Introduction & Motivation

Modern wireless social networks are powerful tools in providing public opinions on social issues. In light of the COVID19 pandemic spread in the US, microblogging networks like Twitter provide a platform for people to express their opinions. These opinions could be interpreted as happiness and displeasure. Public uncertainty makes it difficult for stakeholders (e.g. the government) to determine the effectiveness of the policies that have been to manage COVID19 affairs While the public reviews on microblogging platforms could alleviate some of this uncertainty, they also pose a greater challenge to analyze the numerous public sentiment messages made on a single topic. Inability to accurately analyze the general public sentiments could ultimately hinder any further development of strategic policies made on COVID19.

Objectives

In our project, we aim to perform natural language processing on Twitter posts and achieve the following

- Analyze the type of sentiments based on features like location, age, time, etc.
- Estimate the overall sentiment of COVID19-related tweets
- Correctly classify COVID19 tweets as extremely negative, negative, positive, or extremely positive

Our project will be done in three phases; (1) Pre-processing and cleaning original twitter texts, (2) Performing data analysis and extrapolation, and (3) Creating a supervised learning classification model.

Significance & Applicability

The analysis portion of the project will allow us to understand the characteristics of a population that may need mental health support and medical assistance. With the classification model product in the last phase of the project, stakeholders will be able to easily analyze the huge dataset of public comments and obtain general sentiments on a topic. This allows them to understand the general population's needs during pandemics or public health crises, and make concise follow-up plans to address them. The classification model will also be applicable to other public-opinionated topics like the U.S. healthcare system.

Dataset

The dataset used will be from Kaggle that contains 4 key features; (1) Location, (2) Date, (3) Original Tweet, (4) Sentiment Label Categories. It contains approximately 45000 samples and can be obtained from the following hyperlink.
https://www.kaggle.com/datatattle/covid-19-nlp-text-classification