ASSIGNMENT 1:
Setting up your project

IL: Data Science for Social Good
Minerva University

Link to repo: https://github.com/zeineb-ouerghi/DS4SG-Project-

Title:

Depression Classification Using NLP with User Tweets

Research Question:

Can we use NLP on Tweets in addition to the user profile's activity ((followers, friends, favorites, retweets, number of posts) to make predictions about the user's mental health state and to what extent are these predictions accurate?

Topic Justification:

Major depressive disorder (MDD), also known simply as depression, is among the most prevalent psychiatric disorders globally, affecting more than 5% of adults worldwide. Depression alone affects more than 300 million people worldwide and is one of the largest single causes of disability worldwide, particularly for women. At its worse, depression can lead to suicide, which causes more than 700,000 deaths per year.

Effective treatment of depression requires early detection so that rapid intervention can be taken to reduce the escalation of the mental disorder. Social media is a valuable tool for this purpose since users' social media footprint reveals information regarding their behaviors, activities, thoughts, and feelings, which can be indicators of their mental health state. Additionally, social medias showcase the behavior of people among networks. By studying

patterns in user content and their engagement with others, we can study how mental illnesses can spread or become reinforced within social networks and implement necessary interventions.

Our goal for this study is to understand how to apply machine learning methods (particularly NLP) to predict users' mental health state. Our dataset allows us to perform binary classification after analyzing users' tweet content and studying other factors that are reflective of their social media usage. We are planning to use sentiment analysis as the main tool. Due to dataset and time constraints, our exploration will be rather simple compared to many published research (as this is a very popular and fruitful topic). In spite of that, we believe the project will be valuable for us in terms of learning data science techniques and applying NLP to a real project. We hope to deepen our understanding of how social media can be used for depression detection and reach better performance with our models on this dataset than existing ones. If time allows, we hope to apply our model to other datasets of labeled Tweets to evaluate its performance and confirm that it can be a valuable tool for depression detection.

Data Collection:

After searching through multiple datasets, we found one on Kaggle that we thought would work well for our purpose. The data is collected using Twitter API containing 20,000 tweets in total, equally divided between tweets from depressed users and non-depressed users. The data is in uncleaned format and has been filtered to only keep the English tweets.

Dataset:

https://github.com/zeineb-ouerghi/DS4SG-Project-/tree/main/Data

Codebook:

https://github.com/zeineb-ouerghi/DS4SG-Project-/tree/main/Codebook