Project 1: Group Formation

Names: Cole Sturza, Alex Book, Will Mardick-Kanter

Motivation:

Why is this an important and interesting problem? What triggered the idea?

We would like to classify Reddit comments by subreddit. Namely, we would like to classify r/Communism, r/Liberal, r/Conservative, and r/Libertarian. This is an interesting problem because the commentary of politics includes many forms of sarcasm, idioms, etc. For example, a commenter on r/Liberal could use the phrase “LOCK HER UP!” in an attempt to mock/taunt those with political beliefs that clash with their own. This is an important problem because it could be used in feature creation for a more complex model (particularly if we do not already know people’s political leanings). An example of a more complex problem could be to detect misinformation targeted at specific political belief groups. As far as what triggered the idea, two of us took NLP last semester and there was no final project in that class, so it would be nice to leverage the information gained from that class in this class’ project.

A description of why it would be useful to use / develop machine learning to solve this problem.

Reddit - especially politically-infused subreddits - is rife with sarcasm, metaphors, and other non-literal wordings. We hope that machine learning might help us surmount the barrier that non-literal speech presents.

Data / Data Plan:

A description of your data.

Reddit comments (only the text itself, any links and emojis will be discluded/ignored).

What are some of the interesting or critical features you have? Are there any features you plan to exclude?

The main feature is the text from the comments. We plan to use word2vec to transform the words in the sentences to vectors. This allows us to infer meaning from the words in our models. Other possibilities include using some pretrained language models to create latent features that would feed into our model (possibly BERT).

Upvotes/Downvotes could also be used as features since these work to make comment sections “self-selecting,” meaning the subreddit’s users vote on which comments they agree and

disagree with most. But since the comments will be stripped down to just text for training, these probably won't be too useful.

The depth of the comment could be used as a feature as well, since original comments will typically be related to the topic of the post, while its child comments will diverge after a few levels and the users begin conversations separate from the post.

Approximately (or exactly) how many samples do you have?

There were over 2 billion comments made on Reddit in 2020, and 1.7 billion in 2019, and we are going to limit our comments to 4 subreddits, so maybe a few million comments.

This entry must include whether you already have the data. If you don't already have access to your dataset you must discuss how you plan to have the data by the time you give your official pitch (September 30).

We do not have the data yet. We plan to use the Python package 'praw' to scrape the subreddits we've chosen. Praw is a wrapper for the Reddit API which makes scraping comments from subreddits, including the depth of comments (how many comments below the original comment) and the sorting of the comments (hot, top, controversial, new), a simple task of calling an API.