**Social Networking & Computational Intelligence (Option 2):  
Overview of Reddit Mental Health Dataset**

Abuchi Godswill Okeke

School of Computer and Information Sciences, University of the Cumberlands

MSAI532 - M51 Natural Language Processing

Dr. Toni Farley

January 29th, 2025

**Dataset link:**

<https://zenodo.org/records/3941387>

**Detailed Description of the data set:**

The dataset pulls posts from 28 Reddit communities between 2018 and 2020, with a focus on mental health discussions. The dataset was used to understand the impact of COVID-19 on mental health support groups (Low et al., 2020).

The data is broken down into 28 different mental health and non-mental health subreddits, including:

1. *15 Mental Health Support Subreddits*: Including communities like r/depression, r/anxiety, r/suicidewatch, and r/ADHD.
2. *2 General Mental Health Subreddits*: r/mentalhealth and r/COVID19\_support.
3. *11 Non-Mental Health Subreddits*: For comparison, like r/fitness, r/relationships, and r/personalfinance.

**Data De-identification and User Protection:**

The researchers collected the dataset using the **Pushshift API**, which provides publicly available Reddit data subject to **Reddit API terms** (Low et al., 2020). To protect user privacy, the dataset does not include personally identifiable information (PII) such as usernames or user IDs. The focus is on the textual content of the posts, and any metadata that could potentially identify individual users has been excluded. Doing this ensures that the data is de-identified, maintaining user anonymity while allowing for meaningful analysis of mental health discussions.

**Potential Data Exploitation Risks**

Despite the de-identification measures the researchers used, there is a risk of individuals being identified through context information within the posts, especially if combined with external data sources. For instance, unique phrases, detailed personal stories, or specific circumstances described in the posts might be traced back to individuals by those familiar with the situations. Also, analyzing sensitive topics like mental health without appropriate ethical considerations could lead to stigmatization or misuse of the information.

**Python Script for Loading Dataset Link:**

<https://github.com/Buchiexplores/MSAI-532/blob/main/week4/load_reddit_mh_data.py>

**Reference**

Low, D. M., Rumker, L., Torous, J., Cecchi, G., Ghosh, S. S., & Talkar, T. (2020). Natural Language Processing Reveals Vulnerable Mental Health Support Groups and Heightened Health Anxiety on Reddit During COVID-19: Observational Study. Journal of medical Internet research, 22(10), e22635.