Analyzing Political Tweets on a Depression Prediction ML Model

Sam Spell, James Tipton

Team Responsibilities:

- Sam: Data cleaning and data analysis presentation.
- James: ML fitting and cross-validation, data analysis and acquisition.

Real world task: Political rhetoric and discussions have seemingly become more polarized recently. In history and while reaching adulthood, being able to vote and be a part of politics is a very important role in a stable and healthy society. This project aims to use machine learning to develop a model to predict depression based on a string of text from twitter. Once this model is developed, it can be used to conduct an analysis on political messages sent online. We will be able to draw out patterns in twitter texts that the machine learning model classifies as showing signs of Depression. Another goal of this machine learning model is to extract patterns of text that can be connected to patterns of political messaging if they exist, and to compare this to a temporal aspect. With the changing view on polarized politics, it will be interesting to test if there is a change in the prevalence of messages classified with "depression" throughout different political times.

Machine Learning Model:

For this project, we will be training the model on depression classified strings, and using that to predict the sentiment of political tweets as strings. In order to do this, we will use a Support Vector Machine (SVM), a supervised Machine Learning model that can classify strings. In order to use the SVM with string data, we will need to do some processing of the data, which can include removing stop words, and turning the string into a word frequency vector.

Datasets:

https://www.kaggle.com/datasets/infamouscoder/depression-reddit-cleaned
https://www.kaggle.com/datasets/infamouscoder/mental-health-social-media
https://www.kaggle.com/datasets/neelgajare/liberals-vs-conservatives-on-reddit-13000-posts

Approach and Novel Additions:

Depression is a prevalent problem in the United States and around the world, as well as politically polarizing with respect to stereotypes and stigma. These stereotypes around treatments for depression can potentially cause people to overlook getting help and finding solutions. This machine learning model will be able to analyze a string of text and classify based on being trained on a depression data set. The combination of mental health disorders and political rhetoric in the media is important to determine if there is a correlation between certain messages in political communications.

Machine Learning Timeline/Procedure:

1) Data acquisition and verification — April 4th
2) Data cleaning, manipulation, and carpentry — April 13th
3) Data feature and label extraction — April 20th
4) Machine Learning fitting and cross validation — April 25th
5) Data Analysis and validation — April 27th
6) Report and project analysis completed — May 4th