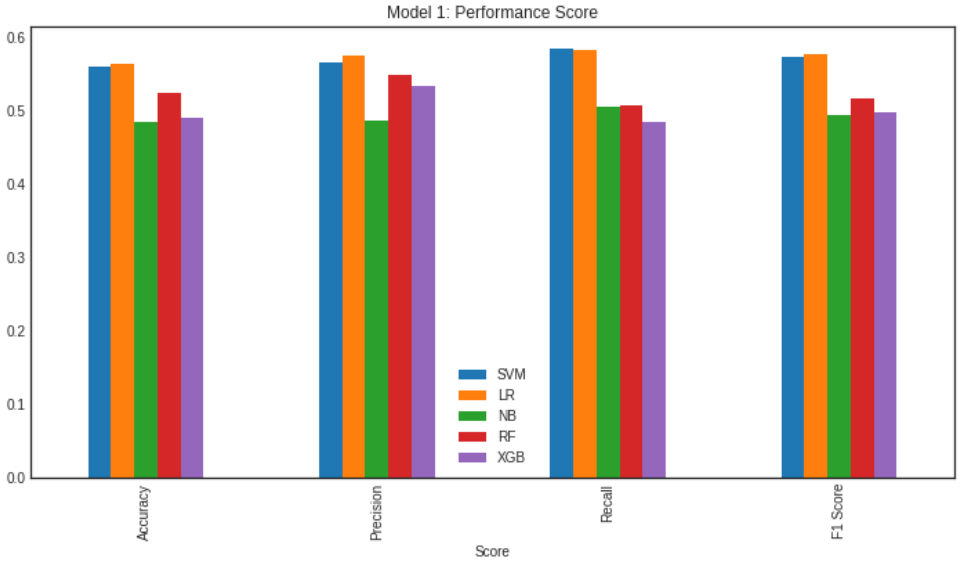

Figure 6.4: Comparison Between the Algorithms

6.2 Results

From the above illustrated data, we can conclude that we got the best result for SVM and Logistic Regression which is 56.09% and 56.53%, while using Bag of Words for feature extraction and Lemmatization with removing non alphabetic characters on the test data set.

SVM is an algorithm that determines the best decision boundary between vectors that belong to a given category and vectors that do not belong to it. It can easily search a classification hyperplane in feature space and for the generalisation capability of the classifier as well. That is a main reason why the SVM can achieves very good results.

Logistic regression is easy to implement, interpret, and very efficient to train. It is very fast at classifying unknown records. It performs well when the dataset is linearly separable. It can interpret model coefficients as indicators of feature importance.

Lemmatization, unlike Stemming, reduces the inflected words properly ensuring that the root word belongs to the language. In Lemmatization root word is called Lemma. A lemma (plural lemmas or lemmata) is the canonical form, dictionary form, or citation form of a set of words. Whereas stemming works faster and just removes or stems the last few characters of a word, Lemmatization considers the context and converts the word to its meaningful base form. This is why we got slightly better results with Lemmatization.

Chapter 7

Future Work and Conclusion

7.1 Future Work

As future works, we intent to apply different deep learning algorithms. This might give us a more better accuracy. We will utilize a different enriched dataset and intend to employ alternative feature extraction methods.

7.2 Conclusion

In this work, we try to get insights into public reaction on COVID-19. There are some analysis from social media data about how people are reacting in this pandemic period. We have made our analysis on Twitter data during this COVID-19 outbreak. With the help of the machine learning algorithms, we can view people's perspective on COVID-19 pandemic. We got the best result for SVM and Logistic Regression which is 56.09% and 56.53%, while using Bag of Words for feature extraction and Lemmatization with removing non alphabetic characters on the test data set.

References

[1] S. S. Aljameel, D. A. Alabbad, N. A. Alzahrani, S. M. Alqarni, F. A. Alamoudi, L. M. Babili, S. K. Aljaafary, and F. M. Alshamrani, "A sentiment analysis approach to predict an individual's awareness of the precautionary procedures to prevent covid-19 outbreaks in saudi arabia," *International journal of environmental research and public health*, vol. 18, no. 1, p. 218, 2021.

[2] M. G. Bhargava and D. R. Rao, "Sentimental analysis on social media data using r programming," *International Journal of Engineering & Technology*, vol. 7, no. 2.31, pp. 80–84, 2018.

[3] M. S. Al-Rakhami and A. M. Al-Amri, "Lies kill, facts save: detecting covid-19 misinformation in twitter," *Ieee Access*, vol. 8, pp. 155961–155970, 2020.

[4] X. Zhang, H. Saleh, E. M. Younis, R. Sahal, and A. A. Ali, "Predicting coronavirus pandemic in real-time using machine learning and big data streaming system," *Complexity*, vol. 2020, 2020.

[5] A. Sarlan, C. Nadam, and S. Basri, "Twitter sentiment analysis," in *Proceedings of the 6th International conference on Information Technology and Multimedia*, pp. 212–216, IEEE, 2014.

[6] https://www.kaggle.com/datatattle/covid-19-nlp-text-classification.

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 Monday 14th March, 2022 at 7:18am.