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. We are collecting data from two sources, Reddit and GitHub We used third party api, moderatehatespeech.com to predict toxicity of the records. Together, all the data will be used to explore the engagement and overall friendliness of the online communities. This can facilitate users while making decision to choose any technology stack. We will further this by creating an interactive dashboard to facilitate users with our findings. We are using streamlit, an open-source Python library that makes it easy to create and share beautiful, custom web apps for machine learning and data science. Furthermore, our initial hypothesis that technical communities can have toxicity associated with them proved to be true. Additionally, we found variance in activity level across the technical stacks community. Therefore, having such analysis open and transparent to a user can aid them in making a decision.

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, 8 pages. 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[14] 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[3]. 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 used toxicity analysis on the collected data using moderatehatespeech.com[9] api. Using the prediction, we explored any possible offensive behavior in these communities and compared against each other for their friendliness. Additionally, we focused on overall activity of the stack using parameters such as active unique users, average turnaround time for issue/question resolution and new submissions. In this project, we are concluding our findings to the research questions and built a user-accessible web dashboard. Currently, we have 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 tool, we aim to provide users with transparency in the adoption new and unfamiliar technology stacks.

2 Data Source

We have kept our analysis constrained to the following technical stacks for this project.

Following are the repositories on GitHub:

- `TypeScript` [20]
- `nixpkgs` [11]
- `rust` [15]
- `flutter` [1]
- `go` [4]
- `kubernetes` [6]
- `swift` [18]

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

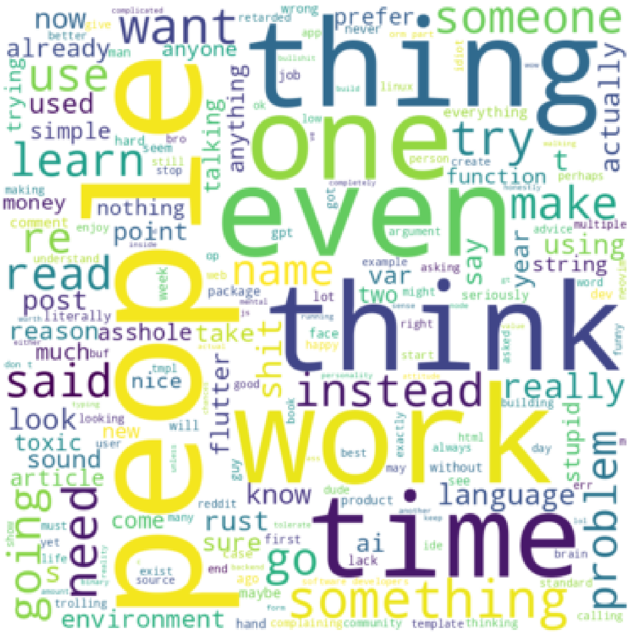


Figure 1. WordCloud generated with toxic comments for the subreddits.

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

- r/typescript [21]
- r/NixOS [10]
- r/rust [16]
- r/FlutterDev [2]
- r/golang [5]
- r/kubernetes [7]
- r/swift [19]

3 Proposed Research Questions

Our goal was to study engagement patterns on the forums and utilize those patterns to aid engineers in adapting the technology stacks. Formally, In the previous project, we intended 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?

4 Our Finding on Research Questions

After collecting data for over 55 days from online communities and performing analysis, we have our findings explained below:

1. **How friendly is any GitHub repository and SubReddit to be used by anyone?**

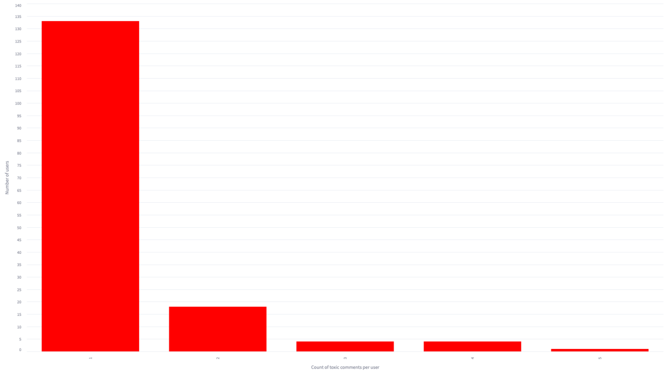


Figure 2. Repeated toxic subreddit users. Count of authors on y axis having number of toxic comments they posted on x axis.

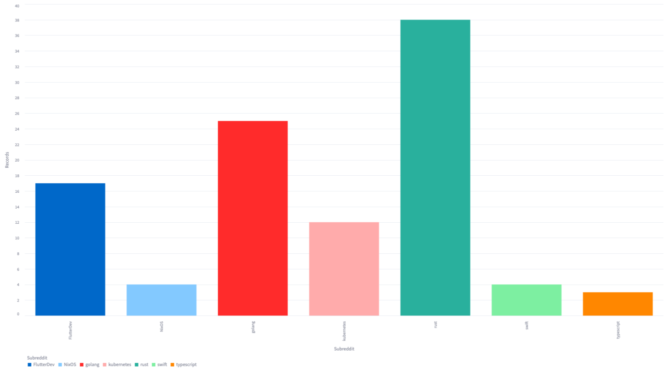


Figure 3. Toxicity per subreddit

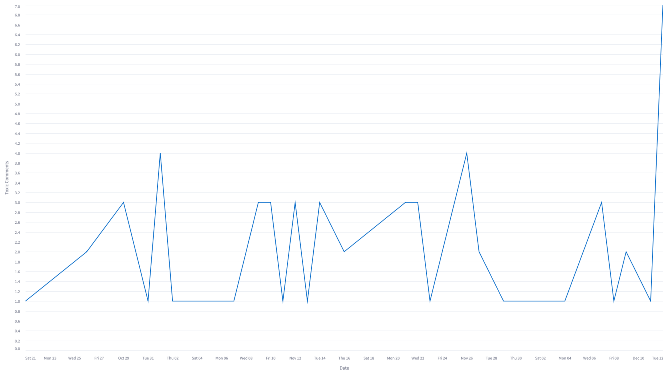


Figure 4. Toxic comments for Rust subreddit

We were able to gather statistics on the toxic comments from reddit communities across all the subreddit. We found 0 records that were flagged by moderatehatespeech on Github. This could be due to the strict moderation rules on Github or less popular choice for a new user as a medium to communicate. We observed few communities with far more and frequent toxic

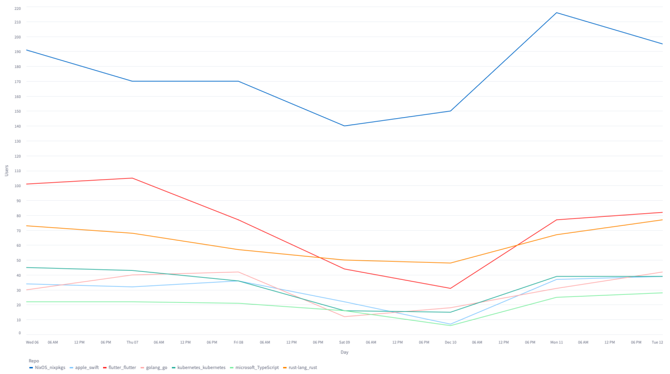


Figure 5. GitHub unique users for repos

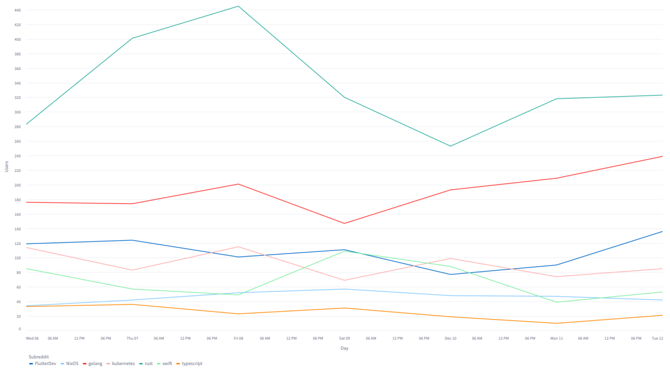


Figure 6. Reddit unique users for subreddit

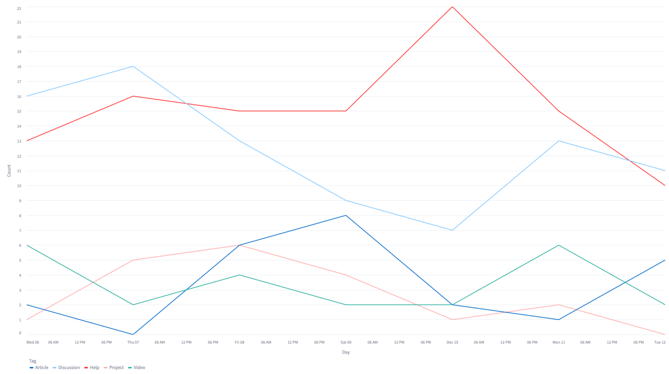


Figure 7. Tag counts per subreddit

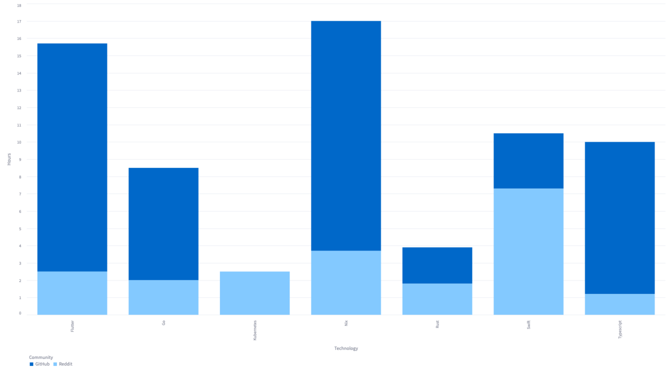


Figure 8. Time for first comment on Subreddit and GitHub repos

comments than others. We created a wordcloud, figure 1 using the flagged comments to give the user an intuitive view of the comments. Additionaly, the toxic comments were not from a repeated offender but from different authors figure 2. The maximum amount of toxic comments from a single author was 5 and most authors had just one comment flagged as toxic during the course of the data collection. We can say that most of the subreddit are friendly to use and all considered GitHub repositories are non-toxic.

2. **How popular is a technology stack and how fast are the issues resolved for a technology stack?** For this research question, our approach was different for github and reddit. On github, an issue is usually labelled as "open" or "close". Hence, we used this duration as one of our metric. For reddit, we considered the last comment on a post as its closing time as the submission had little to no contribution from the community. We also tracked the first responder time, which we define as the first comment on a post/issue. We compared the time and this analysis can help a user to decide the medium for his question. We also used labels from reddit submissions used by authors to categorize their posts as a metric on what type of posts are frequent on reddit community. We found most posts had not label, followed by "help", "discussion" and "project" fig. 7 . We believe these analysis can be used to summarize the popularity of a stack among common users. Apart from observing overall new records, We also consider unique user engagement per day in a community on reddit and github, which can be a big factor to its popularity. We can conclude that this analysis can help understand which technologies have good support on social media platforms.
3. **Can this summarization help decide the users to decide a technology stack?** As we discussed in the proposal, this research question requires us to actually have data from users after looking at our tool. Without surveying others, the answer to this research question would be incomplete. However, from the analysis so far, we can fairly conclude that such a summarization can definitely help a new user to decide the stack as well forum for his question. The tool can be used to compare similar stacks, where community discussions are actually a tiebreaker for a user to decide. In such situations, such a tool can definitely help the user.

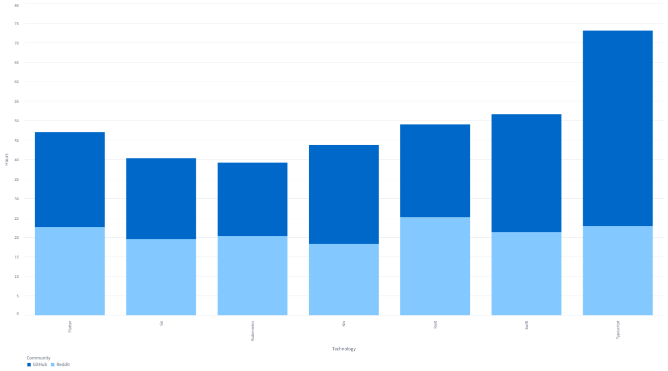


Figure 9. Time of last comment on Subreddit and GitHub repos

5 Interactive Dashboard

In this project, We created a web dashboard that can be accessed over a browser. Using the dashboard, any user will be easily able to make conclusions regarding the subreddit and a GitHub repository for user engagement, user interactions, and toxicity. Following are the functionalities of the dashboard:

5.1 Home Page

At the home page of our dashboard, a user will be able to gather insights about the communities which can be filtered based on date range and stack. They will be able to get graphical information regarding the number of new post/comments on the forum. Additionally, they can also get information regarding unique user engagement per day as per the filters. We provide the user with the information about the type of the posts on reddit.

5.2 Live Data

This page of the webpage have the functionality to get live updates on the traffic in the online forums. The page allows the user to chose the window of "1 hour" or "1 day".

5.3 Toxicity

This page provides the user with the information about the toxicity on reddit. We provide the user with intuitive information in the form of a wordcloud and charts. The wordcloud provides an intuitive representation of the words used in the flagged comments. The user can select the date range as well as the subreddit for a custom search. We also focus on the frequency of toxic comments per user. This helps in analyzing if the comments are specifically tied to a subset of users who exert this behaviour. Additionally, we allow user to have a look on toxic comments.

5.4 Activity

In this section of the dashboard, Users can compare the average time for a github issue resolution, last comment of reddit and first comment on a submssion/issue. Users have the ability to select a date range.

The user will have a high visibility of both data sources. Formally, our objective is provide the user with analysis to answer question 1 and question 2. We use the interactive dashboard to support our answers. For completely addressing question 3 requires additional data, particularly, data from users using the charts and their feedback of the summarization. However, we tried to support our conclusion for question 3 using the charts as much as possible.

6 Tools used to build the dashboard

We used python and use streamlit library to create the web dashboard. Streamlit is an open-source Python library that makes it easy to create and share custom web apps for data science. The plots will change according to the user input. Matplotlib [8], nltk [12], postgres[13] and wordcloud [22] libraries are used to plot the graphs.

7 Conclusion

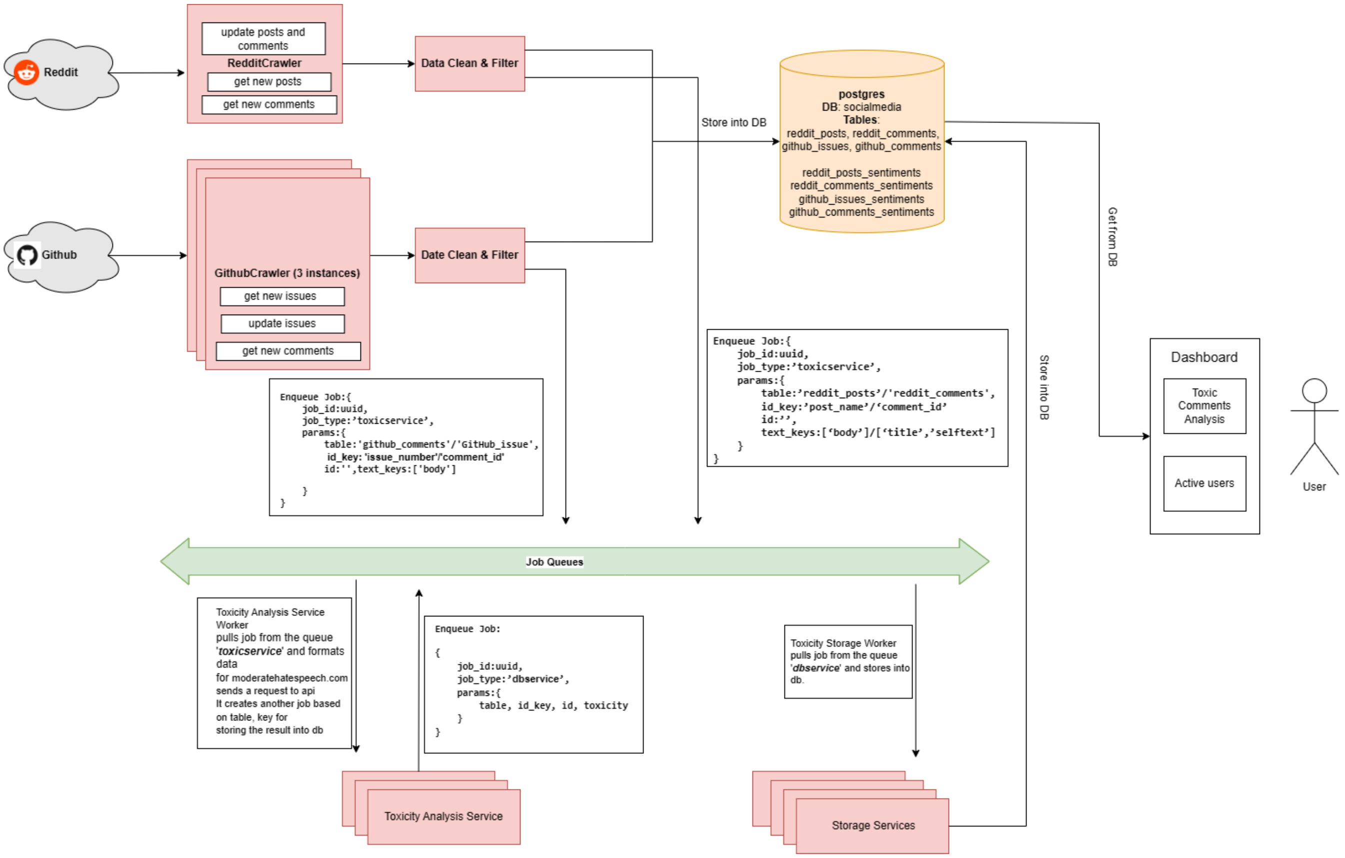
With this project, we are towards the conclusion of our analytical approach to make a technical stack decision. A dashboard was built to showcase and analyze the data collected from Subreddits and GitHub repositories. Different interactive graphs were present in the dashboard using which the user can gauge the subreddits and GitHub repositories on various parameters.

8 Future Work

One of the aspects that we can explore in the future would be the inclusion of more technological stacks and be ability to compare any community with another to make conclusions on the proposed research questions. Since this will require a lot of changes in the collection system including api rate limits, it is out of the scope of our current analysis but can be an interesting approach for the future.

References

[1] 2023. Flutter GitHub. https://github.com/flutter/flutter
[2] 2023. Flutter subreddit. https://www.reddit.com/r/FlutterDev/
[3] 2023. GitHub REST API. https://docs.github.com/en/rest?apiVersion=2022-11-28
[4] 2023. Go GitHub. https://github.com/golang/go
[5] 2023. Go subreddit. https://www.reddit.com/r/golang/
[6] 2023. Kubernetes GitHub. https://github.com/kubernetes/kubernetes
[7] 2023. Kubernetes subreddit. https://www.reddit.com/r/kubernetes/
[8] 2023. Matplotlib. https://matplotlib.org
[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. NLTK. https://www.nltk.org/index.html
[13] 2023. Postgres Database. https://www.postgresql.org/



```
Enqueue Job:{
    job_id:uuid,
    job_type:’toxicservice’,
    params:{
        table:'github_comments'/'GitHub_issue',
         id_key:'issue_number'/'comment_id'
        id:'',text_keys:['body']

    }
}
```

```
Enqueue Job:{
    job_id:uuid,
    job_type:’toxicservice’,
    params:{
        table:’reddit_posts’/'reddit_comments',
        id_key:’post_name’/‘comment_id’
        id:’’,
        text_keys:[‘body’]/[‘title’,’selftext’]
    }
}
```

Toxicity Analysis Service Worker
pulls job from the queue '*toxicservice*' and formats data
for moderatehatespeech.com
sends a request to api
It creates another job based on table, key for
storing the result into db

```
Enqueue Job:

{
    job_id:uuid,
    job_type:’dbservice’,
    params:{
        table, id_key, id, toxicity
    }
}
```

Toxicity Storage Worker pulls job from the queue '*dbservice*' and stores into db.

Figure 10. Proposed System Design, We have extended our initial design with a lightweight web service built using streamlit[17]

[14] 2023. Reddit Data API Wiki. https://support.reddithelp.com/hc/en-us/articles/16160319875092-Reddit-Data-API-Wiki

[15] 2023. Rust GitHub. https://github.com/rust-lang/rust

[16] 2023. Rust subreddit. https://www.reddit.com/r/rust/

[17] 2023. Streamlit. https://docs.streamlit.io

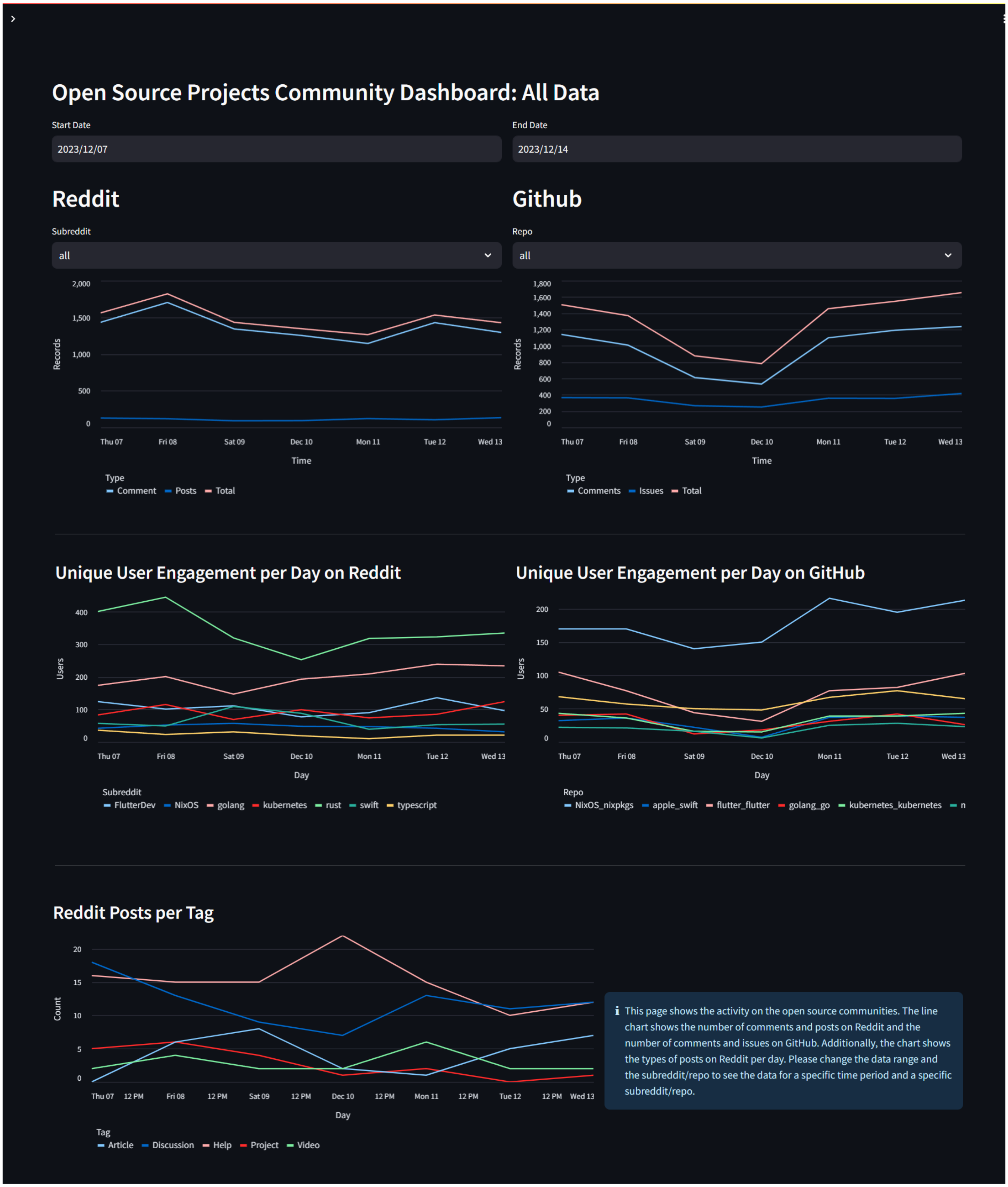
[18] 2023. Swift GitHub. https://github.com/apple/swift

[19] 2023. Swift subreddit. https://www.reddit.com/r/swift/

[20] 2023. TypeScript GitHub. https://github.com/microsoft/TypeScript

[21] 2023. TypeScript subreddit. https://www.reddit.com/r/typescript/

[22] 2023. Wordcloud. https://pypi.org/project/wordcloud/



Open Source Projects Community Dashboard: All Data

Start Date: 2023/12/07

End Date: 2023/12/14

Reddit

Subreddit: all

Github

Repo: all

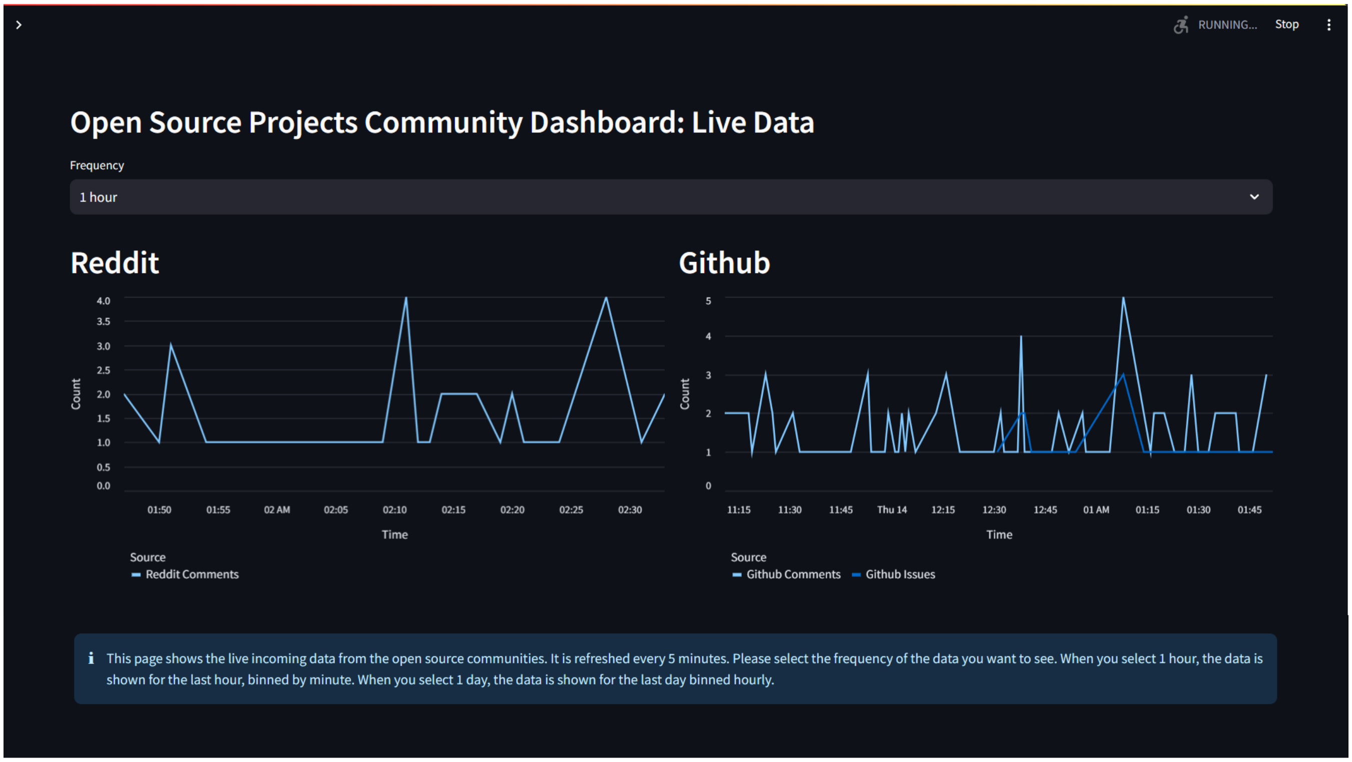
Unique User Engagement per Day on Reddit

Unique User Engagement per Day on GitHub

Reddit Posts per Tag

ℹ This page shows the activity on the open source communities. The line chart shows the number of comments and posts on Reddit and the number of comments and issues on GitHub. Additionally, the chart shows the types of posts on Reddit per day. Please change the data range and the subreddit/repo to see the data for a specific time period and a specific subreddit/repo.

Figure 11. Data overview dashboard

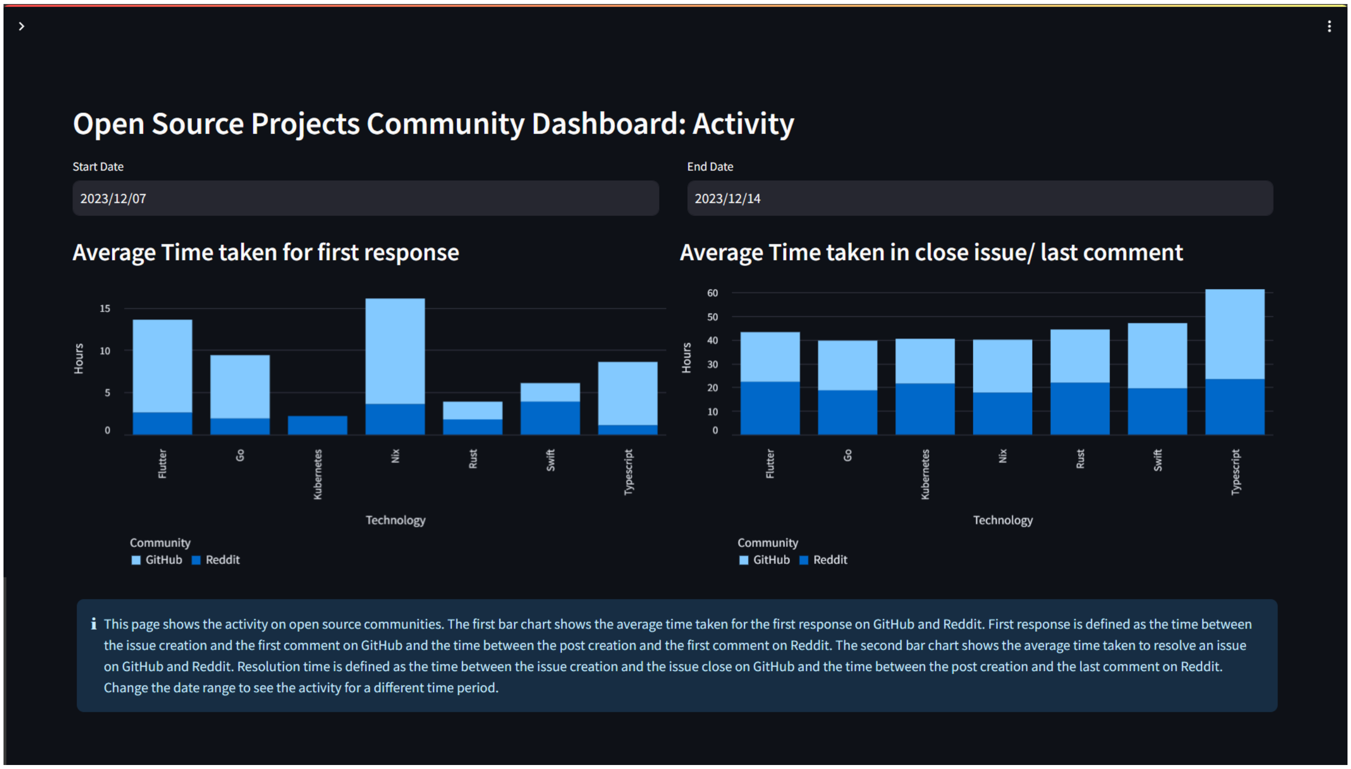


Open Source Projects Community Dashboard: Live Data

Frequency: 1 hour

This page shows the live incoming data from the open source communities. It is refreshed every 5 minutes. Please select the frequency of the data you want to see. When you select 1 hour, the data is shown for the last hour, binned by minute. When you select 1 day, the data is shown for the last day binned hourly.

Figure 12. Live Data dashboard



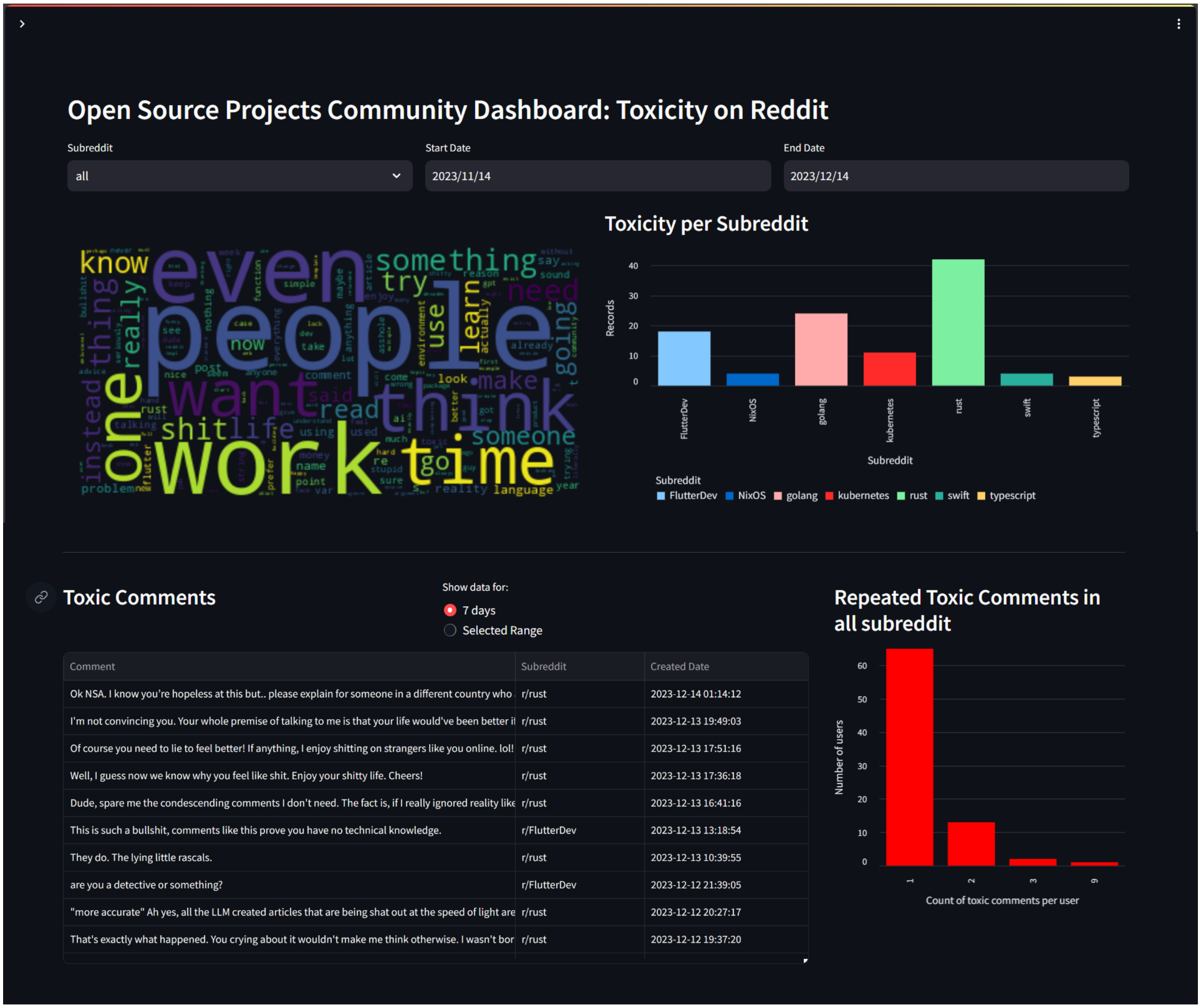
Open Source Projects Community Dashboard: Activity

Start Date: 2023/12/07

End Date: 2023/12/14

This page shows the activity on open source communities. The first bar chart shows the average time taken for the first response on GitHub and Reddit. First response is defined as the time between the issue creation and the first comment on GitHub and the time between the post creation and the first comment on Reddit. The second bar chart shows the average time taken to resolve an issue on GitHub and Reddit. Resolution time is defined as the time between the issue creation and the issue close on GitHub and the time between the post creation and the last comment on Reddit. Change the date range to see the activity for a different time period.

Figure 13. Tracing activity dashboard



Open Source Projects Community Dashboard: Toxicity on Reddit

Subreddit: all

Start Date: 2023/11/14

End Date: 2023/12/14

Toxic Comments

Show data for:

● 7 days
○ Selected Range

Comment	Subreddit	Created Date
Ok NSA. I know you’re hopeless at this but.. please explain for someone in a different country who	r/rust	2023-12-14 01:14:12
I'm not convincing you. Your whole premise of talking to me is that your life would've been better it	r/rust	2023-12-13 19:49:03
Of course you need to lie to feel better! If anything, I enjoy shitting on strangers like you online. lol!	r/rust	2023-12-13 17:51:16
Well, I guess now we know why you feel like shit. Enjoy your shitty life. Cheers!	r/rust	2023-12-13 17:36:18
Dude, spare me the condescending comments I don't need. The fact is, if I really ignored reality like	r/rust	2023-12-13 16:41:16
This is such a bullshit, comments like this prove you have no technical knowledge.	r/FlutterDev	2023-12-13 13:18:54
They do. The lying little rascals.	r/rust	2023-12-13 10:39:55
are you a detective or something?	r/FlutterDev	2023-12-12 21:39:05
"more accurate" Ah yes, all the LLM created articles that are being shat out at the speed of light are	r/rust	2023-12-12 20:27:17
That's exactly what happened. You crying about it wouldn't make me think otherwise. I wasn't bor	r/rust	2023-12-12 19:37:20

Figure 14. Toxicity dashboard