Sentiment Analysis of Reddit's top/bottom 5 subreddits regarding Popularity

By:  
Nick Mariasi

Tristan Mateo

1. **Introduction**

**A) Background**

Although there are many popular social media platforms, none seem to be rich in information and social interaction than Reddit.com. Reddit is a complex discussion and content sharing website made up of hundreds of communities called “subreddits”. The content of these subreddits can range from comedy, to politics, to food, to travel and just about anything people could have an interest in. Content on Reddit is also rated with “upvotes” and “downvote” which are essentially “likes” and “dislikes” respectively. . In an addition to just being an expansive website, Reddit is a notoriously difficult platform to gain any sort of popularity in. Because of this, Reddit is difficult for marketers to utilize to its fullest potential. A true sign of a post’s popularity is if it makes it to the front page of Reddit where all the trending posts go. To get on the trending page, a Reddit user should engage with other users in such a way that their posts are not repetitive and boring; that they are contributing something meaningful to conversations and other posts; and that the content posted by the user is relevant and relatable. Being mindful of social interactions is key when it comes to popularity on Reddit

**B) Problem Statement/Objective/Hypothesis**

The objective of this project is to look at posts from popular online discussion website, Reddit.com, and analyze what makes a post popular vs what makes a makes a post receive less attention. We predict that posts with the most popularity will be ones that contain more positive words as opposed to ones that contain negative words.

**C) Significance**

With social media, communication and access to news and information are more readily available. The advent of social media has opened new avenues in marketing, commerce, and even revolutionized the entertainment industry due to how easily accessible social media platforms are. By taking a number of posts from popular and obscure subreddits, and text mining them, we can deepen our understanding of what constitutes popularity and what doesn’t amongst online communities. This understanding can help improve advertising and commerce as predicting whether something will be well received or not (i.e.- popular) advertisements and products can appeal to intended or wider demographics

**II. Materials and Methods**

All the data from Reddit will be pulled using Reddit’s API and the Python coding language. The data will include titles of the top posts in specified subreddits. Results from the top subreddits will be compared to that of the bottom subreddits in an attempt to normalize the data.

Both the Vader sentiment analysis tool in addition to the Textblob tool was initially used to accumulate the sentiment data. However, TextBlob was used for all the analysis done in this paper and the reasoning for this will be discussed in the results section. It’s important to note for future research that one key difference between the two is that TextBlob simply returns a positive/negative/0.0 value while Vader returns probabilities for the positive/negative/neutral variables respectively.

Seeing as Reddit is very time sensitive with their posts (new posts float to the top all the time), we have opted to collect data every consecutive day at 12pm give or take about 10 minutes or so. Each data set is time stamped with the data using the dateTime library so that files can be easily searched/sorted. Currently, the data is output into a file with some guiding text and some formatting to make it easy to read and transfer the data to excel to form the graphics present later on in the results section of this paper.

Additionally, initially we opted to select our subreddits based on the top 10 subreddits and the bottom 10 of the list of top 100 subreddits as of the time of writing this paper. After some consideration, we have opted to include the r/Trump subreddits for reasons explained later in the results section.

To ensure that we have consistent results that are reliable, we have opted to run the test for a full, consecutive week. This week fell on the last week of November and the first week of December so November 27th to December 3rd.

Furthermore, a short C++ program was written to analyze the files and keep count of traces of ‘0.0’ so that we could get a percentage of how many posts where deemed ‘neutral’ by TextBlob. This could have just as easily been implemented into the Python program, but was an afterthought once the data was collected so adding it would have been meaningless.

Initially, we wanted to keep a record of all the titles the program captured so we have 2 additional files for each day: top.txt and bottom.txt. These files contain all of the titles of each post for each day and was going to be used to determine what words where used most often. After analysis though, results didn’t turn out quite like we expected, and can be found in both the results and limitations section of our paper.

Lastly, and most importantly, it’s imperative to understand how the results came to be. With knowing that results from the TextBlob analysis can only be from (-1) to 1, resulting numbers of -2.478 can be confusing. What we have done is basically added up each positive and negative score from each post to get a net sentiment score. What this means is that the resulting number is either the result of a lot of positive/negative scores or a few extremely positive/negative scores. We feel this gives a good indication of overall sentiment in that subreddits at the time at that day.

GitHub Link: https://github.com/NickMariasi/RedditDataMining

**III. Results**

As of now, all of the results are the result of using the TextBlob library. This is because while Vader gives more data in terms of positive/negative/neutral/compound categories, for this project only the compound value would have been of much use, and that’s all TextBlob outputs. You can think of the compound score as a normalized, weighted composite score. With just this value, it would be possible to determine the overall sentiment much easier. Now while we could have just taken that value using Vader and throw out the rest of the results, we found that the Vader tool simply returns far too many neutral results to posts found on Reddit. This may be attributed to the fact that the tool was designed with the intent of performing sentiment analysis on something more personal like Instagram post where things such as emoticons (which the tool takes into account) are used far more often than on Reddit.

Results regarding how many posts were neutral where surprising to say the least. We decided to add the r/Trump subreddits simply on the basis that we thought more polarizing titles would be present there. However, results show that a large amount of posts on Reddit are considered neutral, even when using the TextBlob analysis. Results can be seen in fig 1, 2, and 3 for the percentage of neutral results found in each subreddit.

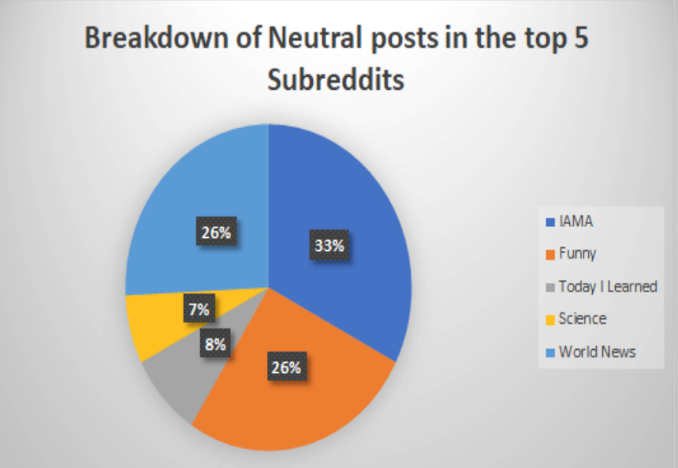
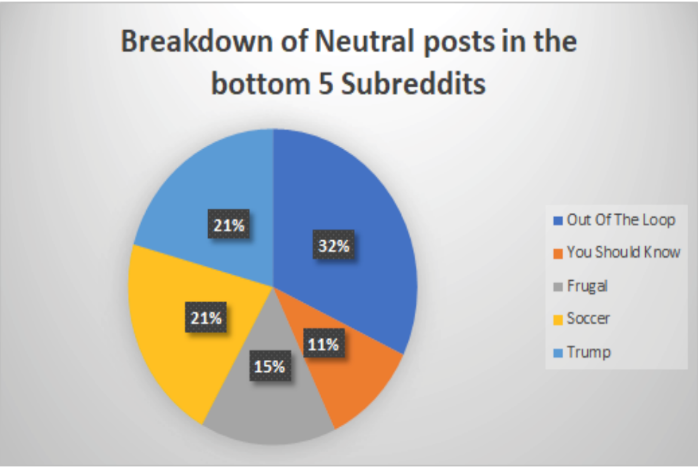


Figure 2 Neutral posts breakdown in the bottom 5 subreddits

Figure 1 Neutral posts breakdown in the top 5 subreddits

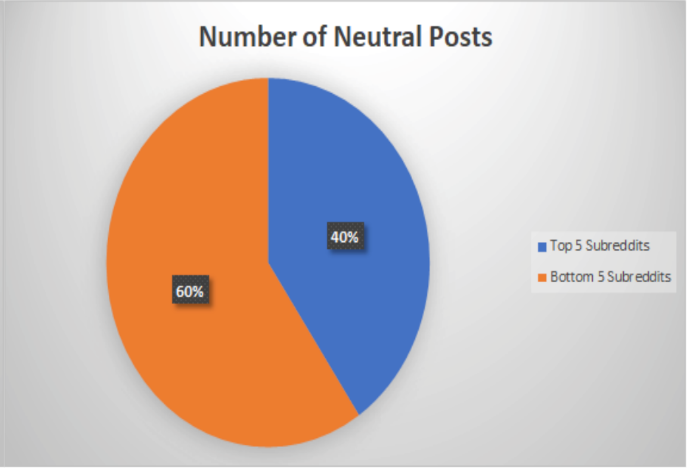
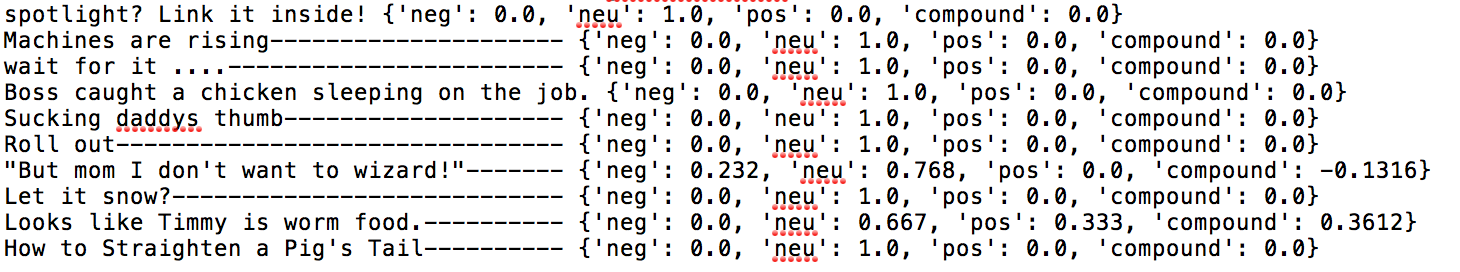


Figure 4 Example data from r/funny that’s too short to definitively label pos/neg

Figure 3 Neutral posts breakdown between top/bottom 5 subreddits

Figure 1 displays the breakdown of the neutral posts found in the top 5 subreddits. You can see that a large majority of the neutral posts come from IAMA, World News, and Funny. While IAMA and World News can most likely just be attributed to the nature of the subreddit (not much polarizing posts found on those subreddits, they are generally questions or neutral articles), r/Funny is interesting. You would think that posts found in r/Funny would be polarizing enough to distinct positive posts from negative posts, however after looking through the data, it’s clear that the titles in these posts are just far too small to make judgement. Data from one day can be seen in figure 4 and a majority of the data is just the same; small snippets of text linked to a picture that is too small to be given any real value other than neutral.

Figure 2 displays the breakdown of neutral posts in the bottom 5 subreddits. From this data, it can be seen that OutOfTheLoop, Trump, and Soccer make up the majority of neutral posts. OutOfTheLoop makes sense, it’s mainly just asking questions which are perceived as neutral, especially in the case of outoftheloop where people are just politely asking for information. This is extremely apparent when you compare it to fig. 5 which shows how much of a subreddit is composed of neutral posts. OutOfTheLoop has a whopping 98.57% of neutral posts that make up its front page. This is something that in hindsight we probably should have expected, but picked up on it too late to change it.

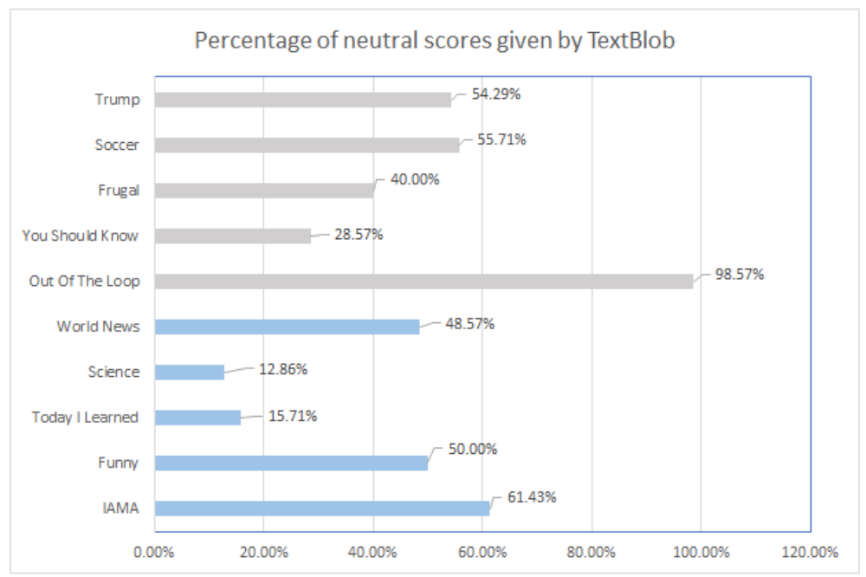
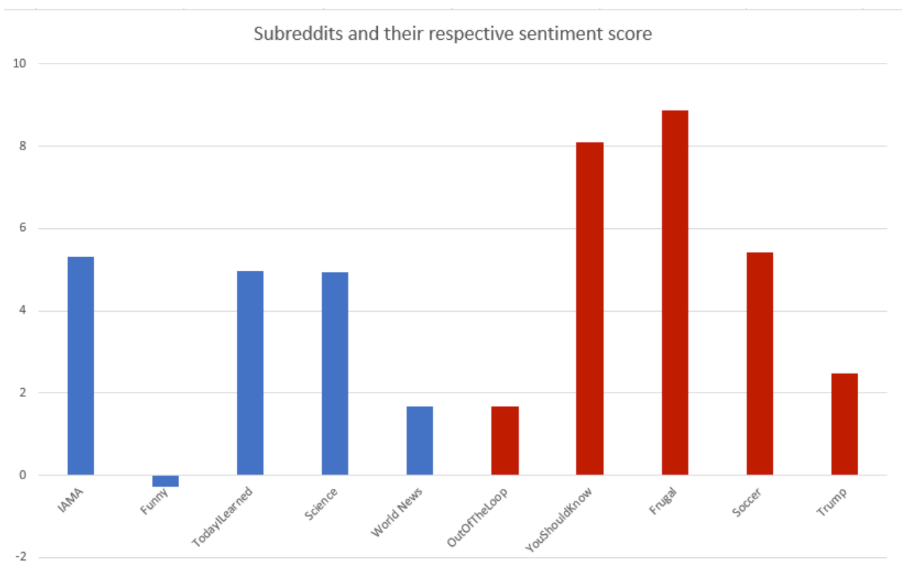
 As for the Soccer subreddit, we sort of expected this result as well, the subreddit doesn’t give off much of an extreme positive/negative vibe which is evident by it’s over 50% neutral posts rating. The Trump subreddit was surprising though, as we added that specifically because we thought that it would be either extremely positive or extremely negative (or bipolar if that makes sense in this context). However, after looking at the results for the posts, it’s fairly clear now as to why there were a lot of neutral results, and that’s because most of the subreddit simply reports news related to Trump like meetings he’s making. Of course, there’s some reports of stuff like bombings which are picked up as negative, but we feel a different, more extreme subreddit would have been better for choosing.

Figure 6 Final sentiment scores based on subreddit, the higher the score, the more positive the posts, the lower the score, the more neutral/negative the posts

Figure 5 Percentage of neutral scores by Textblob by subreddit

Overall, 46.57% of all posts, or 326 out of the 700 posts we analyzed were picked up as being neutral. Of this, 60% of the neutral posts originated from the bottom 5 subreddits while 40% of the neutral posts originated from the top 5 subreddits as seen in fig. 3. These results imply that the top 5 subreddits contain more polarizing titles that can be classified as either positive or negative. This in all likelihood means that the top 5 subreddits contain more posts that users would interact with which probably explains why they are in the top 5 subreddits overall.

Moving to the results that mean the most, you can see that in fig. 6, the only subreddit to dip into the negative value is ironically the r/funny subreddit. This is a result that we did not expect at all and thought that quite the opposite would occur. However, after looking at the data, it’s clear as to why this has happened. Most of the posts on the funny subreddit are sarcastic in their nature, which makes them funny. Now while humor is subjective, we’ve found that most of the humor on Reddit (at least in this week) are posts that are sarcastic in a negative tone. For example, figure 7 displays 3 posts that are seemingly negative in the context of the computer, but when clicked on, link to a picture that most would deem funny and positive.

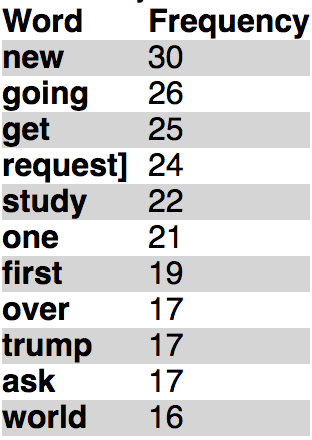
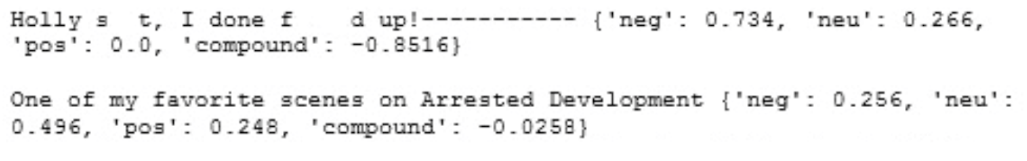
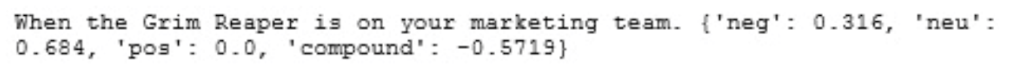
Other than that, looking at the graph it’s evident that both the top 5 and bottom 5 share a subreddit (world news and out of the loop) that has a low positivity rating and high neutral post count. Also, it’s evident that the posts on the bottom 5 of the top 100 subreddits are more positive than that of the top 5. Some might make light of the number of neutral posts in the bottom 5 subreddits being higher which might inflate positive ratings. However, with only a 10% difference, even accounting for that would not make that much of a difference and the bottom 5 subreddits would still yield a higher positivity rating. Overall though, it’s obvious that the majority of subreddits in this list are overwhelmingly positive in their posts and some are left fairly neutral as we’ve found that that’s just the nature of their subreddit.

Figure 7 Posts marked negative although the actual content is supposed to be funny/positive

Lastly, we would like to note that we did get information regarding words used most often in the titles of these subreddits posts. However, this data is not that useful, and we questioned omitting it overall. As expected, the majority of words used are words that pertain to a specific subreddit. For example, President and Trump showed up a lot in the most used words just because we included the Trump Subreddit. Other ‘words’ such as ‘Request]’ also showed up because that’s just the nature of titles in OutOfTheLoop. This leaves us with data ultimately not useful, but we would still like to include it just for the sake of completeness. Because of this, you can find the results in figure 8. We came to these results by simply putting the cumulative list of titles into wordcounter.com and excluding small words like “the,” “it,” etc.

Figure 8 Frequency of certain words used in titles throughout the study

**IV. Discussion**

1. **Analysis**

We expect that the most popular posts on Reddit will be the ones that contain the most amount of positive words. When it comes to the least popular subreddits in the top 100, we predict that the results will most likely stay the same, with the most popular posts containing positive words.

**B. Related Work**

Success on Reddit is heavily dependent on user posts as the platform is very content driven. A study was conducted that analyzed the relationship between the intrinsic quality (measured by “upvotes” and downvotes”) of articles and popularity on Reddit. Using a simple Poisson regression model, it was found that for the most part, articles with the most upvotes received a moderate amount of attention on Reddit [1]. Along with quality of the content, a study has shown that it is possible to predict popularity in subreddits by analyzing the topics presented as well as the words used in posts. However, this is only possible to some extent as it was revealed that since identifying topics is more difficult in more general subreddit communities as these larger communities entertain a more diverse array of topics [2]. A study was done analyzing the activity on the lesser known social networking forum, Slashdot. Although Slashdot doesn’t seem to have a structured community, a unique feature of the site allows user to rate comments (1-5; 5 being the best score) on how helpful or insightful on it was. Through clustering data and creating radial tree representations, it was found that good writers generally received a high score on their comments. The number of replies in response to these writer’s comments depended on the comments’ overall quality (rating between 1 and 5) [3]. Popularity can also be determined through analysis of text though; specifically, what certain words are used (or not used) in addition to how many words there actually are. In an experiment done on Flickr, it was determined that posts with tags such as Nikon, Canon, and 2008 among other words (those are the top three from the table) got the most attention and where the most widely used[4].

**C. Limitations**

One important limitation is the fact that the sample size is perhaps a bit too small.

ideally, this project would be run for a much longer time to get more valid results.

Another important limitation to note is that fact that a lot of reddit’s posts on the top subreddits are very time sensitive. A lot of what is posted is around current events which could potentially sway the data as most of this data would be useless in the future. However, these words may have something in common and could be used in future posts that become popular around different current (future?) events.

**Bibliography**

[1] G. Stoddard, *Popularity Dynamics and Intrinsic Quality in Reddit and Hacker News.* Northwestern University, 2015.

[2] T. M. Rohlin, *Popularity prediction of Reddit texts*. San José State University, 2016.

[3] V. Gomez, A. Kaltenbrunner, V. Lopez, *Statistical Analysis of the Social Network and Discussion Threads in Slashdot.* University of Pompeu Fabra, 2008.

[4] T. Yamasaki, S. Sano, and K. Aizawa, “Social Popularity Score: Predicting Numbers of Views, Comments, and Favorites of Social Photos Using Only Annotations,” in *Proceedings of the First International Workshop on Internet-Scale Multimedia Management*, New York, NY, USA, 2014, pp. 3–8.