Project Proposal

Math189Z - Covid-19: Data Analytics and Machine Learning

Nico Espinosa Dice

April 23, 2020

1 Project Overview

For my project, I will conduct sentiment analysis on Tweets regarding the Covid-19 pandemic. In doing so, I hope to gain an understanding, and possibly a predictive method, of determining how the severity of the virus spreads geographically, using the sentiment of Tweets from that region.

The goal of sentiment analysis is to determine the attitude and emotion present in a given form of communication. Sentiment analysis applies natural language processing techniques in order to do so.

2 Data

I will be using a dataset containing over 150,000,000 Tweets regarding the Covid-19 pandemic. The website below also contains a cleaned version of the dataset.

- Twitter Dataset of Tweets Related to Covid-19
- Github: Covid-19 Twitter Dataset

3 Methods

I have found academic papers, websites, and posts that can help guide my approach towards this problem.

- Sentiment Analysis on Twitter Using Naive Bayes Classification
- Comprehensive Hands on Guide to Twitter Sentiment Analysis with dataset and code
- Sentiment Analysis of Twitter Data
- Sentiment Analysis of Twitter Data
  - While named the same, this paper is different than the one listed above.
- Covid-19 Tweets Dataset and Statistics

I believe that this project offers an opportunity to apply the probabilistic perspective that I have learned in Math189R and Math189Z. (Bayes' theorem is especially applicable to this problem). I will likely follow the Naive Bayes classifier approach.

A Naive Bayes classification algorithm is a type of probabilistic classifier that uses Bayes' theorem for predictions. By applying Bayes' theorem, the algorithm is able to make a prediction based on previous “odds” or probability estimates of past events.

4 Team

Currently, I do not have any partners for this project. I am open to others joining this project or continuing alone.