UNIVERSITY OF
South Carolina

Project Report

Title: Question Answering and Understanding Sentiments of People on COVID-19

CSCE 771: Computer Processing of Natural Language

By: Kunal Ajay Dhavale

Instructor: Dr. Biplav Srivastava

Abstract

The COVID-19 pandemic has become a worst nightmare come true for every single person on this planet. Working from home, wearing masks, maintaining social distancing, etc. are now the new normal. Several people tweeted about COVID-19. There were over 628M tweets regarding COVID-19[1]. With such massive amount of data available, our first objective is to understand sentiments of people and using that the government can formulate plans for the betterment of the country. Our dataset of tweets contains 8426 tweets relate to COVID-19. 6402 are labelled as positive tweets, 2024 are labelled as negative tweets. Nearly 24% tweets in the dataset are negative. We trained our ML model on the dataset. We then compared the result by passing the same input(tweet) to our ML model and VADER (Valence Aware Dictionary and Entiment Reasoner) Sentiment Detector. Another aspect of this project focuses on Question Answering(Q/A). We used Machine Learning models for training the dataset on both Sentiment Detection and Question Answering. The aim of Q/A is to answer specific questions related to number of cases, deaths, mask support in South Carolina counties. Accuracy for both the systems were satisfactory. Q/A system focuses on the keywords on the input question in order to return the result. Even by using simple ML models we can get high accuracy and perform complex tasks.

This report describes the implementation and evaluation of the system that was built. Implementation part includes the methods used for building the system, dataset references, etc. Evaluation includes the actual result of the system after training it on the dataset using ML models.

Acknowledgements

This project has taken considerable amount of time and resources. I would like to thank my course instructor **Dr. Biplav Srivastava** for his time, patience, guidance and for allowing me to work on the topic of my interest. His encouragement motivated me a lot.

I would also like to thank all those people who encouraged me and provided me with important advice.

Contents

Introduction

This document is a report for individual project “Question Answering and Understanding Sentiment of People on COVID-19”. This project is an attempt to create a system which can 1)predict whether a sentiment of a tweet is 0(negative) or 1(positive), 2) Answer specific questions such as “Cases of COVID-19 in counties, number of deaths in counties, mask support in counties”. The Q/A system has a dataset which consists of sample questions asked like “Hello. Can you please tell me covid cases in Charleston county?”. The data we got from NY Times for the covid cases. Manipulation was done on dataset to fit it into the project scope. For the sentiment detection our dataset consists of 8426 labelled tweets. The aim is to train our Machine Learning (ML) model on the training set and then assess the performance of the system.

There are many approaches in order to tackle this problem. Lot of cumbersome ways are out there; we are following ML model approach to get to a desired solution. For people who want to start their journey in Natural Language Processing, ML model-based approach is really a great kick start. Later then we can explore the onerous methods once our baseline is cleared about the topic.

Problem

The motivation to build a project by keeping COVID-19 as the baseline is that the pandemic is a hot topic in the world now. There are not many systems which focus exclusively on this topic. There are millions of tweets related to this topic. Due to which we get humungous information. By analyzing the emotions of the user who tweets, the government can propose new plans which can be lucrative for both the government and citizens. But our project just focuses on predicting whether the input tweet is positive/negative. Usually while performing sentiment detection, neutral sentiment is also taken into consideration. I purposefully omitted the neutral emotion and as far as this project concerns, we will have only two emotions i.e. positive and negative. The reason to omit the neutral sentiment is because this topic is soo sensitive, that the tweets mostly fall into the positive or negative category. So, the dataset has been modified a bit by removing the neutral tweets which were nearly 1-2% of the total tweets in the dataset. The prime focus is on the positive or negative emotion. Positive emotions represent joy, general statement, a descriptive fact about something, etc. Negative sentiment represent apprehension, anxiety, castigating someone, reprimand, admonish, etc. Also, there is an attempt to answer specific questions related to COVID-19. The Q/A system works on the principle on key word-based system. Which means, from the input it gives the output based on the key words in the input.

Related Works

Below is the list of works which are related to this project:

I. **CAiRE - COVID: A Question Answering and Query-focused Multi-Document Summarization System for COVID-19 Scholarly Information Management [2]**
II. **Answering Questions on COVID-19 in Real-Time [3]**
III. **Sentiment Analysis and Emotion Understanding during COVID-19 Pandemic in Spain and Its Impact on Digital Ecosystems [4]**

From the above related works, I got the motivation to build this project. Although our project focuses on simple ML based approach, but the results provided by the system are promising. I will consider this as a kick start to my journey towards Natural Language Processing and there is a lot to learn from this. This project will help me to enter this field of study and eventually tackle cumbersome topics.

Data Sources

Understanding Sentiment

http://www.kgswc.org/hackathon-2020/

This dataset is from a hackathon organized by AI Institute of South Carolina. The dataset consists of text and emojis, but while passing the input to the ML model, we will be cleaning our text as we are not focusing on emojis as far as this project concerns. Dealing with emojis could be considered as a future work for the project. Manipulation has been done on dataset to fit it into project scope.

Question Answering

1. Mask Support in counties
   https://github.com/nytimes/covid-19-data/blob/master/mask-use/mask-use-by-county.csv

2. Covid cases in counties and Number of deaths
   https://github.com/nytimes/covid-19-data/blob/master/live/us-counties.csv

   The dataset was collected on 13th September 2020. The data may be old, but the focus of the project is to show how the system works efficiently by providing the accurate results. Manipulation has been done on dataset by me in order to fit the dataset into the project scope.
   As it is the question answering system, the dataset contains sample questions and answers to the questions. The answers are fixed for e.g. “Cases in Charleston county”, “Hey man, Tell me Charleston cases of covid”, both these will result in the same answer as only the way of asking is different but both question have same meaning.

What about McCormick.	147
maximum cases of covid 19 in which county.	charleston 12956
Bamberg.	493

Fig 1.1: Dataset of covid cases

Fig 1.1 gives us idea about how the entire dataset if organized. The first column are the questions and the right column contains answers. This dataset would be passed to the ML model for training purpose.

Tweets	Sentiment
Our government has FAILED! 🤬 #TrumpFailedAmerica #RIPGOP #TrumpLiesAboutCoronavirus https://t.co/5isyRCU7RC"	0
"Fight #covid19 👊 with an updated version of your #DigitalMarketingPlan in Google Sheets so your team can work on it from home. Great for cross-team collabs and ⚡real-time⚡updates. https://t.co/uz10mK1Ks7 @UserMention	1

Fig 1.2: Dataset for Understanding Sentiment

Fig 1.2 gives us idea about the dataset used for the sentiment detection. The Tweets column contain the input for the ML model. The sentiment column describes the sentiment of the tweet. 1-Positive, 0-Negative.

Method/ Solution Steps/ Algorithm

The core method of this project is the usage of Machine Learning (ML) models. ML is application of Artificial Intelligence which provides the system an ability to learn automatically from the experiences[5].

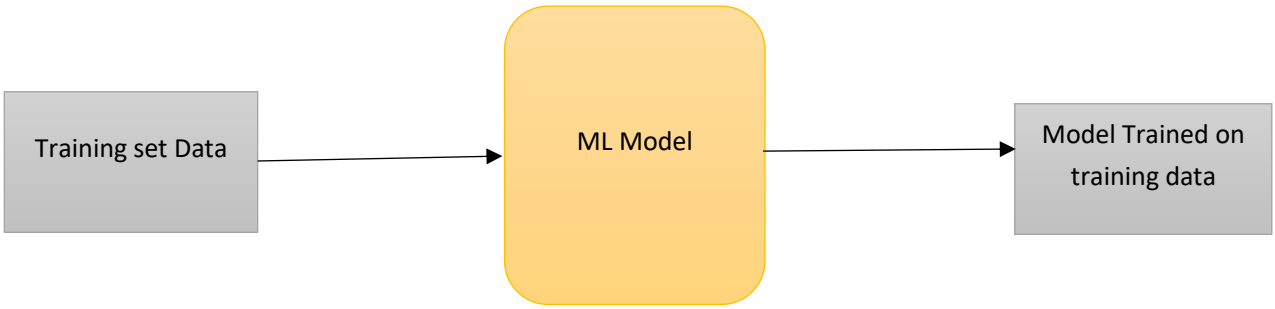

Fig 2.1: ML Method for the system (Training)

For both Sentiment Detection and Question Answering, ML model approach was decided to get the desired results. In Fig 2.1, we can see basic working mechanism of usage of ML models. We pass the training set data to the ML model in order for the model to get trained on our training data and later the performances are evaluated when the model is tested on test data.

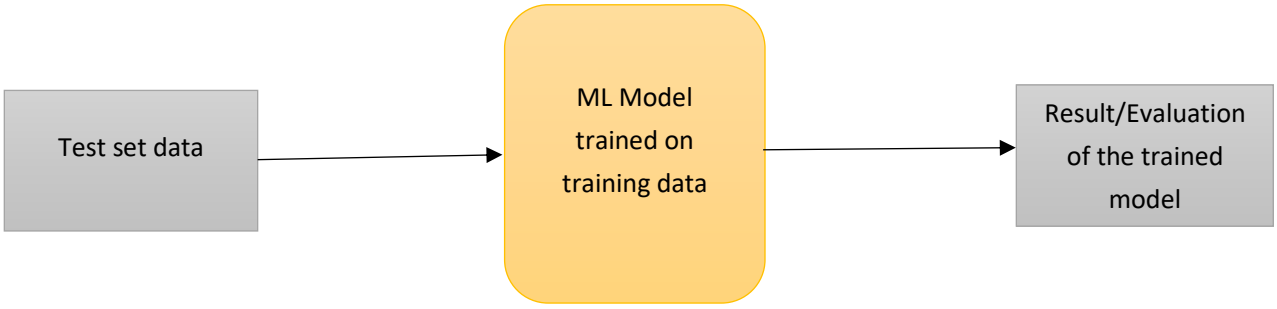

Fig 2.2: ML Method for the system (Testing)

In Fig 2.2, we can see how the test set data is passed to the ML model trained on training set data to perform evaluation of our ML model. This is the underlying structure for our project.

For both Q/A and Sentiment Detection, we follow same procedure right from performing cleaning of the text and passing the cleaned data to ML model and performing the evaluation.

One notable difference is that, for understanding the sentiment we perform additional cleaning of text as the dataset contains emojis and we are not dealing with emojis at this stage.

The cleaning of dataset consists of : converting the text to lowercase, replacing the punctuation with blank space, removal of stopwords which does not play a major role in the final result, using lemmatization/stemming techniques. Once the data is cleaned, corpus is created. It is followed by creation of Bag-of-Words model. In this model, the text is simply represented as a bag i.e. multiset of its words[6]. It simply describes the number of occurrence of words in a corpus/dataset. Then the dataset is split into the training set and test set. The training set is passed to the ML model for training purpose. Once the model is trained, we evaluate the model on our test set.

Fig 3 is an illustration that explains the approach to solve the problem using ML model.

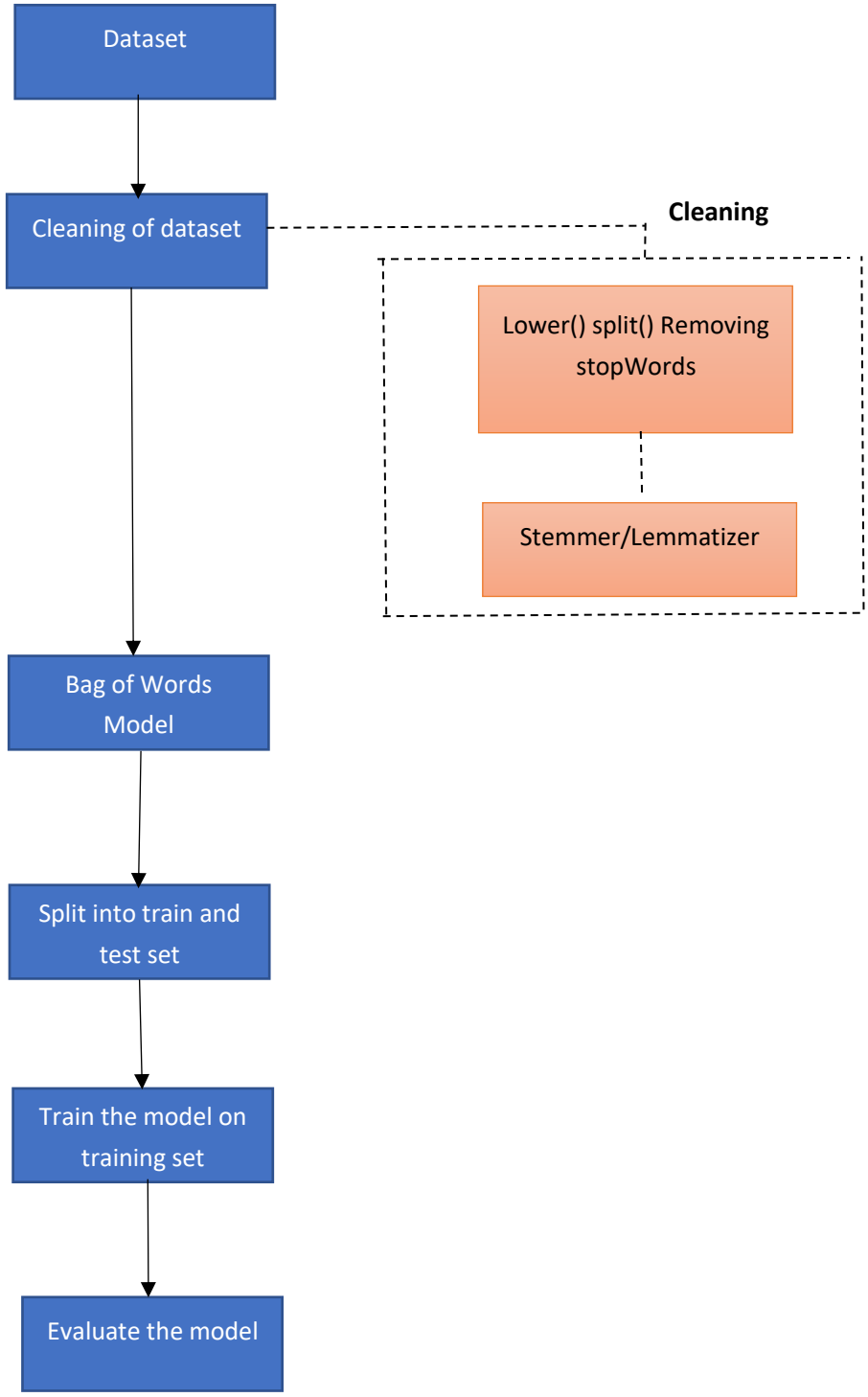
Dataset
Cleaning of dataset
Cleaning
Lower() split() Removing stopWords
Stemmer/Lemmatizer
Bag of Words Model
Split into train and test set
Train the model on training set
Evaluate the model

Fig 3: Method

Demonstration

```
press 0 to exit and 1 to continue
Tweet your thoughts: This is really disturbing. I never expected such things happen to me!!! Kill me god now ;/
 Prediction of ML model :  0
Negative

Vader:Negative
 Do you want to continue 1
Tweet your thoughts: Great news. Go INDIA!! We are soon to be a superpower!!!
 Prediction of ML model :  1
Positive

Vader:Positive
 Do you want to continue 0
```

Screenshot 1: Basic Working of Detecting the Sentiment

```
press 0 to exit and 1 to continue
Tweet your thoughts: I dont want to die. Please dont kill me corona
 Prediction of ML model :  0
Negative

Vader:Positive
 Do you want to continue 1
Tweet your thoughts:  My model is worst. Still I love it
 Prediction of ML model :  1
Positive

Vader:Neutral
 Do you want to continue 1
Tweet your thoughts: I am really very much happily sad
 Prediction of ML model :  0
Negative

Vader:Positive
 Do you want to continue 0
```

Screenshot 2: Our Model Performing Better than VADER

```
Menu
 1: Cases in county  2: Number of deaths 3: Mask support by population
1
Enter question: Hey bro tell me the covid cases in charleston
Your Answer is:  12956
The accuracy of the ML model is  0.95625
Would you like to continue 1-yes 0-no
Enter your choice 1
Enter question: cases confirmed in abbeville
Your Answer is:  366
The accuracy of the ML model is  0.95625
Would you like to continue 1-yes 0-no
Enter your choice 1
Enter question:  Where is least covid
Your Answer is:  McCormick
The accuracy of the ML model is  0.95625
Would you like to continue 1-yes 0-no
Enter your choice 1
Enter question: Maximum covid is where
Your Answer is:  charleston 12956
The accuracy of the ML model is  0.95625
Would you like to continue 1-yes 0-no
Enter your choice 0

Do you want to ask any other information 1=yes 0=no0
```

Screenshot 3: Q/A for number of cases

```
Menu
 1: Cases in county  2: Number of deaths 3: Mask support by population
2
Enter question Can you please do me a favour and tell me how many casualties occured in abbeville
Your Question is  Can you please do me a favour and tell me how many casualties occured in abbeville
 Your answer is :  0
Accuracy of model is  1.0
Would you like to continue 1-yes 0-no
Enter your choice 1
Enter question Most deaths of pandemic
Your Question is  Most deaths of pandemic
 Your answer is :  Richland 62
Accuracy of model is  1.0
Would you like to continue 1-yes 0-no
Enter your choice 1
Enter question very least deaths
Your Question is  very least deaths
 Your answer is :  Abbeville Bamberg Chester Hampton Oconee Saluda Union 0
Accuracy of model is  1.0
Would you like to continue 1-yes 0-no
Enter your choice 0

Do you want to ask any other information 1=yes 0=no0
```

Screenshot 4: Q/A for number of deaths

```
Menu
 1: Cases in county  2: Number of deaths 3: Mask support by population
3
Enter question what is the mask support in Darlington
Your Question is  what is the mask support in Darlington
 Your answer is :  54.800000000000004  %
 Accuracy score of model is  0.922077922077922
Would you like to continue 1-yes 0-no
Enter your choice 1
Enter question  how much people are for masks in charleston
Your Question is   how much people are for masks in charleston
 Your answer is :  65.7  %
 Accuracy score of model is  0.922077922077922
Would you like to continue 1-yes 0-no
Enter your choice 0
Do you want to ask any other information 1=yes 0=no0
```

Screenshot 5: Q/A for mask support

Screenshot 1 and 2 represent the demonstration for the understanding the sentiment. In 2, we can see how our model performs better than VADER in identifying the sentiment of a tweet.

Screenshot 3,4,5 represents the demonstration of Q/A. From this we get an idea about the types of questions answered by the system.

Evaluation

For evaluation of our trained ML model we use Accuracy Score, Classification Matrix, Confusion Matrix. For the Q/A, we have used only accuracy score as a evaluation method. For Sentiment Detection, we considered Accuracy Score, Confusion Matrix and Classification Matrix.

```
[[ 455   73]
 [  53 1526]]

Accuracy of the model is  0.9401993355481728
              precision    recall  f1-score   support

           0       0.90      0.86      0.88       528
           1       0.95      0.97      0.96      1579

    accuracy                           0.94      2107
   macro avg       0.93      0.91      0.92      2107
weighted avg       0.94      0.94      0.94      2107
```

Image 1.1: Evaluation of Sentiment Detection

In Image 1 we can see the evaluation of our Sentiment Detection ML model. Our model has an accuracy of 0.940 which is 94%. Even with the usage of simple ML model method we achieved an accuracy of 94%. Accuracy of Sentiment Detection system can allegedly never be 100% because same sentiments can be perceived in different manner. Even if the data is labelled, it is nearly impossible to achieve an accuracy of 100% for Sentiment Detection system.

We have also compared our model to VADER Sentiment Detector. We passed the same tweet to our model and to the VADER Sentiment Detector and compared the output.

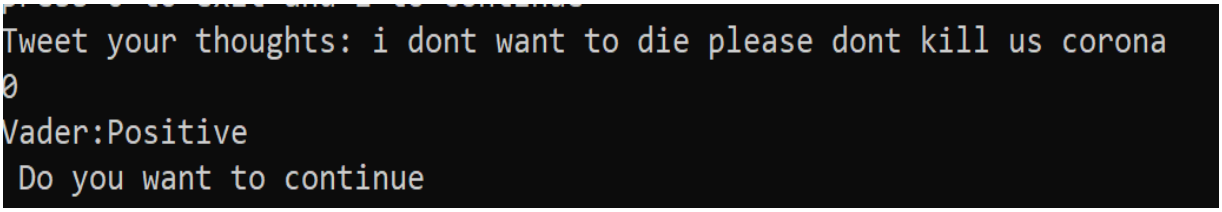

Image 1.2: Comparison of our ML model to VADER

In Image 1.2, we can see that clearly there is a conflict of opinion between our system and VADER. The input is " i dont want to die please dont kill us corona". Our model has predicted it to be a 0, which is negative, but VADER result says that the sentence is positive. Our system and VADER are much similar in performance, but it can be considered that in some cases our model performs better than the VADER.

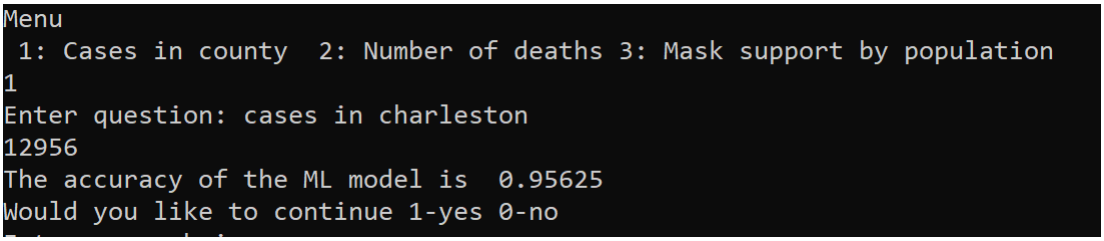

Image 2.1: Evaluation of Cases

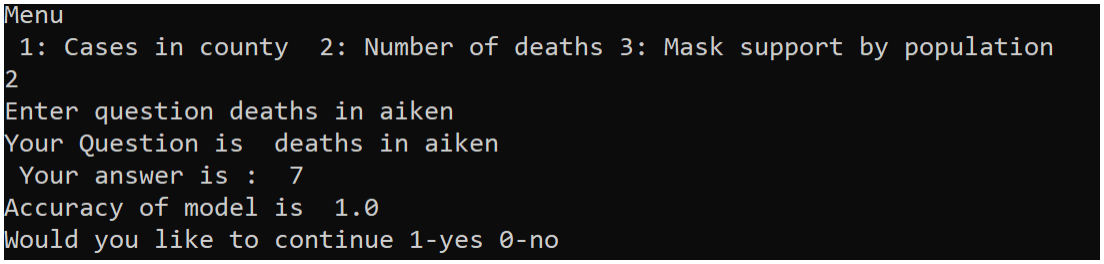

Image 2.2: Evaluation of Deaths

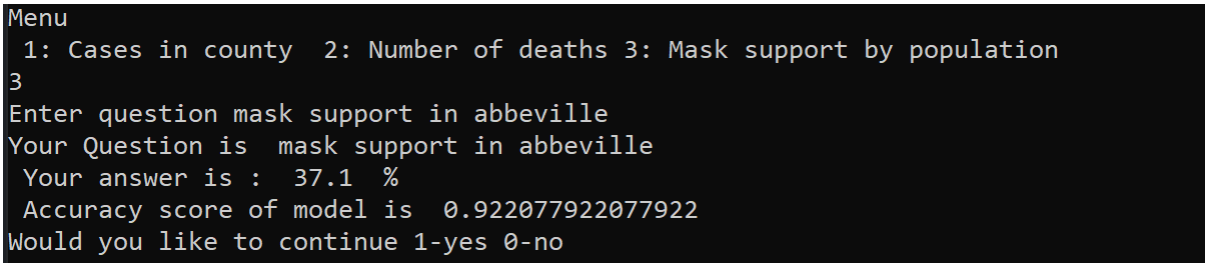

Image 2.3: Evaluation of Mask Support

Images 2.1,2.2,2.3 are the evaluation of the Question Answering system. 2.1 is the evaluation of the ML model used to answer questions related to number of cases in counties. 2.2 is the evaluation of ML model to answer questions related to number of deaths and 2.3 is the evaluation of model to answer questions related to mask support in counties. The accuracies are 95%, 100%,92%. For providing answers to number of deaths the model gives correct answer every time. But with number of cases and mask support we can see the model is not 100% accurate. The reason is that as the dataset contains types of questions. So, there are many ways to ask a single question because of which the model may input incorrect result. As our approach is simple using ML models, this can be considered as a drawback of the system that it is limited to specific domain of input data.

Significance of Work / Discussion

The motive behind building this project is to demonstrate that we can perform NLP using ML model approach we can achieve some promising results. Working mechanism of this project is on a rudimentary level, there is a lot of hope to improve this project by adopting advanced techniques to make the system even more efficacious. Our aim was to have a dataset for question answering and tweet sentiment detection/understanding and train it on ML model and evaluate the performance of our model. In the Evaluation of the model, we observed that our model provided us promising results. The accuracy of each model is over 90%. This project is a great kickstart for all of them who want to take a step towards their NLP journey.

For the Q/A part of the project, we were able to answer the questions correctly. Same for the sentiment detection, we were able to identify the sentiment behind the tweet accurately. This model could be applied to any dataset (of a specific format) and we can then quantify what is the sentiment of the population.

Future Plan

There is a lot of scope for improvement in this project. Currently we are using ML model-based approach for Q/A system. The plan is to use Natural Language Interface to Database. Using this approach, I anticipate that the results would be even more accurate. Also, the system would be able to answer questions related to meta data instead of just the available labeled data in the dataset. Deep learning could even enhance the current sentiment detection system. Usage of Deep Neural Networks will enhance the power of our system resulting in more accurate predictions.

Reference

[1] https://www.tweetbinder.com/blog/covid-19-coronavirus-twitter/

[2] https://arxiv.org/pdf/2005.03975.pdf

[3] https://arxiv.org/pdf/2006.15830.pdf

[4] https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7432069/

[5]https://expertsystem.com/machine-learning-definition/#:~:text=Machine%20learning%20is%20an%20application,use%20it%20learn%20for%20themselves.

[6] https://en.wikipedia.org/wiki/Bag-of-words_model