Group 6 Project Memo

Alex Bloyer, Hugo Simonet, Matthew Aydin, Will Liu

Our group plans to gather data using the public reddit API to conduct sentiment analysis of user comments. Our dataset will include information from the top posts on several subreddits of interest: post comments, user karma, upvotes/downvotes, and sentiments scores. We will be obtaining this information using the Python Reddit API Wrapper (PRAW), creating a read-only instance of Reddit that can be used to obtain submission from subreddits, comments on submissions, and the redditor accounts responsible for the comments. We would like to gather several thousand comments from posts across a variety of subreddits. This equates to several thousand observations in the dataset with 3 predictors (# of upvotes/downvotes, author karma, subreddit of origin) so far. Once the data is gathered, we would like to create additional predictors such as the frequency of certain words or other statistics relevant to the specific comment and subreddit of origin. The dataset will have numerical and categorical variables alongside text data. We do not anticipate missing data due to the structure of the reddit API and the methods in PRAW.

We are primarily interested in predicting the sentiment of a redditor's comment: positive or negative. By analyzing the patterns of sentiment across the subreddit of origin, we hope to make generalized claims about the broader subreddit's tendency towards positive or negative sentiment and topics. The response variable will be sentiment, which can either take the values positive, or negative. Our primary goal is to create a descriptive model that can predict the sentiment of a post from the comment text. Beyond this, further analysis is desired but unconfirmed.

We plan on spending week 4 setting up reddit API access, installing PRAW, and further developing our research questions. Week 5 will consist of collecting the data from the defined subreddits of interest and creating our dataset. In week 6, we will perform sentiment analysis modeling, ideally using naive bayes or logistic regression to assign a probability of being a positive sentiment. Week 7 will consist of analysis & visualization of the data. Finally, in weeks 8 and 9, we will prepare the final report and presentation.

Our biggest obstacle is likely to be within the natural language processing aspect of this project since none of us have prior experience. Specific questions are likely to arise in several weeks, which we will hopefully address in office hours.