Analytical Approach In Making Technical Stack Selection

Apoorv Yadav
Binghamton University
Binghamton, USA
ayadav7@binghamton.edu

Akash Rasal
Binghamton University
Binghamton, USA
arasal2@binghamton.edu

Abstract

Social Media is an integral part of our lives. It is omnipresent and influences our day-to-day decisions. We can collect a lot of this data and perform analysis on it. Using this data, we can gain insights on various situations. Web repository platforms are easy to use and manage open-source projects. Posting issues and getting them resolved is a part and parcel of open-source projects. On the other hand, for development related issues, support, any interesting findings and geeky discussions, users use social media platforms. While selecting a technology stack for a project or an organization, it is really important to gauge trust worthiness of the technologies. Along with this, it is also important to check on toxicity of a technology community before choosing it to identify its usability. We are collecting data from two sources, Reddit[12] and GitHub[4]. We use third party API, moderatehatespeech.com [9] to predict toxicity of the records. We used the reddit and github activity data to gather insights on the engagement on our data sources. These insights have visible trends for the user interactions and therefore might help a new user to decide his next stack. Furthermore, this project can be extended to any number of technical communities. Together, all the data is used to explore the engagement and overall friendliness of the online communities. It helps to estimate the number of issues or naivety of technologies. This can facilitate users while making decisions to choose any technology stack.

ACM Reference Format:
Apoorv Yadav and Akash Rasal. 2023. Analytical Approach In Making Technical Stack Selection. In *Proceedings of Project Proposal (CS 515 '23).* BU, Binghamton, NY, USA, 10 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

CS 515 '23, Fall 2023, Binghamton, NY, USA

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM…$15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 Introduction

With user adoption of different frameworks, new technologies and languages, online communities are increasing. Online communities help resolve issues and have healthy discussions. Subreddits are very important when resolving an issue while working on a technology/developing an application. Users post their application development issues on subreddits where other users from the globe help resolve it. They often ask follow-up questions to get a clear understanding of the issue and try to provide a solution. The comments can be upvoted, downvoted and can be shared. The open-source community enjoy helping each other over social media platforms. For open-source projects, a repository of the code is maintained, and users could run the code and contribute by posting issues, resolving the issues and keeping the project building. Many open-source projects are developed together using version management platform GitHub. In GitHub, an issue can be raised in a repository using the issue section. Comments can be added on issues, reactions can be given to a comment, and issues can be closed. An open-source community can be analyzed for its engagement, comments on issues and closure of issues.

We are using toxicity analysis on the collected data using moderatehatespeech.com[9] API. Using the prediction, we explore any possible offensive behavior in these communities and compare it against each other for their friendliness. Additionally, we focus on overall activity of the stack using parameters such as active unique users, average turnaround time for issue/question resolution and new submissions. Currently, we choose only a subset of technologies however, we believe that this approach can be used to compare toxicity of any technology stack having presence on reddit and GitHub. With our analysis, we aim to aid users in the adoption of new and unfamiliar technology stacks.

2 Data Source

For gathering data, we have selected few GitHub repositories and their Reddit counterparts. The selection criteria that we used were primarily the popularity of the technology, contributions on the GitHub repository and active community on Reddit.

Following are the repositories on GitHub:

- `TypeScript` [17]
- `nixpkgs` [11]

- `rust` [13]
- `flutter` [2]
- `go` [5]
- `kubernetes` [7]
- `swift` [15]

Following are the most popular subreddits (in terms of size) on Reddit:

- `r/typescript` [18]
- `r/NixOS` [10]
- `r/rust` [14]
- `r/FlutterDev` [3]
- `r/golang` [6]
- `r/kubernetes` [8]
- `r/swift` [16]

3 Proposed Research Questions

Our goal is to study engagement patterns on the forums and utilizing those patterns to aid engineers to adapt the technology stacks. Formally, we intend to focus on below research questions:

1. How friendly is any GitHub repository and SubReddit to be used by anyone?
2. How popular is a technology stack and how fast are the issues resolved for a technology stack?
3. Can this summarization help decide the users to decide a technology stack?

We attempted to gather insights from the data collected to focus on the above research questions. Using activity data as well as text based analysis like toxic speech recognition and semantic analysis, we were able to throw light on how safe and friendly each of the community is. This can be observed in fig 6, fig 5, fig 8 and fig 9. In terms of evaluating a communities popularity, we utilized unique user count who engaged per day, number of submissions, upvotes, first response time, active time of a post, issue resolution time and number of comments. We define first response time as the time elapsed between the creation of a post/issue and its first comment. Similarly, we define the active time of a reddit post as the time between the first comment and the last comment.

4 Related Work

Over the past few years, there has been a surge in the number of Reddit users. Especially, during the pandemic 2020, number of active Reddit users has increased drastically[19]. Being a popular social media platform, Reddit is no exception to toxic users and cyber bullying[22]. When using a subreddit, it is highly important that we choose and use a accommodating subreddit. Also, when we consider technology related social media like GitHub, there are very few users who actively contribute to a repository [20]. There is a need to analyze toxicity and total active users on Reddit and GitHub which can facilitate the decision to choose a technology.

Sr. No	Dataset	Number of Records
1	GitHub Issues	14,462
2	GitHub Comments	37,865
3	Reddit Posts (excluding r/politics)	5,158
4	Reddit Comments (excluding r/politics)	52,549
5	Reddit Posts (r/politics)	2,429
6	Reddit Comments (r/politics)	314,988

Table 1. Total Data Collected

5 Implementation

To analyze toxicity of an issue/post, we sent data to an API (moderatehatespeech.com) which analyzed the text and report the toxicity on a scale of 0 to 1 where 0 being least toxic and 1 being the most toxic. Using toxicity for each data point in the respective subreddit and repository, we deduced the average toxicity of subreddit and its corresponding GitHub repository. This analysis will help us interpret friendlier tech communities.

With an aim to check and see responsiveness of a technology stack, we analyzed post time and first comment and last comment time for subreddits. Similarly, for GitHub repositories, we made use of issue creation time, first comment time and issue close time for each issue. With our assumption that there will be some variations in the complexity of a post or an issue which will affect the response times, we still think it will be an important parameter to compare online communities for each technology stacks. Identifying unique users for each technology community and comparing it with other communities will help us study engagement. Further analysis is in section 6.

6 Result Discussion

6.1 Total data collected

Table 1 shows the total data collected for all subreddits, and its corresponding GitHub repository. In total, 110,034 data records were collected collectively for all subreddit(excluding politics) posts, comments and GitHub issues and comments. 2,429 posts and 314,988 comments were collected for r/politics subreddit.

The Figure 2 represents time on the X-axes and count on the Y-axes. Here, we can see that over the weekends, there is a drop in the number of comments and issues on GitHub.

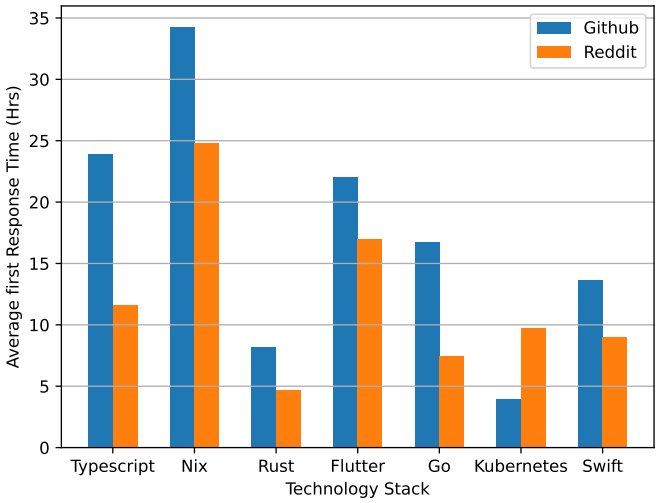

Figure 1. Reddit and GitHub average First response time, i.e., time elapsed between the submission and its first comment

6.2 Average First Response Time

In figure 1, on the X axis we have different subreddits and on the Y axis, we have average first response time. The above data is calculated over 40 days. From this, we can see that the NixOS community takes the most time to respond on any platform. Tech communities like RUST, and Swift respond similarly on both the platforms. We also observe that for almost all the technology stacks, response time on subreddits is smaller than their corresponding GitHub repositories.

6.3 GitHub Issue average Close time

In figure 3, repositories are on the X axis and average issue close/resolution time is on the Y axis. Among the 14462 issues collected, 9,883 issues are closed while creating this report and average time is calculated for the duration of issue creation and issue close time. Although we consider that some issues might be complex than other and it drastically affects the outcomes, we see that Swift repository has the least average issue closing time and Go lang has the highest issue closing time.

6.4 Reddit Average Post Active Time

In figure 4, subreddits are on the X axis and average active time is on the Y axis. We define the active time as the time elapsed from the first comment on the post to the last comment yet on the post. This can be considered in terms of how long a particular submission stay popular or relevant in the feeds of the subscribers. Longer duration might relate to better resolutions to a query.

6.5 Toxicity Analysis

6.5.1 GitHub issues and comments. We analyzed 14,462 issues and 37,865 GitHub comments over 40 days, and we couldn't find any significant toxic behavior. Only 5 comments were flagged by moderatehatespeech.com API. We predict that as GitHub is a professional platform, users are reluctant to be toxic on GitHub. Although we see less to no toxicity on GitHub, we cannot conclude that users are not toxic on GitHub by analyzing only 7 repositories. More exploration can be done on other popular repositories for toxicity.

6.5.2 Reddit toxicity analysis. Figure 5 shows toxicity for 150 comments on subreddit posts. X axis represents subreddits and Y axis represents the number of toxic comments. As discussed earlier, moderatehatespeech.com API was used to categorize comments. It is found that NixOS has the least toxic comments while Rust subreddit has the highest toxic comments. Although very few comments are flagged as toxic, we can see that users are more toxic on reddit than on GitHub. Also, we can observe that NixOS, Kubernetes and Swift are drastically less toxic than the other 4 subreddit technologies.

In figure 6, we plotted a word cloud of the data collected from toxic categorized Reddit comments. From this, we can gauge those words like stupid, idiot, know, and go are used more in toxic comments.

Fig 7 shows words and their occurrences (on X axes and on Y axes respectively). We can see that go, know, and people have high frequency in toxic comments.

6.5.3 Reddit comments sentiment analysis. Figure 8 gives us a perspective of positive, negative and neutral comments. The analysis was performed using Vader Sentiment Analysis [21], which describes if the comment is on a positive side or is negative. An example of a positive comment is,

> *"Sounds like a good resource, Thank You"*

An example of a negative comment is,

> *"If you want to build Android App Store, consider dart/flutter instead, but if programming for the Apple ecosystem only stick to Swift."*

Similar analyses 9 was done on github comments and it was found that there are many positive comments when compared to negative comments.

6.6 Activity per day

Figure 11 describes total activity in terms of new submissions (post/comment) on both the sources of all stacks daily. X axes represent date and Y axes represents number of new submissions. We can see that trends of comments and posts coincide and are closely related. The dip on 4th-6th November is due to little/no data collection as the VM of our data collection system was down. We observe that GitHub issues and comments are highly co-related. Also, the dips occur during weekends which signify that the tech community is not highly active over the weekends on GitHub. Rust has the most unique active users while TypeScript and NixOS have the least unique active users per day fig 14.

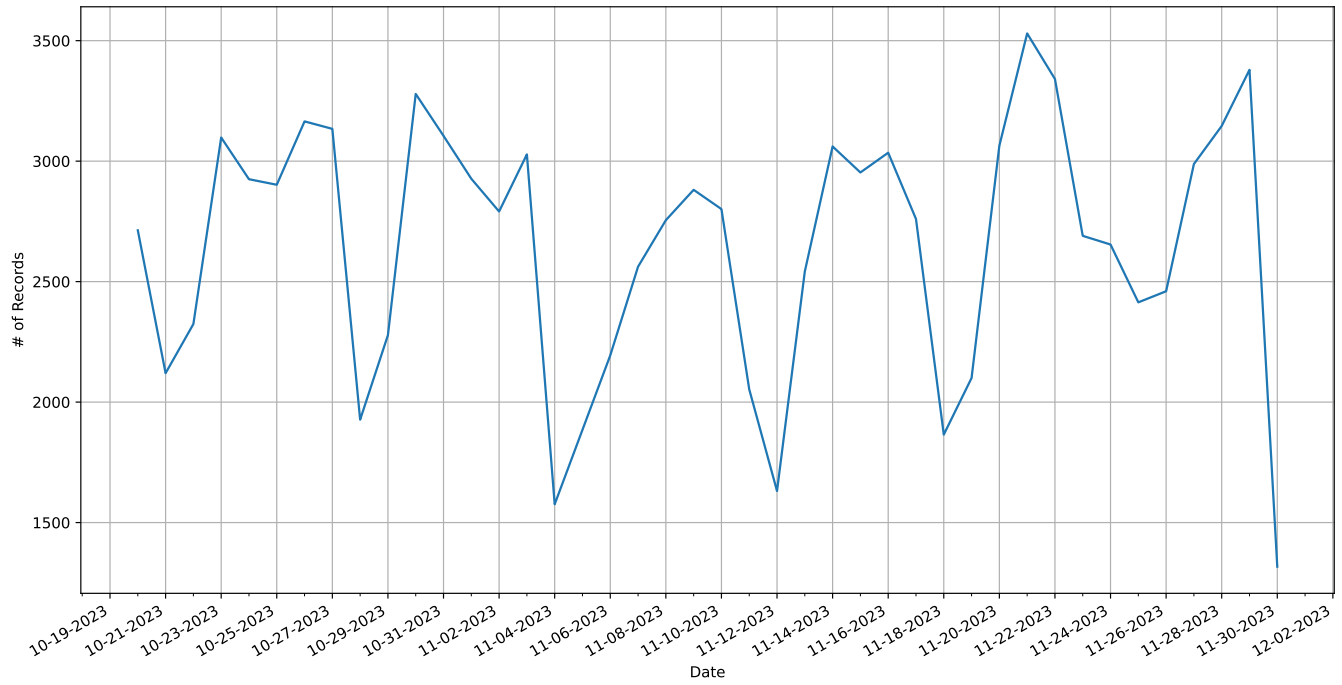

Figure 2. Total Data Over Time as of 30 Nov 2023 01:45 pm UTC

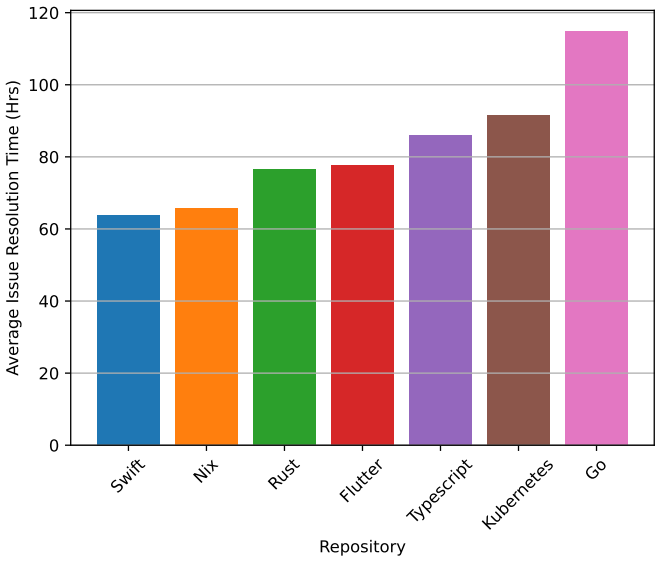

Figure 3. Average GitHub issue close time

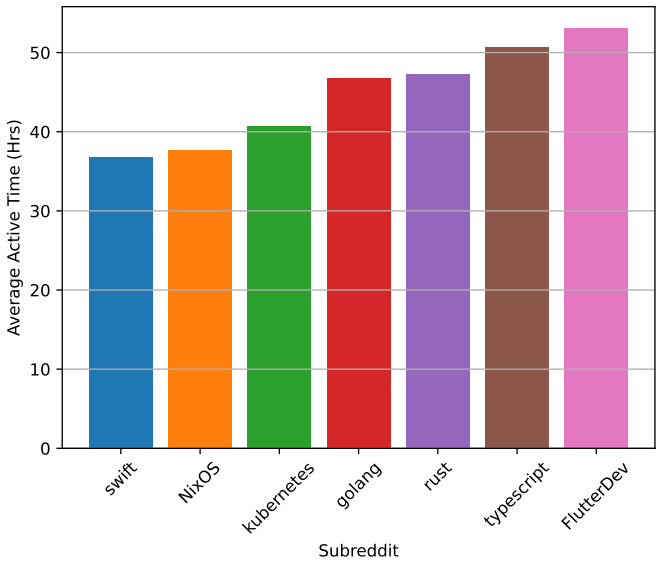

Figure 4. Average Reddit active time

We also plotted CDF on number of comments on a post on github vs reddit overall to compare the interaction from the community in these stacks fig 10.

7 System Design

In this project, we extend our existing system Figure 19 to get data from the moderatehatespeech.com api and to store it in new tables. We are using job processing system faktory [1] to process this in more effective and fault proof manner. Since the api is susceptible to downtime, using job queue system will supposedly ensure to process as much data as possible. After a comment/post is inserted successfully in the database, we inserted it into job queue for getting sentiments. After the data is fetched, it is inserted into source specific

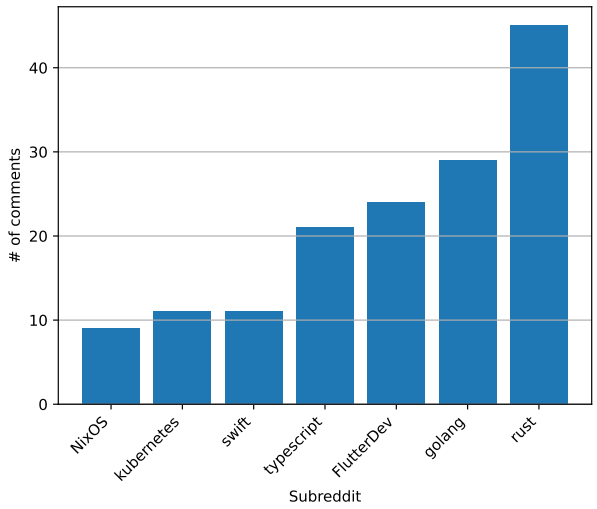

Figure 5. Reddit comment toxicity analysis

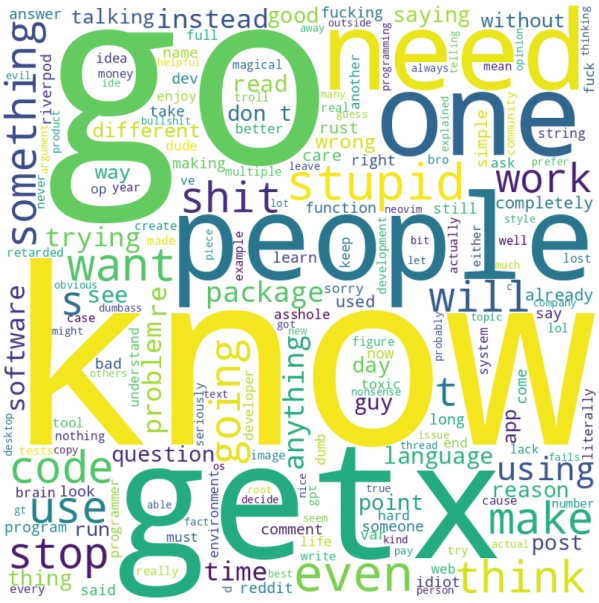

Figure 6. Word cloud of toxic reddit comments

queues to be picked up for being stored in the database. We attempt to make this system as generalized and extendable as possible to accommodate additional data sources in the future.

8 Data Collection: r/politics

As per the additional requirements of the project, we collected data from r/politics subreddit from Nov 1 to Nov 14. We collected the posts as well as comments with a frequency of 1 minute. We plotted the required charts for number of posts vs time (daily) fig 12 and number of comments vs time

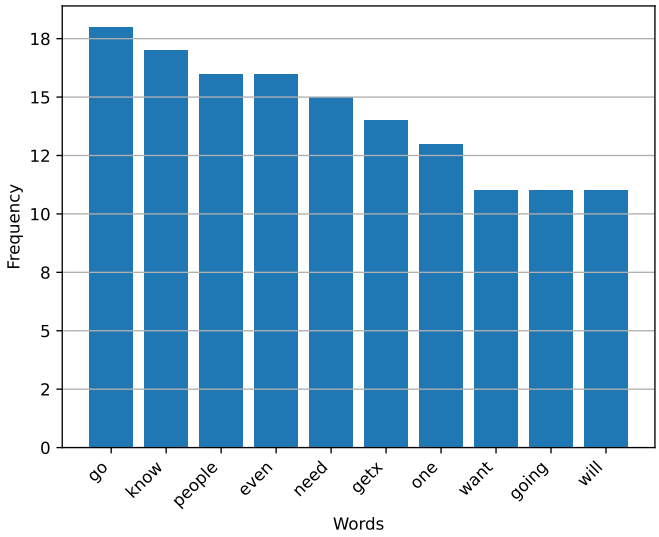

Figure 7. Most Frequent words in Reddit toxic comments

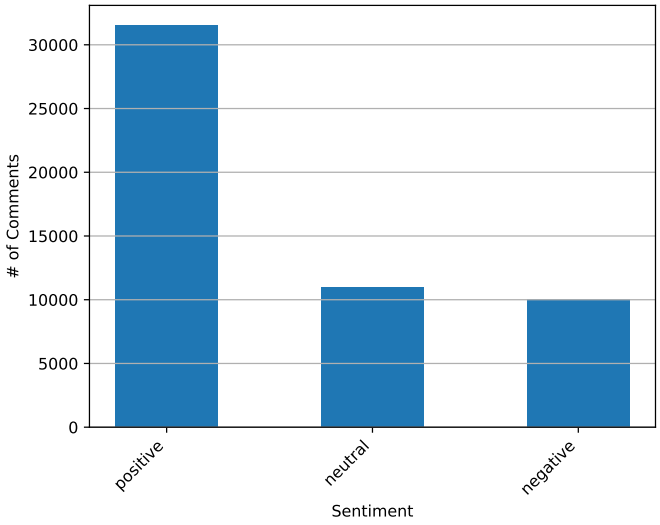

Figure 8. Sentiment on non-toxic Reddit Comments

(hourly) fig 13. There is a dip in the count from 4th to 6th Nov due to the unavailiblity of the VM.

9 Conclusion

The system to collect data and analyze data from Reddit and GitHub was built. User engagement, user toxicity analysis in posts, issues, comments was made for all chosen subreddits and GitHub repositories. A comparison was made between different technology stacks, their users and the toxicity in posts.

From our observation, GitHub has little to no toxicity, takes comparatively more time to respond to issues. From the response times and issue closing times, we can see that

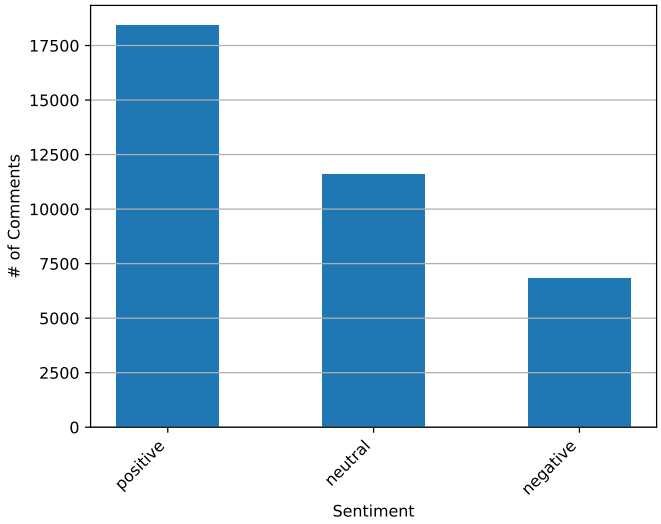

Figure 9. Sentiment on non-toxic Github Comments

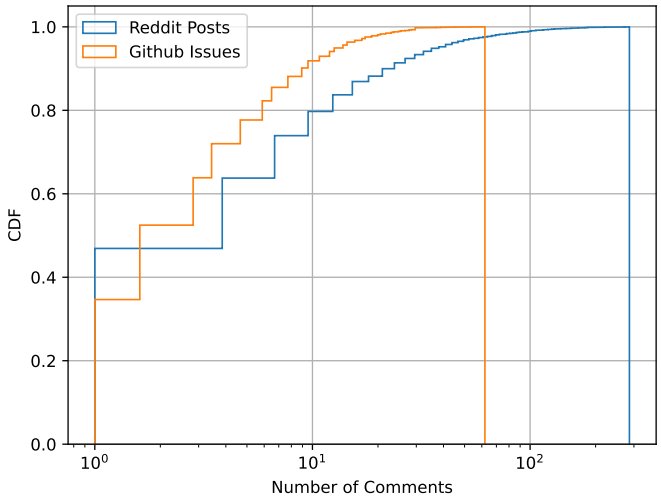

Figure 10. Comments on Posts CDF distribution

Kubernetes is popular among the selected technology stack on GitHub. We observed that response time for Reddit is low as compared to GitHub but Reddit has higher toxicity. NixOS is the least toxic technology and can be termed as friendly. Also, from our analysis, we can conclude that longer issues/post descriptions on both GitHub and Reddit are preferred by users to solve issues.

Apart from this, an analysis was done on r/Politics subreddit for the number of posts and comments.

10 Future Work

One of the aspects that we can explore in the future would be how does the structure of reddit post affect the engagement on the post. As a user in technical forum, a post can contain URLs, media, and code. Usually, these are helpful in providing the community more details about the post and hence might affect the engagement on it. Since, this will require lot of changes in the collection system, it is out of scope of our current analysis but can be an interesting approach for the future.

References

[1] 2023. Faktory. https://pypi.org/project/faktory/
[2] 2023. Flutter GitHub. https://github.com/flutter/flutter
[3] 2023. Flutter subreddit. https://www.reddit.com/r/FlutterDev/
[4] 2023. GitHub REST API. https://docs.github.com/en/rest?apiVersion=2022-11-28
[5] 2023. Go GitHub. https://github.com/golang/go
[6] 2023. Go subreddit. https://www.reddit.com/r/golang/
[7] 2023. Kubernetes GitHub. https://github.com/kubernetes/kubernetes
[8] 2023. Kubernetes subreddit. https://www.reddit.com/r/kubernetes/
[9] 2023. ModerateHateSpeech.com. https://moderatehatespeech.com/
[10] 2023. NixOS subreddit. https://www.reddit.com/r/NixOS/
[11] 2023. Nixpkgs GitHub. https://github.com/nixos/nixpkgs
[12] 2023. Reddit Data API Wiki. https://support.reddithelp.com/hc/en-us/articles/16160319875092-Reddit-Data-API-Wiki
[13] 2023. Rust GitHub. https://github.com/rust-lang/rust
[14] 2023. Rust subreddit. https://www.reddit.com/r/rust/
[15] 2023. Swift GitHub. https://github.com/apple/swift
[16] 2023. Swift subreddit. https://www.reddit.com/r/swift/
[17] 2023. TypeScript GitHub. https://github.com/microsoft/TypeScript
[18] 2023. TypeScript subreddit. https://www.reddit.com/r/typescript/
[19] Logan Coe Bria Evert Kalifa Ford Tyler Frank Dr. Daryl Green*, Dr. Jack McCann. 2022. The Rise of Reddit: A Case Study on Persuasive Technology. *American Research Journal of Business and Management* 8, 1 (2022), 24–28. https://doi.org/10.21694/2379-1047.22006
[20] Michael Färber. 2020. Analyzing the GitHub Repositories of Research Papers. In *Proceedings of the ACM/IEEE Joint Conference on Digital Libraries in 2020* (Virtual Event, China) *(JCDL '20)*. Association for Computing Machinery, New York, NY, USA, 491–492. https://doi.org/10.1145/3383583.3398578
[21] C. Hutto and Eric Gilbert. 2014. VADER: A Parsimonious Rule-Based Model for Sentiment Analysis of Social Media Text. *Proceedings of the International AAAI Conference on Web and Social Media* 8, 1 (May 2014), 216–225. https://doi.org/10.1609/icwsm.v8i1.14550
[22] Alexey N. Medvedev, Renaud Lambiotte, and Jean-Charles Delvenne. 2019. The Anatomy of Reddit: An Overview of Academic Research. (2019), 183–204.

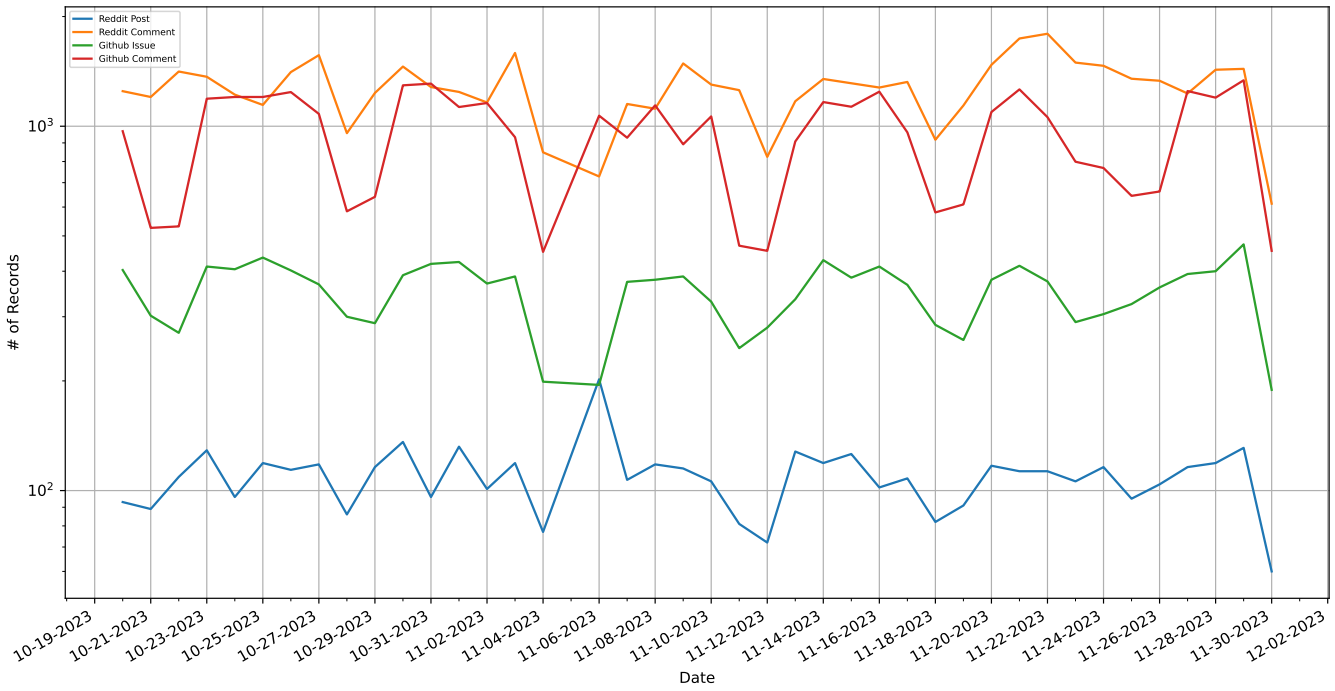

Figure 11. Data from all the sources

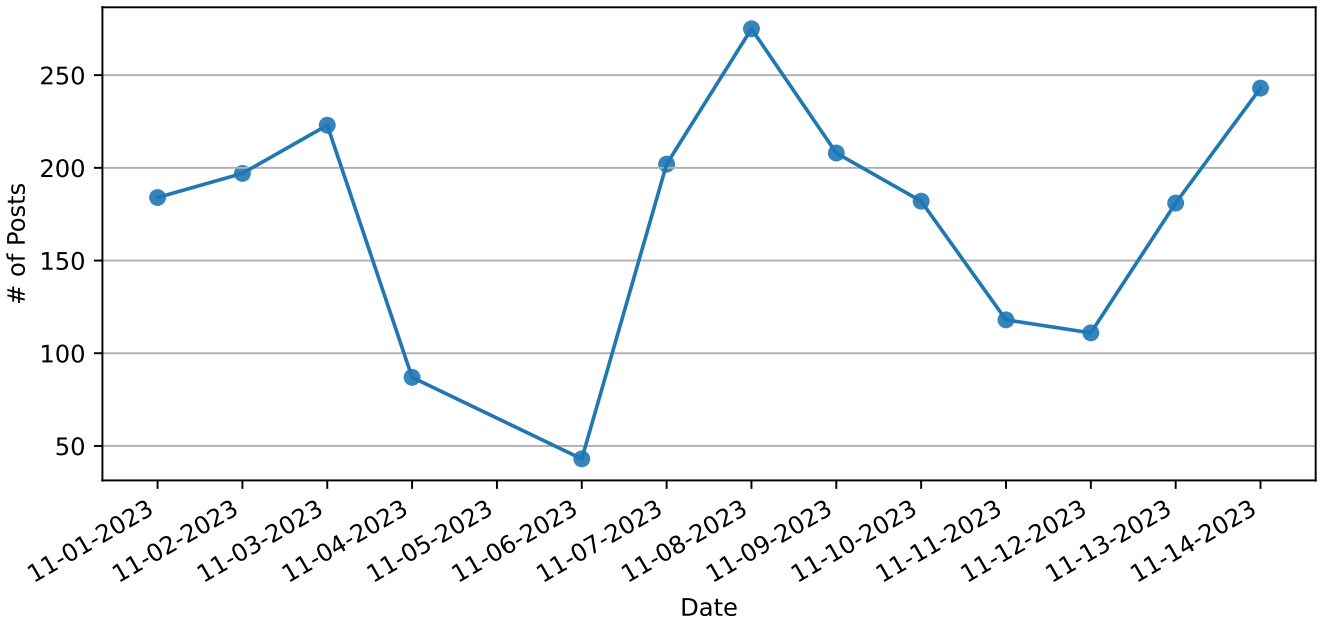

Figure 12. Posts in r/politics binned daily from Nov 1 2023 to Nov 14 2023

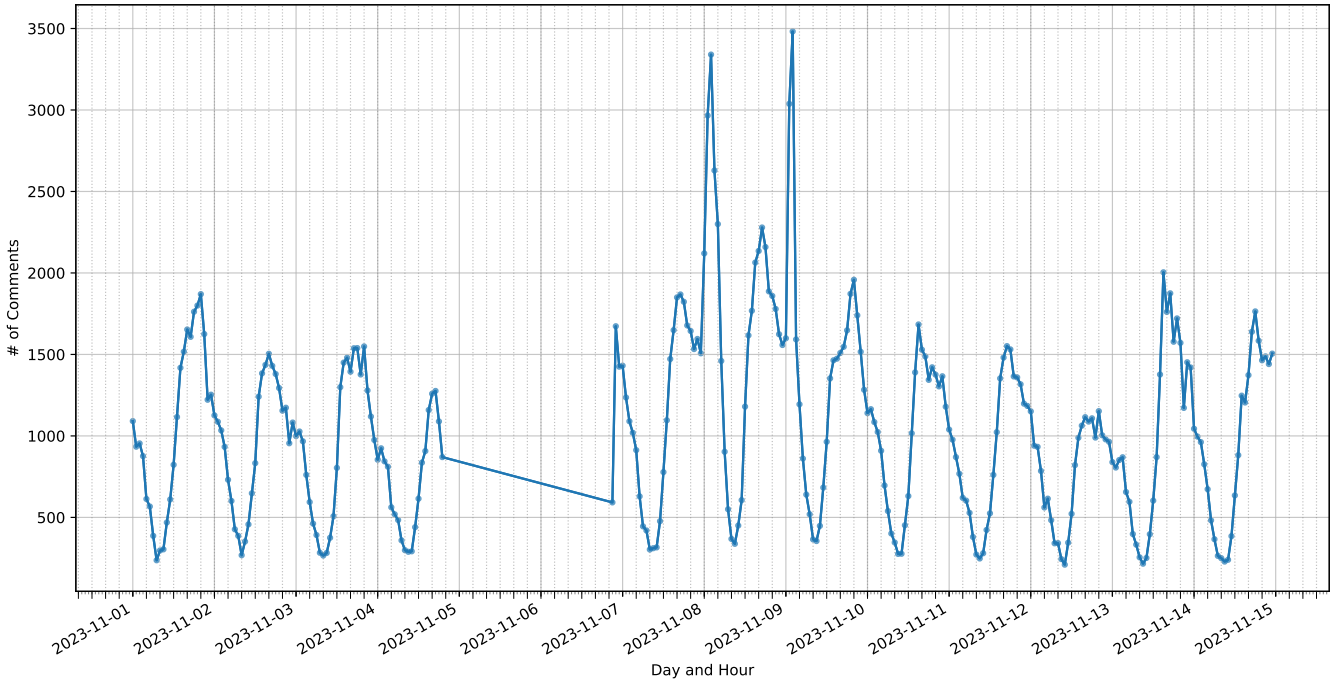

Figure 13. Comments in r/politics binned hourly from Nov 1 2023 to Nov 14 2023

Appendix

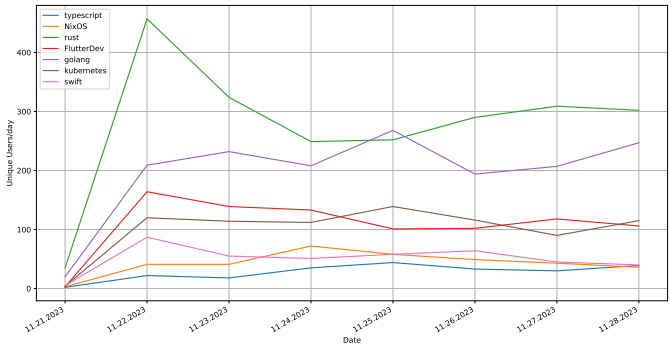

Figure 14. Unique users per day

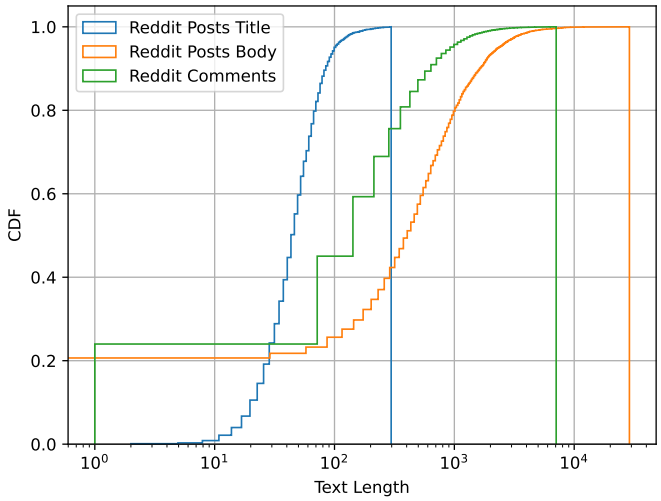

Figure 15. CDF Distribution of word length of reddit posts and comments

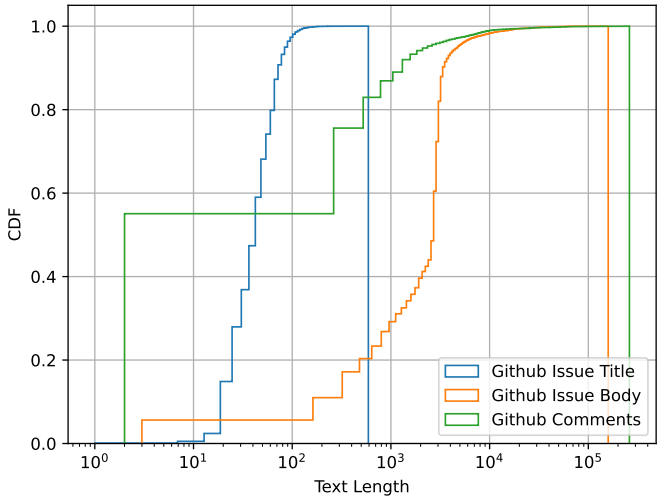

Figure 16. CDF Distribution of word length of github issues and comments

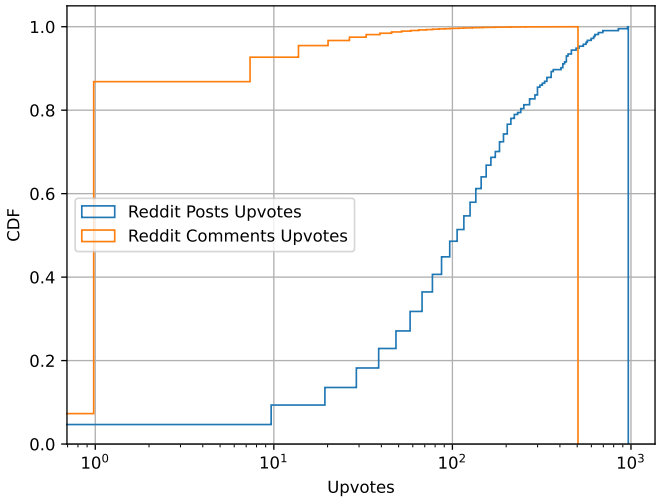

Figure 17. CDF Distribution of upvotes on Reddit posts and comments

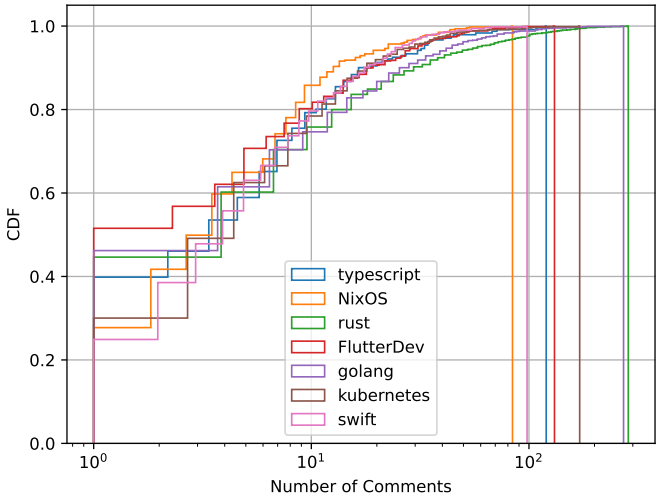

Figure 18. CDF Distribution of Number of Comments on a post on Reddit. It shows that all seven subreddits have similar trends on how many comments a particular post can get

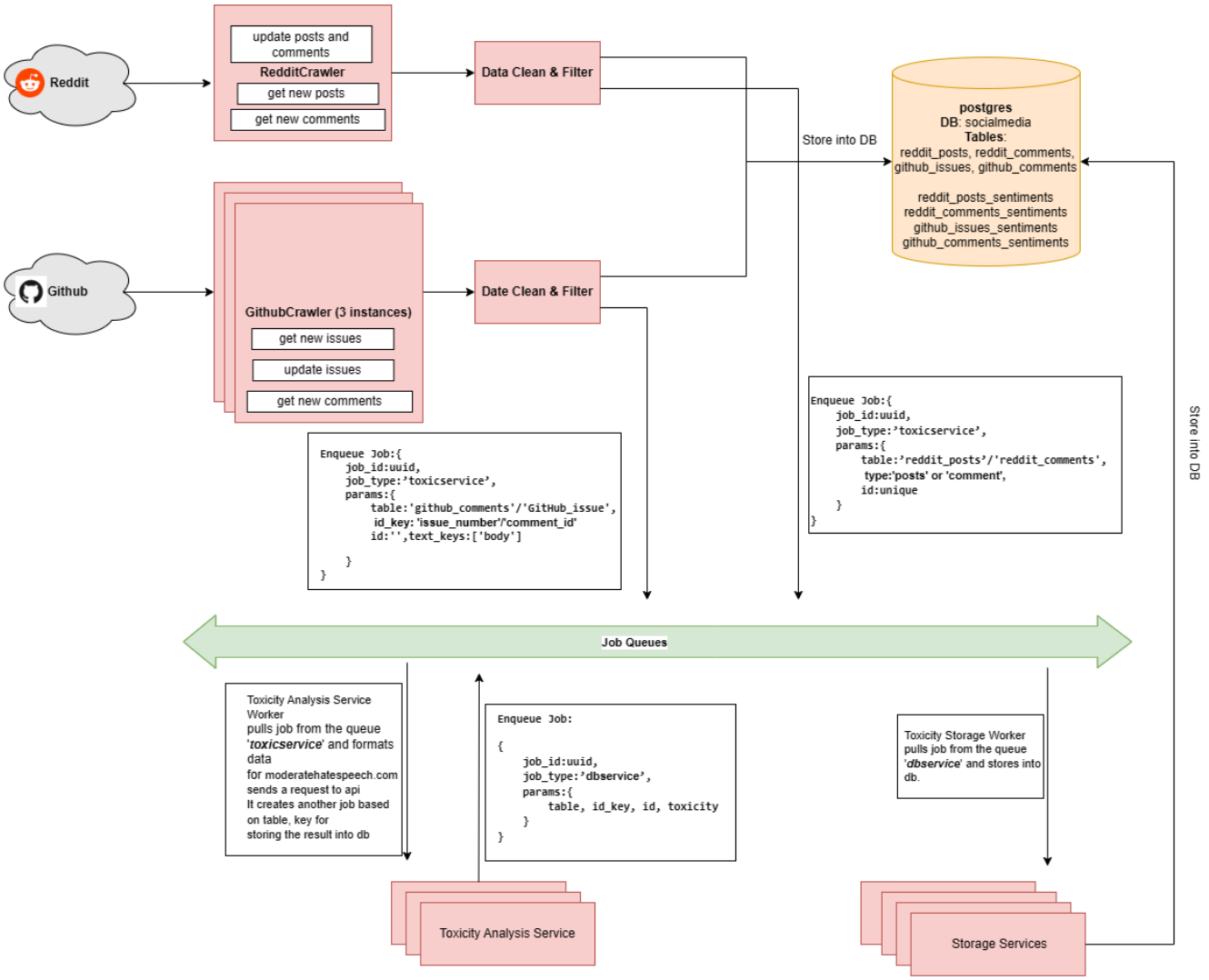

Figure 19. System Design Architecture for project 2.