Will Carnevale 8.20.2021

Business Project Write-Up: Twitter Sentiment and COVID

Abstract

The goal of this project was to pitch the US Federal Government a data science solution to mitigating large-scale viral outbreaks like COVID. The data science solution proposed was to perform natural language processing (NLP) on a large dataset of Twitter posts and feed the results into a classification model, which could classify posts into more granular topics. This Topic analysis would work in conjunction with a time series analysis of COVID data to predict areas of vulnerability that may be at risk for being ground-zero for a large viral outbreak. In order to demonstrate the importance and viability of the project, I used data on a set of Twitter posts related to the COVID pandemic from 2020 to explore the differences between posts with different sentiments. Notably, Tweets that have negative sentiment tend to mention supermarkets. After exploring and analyzing the data in Google Sheets, I visualized some preliminary insights with Tableau.

Design

Opportunity

The onset of the COVID pandemic has awakened the world to an alarming reality: we are ill-equipped to respond swiftly and with force when a novel virus begins to infect all of humanity. Part of the issue is not knowing where a response is needed, and also not being able to respond quickly enough for containment. The Problem is: how can we efficiently gather actionable information to make informed decisions on allocations of resources?

Impact

The desired impact of this project is to utilize twitter posts to predict areas of vulnerability for viral (COVID) outbreaks so that we can quickly bolster the medical infrastructure to support increased caseloads. Further desired impacts include determining an appropriate allocation of aid so that less resources are wasted on areas that may not need the support.

Impact hypothesis

Performing NLP on Twitter data and combining with a time series analysis of COVID

caseload data will add predictive power to governing bodies to make quicker decisions with regard to allocating resources.

Assumptions

Twitter sentiment holds insights that can be used to predict areas of vulnerability for viral outbreaks.

Data

For the preliminary exploratory data analysis (EDA), I used two datasets.

The first dataset is a dataset of tweets (International) cataloged on the website *Kaggle* as a text classification challenge. It is open source (provided you have an account for *Kaggle*) and can be found here: https://www.kaggle.com/datatattle/covid-19-nlp-text-classification. The raw data consists of nearly 3800 rows and 6 columns. Each row represents an individual tweet and the columns include UserName (enumerated as a list of numbers for privacy), ScreenName (same as UserName?), Location (city where tweet was sent), TweetAt (date that tweet was sent), OriginalTweet (the Tweet itself), and Sentiment.

The second dataset is a time-series dataset of COVID cases and deaths (in the United States only) provided by Johns Hopkins. It is also open source and can be found here: https://www.kaggle.com/codebreaker619/john-hopkins-covid-19-case-tracker-dataset. The raw data consists of nearly 20000 rows and 13 columns. each row is the number of new Covid cases reported in a given day at a city-level location. The columns for this dataset include state, date, total population, cululative cases, cumulative cases per 100000, cumulative deaths, cumulative deaths per 100000, new cases, new cases 7 day rolling average, new deaths, new deaths 7 day rolling average, new deaths per 100000, and new cases per 100000.

Algorithms

In Google Sheets, I cleaned the data, removed errant data, and aggregated certain data. I then performed some preliminary exploratory analysis and visualized my preliminary insights in Tableau.

Tools

- Google Sheets for EDA
- Tableau to visualize data