Comparative Analysis of Contributor Productivity on Reddit: A Case Study of r/emacs and r/vim Subreddits

Author: Mohamed Ashraf Ouf

Date: 18/12/2023

Introduction

This report presents a comprehensive analysis of contributor productivity within two Reddit subreddits, r/emacs and r/vim, over a three-month period from October 1, 2023, to December 15, 2023. my objective was to understand the behavior of contributors and differences in contributions between these communities.

Methodology

Data for each subreddit was scraped using the Reddit API, adhering to the platform's guidelines. We extracted the following metrics for each post to quantify productivity and engagement:

1. **Number of Upvotes (Ups)**
   - **Reason:** Upvotes are a direct measure of community approval and popularity of a post. A higher number of upvotes generally indicates content that resonates well with the subreddit audience.
2. **Upvote Ratio**
   - **Reason:** The upvote ratio is the percentage of upvotes relative to total votes (upvotes plus downvotes). This metric gives a normalized view of community sentiment towards a post, useful for comparing posts with different numbers of total votes.
3. **Number of Comments (Num_Comments)**
   - **Reason:** The number of comments serves as an interaction metric, indicating the level of engagement and discussion a post generates. A high number of comments suggests that the post sparked conversation and is of interest to the subreddit community.
4. **Post Score**
   - **Reason:** The post score (calculated as upvotes minus downvotes) is a Reddit-provided metric representing the overall net approval of a post. This score is indicative of the community's net sentiment towards a post and is commonly used to rank content on Reddit.
5. **Post Length in Characters (Post_Length_Chars)**
   - **Reason:** This metric measures the amount of effort and detail a contributor put into the post. Longer posts might suggest more in-depth discussion or analysis, which could be characteristic of certain subreddits.

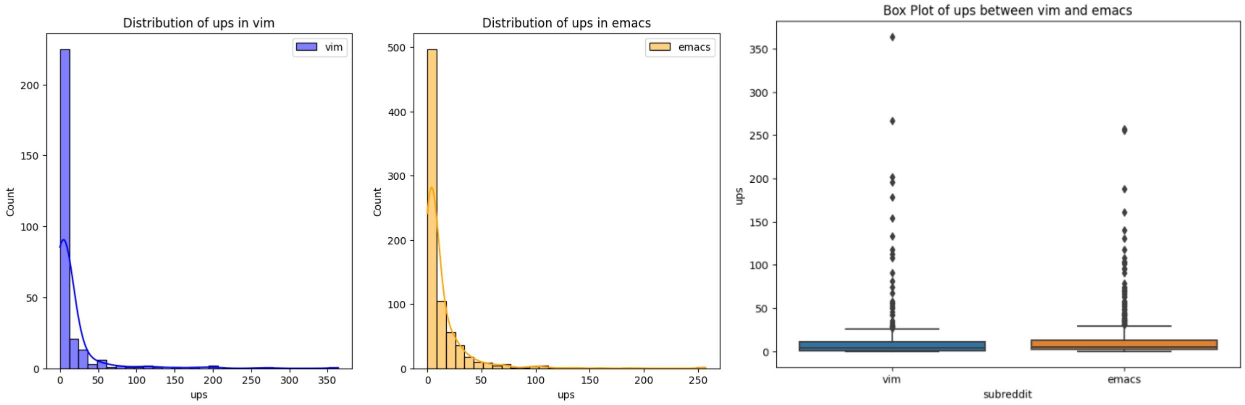
6. **Post Length in Words (Post_Length_Words)**

    - **Reason:** Similar to character length, word count provides another dimension to quantify post content. It helps in understanding whether posts are concise or elaborate, a factor that might correlate with the subreddit's nature or topic complexity.

Results

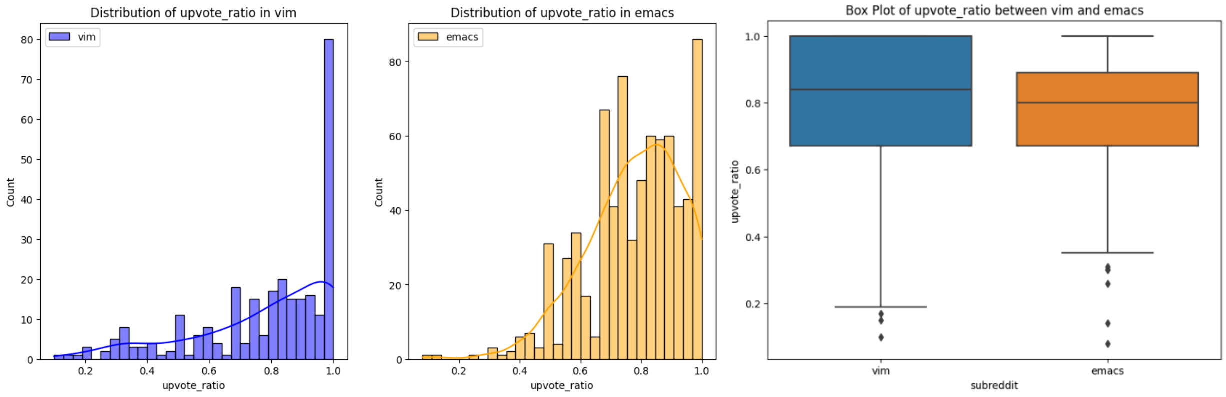
1. Data Distribution Analysis

Graph 1: Histogram and Box Plot of Upvotes for r/emacs and r/vim



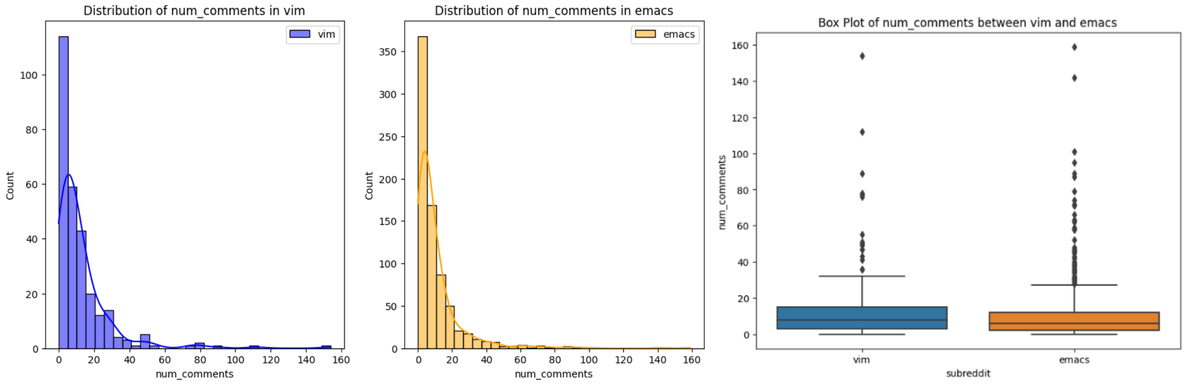
Both distributions are right skewed, indicating a higher frequency of posts with fewer upvotes and fewer posts with a high number of upvotes.

Graph 2: Histogram and Box Plot of Upvote Ratios for r/emacs and r/vim



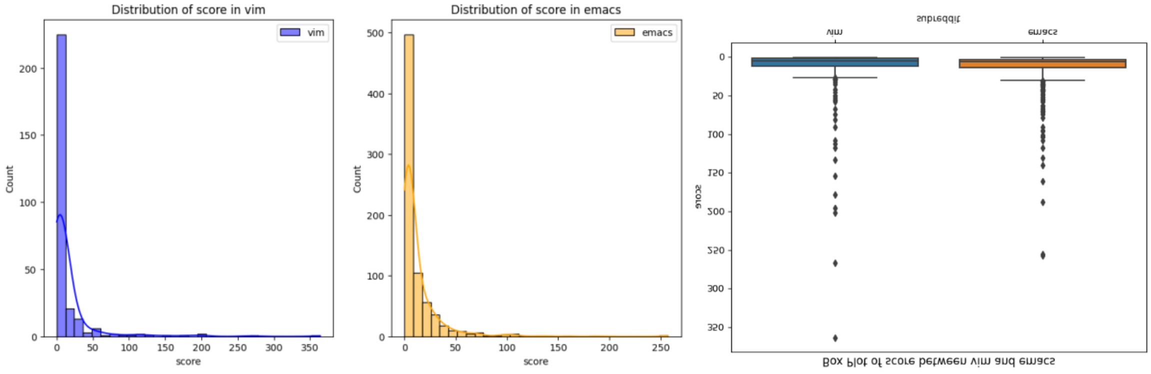
The graphs are bimodal for both groups, with peaks around 0.2-0.3 and 1.0 (perfect upvote ratio), suggesting a common pattern where posts either have a low upvote ratio or are highly favored.

Graph 3: Histogram and Box Plot of Number of Comments for r/emacs and r/vim [Insert Histogram and Box Plot for Number of Comments]



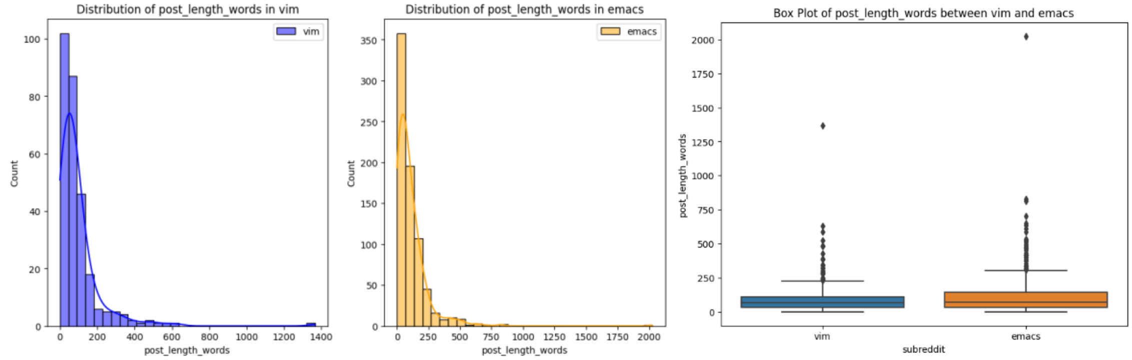
these distributions are right-skewed. Both groups have a large number of posts with few comments and a much smaller number with a high number of comments.

Graph 4: Histogram and Box Plot of Scores for r/emacs and r/vim



Both groups exhibit a right-skewed distribution with most posts having lower scores.

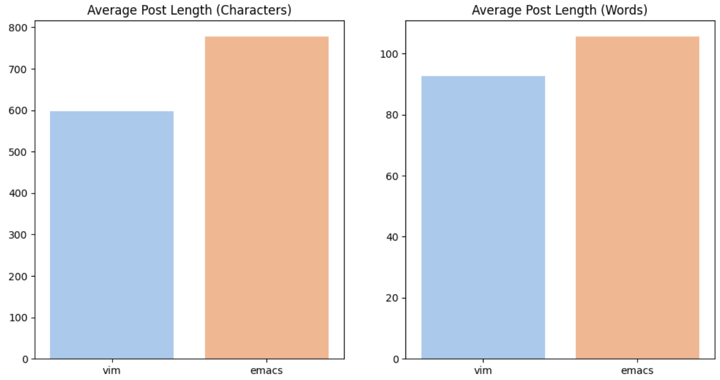
Graph 5: Histogram and Box Plot of Scores for r/emacs and r/vim



The distribution is right skewed for both groups, consistent with the previous observations, indicating that posts with fewer words are much more common than longer ones.

2. Additional Analysis

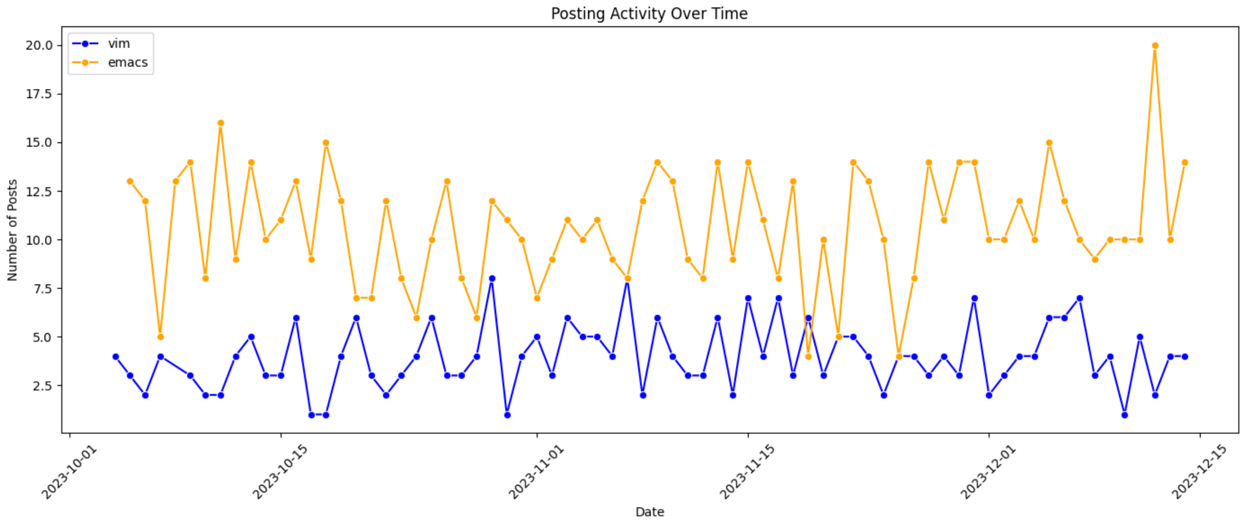
Graph 6: Comparative Analysis of Average Post Length in Characters and Words



The word and character counts for the subreddit are similar, but Emacs tends to have slightly higher values than Vim across all plots, suggesting a consistent pattern of higher interactions in Emacs posts.

3. Time Series Analysis of Posting Activity

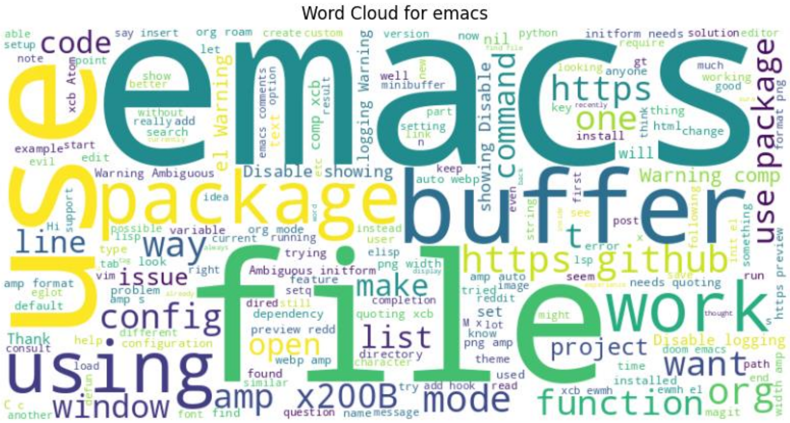
Graph 7: Time Series Plot of Posting Activity Over Time for r/emacs and r/vim



The time series shows that emacs subreddit is more active than vim.

4. Content Analysis

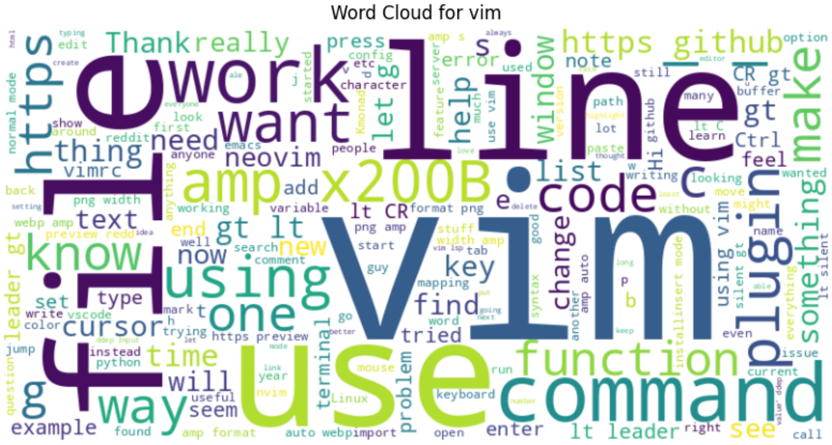
Graph 8: Word Cloud for r/emacs



Word Cloud for emacs

The word cloud for r/emacs features prominent terms such as 'buffer', 'function', and 'package', which are indicative of the technical discussions that revolve around software customization and usability. The prevalence of the word 'work' suggests that efficiency or workflow optimization is a key concern among users.

Graph 9: Word Cloud for r/vim



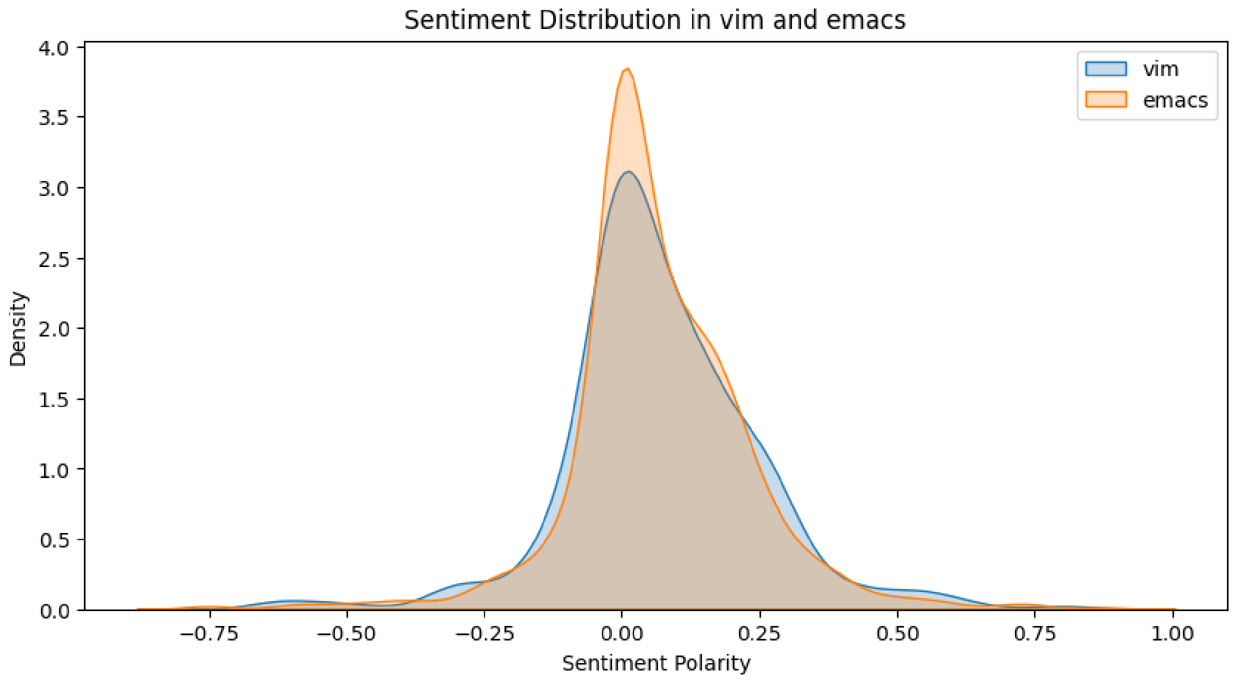
Word Cloud for vim

n the r/vim word cloud, 'line' and 'command' stand out, highlighting the community's engagement with Vim's editing capabilities and command-line interface. The frequent appearance of 'use' underscores a focus on practical usage tips and tricks.

The word clouds for r/emacs and r/vim both emphasize common terms like 'use', 'file', and 'function', reflecting a shared emphasis on utility and programming within the text editor communities. The unique terms in each cloud reveal the distinct characteristics of the discussions in their respective subreddits.

5. Sentiment Analysis

Graph 10: Sentiment Distribution in r/emacs and r/vim



The sentiment distribution graph exhibits similar patterns for both the r/emacs and r/vim subreddits, indicating that the overall tone of posts tends to be comparable across both communities. However, the r/emacs subreddit shows a slightly higher peak density at 4, compared to the maximum of 3 for r/vim. This suggests that posts in r/emacs may, on average, have a marginally more positive sentiment or a higher concentration of posts with similar sentiment scores. Despite this difference, the close proximity of the distributions implies that the emotional tone in discussions is generally alike between these text editor-focused forums

Conclusion

The comparative analysis of the r/emacs and r/vim subreddits over a three-month period has offered valuable insights into the behavior and engagement of Reddit contributors. Key metrics such as upvotes, comments, and sentiment analysis indicate that both communities are similarly active, with a tendency for a few posts to receive most interactions. Notably, r/emacs displays slightly higher engagement and a marginally more positive sentiment. Word clouds highlighted common technical discussions but also distinct topics that cater to each subreddit's unique culture. In essence, while r/emacs and r/vim are alike in many respects, they each retain their distinctive identities shaped by their contributors.

Future Work

In my efforts to extract Reddit data, direct use of the Reddit API proved to be the most effective method, particularly for its accessibility and ease of use. However, this approach is constrained by limitations in how far back one can retrieve historical data, which may impact the comprehensiveness of long-term trend analysis.

An alternative approach considered was the utilization of the Pushshift API, renowned for its capacity to access extensive historical Reddit data. Regrettably, challenges arose when Pushshift restricted the issuance of new access tokens, and attempts to connect resulted in authentication errors, as experienced through the response {"detail":"Not authenticated"}.

Looking ahead, there are plans to overcome these hurdles and further explore Pushshift as a viable data source. The intention is to find new methods to tap into its rich repository or perhaps seek updates on its token distribution policy, thereby expanding our capacity to extract a broader range of data.