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 and correlation 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 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 hypothesis that posts with less words will have a greater popularity. Time of creation, we hypothesis, 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. TODO

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 and graph creation. Data used for this analysis came from the open-source data library SocialGrep. The main data was presented in csv format making it easier to work with. The main data sets we used were: “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”, and “Six Months of GME on Reddit”. Each data set had a set of posts and set of comments, throughout the process we used one or both sets to run our analysis. TODO

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 this was not the case, we found that the number of comments a post had didn’t consistent result in popularity but in every dataset, there was a length of comment that had a higher average popularity. The more comments a post had did not result in a higher popularity, the data sets all resulted in the most popular posts having “medium” length, 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 and Sentiment:

TODO

Conclusion:

* 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.
* Which words are most frequently associated with a high, moderate, and low score?
  + TODO

Challenges, Issues, and Improvements,

At the beginning of our project the main issue was getting usable 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 data sets 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. TODO