Classification of COVID-19 Related Tweets

Performance Comparison between different ML model and NLP techniques

Nisha Poudel
CSCE
University of Nebraska Lincoln
Nebraska, USA
npoudel2@unl.edu

Sujan Shrestha
CSCE
University of Nebraska Lincoln
Nebraska, USA
sshrestha11@huskers.unl.edu

Boyang Hu
CSCE
University of Nebraska Lincoln
Nebraska, USA
boyang.hu@huskers.unl.edu

***Abstract*—The year 2020 has truly marked itself as the darkest year in the calendar as the novel COVID-19 virus took the world by storm. Many people died, business slowed down, health conditions deteriorated, food scarcity became prevalent , unemployment rose this year and these are only the tip of the iceberg of the damages that COVID had done. Since the beginning of the year 2020, the number of cases related to COVID-19 has been continually growing. According to the latest report from CDC, there are over 35.9 million cases of COVID worldwide to this date and it's still growing exponentially. CDC also mentions that 1.05 million people have died since the outburst of COVID-19. The only thing that kept this world going was "social media". The use of social media rose exponentially during COVID crisis. People started tweeting and sharing their experience, stories, pictures, symptoms, thoughts, conditions, emotions, even fear and panic during the crisis. In this paper, we attempt to mine those tweets from the internet, by creating a model that can clearly distinguish between COVID related and non COVID related tweets.**

***Index Terms*—KNN, binary classification, data processing**

I. Introduction

'Chaos theory' suggests that even a random-looking data can have high valued information stored within it. Similarly, the COVID-19 related tweets may have some hidden valuable information stored within them. It could be the key to control the COVID-19 crisis. We believe that information on COVID-19 can not only come from databases generated by hospitals but also from tweets tweeted by people who have experienced the pandemic first hand. So, it is essential to predict any COVID-19 related tweets.

II. Motivation

By classifying tweets into COVID related and unrelated, we primarily hope to learn any new information on COVID, we hope to learn how differently people are reacting to the COVID-19 crisis from the tweets. We also hope to learn information on people's sentiment, emotions, rage, or anger in the context of COVID. The findings from our study could be further be rationalized using sentiment analysis on the tweets. In addition to that, we hope to discover any correlation among COVID-related tweets and features such as geographic location, the posted date of the tweet etc. which might help to track how different regions are handling the crisis. These information could later prove to be extremely beneficial in developing effective strategies to combat any pandemic.

III. Problem

Applying machine learning to analyze COVID-19 is a new hot topic; Currently, there aren't many good models that can distinctly identify COVID and non-COVID tweets. In addition to that, there is not enough data on COVID-19 that can support any model. Hospitals, health organizations have been working day and night to gather data on COVID, but it's just not enough. And many researches are currently on hold due to the lack of new dataset on COVID-19.

IV. Data Summary

We are using Kaggle, the world's largest data science library, to get our data. We use 3 different datasets in this project. Two of the datasets have tweets only related to COVID and the other has tweets not related to COVID. The reason to use 3 datasets in this project is primarily because we couldn't find any dataset that is large enough for the proposed research and has all the necessary features. In addition to that, twitter only released a few datasets containing only COVID-19 related tweets. So, it is up to the researchers to find other twitter datasets with similar features but are not related to the COVID-19. There are 13 total features in the selected datasets but we decided to use 6 features in this research. The features are $user\_location$, $user\_description$, $user\_followers$, $user\_verified$, $date$, and $text$ (actual tweets).

We plan to generate a dataset that contains few thousands of tweets that either COVID-19 related or not. If there is no relevant data publicly available, we will leverage Twitter API and download and label our own dataset.

V. Methodology

We use Google Colab notebook as our coding environment and Python for programming. SVM (Support Vector Machine Classifier) and NBC (Naive Bayes Classifier) are widely used and regarded as few of the best text classification algorithms. Before we run our models, we run exploratory data analysis to get some sense about our data. We then use different feature extraction techniques to pre-process and clean our data and retrieve valuable features. The techniques to be used are text normalization using Lemmatization, tokenization, Word-Embedding-by-Word2vec [1], feature vectorization, removal of stop words etc. We use Natural Language Toolkit library

offered by python for this purpose. Our intent to pre-process our data is to get a faster convergence and optimal results.

VI. Conclusion

This project aims to develop an ML model to classify tweets between COVID-19 related automatically. We run our model performance in accordance with advanced Natural Language Processing techniques and tune the hyper-parameters to find the best model that fits this tweet classification problem.

References

[1] "Word-embedding-by-word2vec," https://github.com/rhasanbd/Word-Embedding-by-Word2vec, Accessed on Oct. 2020.