# Reddit Account Karma Analysis

Uday Lingampalli
San Diego, California
ulingampalli@ucsd

***Abstract*—The social media like platform of Reddit involves a value called ”Karma” on every user's account, serving as a measure of how other users support and react to this user's content posted on the site. I explored general trends in karma, as well as the topics, sentiment, and time of user content and how it affected the reaction it received from others, and the influence it created on karma overall. Finally, I create a k-NN regression model to predict karma based on the aforementioned information, with a moderate degree of success.**

***Index Terms*—sentiment,**

## I. Introduction

Founded by Steve Huffman and Alexis Ohanian in 2005, Reddit is primarily a social media site focusing on news. The users, called Redditors, can create posts, and express their opinions on other users' posts via comments, and can reply to others' comments with their own comments as well [1]. Posts are not just posted anywhere - users post them in subreddits, or a community of Redditors where content is based around a certain topic, e.g. r/news, where r/ is the subreddit header that precedes every subreddit, and news is the topic. Users also have the option to view popular posts from an aggregation of all subreddits [2].

Aside from comments, Redditors have another way to interact with other users' submissions - a term I will use to describe both posts and comments - upvoting and downvoting. An upvote indicates that you enjoyed the content, and downvoting indicates you disliked it. The score of a submission is the number of upvotes the submission recieved minus the downvotes, and is displayed below the submission, serving as a measure of the community's reaction to it [3]. A user's karma, which can be viewed on their account page, is the sum of their comment karma and link (post) karma, each of which represent all of the user's scores of that type of submission. It is not quite the sum of these scores, the exact equation is unknown, but it is somewhat similar to the sum [4].

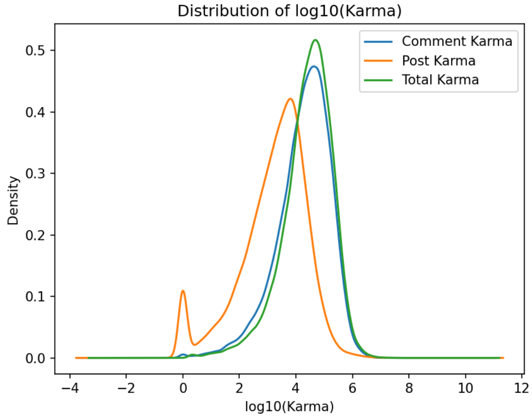

Fig. 1.

Karma can then be considered to be a measure of ”anonymous” popularity of a user - in that people don't necessarily recognize them, but their content has accumulated a lot of support and agreement from others. What, then, factors into a user's karma? Certainly, karma is akin to a sum of scores, so the longevity of an account matters, as well as the content it posts, the subreddit on which they are posted (comments on subreddits with more members will naturally receive more attention, and therefore more potential for upvotes), and even the time at which their content is posted. In this study, I will analyze how these various features are related to the karma of an account.

## II. Data Collection

In order to have a dataset of Reddit users, I used PRAW (Python Reddit API Wrapper) to retrieve the 1000 posts with the highest scores across all subreddits. PRAW retrieves posts with a portion of the total comments with an option to load more. I scraped all of the distinct authors of these initial comments for the 1000 posts, ignoring deleted and suspended users, giving me 32718 accounts. From these accounts, I gathered the year their account was created, their total post karma, total comment karma, and total karma.

Additionally, I also collected each accounts' 10 submissions with the highest scores, the respective scores, the sum of these scores, the content of the submission, and the date of each submission, each as a list under their respective column. Finally, I used the SentimentIntensityAnalyzer function from the NLTK (Natural Language Toolkit) library on the content of these submissions. This gave each submission a sentiment score from -1.0 to 1.0, where -1.0 represented very negative, and 1.0 represented very positive. I assigned each submission as 'Negative' if the sentiment score was less than -0.25, 'Positive' if the sentiment score was greater than 0.25, and 'Neutral' otherwise, and saved this feature as well.

## III. Data Analysis

After collecting the data, the following visualizations express key information about the data, relevant to how the aforementioned features affect an account's karma. First, I analyze account information, then the data retrieved from their top 10 submissions, and finally the additional information obtained by conducting sentiment analysis on these submissions.

### A. Account Analysis

*Figure 1:* As we can see from this kernel density estimate plot of the logarithms of Comment Karma, Post Karma, and Total Karma, Comment Karma follows the curve of Post Karma extremely closely, while Total Karma does not. This indicates that karma is more frequently obtained from comments rather than posts. We also see that accounts with lower karma tend to have more of it from posts.

*Figure 2:* This bar chart displaying the number of accounts created per year reveals that most accounts were created between 2011 and 2020, with a significant drop-off before and after that. This also indicates Reddit activity may have been higher than average overall during this period.

*Figure 3:* This bar chart displaying the average Comment Karma, Post Karma, and Total Karma per accounts created each year, shows a general trend that from 2008 onwards, older accounts will have more Comment and Total Karma. Notably, Post Karma does not follow this trend as much, indicating a general infrequency of creating posts across the dataset.

### B. Comment Analysis

*Figure 4:* The Top10Ratio is the sum of the scores of the collected comments per user, divided by the Total Karma of that user. In essence, it conveys how much a user's Total Karma is due to a few popular submissions. The graph reveals that from a Total Karma from 100 to 10000, these users mostly have a few influential submissions, but beyond that, the contribution their top 10 submissions have on their karma decreases.

- Note: You may notice that this ratio can be over 1. This is due to the fact that comment and post karma is not exactly the same as the sum of submission scores, it is slightly less, and downvoted submissions will decrease karma without changing the sum of the scores for the top 10 comments.

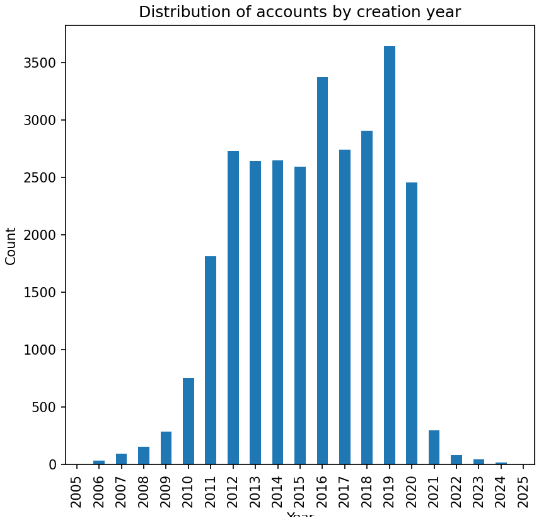

Fig. 2.

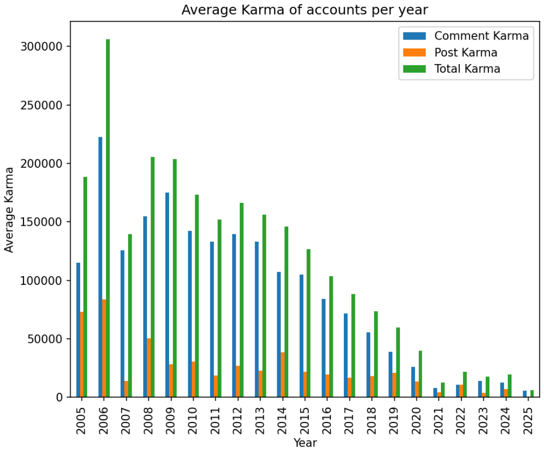

Fig. 3.

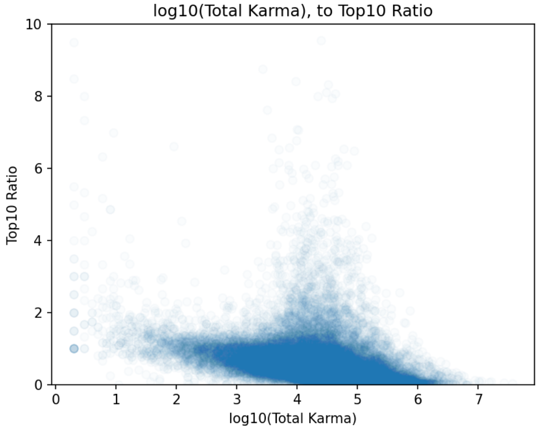

Fig. 4.

*Figure 5:* This graph shows the number of submissions over time. Considering the collected submissions are each user's

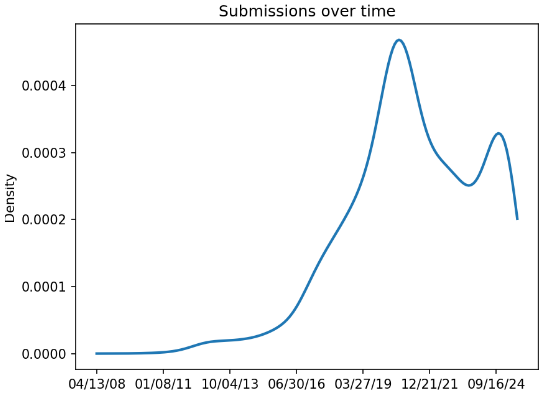

Fig. 5.

highest upvoted submissions, we can see that most of these submissions were posted between 2016 and 2021, with a drop off before and after that.

### C. Sentiment Analysis

*Figure 6:* This bar chart shows that Neutral submissions were the most common, followed by Positive submissions and lastly Negative submissions.

*Figure 7:* This bar chart also shows that Neutral submissions had the highest scores on average. However, counterintuitive to Figure 7, Negative submissions had a higher average score than Positive submissions.

*Figure 8:* For the 10 subreddits with the highest total scores, this graph reveals the breakdown by sentiment, revealing that the influence the sentiment of a submission has on the score is highly influenced by the subreddit it is posted in.

*Figure 9:* This graph reveals the amount of submissions by sentiment over time. Although they follow each other very closely, it is notable that 2020 had more Positive and Negative submissions than Neutral, indicating more polarized opinions during this time period, and more Neutral opinions afterwards.

Overall, I learned many relationships between the features of the data. Importantly, I discovered how comment karma mostly composes total karma, Reddit activity over the past 17 years in both account and comment/post creation as well as sentiment, and the impact of the subreddit and sentiment of comments/posts on their reception. My specific data analysis approach relied on trying to find relations between all features. The advantages of this approach was that it quickly allowed me to identify simple relations between features, and which features impact others, but the primary disadvantages were that there was no opportunity to identify more complex relations between more than two features, nor specifically quantifying the relationships between features when possible. An assumption I make with this approach is that there are no confounding features and that features directly influence each other, which may not be the case.

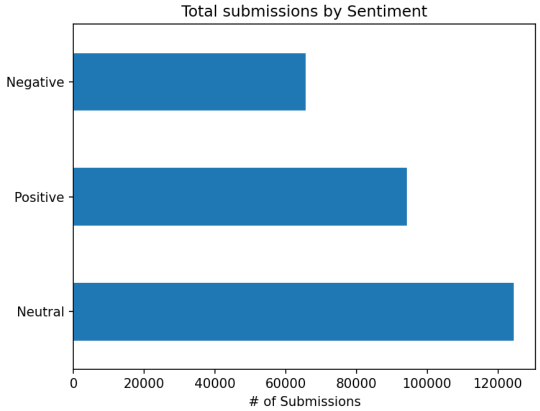

Fig. 6.

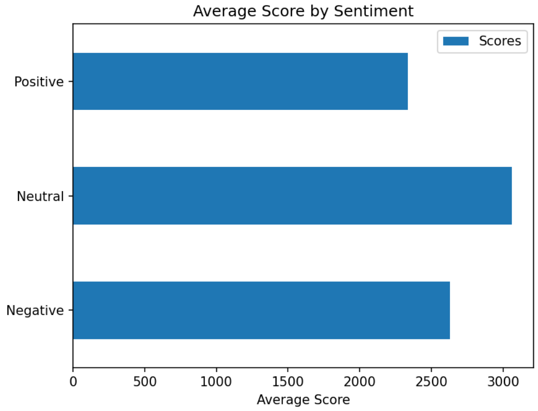

Fig. 7.

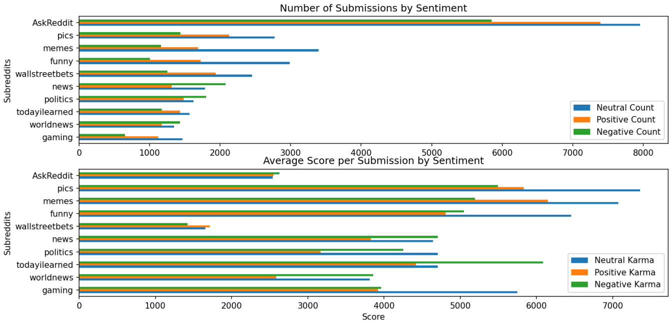

Fig. 8.

## Karma Model

Next, these findings regarding the data were used to construct new features, to create a model that could predict the karma of an account, without knowing the comment karma or the post karma respectively. Due to the fact that recording data per each of the top 10 comments would be a lot of features, and the need to quantify features like sentiment, I decided to engineer four new features. The purpose of these features would be to reduce the dimensions of the data, as well as extracting numerical information from qualitative features. The new features created were:

- *SubSentAvg*: This variable took the sentiments and subreddits of each user's top 10 submissions, and returned a value which represented the weighted average of the

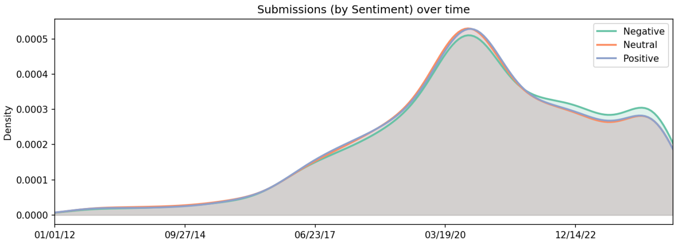

Fig. 9.

count of submissions of the respective sentiments in the respective subreddits of all collected comments. The intent of this feature was to quantify sentiments and subreddits of a user's comments.

- *SubTotAvg*: This variable took the subreddits of each user's top 10 submissions, and returned a value which represented the weighted average of the count of submissions in the respective subreddits of all collected comments. The intent of this feature was to quantify the average popularity of the subreddits each user participated in.
- *dateSentAvg*: This variable took the sentiments and dates (excluding day) of each user's top 10 submissions, and returned a value which represented the weighted average of the count of submissions of the respective sentiments on the respective dates of all collected comments. The intent of this feature was to quantify the average level of activity of particular sentiments on all the days a user's top 10 comments were posted.
- *dateTotAvg*: This variable took the dates (excluding day) of each user's top 10 submissions, and returned a value which represented the weighted average of the count of submissions on the respective dates of all collected comments. The intent of this feature was to quantify the average level of activity on all days a user's top 10 comments were posted.
- *scoreSentAvg*: This variable took the sentiments, subreddits, and scores of each user's top 10 submissions, and returned a value which represented the weighted average of the scores of submissions of the respective sentiments in the respective subreddits of all collected comments. The intent of this feature was to quantify the average community response to posts and comments that match the subreddits and sentiments of the user's top 10 comments.

My final processed dataset included these variables, as well as changing the creation variable to instead track how long the account had existed, and the sum of the scores of their top 10 submissions. After trial and error with multiple models, I standardized all the columns, and settled on k-NN regression algorithm. With n=49 neighbors, I achieved an $R^2$ coefficient of 0.27. However, when considering only accounts with less than 100000 total karma, around 75 percent of the total dataset, I achieved a $R^2$ coefficient of 0.37 with n=29 neighbors. This indicates that my dataset does not contain all the information needed to effectively predict karma for accounts with a large amount of karma.

## Conclusion

My analysis on what impacts the karma of Reddit accounts yielded several notable findings. The most important findings were the relative importance of comments over posts, how the unique combination of the sentiment and subreddit of a submission influences its score, and submission trends over time, particularly 2019-21 resulting in less Neutral posts relative to other sentiments, and generally increased activity during those years. However, the significant improvement in the ability of k-NN model to predict karma for lower karma accounts vs. higher karma accounts while still being applied to the majority of the dataset indicates missing information that may be able to further improve the accuracy of the model, although there is a valuable conclusion even in being able to these features to somewhat useful predictions for low karma accounts. To improve this analysis, further information that could be collected include the total number of submissions (though this would have to be counted one by one due to this lack of other options), further analysis on the times of the day when submissions are posted, and simply collecting more submissions from each user for a more thorough analysis, all of which can also be implemented in future iterations of this analysis. If additional functionality to the Reddit API to scrape more features is added, I may consider revisiting this analysis to see what new relationships I can glean, and if my model can be improved.

## References

[1] M. Moradian, "The History of Reddit — Honor Society," www.honorsociety.org, Aug. 17, 2020. https://www.honorsociety.org/articles/history-reddit

[2] "Reddit.com Guide: Understanding Subreddits — Honor Society - Official Honor Society® Website," Honorsociety.org, Aug. 17, 2020. https://www.honorsociety.org/articles/redditcom-guide-understanding-subreddits (accessed Aug. 23, 2025).

[3] [1]"How Reddit Voting Works — Honor Society - Official Honor Society® Website," Honorsociety.org, Aug. 17, 2020. https://www.honorsociety.org/articles/how-reddit-voting-works (accessed Aug. 23, 2025).

[4] [1]A. Dutta, "Reddit Karma Ultimate Guide: Everything You Need To Know," Socinator, Apr. 25, 2024. https://socinator.com/blog/ultimate-guide-on-reddit-karma/ (accessed Aug. 23, 2025).