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, r/HealthAnxiety and our general population anxiety data will come from 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 CDC and WHO.

Additionally, we will use unsupervised k-means clustering to cluster Reddit posts that have been identified as containing language that is related to anxiety. We will then go back and assign labels onto 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

- Reddit
    - Contains: Text information from title and body of reddit posts, from 2019-2021
    - Subreddits: r/Nursing, r/Teachers, r/Health Anxiety
    - We will scrape Reddit posts with text information from the time period of January 2019 to December 2021. We understand that this time frame is not a definitive or comprehensive scope in terms of the pandemic's effects and that it does not necessarily include the "end" of the pandemic; however, we are most interested in the effects brought on by the onset of the COVID-19 pandemic, and our selected time frame encompasses at least one year of active quarantining and public health crisis and at least one year prior to the onset of the pandemic for baseline comparison.
- CDC Household Pulse Survey & NHIS
    - Contains: CDC HPS contains survey information on Americans for anxiety levels/symptoms from 2020 to 2024, NHIS supplements these numbers for the year of 2019
    - There is some concern because NHIS and CDC HPS are separate studies, plus the discontinuity between these two studies happens to be right around the onset of the pandemic. However, CDC operates both of these studies and recommends using NHIS as a baseline for comparison on the HPS data.
- WHO Global COVID-19 Trends and Impacts Survey

    - Contains: Worldwide survey data on anxiety and financial worry from January 2020 to March 2022
    - The dataset does not extend to the end period of our scraped data, but it should provide a point of comparison for our scraped reddit data
- NCHS Indicators of Anxiety or Depression Based on Reported Frequency of Symptoms During Last 7 Days
    - Contains: Indicator (depressive disorder or anxiety disorder symptoms), subgroup, state, date, value from April 2020 to January 2022
    - The dataset does not begin at the starting timeframe of our scraped data, but it should provide a point of comparison for our scraped reddit data

Data Cleaning and Wrangling

- For Reddit Data:
    - Cleaning: stripping punctuation, letter case, tokenizing
- For CDC Data:
    - Reformatting Datetime data
- For WHO Data:
    - Reformatting Datetime data
- For NCHS Data:
    - Clean and sort data - remove depressive disorder, reformatting datetime data, group on time

We will merge these datasets on date so that we can have a comparative visualization of each of these sources of data.

Data Analysis and Visualization

Dependent Variables:

- Anxiety Levels: Measured using sentiment analysis and keyword frequency in Reddit posts.
- Types of Anxiety: Clusters derived from topic modeling (e.g., health-related anxiety, job-related anxiety).
- Rate of Anxiety Increase: Change in anxiety-related language over time in each subreddit.

Independent Variables:

- Worker Group (Essential Reddit Users vs. Anxious Reddit Users vs. General Population): Data from r/Nursing and r/Teachers vs. r/HealthAnxiety vs. CDC/WHO reports.
- Time Period: Pre-COVID vs. early-COVID phases (approximately January 2019 to December 2021 where available).

- Anxiety-Related Keywords and Phrases: Used to classify posts into anxiety-related categories.
- Other potential independent variables:
    - Demographic information: Age, gender, residing state, race

Data Analysis Methods:

- **Sentiment Analysis & Clustering**
    - Using unsupervised k-means clustering, we will group posts that contain anxiety-related language.
    - These clusters will then be labeled and analyzed for frequency trends over time.
    - Reddit posts will be classified as having positive, neutral, or negative sentiment, with a focus on anxiety-related negativity.
    - We will track changes in sentiment over time within the subreddits.
- **Comparative Analysis**
    - We will compare the rate of anxiety increase between essential worker subreddits, health anxiety subreddits, and the general population using CDC, WHO, and NCHS reports.
    - Statistical techniques such as t-tests, ANOVA, and/or regression will determine significant differences between groups.

Visualization:

Potential Data Visualization Options:

- **Time-Series Graphs**:
    - Anxiety levels over time in r/Nursing, r/Teachers, and general population data.
    - Changes in sentiment scores pre- and post-COVID onset.
- **Word Clouds & Heatmaps**:
    - Most frequent anxiety-related words in essential worker discussions.
    - Distribution of anxiety topics over time.
- **Scatter Plots**:
    - Correlation between anxiety-related keywords and major COVID-19 events (e.g., lockdowns, case surges).

Responsibilities

Current roles:

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

Future roles:

- Data labeling: Amrita and Mia
- Label Review: Anita and River
- Data visualization: Anita and Mia
-
- Presentation and slides: Amrita and Anita

- Video: Mia and River
- README: Anita
- Final Report: Everyone

Workplan (optional): If helpful for your group, create a task list with a timeline specifying when each task should be completed before the presentations and final submission.