CSCE 590 Final Project Report

Project Title: COVID-Related Tweets Analysis Before and After Vaccine Release to Public During COVID-19 Outbreak

Introduction

COVID-19 is a serious global health problem that persists worldwide, it has drastically challenged healthcare systems worldwide with currently 265 million cases and 5.26 million deaths since March 13, 2020 (Worldometers, 2021). During the epidemic, social media is a common way for people to share their emotions and opinions (Zhu et al., 2020). In this research work, I'll collect English language tweets during the COVID outbreak and find out the differences in people's focus and shifts in emotions when the vaccine come out. The results could provide instructive information for the health provider, medical practitioners, and the public.

Data Pre-processing

In this project, I collected Twitter data using the Twint Python package, an advanced open-source Twitter scraping tool that allows for scraping tweets from Twitter rather than using Twitter's official API (Pratama, 2020), which provided the keyword search in the history data of Twitter. Based on the CDC report (Centers for Disease Control and Prevention, 2021), the period before the vaccine release defined as the time range from 03/01/2020 to 11/30/20220. And the period after the vaccine release defined as the time range from 12/01/2020 – 08/30/2021. The keywords set including the most common words regarding the COVID-19 topics, which including COVID, CORONA, VIRUS, CORONAVIRUS, and EPIDEMIC. For each period, I collected 10000 tweets before the cleaning.

After collect the tweets from different period, I cleaned and pre-processed the dataset to make sure they can fit into the machine learning models and improve the performance of machine learning models. Following tweets preprocessing strategies used in previous literature (Irfan, 2020), the initial step was to remove all tweets are not writing by the English language and the length of tweets less than five words. After that, I removed all special characters, URLs, hashtags, emojis, mentions, punctuations, ticks with the next character, numbers, and over spaces. To remove the meaningless words from the tweets, I applied NLTK library used in previous research (Loper and Bird, 2002) to recognize and remove the stop words. Then I converted tweet text to lowercase for all characters to avoid case sensitivity and tokenize the cleaned tweets into separate words and convert them into numerical vectors as inputs of machine learning and deep learning models by the function called Term Frequency-Inverse Document Frequency (TF-IDF), which is a popular method to convert the textual value to numerical value (Zhang et al., 2011).

Methodology

In this study, I first applied word frequency analysis to output the word cloud to distinguish the difference of word frequency between the two periods. And build classification models to compare the performance and cross validate the reliability of feature analysis results. To find the most important words which are relevant to the period classification process, I applied random forest to analyze the feature importance during the before/ after vaccine prediction process. And output the ranking of important features.

Results

The word cloud output as shown in the figure 1 and 2. From both figure we could find that the trends for infection, related symptoms, health impact is most common COVID topic in the Twitter community before vaccination. After the vaccine related, unvaccinated cases, unmask, and how to get free vaccine is more popular in the community.

Timeline

Description automatically generated

Figure 1. Word could for the period before the vaccine release

Text

Description automatically generated

Figure 2. Word could for the period after the vaccine release

Then I compared the performance of different machine learning models and find that random forest achieved the best performance as 86% accuracy (Figure 3).

Chart, bar chart

Description automatically generated

Figure 3. Machine Learning Models Performance Comparison

The we use the feature importance calculation function provided by the random forest model to output the importance ranking among all tweets I collected, the results as shown in the figure 4.

Chart

Description automatically generated

Figure 4. Feature Importance Ranking by Random Forest

From the feature important analysis, we grouped most important words into these two groups based on the literal meaning, not consider the context of tweets, and the words from these two categories could happen in the same tweets. These three groups including (1) COVID-related common words, including COVID, vaccinated, unvaccinated, and vaccine. (2) The society orientation words, including trump, florida, floridans, meeting, children, people, kids, and work. (3) The health-related words, including relief and hospitals.

Discussion

In this study, we collected the tweets before and after the COVID-19 vaccine release and analyze the word frequency after the data cleaning. Then build different machine learning classification models based on the popular supervised machine learning algorithms and output the feature importance ranking from the best performance model: random forest. The current finding could help us find the difference of public focusing before and after the vaccine release better and provide the fundamental results for the further study. In the further study, we could focus on the data annotation and the sentiment analysis to understand the dataset better.

Reference

1. Worldometers. (2021). COVID-19 CORONAVIRUS PANDEMIC. https://www.worldometers.info/coronavirus/?utm\_campaign=homeAdvegas1?%22 [Accessed Nov 20, 2021].

2. Zhu, B., Zheng, X., Liu, H., Li, J., & Wang, P. (2020). Analysis of spatiotemporal characteristics of big data on social media sentiment with COVID-19 epidemic topics. Chaos, Solitons & Fractals, 140, 110123.

3. Centers for Disease Control and Prevention. (2021). CDC Museum COVID-19 Timeline. <https://www.cdc.gov/museum/timeline/covid19.html> [Accessed Nov 20, 2021].

4. Pratama, A. (2020). How to scrape Tweets from Twitter with Python Twint, www.medium.com [Accessed Nov 20, 2021].

5. Irfan Alghani Khalid. (2020). Cleaning Text Data with Python. https://towardsdatascience.com/cleaning-text-data-with-python-b69b47b97b76 [Accessed Nov 20, 2021].

6. Loper, E., and Bird, S. (2002). Nltk: The natural language toolkit. arXiv preprint cs/0205028.

7. Zhang, W., Yoshida, T., & Tang, X. (2011). A comparative study of TF\* IDF, LSI and multi-words for text classification. Expert Systems with Applications, 38(3), 2758-2765.