Subject: Project Proposal
Team members: Fan Zang, fanzang2; Chun Yang, chuny2; Yanlin Liu, yl128
Captain: Yanlin Liu

Sentiment Analysis on COVID-19 Tweet

Introduction

We plan to build a sentiment analysis model which focuses on posts related to COVID-19 on Twitter. More specifically, we hope to explore the underlying sentiment of public opinions about COVID-19. Our sentiment analysis mode should be able to classify each tweet into three categories: positive, negative, and neutral.

Background

Except for news and CDC data, public opinion on social media is also an important research topic. The COVID-19 pandemic affected everyone's life and generated unprecedented challenges for society. We are interested in how people reacted to Covid-related issues on Twitter during a certain time period, and how their opinions changed over time.

Goals

We want to understand how people feel about a specific topic, this model can be later reused for different topics. We could also examine the feeling of the public towards not only the pandemic, but also the government strategies and vaccines. We want to generate a model that would help us understand the public opinions on twitter more intuitively with the help of information retrieval. Hopefully, our model can:

1. Allow a quick sentiment analysis of user input messages.
2. Display the trend of public opinion about COVID-19 during a user-selected period.
3. Analyze the sentiment of public feedback about different types of COVID vaccines.

Outcomes

We expect to explore the following outcomes:

1. Use the test dataset from SemEval to calculate the accuracy, precision, and recall of our model.
2. Calculate sentiment results of public opinions during a certain period of time
3. Compare the sentiment results with official news reports and CDC statistics.

Methodology

We plan to develop a scraper using Selenium Web Driver & Tweepy API to retrieve tweets related to COVID-19 cases and vaccines. Then we plan to use the bag of word model to vectorize tweets and use the open dataset from SemEval to train a logistic regression model for sentiment prediction.

We will use the following programming languages and frameworks:

Backend: Python, React, AWS
Frontend: Python, HTML, CSS, Javascript

Workload and task schedule

Week	Task	Workload
Week 10	Research on previous papers	10 hrs
Week 11	Retrieve and clean data from Twitter	10 hrs
Week 12	Analyzing the data	10 hrs
Week 13&Week14	First draft & Front end UI constructed	15 hrs
Week 15&Week16	Model evaluation & final draft	15 hrs