Problem Statement

How can we use data from Twitter to assess public attitudes toward the COVID-19 pandemic on a five-point scale within the next three weeks?

Context

For the past 22 months, the COVID-19 pandemic has wracked the globe. In these trying times, many people have turned to Twitter to express their fears, hopes, and frustrations regarding the pandemic. Their tweets provide a possible metric for assessing public attitudes toward the pandemic and the measures put in place to combat it. Given the large volume of tweets produced each day, however, it is not feasible to read and analyze every tweet by hand. A method for automatically assigning sentiment scores to tweets about COVID-19 would be extremely helpful.

Criteria for Success

The primary criterion for success will be the development of an algorithm for automatically assigning sentiment scores to tweets about the COVID-19 pandemic.

Scope of the solution space

The solution space consists of tweets referencing the COVID-19 pandemic.

Constraints within the solution space

I foresee two constraints within the solution space. First, sentiment is culturally and individually conditioned, so there's always the possibility of bias. Second, the majority of the tweets within the dataset come from the very beginning of the pandemic when emotions were highly polarized. Some saw the pandemic as the harbinger of the apocalypse, while others believed that it would be over within a few months. It is possible therefore that the dataset contains a larger proportion of extremely negative and extremely than we would expect 22 months into the pandemic.

Stakeholders to provide key insight

The creator of the dataset, Aman Miglani, might be able to provide insight into how he assigned sentiment scores to each tweet.

Key data sources

The data for this project will come from the Corona Virus Tweets NLP - Text Classification dataset (https://www.kaggle.com/datatattle/covid-19-nlp-text-classification/tasks?taskId=5183), a collection of 44995 tweets about the ongoing COVID-19

pandemic. This dataset contains the original text of each tweet, its date, the location of its sender (where known), and a manually assigned sentiment score ranging from extremely negative to extremely positive.

Deliverables

The deliverables for this project consist of GitHub repository containing the code for the project as well as a slide deck and project report summarizing the modelling results and key findings.