Simon Fraser University

CMPT353: Computational Data Science

Analysis of Reddit Post Popularity
What Makes a Reddit Submission Good or Bad?

Authors:
Arda Cifci - 301386128
Ryan He - 301237957

Instructor:
Greg Baker

August 4, 2023

Abstract

This study explored factors influencing the popularity of Reddit posts by analyzing a 2016 dataset. Our findings revealed that simpler language, longer posts, active subreddits, and post timing significantly affected a post's popularity. The sentiment and number of comments a post had also played a key role. However, given the constraints of data volume and the complexities of online language, future research could further enhance our understanding of Reddit post popularity.

Contents

1 Introduction

1.1 Problem

As one of the largest and most popular online forums, Reddit is a massive and diverse online community and a trove of conversations and content. What, then, sets a top post apart from the rest? What unseen elements lead a user toward Reddit fame? With these captivating questions as our guide, our project's objective is to explore what makes these Reddit submissions good or bad, and more specifically, we wanted to decipher the underlying factors that propelled these Reddit posts to the top.

1.2 Refining the idea

We refined this overall problem to focus on answering several key questions about the characteristics of what could make a Reddit submission popular:

- Does the complexity or simplicity of the language used in the submission sway its popularity?
- Do the number of comments or the length of the submission hold any bearing on its popularity?
- Could the vibrancy of the subreddit influence its popularity?
- Does the submission's sentiment influence its popularity?
- Does post-timing play a role in a submission's popularity?

To answer these questions, we decided to use the score of the submissions as the main metric to decide whether the post was good or bad. This enabled us to center our analysis around whether or not these characteristics entailed any perceivable differences in the scores of the posts. In the ensuing sections, we will explore our data-gathering and cleaning process, our analysis methodology, our findings, and ultimately, our conclusions to unveil factors that determine the success of a Reddit post.

2 Data

2.1 Gathering

Our data consisted of Reddit submissions from 2016 as part of the full Reddit Submissions corpus on the SFU cluster. Similar to the repartitioned Reddit submission corpus, our data gathered is organized by month. We utilized PySpark to randomly select 30 percent of data from each month to ensure a representative sample while maintaining a manageable size that we can analyze with. This gathered data was temporarily stored in JSON format before the cleaning began.

2.2 Cleaning and Processing

For the data cleaning process, we employed a series of functions, as seen in gather_clean.py. Utilizing built-in PySpark functions and pyspark.sql imports, we transformed and cleaned our raw gathered data as follows:

- `one_word`: The raw data contained many leading and trailing spaces in both the 'selftext' and 'title' fields of each post. This function trims these spaces, ensuring accuracy in our subsequent analyses that depend on these fields. By removing these

spaces, we ensured accurate word counts and readability scores, which are crucial for our analysis.

- `filter_unwanted_data`: This function was crucial in filtering out rows with missing or irrelevant data. It removed any rows where essential fields such as 'score', 'num_comments', 'ups', 'created_utc', 'subreddit', 'author', 'title', 'selftext', and 'subreddit_id', etc., were null or contained placeholders like '[removed]' or '[deleted]'. Furthermore, this function was instrumental in refining our dataset by excluding any posts marked as 'over_18' and non-text-based posts ('is_self' = False). We decided to remove posts marked as 'over_18' because such posts often contain mature or explicit content that might not be suitable for all Reddit users and could skew our analysis. By focusing on posts accessible to all users, we ensure our analysis represents the broader Reddit community.
- `fix_date`: The raw data contained timestamps in Unix Epoch format, which is challenging to interpret for humans. As our analysis included exploring the impact of post timing, this function was critical for transforming each post's timestamp ('created_utc') into a more readable timestamp and date format.
- `select_columns`: After our initial filtering, we needed to keep only the fields necessary for our analysis. This function pruned our dataset to include only these relevant fields, resulting in a more focused and manageable dataset.

Finally, we limited our cleaned data to 25,000 rows per month for a total of 300,000 rows of Reddit posts and their respective features. We stored this dataset in JSON format with gzip compression for more efficient storage and retrieval. This comprehensive cleaning process enabled us to transform our raw Reddit into a clean and structured dataset.

3 Techniques

Our data analysis process comprised several statistical tests that aimed to understand the influence of various features on the popularity of Reddit posts. Here are the key areas we focused on:

- **Readability Analysis**: We used the T-test to compare the mean scores of posts with high versus low readability. This was determined by the 'selftext' and 'title' using the Flesch Reading Ease (FRES) score and Dale-Chall readability formula. We assumed our dataset to be "normal enough" for T-test due to the Central Limit Theorem.
- **Subreddit Popularity Analysis**: We used the Mann-Whitney U Test and ANOVA to examine the influence of subreddit popularity on post scores. The data were divided into three categories: high, medium, and low subreddit popularity, to better represent the Reddit community.
- **Post-Length Analysis**: We applied the Mann-Whitney U Test and ANOVA to assess the impact of post length on scores, dividing the data into high, medium, and low post lengths to capture the diversity across Reddit.
- **Date Analysis**: To explore whether or not the time posted of the submission affects the score, we used linear regression to calculate a linear least-squares fit line between the hour posted and the average score for each hour. We ensured that the data met the linear regression requirements by calculating the residuals and ensuring they were normal enough to proceed with the regression by looking at a histogram of the residuals. Using a linear fit line, we can see the trend and

correlation between the two data sets and answer our question.

- **Number of Comments Analysis**: The Mann-Whitney U Test explored the correlation between the number of comments and post scores by comparing score distributions for posts with high and low comment counts.
- **Sentiment Analysis**: We utilized a chi-square test to determine whether or not the categories are independent (i.e it doesn't matter what category you're in; the proportions will be the same). The only requirement needed for this test is that each category needs at least 6 observations which we easily fulfill. We evaluate each submission sentiment using the VADER sentiment analysis library and categorize the compound sentiment score into 'positive', 'negative', or 'neutral'.

4 Findings

4.1 Readability Analysis

In the T-Test, the null hypothesis states that there is no significant difference between the groups. However, by comparing the mean scores between Reddit posts with high versus low 'selftext' and 'title' readability scores, we found p-values for all four T-tests well below 0.05. This result led us to reject the null hypothesis, indicating a significant difference between the groups. The T-test results unveiled an inverse relationship: as readability scores decreased, the score of the post increased.

As depicted in *Figure 1*, posts with lower readability scores consistently received higher mean scores. This suggests that Reddit users tend to favor posts written in simpler language.

4.2 Subreddit Popularity Analysis

The null hypotheses for the Mann-Whitney U Test and the ANOVA state that there is no significant difference between the means or distributions across data sets. However, the results of both tests (a p-value of 0.0 for the Mann-Whitney U Test and 2.99e-67 for the ANOVA test) led us to reject these null hypotheses. The findings demonstrated significantly different distributions between high, medium, and low-popularity subreddits.

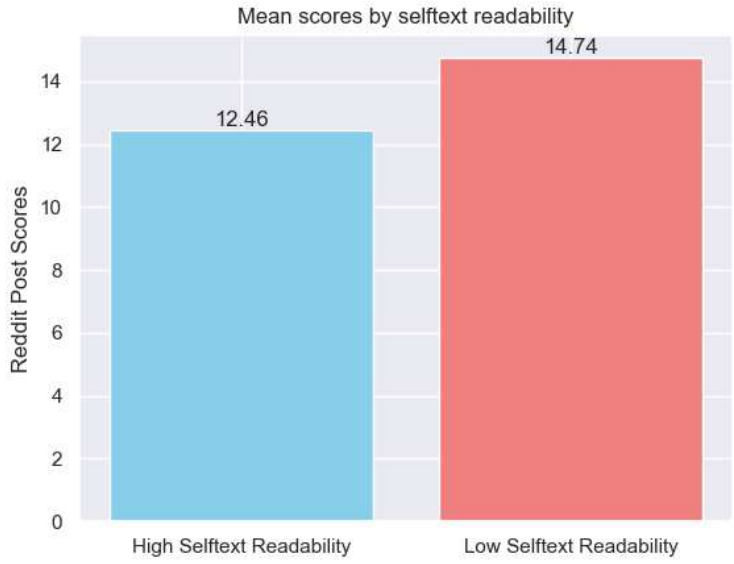


Figure 1: Selftext readability scores

As depicted in *Figure 2*, Reddit posts from high-popularity sub-

reddits consistently showcased higher mean scores. This suggests that posts originating from more popular subreddits are more likely to gain higher popularity.

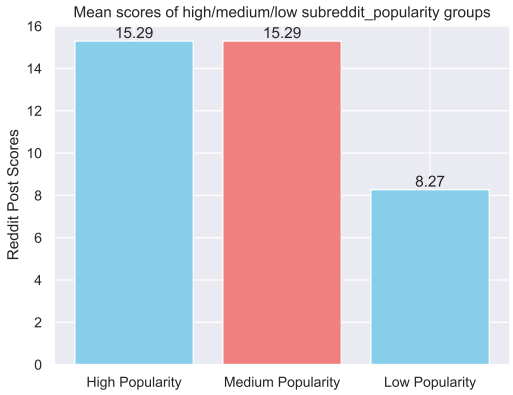


Figure 2: Subreddit popularity ANOVA

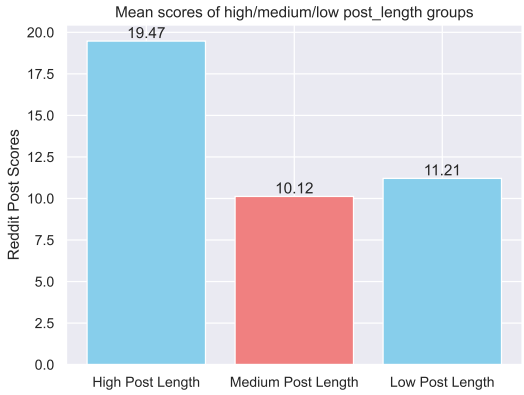


Figure 3: Post Length ANOVA

4.3 Post-Length Analysis

The null hypotheses for the Mann-Whitney U Test and the ANOVA state that there is no significant difference between the means or distributions across data sets. However, the results of both tests (a p-value of 0.0 for the Mann-Whitney U Test and 6.40e-58 for the ANOVA test) led us to reject these null hypotheses. The findings revealed significantly different distributions between high versus low word counts posts.

As depicted in *Figure 3*, Reddit posts with a higher number of words consistently scored significantly higher mean scores than posts with medium or low word counts. This suggests that lengthier posts, potentially more curated and well-written, appeal more to their respective communities.

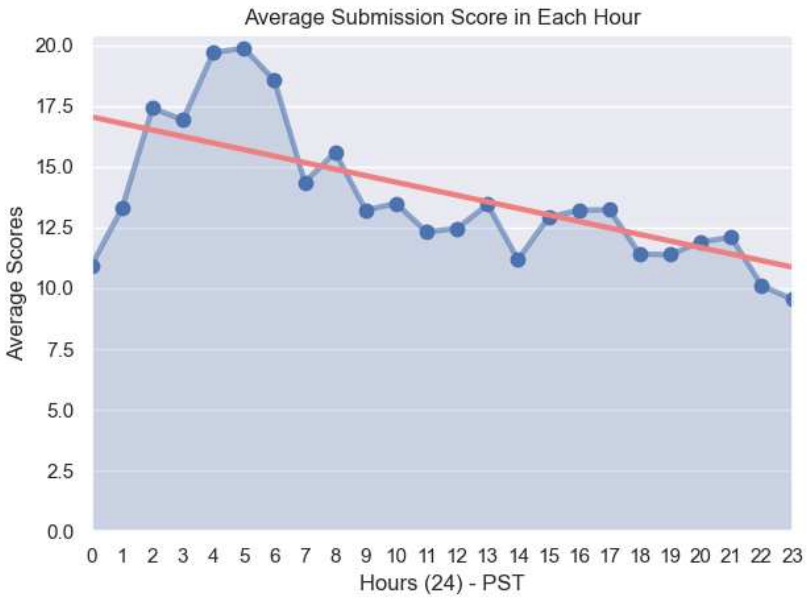


Figure 4: Average submissions by hour

4.4 Date Analysis

The null hypothesis for linear regressions states that the slope of the line is equal to zero. With our fit line having a p-value of 0.00047, it is safe to assume that the relationship between the hour posted and the average score for the hour has some sort of upward or downward correlation. Looking at the correlation coefficient, the r-value we get from the linear regression is -0.65 (rounded). This shows that there is a fairly strong negative correlation between the average scores per hour and the hour posted.

As seen in *Figure 4*, Reddit submissions posted between 4 AM to 6 AM PST have a noticeably higher overall score than the other hours. This suggests posting early in the morning (at least for people in the PST timezone) results, on average, a higher score.

4.5 The number of Comments Analysis

The Mann-Whitney U Test's null hypothesis posits that distributions are equal across two data sets. However, the test's results, with a p-value of 0.0, led us to reject the null hypothesis, revealing significantly different distributions between posts with high versus low numbers of comments.

As depicted in *Figure 5*, Reddit posts with a high number of comments showcased a significantly higher mean score of 30.82 compared to the mean score of 3.75 from posts with a low number of comments. This suggests that posts with more active commentary are generally more popular within the Reddit community.

4.6 Sentiment Analysis

The null hypothesis for a chi-square test states that the categories are independent. i.e., it doesn't matter what category you're in; the proportions will be the same. However, the result of a significant p-value of 1.78e-161. As a result, we reject the null hypothesis in favor of the alternative hypothesis - the two categories are not independent. As such, the chi-square test suggests that high and low scores are dependent on their sentiment category ('positive', 'negative', 'neutral').

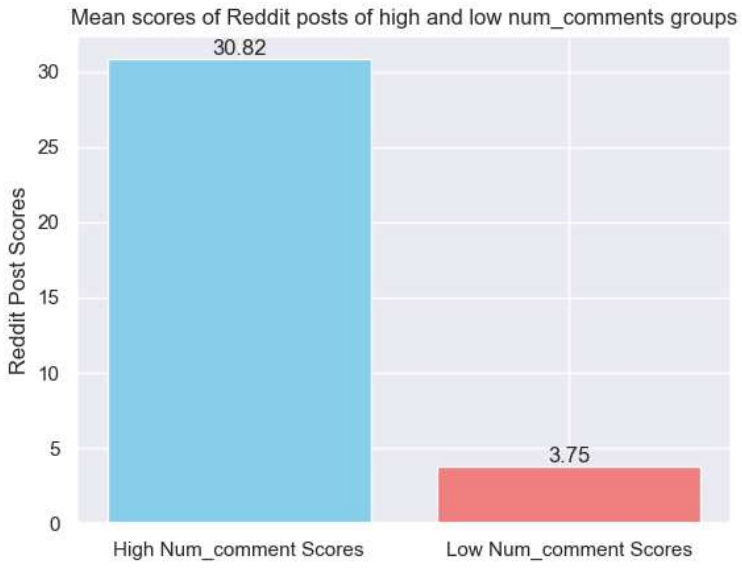
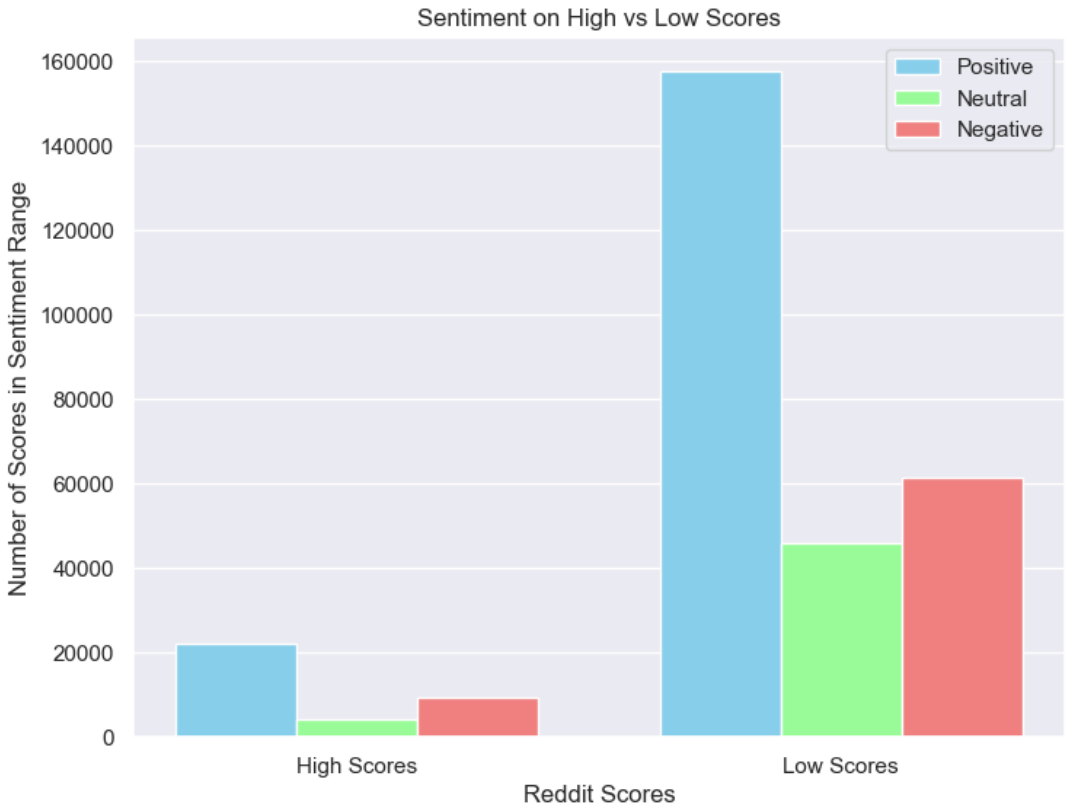


Figure 5: Number of comments

As seen in *Figure 6*, the sentiment category of a post, whether positive, neutral, or negative, appears to impact whether the post will receive a high or low score.



	Positive	Neutral	Negative
High Score	21949	4167	9306
Low Score	157522	45832	61224

Figure 6: Sentiment Scores

Drawing on our chi-square test results, we can conclude that Reddit users' engagement with posts is, in part, influenced by the emotional tone of the content. Consequently, creators who pay attention to the sentiment of their posts may have an edge in eliciting higher engagement from the Reddit community. Further research could explore how different subreddits respond to varying sentiment categories, potentially uncovering nuanced patterns of user engagement across the diverse Reddit platform.

5 Limitations

- Data Volume: The sheer volume of Reddit data made comprehensive processing and analysis challenging within the given timeframe. We had to rely on a random data sample from 2016, which might not reflect current Reddit trends entirely.
- Readability and Sentiment Analysis: While our readability and sentiment analyses were insightful, the readability and sentiment scores depend on the specific formulas used. These may not fully capture the nuances of online language or certain post contexts.
- Limited (Text-Based) Metrics: Due to time constraints, our analysis focused on a few key metrics. However, Reddit post popularity is likely influenced by other factors that we were unable to explore, such as poster reputation, multimedia elements (images, videos), and interactions with other social media platforms.
- Future Improvements: Given more time, we would have incorporated a broader set of metrics into our analysis, explored more machine learning techniques for prediction, and delved deeper into the temporal aspects of post popularity.

6 Conclusion

The objective of our project was to unearth the factors that contribute to the popularity of Reddit posts, as measured by their scores. Our results unveiled an intriguing facet of user engagement: simplicity is paramount. Posts written in more straightforward language—those with lower readability scores—garnered more popularity. This suggests that content that is accessible to a broader audience resonates more on Reddit.

Additionally, we discovered that the length of a post and the number of comments it elicits significantly influence a post's popularity. These findings underscore that Reddit users across various communities appreciate in-depth, substantive content and vibrant discussions.

The popularity of the subreddit from which a post originates also emerged as a crucial factor. Posts from highly active and populous communities were generally more popular, highlighting the significance of community engagement in Reddit's platform structure. Similarly, the timing of a post can also significantly impact its popularity. Submissions that were posted between 4 am to 6 am pst on average had higher scores, showing that posting in early mornings (pst) may give a sizable headstart for your post's popularity.

Finally, our sentiment analysis revealed that the overall tone of a post can influence its score as the relationships between high and low scores with positive, negative, and neutral sentiment depend on each other. This finding reminds us that words often carry more than just information. Words can also carry emotions that will resonate with readers, influencing their reactions and engagement with a post.

In conclusion, our project utilized the power of data science and statistical tests to delve into the intricate workings of Reddit, one of the most popular social media platforms. Although we were only able to explore a small fraction of the vast ecosystem of Reddit, the findings provide valuable insight into the dynamics of user engagement and the different factors contributing to the popularity of Reddit posts.

7 Project Experience Summary

7.1 Arda

- Utilized PySpark to extract, clean, and organize vast unstructured Reddit datasets, transforming them into a usable format for analysis.
- Performed comprehensive data analysis using statistical methods such as Chi-Square, Linear Regression, and Natural Language Processing (NLP) to identify potential correlations between posting times and the resulting Reddit scores.
- Ensured high-quality results by debugging and refactoring code to improve efficiency, readability, and maintainability.
- Gained valuable experience in working with big data, understanding the importance of data cleaning, and learning how to analyze and interpret results.

7.2 Ryan

- Utilized pyspark to gather, clean, and organize a vast amount of unstructured data from Reddit into a usable format for analysis.
- Performed comprehensive data analysis using statistical methods such as ANOVA, Mann Whitney U Test, Chi-Square Test, and T-Tests to identify potential correlations between the features of Reddit posts and their scores.
- Gained a strong understanding of data analysis and interpretation, as well as significant hands-on experience with statistical tests and Python programming.
- Developed strong problem-solving skills and learned the importance of data cleaning, data preprocessing, and data visualization in data analysis.