# Foundations of Data Science Capstone Project Proposal

## Social Media Moderation with Sentiment Analysis

### Problem

Online communities encourage user participation in the form of comments. Traditionally, users vote on comments. However, this does not provide context on the overall value of the comment, as not all popular comments add to the user experience. Currently, there is not a classification system that categorizes comments into “positive”, “neutral” or “negative”. This project would aim to perform sentiment analysis on comments from Reddit, an online forum, to identify users who are positively and negatively contributing.

### Client

My client is Reddit, one of the internet’s most popular forums for user’s to post and share information. Reddit is made up of various communities known as sub reddits, centered around specific topics. Users vote posts and comments up or down based on their opinion and earn “karma” for their participation. However, despite earning karma for popular posts or comments, if there are too many negative posts in a community people will be disincentivized to share and reddit will lose revenue.

Therefore, it is useful to identify users and communities who have consistently have negative posts. By classifying comments by sentiment and identifying negative posts, reddit can ensure that it removes or flags users who are consistently destructive to the community and rewards users who provide positive participation, not just those who are the most popular.

### Data

Data will be collected from the reddit API endpoint at https://www.reddit.com/dev/api. Data for one year (May 2015 to May 2016) will be gathered by using rCurl to download comments, users and related data in json and then convert the data to data frames using rJson. Data will also be reviewed using CSV. Alternatively, a pre-existing data set on BigQuery exists at https://www.reddit.com/r/bigquery/comments/3cej2b/17_billion_reddit_comments_loaded_on_bigquery/

### Approach

After loading the data consisting of the user, comment, and score, a column will be added for the sentiment and the data will be cleaned to remove unnecessary characters. The data will be split into a training set (60%), a validation set (20%) and finally tested on new data (20%). Each comment in the training set will be assigned a sentiment score based on the sum of the words used using a sentiment lexicon. A prediction model will be developed in R to help solve the classification problem focusing on naïve Bayes.

### Deliverables

Deliverables will include all relevant code in R Markdown format, as well as a slide deck outlining the background and approach.