CS410 Final Project Proposal: 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)

Background and motivation:

Within the most recent years, there has been much research on sentiment analysis on various social media sites including Facebook and Twitter. Public sentiment is an integral portion for many different domains such as Marketing, Politics, Education, etc. Amongst current Reddit datasets, most fall under the domain of finance (e.g. a dataset on r/wallstreetbets). We propose to create a Reddit dataset with manual labels specifically geared towards news (data from r/worldnews and r/usnews) and create a baseline sentiment analysis algorithm. By hosting both as a competition task, we can leverage LiveDataLab to run various experiments ultimately pushing the accuracy of our predictions. The dataset can be used to help anticipate public sentiment on ensuing novel news topics while the baseline model serves as a benchmark for future work.

Steps:

In order to proceed with this project, there are several things we need to do. First, we need to choose at least one subreddit, and possibly multiple with a similar theme that can make a coherent data set. Then, we need to compile and scrape the data. In this step, there are several parameters for us to decide on, such as how much data we should scrape, what time period to use, as well as which comments in the articles we choose to keep or omit. Once we have compiled our data set, we need to manually judge and label the data. Finally, we will build a baseline sentiment analysis that other submissions will be compared against.

Technical Implementation:

There are a variety of techniques and libraries that will be needed to be used for this project. The main language is Python 3, as it contains a lot of support for web crawling and also for sentiment analysis.

A basic web crawler will be written to extract the top N threads of a popular subreddit. The scraped data will have the necessary text data for sentiment analysis, and we will also add a section for manual judgement. The team will split the articles into three parts and manually analyze each for a positive, negative or neutral sentiment.

A parser will be written to analyze each article's headline and contents. The parser will also be used to extract the thread's comments, as well as establish upvote/downvote thresholds which determine if the comment will be put in the dataset for text analysis.

In addition to providing the base data set for the leaderboard competition, the team will also establish two baselines for submission acceptance. Python contains a variety of sentiment analysis and NLP libraries. The most popular one is NLTK, but a strong alternative is the TextBlob library. The team will use NLTK due to its tendency to work better with labelled data,

which our team will provide via manual analysis. A stretch goal will be to use TextBlob and do a comparison on the accuracy of the two models.

Schedule:

In terms of a timeline, we aim to have a working scraper by October 31, which we can run on the subreddit(s) we choose. Then, we will scrape the data and start manually labeling the data. We aim to have this step done by November 15, when we need to submit the progress report. After that, we will build our baseline sentiment analysis, and depending on how much time we have left over, we can fine-tune the baseline or scrape and label additional data.