Final Report - Reddit Comment Analysis

Stella Maltcheva
Student Number: 301602887

CMPT 353 - Computational Data Science
Instructor: Greg Baker

Department of Computing Science
Simon Fraser University

August 2025

Introduction

Online communities like Reddit provide a rich and dynamic environment for social interaction, but understanding what drives user engagement remains a challenge. This project aims to address that challenge by analyzing a large subset of Reddit comments to uncover key patterns and features that contribute to higher visibility and interaction. The ultimate goal is to translate these findings into practical, evidence-based suggestions that Reddit users can apply to boost the scores of their comments. By identifying how to post for maximum impact, this analysis seeks to help users feel more connected to their communities, increase their participation, and foster a more engaged, interactive environment on the platform.

To accomplish this, I analyzed a dataset of comments from nine popular subreddits, which are categorized into three distinct groups: general discussion (`AskReddit`, `relationships`, `todayilearned`), news/politics (`politics`, `worldnews`, `SandersForPresident`), and gaming/ entertainment (`gaming`, `movies`, `leagueoflegends`). By examining engagement patterns across these categories, I hoped to determine whether the most effective strategies for attracting attention are universal or if they vary significantly depending on the community's topic. The analysis will focus on how a comment's score is influenced by factors such as the time of day it is posted and the sentiment expressed within its text. Ultimately, the goal of this analysis is to find insights about user engagement throughout the day and week in different categories and how it may be influenced by category, sentiment score, and more.

Data Collection

The data for this project consists of Reddit comments and stored on the Simon Fraser University (SFU) cluster, covering the months of September through December 2024. This raw data contains a comprehensive set of fields for each comment, including the author, text body, score, timestamp, and subreddit.

Due to the immense volume of Reddit data, creating a computationally manageable and meaningful subset was a challenge. Firstly, I extracted a 1% sample of all comments from November and December 2024 and extracted a table of all the subreddits, ordered by number of comments and including average score. I chose nine popular subreddits that fit into our three categories (general discussion, news/politics, and gaming/entertainment) I made sure they were popular so that there was plenty of comments from a diverse number of users.

After selecting the target subreddits, a more focused extraction was performed to retrieve all comments from these specific subreddits for the entire September-to-December timeframe. To further reduce the dataset size while maintaining representativeness, after filtering for the 9 subreddits, I took 1% of the data and further reduced the comments to 5000 per subreddit at random if they were still too large. This allowed me to have a reasonably small data set to analyze and extract from the cluster.

This process resulted in a final dataset of just under 40, 000 comments. Of course, this resulted in some limitations. Rare engagement patterns might be missed due to the sampling. However, the resulting dataset is diverse enough to support a robust comparative analysis across the chosen categories.

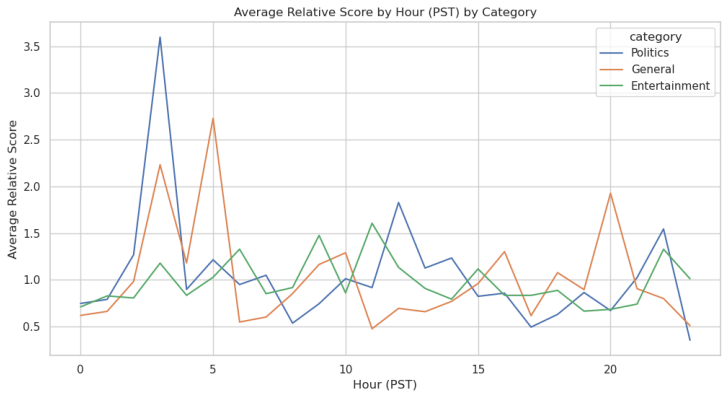
Analysis

Using PySpark, I cleaned the dataset by removing deleted or removed comments and mapped subreddits into broader categories (General, Politics, Entertainment). I extracted time features by converting UTC timestamps to Pacific Time, enabling analysis by hour of day and day of week. For sentiment analysis, I applied the VADER sentiment analyzer from ntlk to each comment, producing a compound score between -1 (negative) and +1 (positive). To account for differences in baseline scores between subreddits, I calculated a relative score by dividing each comment's score by the average score of its subreddit like in Exercise 11. I then used t-tests to compare sentiment or relative scores between specific groups (e.g., Thursday vs. Monday, early morning vs. evening hours, Politics vs. other categories) and visualized trends using Seaborn plots for hourly, daily, and sentiment-based comparisons.

At first, I thought that the word length or comment length would possibly be correlated with an increase in comment score. However, when plotting the comment length against the average score and word length against average score, I found no significant correlation. The average comment lengths were around 163.9, 194, and 186.3 characters for entertainment, general, and politics categories respectively).

Average Relative Score by Hour

The figure below shows the average relative score for each category by hour of the day (PST).

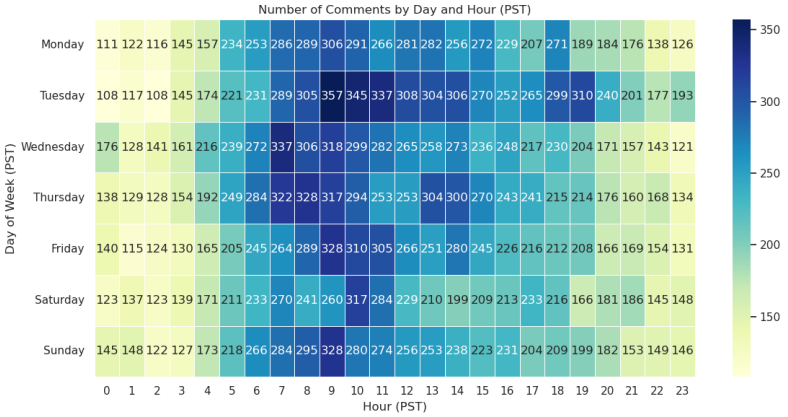


Politics comments has a very large peak around 3 AM. General discussion peaks around 5 AM, while Entertainment maintains relatively stable engagement throughout the day. These patterns suggest that user activity and engagement vary by category throughout the day slightly for the General and Politics categories, although remain relative steady for the Entertainment category. To formally test peak vs low hours, independent two-sample t-tests can be performed. For example, testing Politics peak hour (3 AM) versus all other hours can determine if the difference in mean relative score is statistically significant. After conducting a few t-tests, even when comparing the politics category at 3 AM and 5 PM, I could find no statistically significant result and failed to reject the null hypothesis. It is important to note that for every analysis done based on time throughout the day, the time varies in different parts of the globe. While it may seem strange that there is a spike at 3 AM or 5 AM, this likely indicates that the users posting the comments with high scores are not in the PST timezone. They are possibly from the east coast of North America, or they could be from the UK for instance. The timezone was originally UTC, and 3 AM - 5 AM in the UK would be 11 AM to 1 PM. This could align with people in that timezone going on a lunch break at work and spending

time on reddit.

Heat Map of Comment Counts

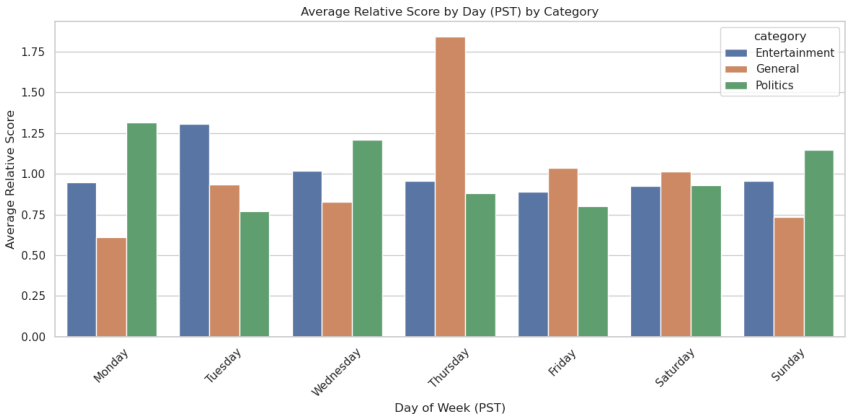
I also created a heatmap of the comments per hour per day, and found that there are most comments between 7 AM and 11 AM, with a peak on Tuesdays as 9 AM.



This would make sense, as people on the west coast of North America are waking up and starting their day around that time, and much of the rest of North America wakes up around that time, and often go to social meida in the morning for news and interacting with others. The original time zone was the UK (UTC), a country that speaks primarly english, as this corresponds to around 5 PM UTC (or UK time) which is right when many people get off work for the day and check social media.

Average Relative Score by Day

The figure below compares the average relative score across days of the week.

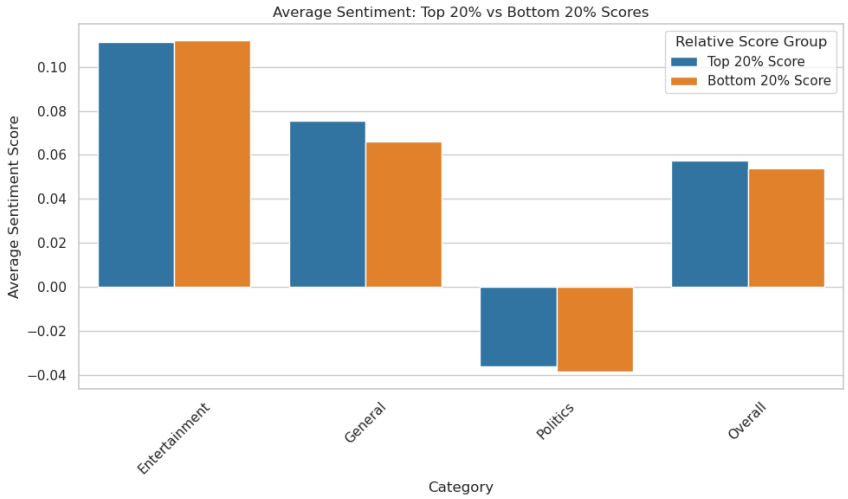


All categories seem to have a mostly steady score throughout the week. The biggest difference seems to be on Thursdays, where there is a large spike in relative scores for the general category. A t-test comparing General on Thursday vs Monday (the lowest average relative score) yielded $t = 0.052$, $p = 0.9589$, indicating no statistically significant difference. Even so, it is interesting to observe that the relative scores slightly differ by category and days of the week.

Average Sentiment: Top 20% vs Bottom 20% Scores

The figure below shows the average sentiment for the top and bottom 20% of comments based on relative score. In an attempt to relate relative scores to sentiment, I compared the top 20% and bottom 20%. I also did this for 10 and 15%. Across all categories, higher-scoring comments tend to have nearly the same sentiment as lower scoring comments.

Despite not being able to relate sentiment to relative scores, I did find that the average sentiment varies among categories. Specifically, the politics and news category has an average negative sentiment while the entertainment category has the highest average sentiment score. The difference in semantic score between the politics/news category and other categories is confirmed by a t-test.



the p-value round to 0.0000, which presents extremely strong evidence that it is different. In this case the null hypothesis stating that all categories have the same sentiment score is rejected. The mean sentiment scores for the categories are as follows: 0.130716 for entertainment, 0.087098 for general, and -0.017649 for Politics. Note that these are all close to zero indicating that most comments have a neutral score and that politics/news generally have significantly more negative comments than the other two categories. The semantic score also has limitations, and may not accurately portray the tone of a comment. Even so, the p-value of 0.000 clearly shows significantly lower semantic scores for comments in the politics/news category.

Conclusion

This analysis investigated patterns in Reddit comment engagement across categories, times, and sentiment. Although the relationship between comment sentiment and relative score was generally weak, several insights emerged. Timing appears to influence engagement as different categories peak at various hours and days, suggesting that posting when a target community is most active can increase visibility. Relative scores are not strongly correlated with comment volume, indicating that additional factors—such as user location, posting time, day of the week, mood, or current events—also play a role.

While high-scoring comments do not consistently show more positive sentiment, the politics/news category generally exhibits slightly negative sentiment compared to other categories. Users seeking to maximize engagement may benefit from aligning their tone with the norms of each community. For instance, participants in politics or news subreddits might moderate their tone when posting in general or entertainment communities to better match audience expectations.

These findings provide actionable guidance on subreddit activity throughout the day and week. Reddit could, for example, highlight peak activity periods (such as Tuesdays at 9 AM PST) or recommend delaying posts until optimal times (when average scores typically peak). Additionally, if a user engages with a subreddit whose average sentiment differs from their usual communities, the platform could suggest minor adjustments in tone to improve engagement. Addtionally, if Reddit had a feature to calculate the average semantic score of large subreddits, it could alert users before posting a comment if the semantics score of their comment significantly differs from the average semantic score of that community.

Project Limitations

Several factors should be considered when interpreting these results. First, time-based patterns are influenced by the global distribution of users. Peaks at early morning hours in PST, like 3 AM or 5 AM, likely reflect activity from users in other time zones, such as the east coast of North America or the UK. Apparent "best times" to post might not align with local Pacific time, and time zone conversions can introduce noise.

Subreddit differences also play a role. Each community has its own culture, norms, and posting frequency, which affects engagement. Strategies effective in one subreddit may be ineffective in another, contributing to variability in relative scores across categories and times.

Dataset limitations exist as well. Only a subset of comments was used, capped at 5000 per subreddit, which may exclude rare engagement patterns or extreme scores. Metadata like post title, thread depth, or parent post popularity were not included, yet these factors can strongly influence engagement.

Broader context also affects comment scores. Trending topics, viral posts, or social events can drive engagement independently of timing or content. Sentiment analysis, while useful, may miss sarcasm, humor, or community-specific language, explaining why sentiment was weakly correlated with relative score despite differences between categories.

Lastly, statistical assumptions, such as independence and normality in t-tests, may not fully hold. Outliers can skew averages, and relative scores normalize across subreddits but cannot fully account for highly active users or unusually popular posts. If I had more time, I would like to look into more categories and subreddits and how the vary over time as well as other things that influece score like current trends and events. If it is possible, it would be great to find out where each user is located as well, so data about the hours of peak activity could be looked at by region.

Project Experience Summary

In this project, I conducted a large-scale analysis of approximately 40,000 Reddit comments from nine subreddits to investigate patterns in user engagement and comment visibility. Using PySpark, I cleaned and preprocessed the data by filtering deleted comments, mapping subreddits into broader categories, and normalizing comment scores relative to subreddit averages to enable fair comparisons. I applied sentiment analysis with VADER from Python's NLTK library to quantify comment tone and explored correlations between sentiment and relative comment scores across different categories, times of day, and days of the week. To visualize engagement trends, I used Seaborn to generate heatmaps, hourly and daily activity plots, and sentiment comparisons, producing actionable insights into optimal posting times and community-specific behaviours. Finally, I interpreted these results in the context of global user distribution, subreddit culture, and platform dynamics, providing evidence-based recommendations to maximize engagement as well as acknowledging the limitations. This project demonstrates experience in large-scale data acquisition within a cluster environment, along with skills in data cleaning, analysis, visualization, and interpretation, and the ability to translate complex findings into actionable insights.

References

Python Software Foundation. (n.d.). *datetime — Basic date and time types.* Retrieved from `https://docs.python.org/3/library/datetime.html`

NLTK Project. (n.d.). *NLTK:: Sample usage for semantics.* Retrieved from `https://www.nltk.org/howto/semantics.html`

Waskom, M. (n.d.). *seaborn: statistical data visualization — seaborn 0.13.2 documentation.* Retrieved from `https://seaborn.pydata.org/`

Suvrat. (2025, April 23). *Sentiment analysis using Python.* Analytics Vidhya. Retrieved from `https://www.analyticsvidhya.com/blog/2022/07/sentiment-analysis-using-python/` SFU. (n.d.). *Sign in - CAS – Central Authentication Service.* Retrieved from `https://coursys.sfu.ca/2025su-cmpt-353-d1/`