Final Project Progress Report: Controversy Detection
Author: Ryan March

1. Which tasks have been completed?

My first task was to complete a technology review focused on state of the art controversy detection procedures so that I could get a baseline for what methods to pursue for my project. Because of this, I was able to perform this task for my Technology Review assignment:
https://github.com/HeapsOfRam/tech_review/blob/main/techreview.pdf
My conclusion from this is that there are many promising methods, many focusing on graph techniques. I think I also want to leverage these types of techniques, while adding my own extensions such as different graph construction procedures and similarity metrics.

The next big hurdle is collecting data. For this, I have downloaded a few datasets, such as:
https://www.kaggle.com/datasets/cerolacia/its-controversial
This dataset unfortunately does not seem to have the features I would like to explore, however; it is hard to correlate tweets with one another and figure out if they are part of the same "discussion". However, I can still perform some analysis based on hash tags.

Aside from this, I have started building (and cloning) web scrapers to assemble a generic text dataset from some online discussions forums that do not have APIs, ie:
https://github.com/Watchful1/Sketchpad/blob/master/postDownloader.py
There are some datasets, like Twitter, that do have APIs associated with them. I have signed up for a Twitter developer account and am working to get the right access so that I can download tweets and discussions via the API and relate them so that I can use them for my analysis. This is going a little slower than expected because of administrative hurdles (ie understanding API, etc), but I have started collecting some data now.

2. Which tasks are pending?

The remaining tasks from here are pending, namely:
- Create hypothesized model
- Test hypothesized model on data
- Verify results, add user interface
- Final Project delivery/video

Because I was able to review prior art in my technology review, the first task listed here is already in progress. I have various ideas about how to change my similarity metrics and graph construction protocols to try to achieve a slightly different flavor of controversy detection. However the remaining tasks have not been started. I plan on furthering my progress on creating a hypothesized model using the data this week, which will help make progress on the first and second points – assuming I can get enough data collected.

3. Are you facing any challenges?

Data collection is a bit of a challenge. I expected to be able to find more available datasets, or to be able to leverage existing datasets. However, many datasets regarding controversy detection are built on news articles, meanwhile I want to go in a different direction and perform controversy detection on online discussions (in forums such as Twitter, Reddit, etc). Luckily, I want an unsupervised classifier so

I do not need labeled data, which seems even harder to find. In the worst case, I will be able to utilize my rudimentary web scrapers to pull in text data for a couple of specific platforms, which is what I am currently working on. I think the data collection process will likely continue throughout the project, so that I can start developing my model on a smaller dataset while aggregating data for more test cases.