Analysis of the popularity of Reddit posts

By: Eva Tarr and Kyle Krawec

Summary

We will be analyzing the popularity of reddit posts through a variety of methods, mainly ones that come from graphing, correlation, and entity analysis techniques. We plan to try to answer questions related to the popularity of a post and what could make a post popular:

* Does the sentiment of a comment result in greater popularity (does negativity gain popularity more than positivity)?
* Does the number of words in a comment result in greater popularity (does user attention span play a role in popularity)?
* Does the time of creation of a post effect its popularity (posting at a certain time of day or posting after the initial “hype” of a topic)?
* Do more popular posts have more comments (do users engage in more popular things)?
* Which types of words are most frequently associated with a high, moderate, and low score?

Hypothesis

The main social computing goal for our research is to better understand the factors that contribute to popularity on social sites. Do users tend to engage in more negative posts? We hypothesize that this is the case and that the negativity on social sites is more welcomed than positive ones. What makes users engage in a post, is it the number of words in a post? We hypothesize that posts with less words will have a greater popularity. Time of creation, we hypothesize, would also play a large role, we believe that the data will show that based on the time of day a post would be more likely to have lots of popularity as opposed to others. This would show the general human habits of users and when they would be using these social sites. We also hypothesize that as a topic becomes “old” it would decrease in popularity. Do users engage(comment) more with already well-liked posts? We hypothesis that user will engage in more popular posts, so the posts with the greater popularity would result in more comments. Do certain categories and types of words correlate with high, moderate, and low scores for Reddit comments? We hypothesize that subreddit comments in the high scoring category will contain more words related to the subreddit topic, the moderate category will contain mostly generic words, and the low category will contain generic and unrelated words. We also hypothesize that high scoring comments in controversial subreddit will contain more inflammatory words, while low scoring words will not and vice-versa. And final we hypothesize that controversial topics will contain a subset of words that are used much more in one category, high, or low scoring comments, when compared to another. For example, a subset of words in high scoring comments may appear more than low or moderate scoring comments, and vice-versa.

Data and Sources

Data used for this analysis came from the open-source data library SocialGrep. It has a large and free collection of various Reddit datasets intended to be used for data science applications. All datasets are clean and available in csv format to aid with analysis. Each dataset is also consistent, meaning all contain identical attributes such as *time of creation*, *title, body, score, sentiment*, etc., three of which (*body, score, sentiment*) are particularly useful for this study.

The datasets used for this study include the following:

* *NoNewNormal dataset on Reddit*
* *One year of Doge on Reddit*
* *June 2022 Bitcoin on Reddit*
* *2022 Freedom Convoy on Reddit*
* *The Reddit COVID Dataset*
* *Ten Million Reddit Answers*
* *Six Months of GME on Reddit*

Each data set contained Reddit posts and comments, throughout the process we used one or both sets to run our analysis.

Technology, Techniques and Methods

The main language used for data analysis and graphing was Python. Python libraries pandas and matplotlib were used to achieve basic analysis, correlation analysis, and graph creation. For word count and frequency, we took advantage of advanced third-party tools such Google’s Natural Language API. With this tool, Google offers a set of pre-trained models to extract insights with natural language understanding for various applications. Features including sentiment analysis, entity analysis, entity sentiment analysis, content classification, and syntax analysis. For the purposes of this study, entity analysis was chosen because word frequency in relation to score is the targeted metric.

What are entities and why is entity analysis important?

To investigate what words, correlate with a high, moderate, and low comment score, natural language processing is an essential pre-processing technique to understand the essence of textual information. NLP tools reduce irrelevant words, grammar, numbers, etc., unrelated to the writer’s intentions and motivations. Many free NLP tools are difficult to implement however and take as significant amount of compute resources to perform accurately, that is why the Google NLP toolset was used. As mentioned, googles entity analysis aided in natural language processing for this study. To define ***entity analysis***, it inspects the given text for known entities (Proper nouns such as public figures, landmarks, and so on. Common nouns such as restaurant, stadium, and so on.) and returns information about those entities. By extracting entities from Reddit comment bodies, we can filter irrelevant information from user comments to gain a deeper understanding or “*spirit*” of the comment itself.

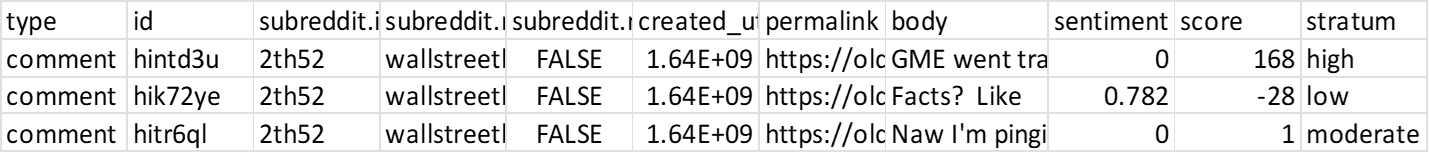
How was the Entity Analysis Performed?

Scores tend to follow a “*mirrored*” Power Law in the positive and negative directions, meaning very few comments scored ***high***, most scored ***moderate***, and very few scored ***low***. For example, as seen below we can define comments sorted by score with a high score between a range of {1000, 50}, a moderate range between {49, -4}, and a low range between {-5, -200}.

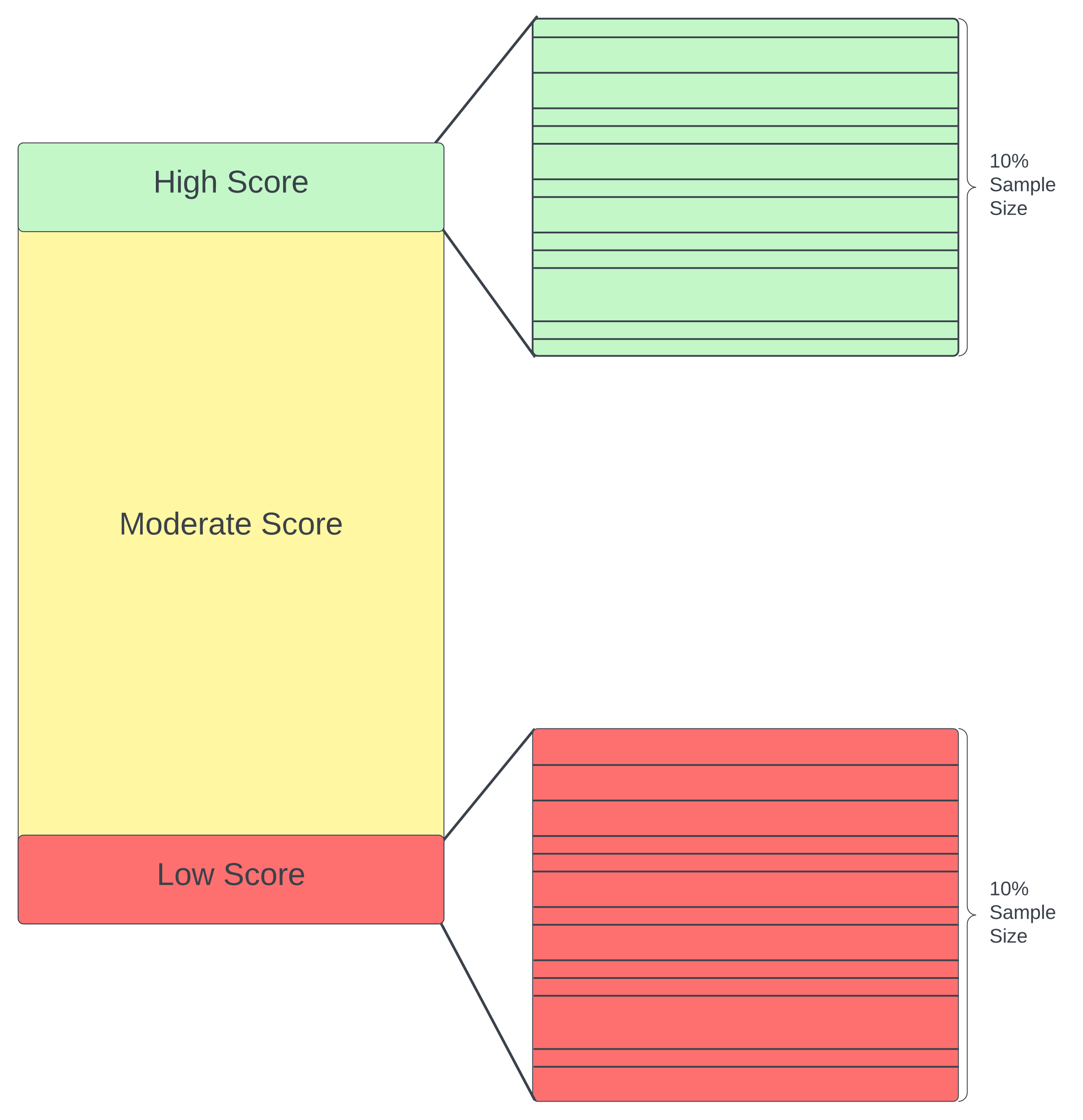
A graph with a line

Description automatically generated

Considering this mirrored power law distribution pattern, data was placed into 3 separate categories or “*stratum*” based on score: *high*, *moderate*, and *low*. The precise stratum boundaries were determined manually by the major trend shifts on scores. For example, the following shows pre-processed comment data that has been organized into the three categories based on score:



The entity analysis was performed using **5%** to **25%** equally random sample of the comments in the *high* and *low* score categories (stratum) as show below:



Using the entities collected from the random samples, a frequency analysis was performed by calculating the frequency of each entity observed in all three strata. Using the frequency of the observed entities within each stratum, we can obtain their normalized frequency to compare the “*strength*” or “*attention”* of the entity used in each stratum.

The normalized frequency was calculated using the following equation for each stratum:

A black text on a white background

Description automatically generated

For example, say we discovered “*apple*” was the #1 word used in the high stratum. If the frequency of “*apple*” was 5332, 34353, and 2331, with a comment count of 15323, 133400, and 12032, in the high, moderate, and low stratum respectively; their normalized values would be: 0.361, 0.257, and 0.194, respectively. Here, we can see “*apple*” occurs 10.4% more times in the high stratum compared the moderate, and 16.7% more times compared to the low.

Using this technique, we can perform cross comparison analysis as well. For example, we can take the top 10 most frequent entities observed in the low stratum and compare their normalized frequencies against that of the moderate and high stratum. We cannot, however, compare the top 10 most frequent entities observed in the moderate stratum because entity analysis was not performed in the moderate stratum due to cost constraints. The priority of this study is to focus on the entities that occur most frequently in the *high* and *low* categories to observe possible differences between comments with the most attention.

Analysis

Popularity and Sentiment:

The tests were performed twice, with different data sets. The overall conclusion is that there is no correlation with the sentiment attached to a comment and its popularity. This is contrary to our hypothesis. The correlation coefficient for both tests came out to be -0.044265 for the “2022 Freedom Convoy on Reddit” comment data set and -0.006067 for the “June 2022 Bitcoin on Reddit” comment dataset. The sentiment for each comment is one that is generated by the SocialGrep database.

A graph with red lines and numbers

Description automatically generatedA graph of a graph with red lines

Description automatically generated with medium confidence

Popularity and Word Count:

The tests were performed on 2 different data sets. We theorized that the correlation to word count would be that popularity would coincide with shorter comments. This however was not the case and we found that the number of words did not have a strong correlation to popularity. We did however find a trend pertaining to social conformity, in most test that we ran we found that there was a large group of comments that all had the same length regardless of its popularity. This could be because of social conformity and that users were following others and making the same length post. The correlation coefficients were 0.0103048 for the “One year of Doge on Reddit” data set and 0.1305555 for the “2022 Freedom Convoy on Reddit” data set.

A graph with red lines and a point

Description automatically generated A graph with red lines

Description automatically generated

Popularity and Creation Time (Month):

The tests were performed on 2 different data sets. We theorized that the popularity of a topic would decrease over a long period of time. The results of our analysis supported our hypothesis. Over multiple datasets the overall average popularity decreased; in this analysis all posts were separated by month. The “One year of Doge on Reddit” was an extreme case as after two months there was a drastic decrease in popularity. The “NoNewNormal dataset on Reddit” wasn’t as drastic but it did have a downward trend overtime.

A graph of a bar

Description automatically generated with medium confidence A graph of a bar graph

Description automatically generated with medium confidence

Popularity and Creation Time (Daily):

The tests were performed on 3 different data sets. We theorized that the popularity of a comment would correlate on the time of day it was made, specifically that higher scores would correlate to the times between work hours (about 6-8 am, 5-12am). Over multiple datasets we found a trend that supported out hypothesis. The averages of popularity tended to be higher in the morning, midday (lunch) and evenings. Through multiple tests the same pattern emerged. We tested using the data sets, “NoNewNormal dataset on Reddit” comments, “One year of Doge on Reddit” comments, and “2022 Freedom Convoy on Reddit” comments.

A graph of a graph

Description automatically generated with medium confidence A graph of a graph

Description automatically generated with medium confidence

A graph of a graph

Description automatically generated with medium confidence

Popularity and Comments:

The tests were performed on 2 different data sets. We theorized that the popularity of a post would correlate with a higher number of comments that a post has. After analysis we found that the higher the number of comments a post had didn’t consistently result in higher average popularity. However, we did find that in every dataset, there was a “number of comments” that had a higher average popularity. The most popular posts had medium “number of comments”, by medium we mean it fell between the having the most and least comments on a post. The data sets used were “NoNewNormal dataset on Reddit” posts and comments and “June 2022 Bitcoin on Reddit” posts and comments.

A graph with red lines and a line

Description automatically generated A graph with red lines

Description automatically generated

Entity Frequency Analysis:

The frequency analysis focuses on two controversial topics, the first being the **Covid *‘antivaxxers*’** and **Trucker Convoy ‘*ottawa*’** subreddits, a large topic the **Gamestop *‘wallstreetbets’*** subreddit, and the more benign **Relatable Humor *‘meirl’*** subreddit. These were chosen to compare the differences between more controversial and low energy subreddits. For each topic comparison, the top 10 most frequent words are chosen in the *high* and *low* stratum. The each of these top 10 are compared with the other two. For example, we see with the following dataset that the *high* strata is compared directly with the *moderate* and *low* to observe which entities are correlated with a high score. Following this we will see the *low* strata is compared directly with the *moderate* and *high* to observe which entities are correlated with a low score.

**Covid “antivaxxers” Subreddit**

A screenshot of a graph

Description automatically generated

In one of the controversial subreddits, we see many of the high and moderate scoring comments contain “work” while the other words do not have much difference. What is also noteworthy is the types of high scoring comments contain words germane to the topic compared to other datasets as we will see. This may indicate that users are interesting in specific information or following arguments that are not verbose and quite dense.

**Trucker Convoy “ottawa” Subreddit**

A screenshot of a graph

Description automatically generated

Another more controversial topic, we see the opposite in terms of generic vs. germane word use. More but not all words in high scoring comments are generic, two to four letter words. There are some patterns however, the entities “people”, “Ottawa”, “police” are found more often in high scoring comments to a great degree. This may indicate that more verbose comments were scored higher indicating that stories and news related material on the topics of Ottawa, police, and people (at the protest) were most discussed and favored.

A screenshot of a graph

Description automatically generated

Again, we see a difference in word use with comments of low score. Low scoring comments tend to use specific, opinionated, and more controversial words compared to that of high scoring comments with very little generic words. Its also noteworthy that the difference and total frequency of the top entities in low scoring comments compared to high is quite small. This indicates that comments using these words are not guaranteed to have a low score, however not following the pattern of high scoring comments will have a greater chance of being rated lower.

**Gamestop “wallstreetbets” Subreddit**

A screenshot of a graph

Description automatically generated

With the subreddit related to a large event, we start to see the pattern we hypothesized. More benign topics tend to have a small to no difference in entity frequency when comparing stratum. In other words, the spread between word frequency in the high to moderate, and the high to low scoring comments. This may indicate that less controversial topics do not tend to influence scores among Reddit users. We will see a slight deviation from this assumption next, however.

A screenshot of a graph

Description automatically generated

Here we see one case of a deviation from the initial assumption, however this is largely to the intense spamming of one entity namely “GME” which is the GameStop stock market ticker symbol. In this instance, there was a small number of accounts spamming (meaning repeated use such as “GME GME GME GME…” for example) this entity which results in many down votes. It is also noteworthy that profane language was also not favored in this subreddit. These finding may indicate that users do not favor spam and profane language when discussing GameStop.

**Relatable Humor “meirl” Subreddit**

A graph of different colored bars

Description automatically generated with medium confidence

Once again, we see the hypothesis that non-controversial subreddits do not contain large difference in the spread between each stratum. In this example, we see than nearly all words in the high scoring stratum tend to be generic three to four letter words. This may indicate that verbose, natural, and clean language is favored.

A screenshot of a graph

Description automatically generated

When comparing the low scoring comments, we see that profane language was not favored among the users. The words used tend to contain longer yet still generic words.

Conclusions

* Does the sentiment of a comment result in greater popularity (does negativity gain popularity more than positivity)?
  + After analysis we found no correlation between sentiment and popularity.
* Does the number of words in a comment result in greater popularity (does user attention span play a role in popularity)?
  + After analysis we found no correlation between word count and popularity, but we did find a trend that could be because of social conformity.
* Does the time of creation of a post effect its popularity (posting at a certain time of day or posting after the initial “hype” of a topic)?
  + Monthly
    - Overtime popularity of a topic decreased as we hypothesised.
  + Daily
    - The time of day that a comment is made does correlate to its average popularity as we hypothesised.
* Do more popular posts have more comments (do users engage in more popular things)?
  + Contrary to our hypothesis the number of comments on a post did not correlate to its overall popularity.
* Do high, moderate, and low scoring comments contain related, generic, and unrelated entities respectively to the subreddit topic?
  + No this has not been observed. Many subreddits contain this pattern or the opposite of it. What was found however, that subreddits devoted to niche topics tend to use words germane to the topic. This does not indicate any relevance to entity frequency and score, however.
* In controversial subreddits, are controversial entities correlated with a high comment score, while benign entities are correlated with a low comments score, vice-versa?
  + In fact, the opposite was the case. The “ottawa” subreddit for example indicates that users do not always favor controversial language in perceived “controversial” topics.
* Is there a larger spread in entity frequency between stratum of controversial topics over that of the benign?
  + This was observed. The largest difference of entity frequency compared between stratum was found more in controversial topics than the others. For example, the spread of high scoring comments over moderate and low were much larger in the “anitvaxxers” and “ottawa” subreddits if we compare them against the more benign “wallstreetbets” and “meirl” subreddits.
* Unexpected findings of the entity frequency analysis?
  + Many times, spamming its not favored such as the entity “GME” in the “wallstreetbets” subreddit.
  + Clearly profanity is negatively correlated with score in half of the studied subreddits.

Challenges, Issues, and Improvements

At the beginning of our project, the primary issue was retrieving useful data. We originally wanted to mine our own data but upon doing research into this, the many restrictions on data mining made us take a different approach. We then wanted to find some open-source data sets that we could analyze that could result in some meaningful insights pertaining to our original question of what factors contribute to a post becoming popular. We found the open-source site SocialGrep, from there we were met with some issues pertaining to what data we could use without paying for a subscription. The main data sets we could use were ones from 2022 and 2021, this was useful however we were hoping for larger and more diverse datasets so we could yield more accurate results. Another issue we found was that posts did not have the SocialGrep generated sentiment, only the comments did, so that limited us to using comments for the sentiment related analysis. Another data related issue was that the posts text body was not provided for any of the datasets. If we had this data, more analysis could have been done pertaining to content and entity analysis could have also been done on posts. While doing analysis we also ran into the issue of slow processing, this resulted in us using smaller data sets so we could do more analysis within the timeframe. An improvement would be to use either cloud resources so computing power could be outsourced or using a different form of analysis besides python that could be more efficient.  
  
With entity frequency analysis, using Google Cloud offered exceptional resources, however costs become considerably high with large datasets and in-depth study. Sampling data was required (due to costs) which resulted in only 5% to 25% of total entities being analyzed and therefore reduced accuracy. Calculating entity frequency proved to be challenging due balancing cost constraints and accuracy when dealing with large datasets. It also provides little context into the reason for word frequency relative to other stratums. We could improve accuracy through developing or using generative large language models, however this may add additional implementation complexity. Another method is to investigate other platforms besides Google that may offer tailored and/or cheaper tools for textual analysis. Another improvement would be to investigate entity sentiment relative to score in addition to entity frequency. This would provide more context in regard to the words used.