PROJECT: Project 1 – Reddit Depression Detector project

INFORMATION: this project was coded in Google Colab. One way to run is to upload the project file into Google Colab and open it as a notebook.

KNOWN ISSUES: getting embeddings from RoBERTa currently is not working (it runs until the session crashes from using all available RAM). As such, evaluating the ACU for RoBERTa is also currently not working. Additionally, evaluating the ACU for the trained LDA model results in several of the symptoms having a NaN score.

OTHER NOTES: Unfortunately, I currently am unable to fix any known or unknown issues (or do much in the way of testing and debugging)–this readme is being written as of 4:00 AM, December 16. Due to circumstances in my personal life outside my control which had secondary effects on my physical and mental health, I was unfortunately not able to start working on the project until the last couple of days, during which time I also had a final for another course. I am sorry and I have tried to do the best I could given the circumstances.

ETHICS DISCUSSION: As mentioned in the assignment handout, this project's contents and themes deal with a variety of mental health struggles and their side effects, including but not limited to self-harm. As such, it is imperative that this project (and other NLP applications along the same lines of this project) handle the topic carefully and with consideration to all involved. There are benefits to applying NLP to depression and, more broadly, mental health treatment, but there are several downsides to doing so as well. Below, I will list and elaborate on a few of each:

- BENEFITS:
    - Applying NLP to mental health treatment may allow medical professionals to spot at-risk users earlier and more accurately, which could provide an opportunity for treatment before any worsening in the user's condition.
    - Applying NLP to mental health treatment could lead to improvements in how depression (and other disorders) are diagnosed, improvement in understanding of how such conditions present themselves, and overall improvements in mental health treatment as a downstream effect.
    - Applying NLP to mental health treatment could help reduce stigma surrounding mental health topics by analyzing and openly discussing mental health symptoms and their presentation in users.
- DOWNSIDES:
    - Although the users involved are technically anonymous, and are technically posting on a public forum (i.e. it *could* be presumed that anything they say is "fair game"), some may be uncomfortable with their data being collected, stored, and analyzed regardless. This is especially the case because of the sensitive nature of the posts.
    - Although applying NLP to mental health treatment in this manner could lead to improvements in understanding and treatment of mental health conditions, it could also lead to exploitation of vulnerable people if the tools are used by bad

actors (e.g. an insurance company may look for ways to deny treatment based on presumed mental health status from data from a user's post history).

- Although the majority of researchers will take steps to ensure their methodology is as sound as possible, it is still possible for their NLP models or their application of the models to have serious errors and thus serious consequences for users or patients on the receiving end. For example, false positives could lead to stigmatization of users who are not actually struggling, and false negatives could lead to the missed opportunity to help someone in need.
- Although data was collected from a wide variety of subreddits, the data is still biased due to being almost entirely sourced from Reddit. One possible source from which bias could arise is Reddit's user base, which skews heavily young, white, male, and Western. The way users from such backgrounds discuss mental health struggles may not be applicable to users from other backgrounds or cultures, which could lead to inaccurate results from the models.