Project Description

Team RAAM is studying the changes in anxiety in essential workers on Reddit in comparison to a general population during the onset of COVID-19. Our main research question is: How did COVID-19 impact the different types and rates of anxiety in certain impacted worker positions in comparison to a more general population? In order to gather the data for our measurements, we will be using three different sources. Our scraped data source will be scraped from the Reddit subreddits r/Nursing, r/Teachers, and r/HealthAnxiety. Our general population anxiety data will come from the CDC and WHO. Once we have scraped our relevant data from r/Nursing and r/Teachers, we will use natural language processing to identify the relative rate of anxiety increase on these subreddits versus the relative rate of anxiety increase in the data from the CDC and WHO.

Additionally, we will use topic modeling, specifically VADER, to cluster Reddit posts that have been identified as containing language that is related to anxiety. We will then go back and assign labels to those clusters and compare/contrast the relative frequency and increase of each type of anxiety. We expect that health-related anxiety will increase, and we also expect to see general anxiety increase in r/Nursing and r/Teachers to be greater than the increases in the general population.

Data Sources

Data Scraping

We wrote a Python file to identify and store links to all posts under the subreddits r/HealthAnxiety, r/Nursing, and r/Teachers between February 2019 and February 2022 under the "top" section. We sent out 100,000 requests and scraped data from the # of posts currently under re-evaluation.

We then built a scraping script to extract detailed information from the collected posts, including titles, post bodies, timestamps, scores, and user details. Additionally, up to 50 comments per post were retrieved, capturing comment text, scores, timestamps, and usernames. The scraping process was designed to minimize missing or inaccurate data while maintaining compliance with Reddit's request policies.

Data Cleaning and Wrangling

All of the reddit data has been successfully scraped. We have scraped from r/HealthAnxiety, r/Nursing, and r/Teachers. It has also been organized into an SQL database since the dataset size was large and a bit confusing (comments per post meant a repetition of posts). These are available on our repo. The data was organized into 3 tables per reddit page: a raw table that contains all the scraped data as-is, a posts table that contains only unique posts (as in the raw table each post is repeated with each new comment), a comments table that contains all the comments, linking them to the unique post_id they were paired with.

The data is cleaned for three things:

```
WHERE comment_body = "No Comment"
OR comment_body = "[deleted]"
OR comment_username = "AutoModerator"
```

This deletion statement looks for an empty, deleted, or bot comment.

Next, for LIWC, we must clean, tokenize, and stem/lemmatize all the textual data before using these documents for topic modeling. After topic modeling, we then must generate an anxiety score for each time unit to track anxiety over our selected period.

Data Analysis and Visualization

We have done exploratory data analysis and visualization for both our Reddit dataset and our supplemental data from the CDC and WHO. These include some preliminary time series graphs and bar graphs. However, we will do more after we will apply VADER to our Reddit posts and comments to track changes in sentiment. We will also apply LIWC to the Reddit dataset.

Responsibilities

Past roles:

- Scrapping Reddit: Amrita and River
- Progress Report 1: Anita and Mia
- Additional Datasets: Anita and Mia

Current roles:

- README file: Anita and River
- Data cleaning: Mia and Amrita
- SQL Database: Mia
- VADER: Mia
- LIWC: Amrita
- Supplementary Data Analysis: Anita and River
- Exploratory Data Visualizations: Anita and Mia

Future roles:

- Data visualization: Anita and Mia
- Presentation and slides: Amrita and Anita
- Video: Mia and River
- README: Anita
- Final Report: Everyone