CS410 Final Project Progress Report: Reddit text data curation and sentiment analysis

Topic:
Leaderboard Competition: Data Set Creation

Team Members:
Ivan Cheung (icheung2@illinois.edu) - Leader, Jeff Zhan (zhan35@illinois.edu), Austin Wang(austinw7@illinois.edu)

What tasks have been completed?
Written to grab submissions and corresponding comments in batch, the web crawler efficiently fetches the metadata necessary for annotations. Based on exploratory analysis, several functions were written to remove irrelevant submissions. For instance submissions with only one upvote were removed as they had no user interaction and generally had no comments. The top n comments and submissions were utilized to find not only conversational ones but also highly ones that were highly interacted with.

After the data was crawled, a basic parser program had to be written to grab the data into a workable format. The parser fetched about 1200 posts with a minimum upvote count of 10, along with the top 10 comments, link, upvote/downvote ratio, and headline. This data will be part of the final submission to show that a data set has been created as per the project topic. We will also feed this data set into our sentiment analysis program that is part of the pending tasks.

What tasks are still pending?

There are about 1200 rows of data in our data set, and we are in the progress of manually annotating them. This will serve as our optimal results that we can compare our sentiment analysis program results with. Each of us will have to annotate all the rows to minimize one's personal bias on whether the news is positive or not. We estimate it will take almost a week for us to go through the entire data set.
In addition, we still have to create the NLTK sentiment analysis tool that will read the data and provide baseline results. This will also include parameter tuning to achieve higher accuracy. We estimate the code, with fine tuning, will take a week of work.
Finally, as mentioned in our proposal, a stretch goal will be to also create a TextBlob algorithm and compare the results of it to the NLTK one.

Challenges:

Currently, there are no major challenges or roadblocks with the project. With that said, however, there is still room for improvement. Firstly, we would like to expand our data set to include additional subreddits such as *r/usnews* and *r/finance*. This would greatly enhance the potential use cases and give a more holistic view. If time permits, we would

additionally like to improve the baseline model, utilizing feature selection and various Natural Language Processing techniques.