Proposal: Assessing Suicide Risk in Social Media Posts Using NLP
Team: Branndon Marion, Louis Wu, Shruti Jain
DS266 Fall 2024

Project Overview

We aim to assess suicide risk in social media posts comparing results from pre and post-COVID periods. Our goal is to develop a classification model that differentiates between posts indicating suicide risk and those related to other mental health conditions such as, but not limited to, depression and anxiety. This task will be challenging due to the subtle linguistic differences between these mental health issues, requiring sophisticated methods beyond keyword-based models.

Motivation

Suicide is among the top causes of death globally, making its early detection critical. Current methods broadly classify mental health issues, but accurately identifying suicide risk is essential for effective intervention. Differentiating suicide ideation from other mental disorders is challenging because these conditions often share similar linguistic patterns. Our model will address this by focusing on these nuanced differences.

Datasets

We will use the **Reddit Mental Health Dataset (2018-2020)**, which includes posts from 28 subreddits, covering pre- and early COVID-19 periods. We have developed our own **web scraper** using the Reddit API to acquire more recent posts. This will enable us to create separate training and testing datasets based on pre- and post-COVID timeframes. We will only mark the posts in r/SuicideWatch as positive risk, but also plan to include a broader set of mental health conditions including but not limited to depression and anxiety as no-risk posts.

Algorithms

We will explore transformer-based models like **MentalBERT** and **DisorBERT**, which are tailored for mental health language. These models will be compared with traditional models like the **base BERT** model as baseline models.

We will use the F1 score to evaluate the performance of each model across the entire testing dataset. We also plan to conduct the following experiments which span both NLP and traditional machine learning:

- Changing the number of trainable layers in the variations of BERT models
- Comparing the accuracy by subreddit groupings (e.g. how does the difference between the accuracy of r/depression and r/fitness vary by model?)
- Evaluating the model's generalizability over time based on the date of the post
- Comparing the accuracy of non-risk posts between users who have also posted on r/SuicideWatch in the same timeframe vs those who have not

References:

- Natural Language Processing of Social Media asScreening for Suicide Risk
- MentalBERT: Publicly Available Pretrained Language Models for Mental Healthcare
- DisorBERT: A Double Domain Adaptation Model for Detecting Signs of Mental Disorders in Social Media
- Natural Language Processing Reveals Vulnerable Mental Health Support Groups and Heightened Health Anxiety on Reddit During COVID-19: Observational Study