SWE LAB TEAM 10
Dr. Sridhar Chimalakonda , Mentor: Ms. Jahnavi K

Harmonize
RELEASE 1: REPORT

By CS21B0 13,43,45,47,53

INTRODUCTION

Harmonize is a Chrome Extension built to make Software Engineering online spaces more peaceful and harmonious than they currently are. It focuses on giving young and budding software engineers a positive environment so that they can grow and build successful careers.

OTHER LINKS/DOCUMENTATION:

GitHub Repository Link: Click Here

WSR: Click Here

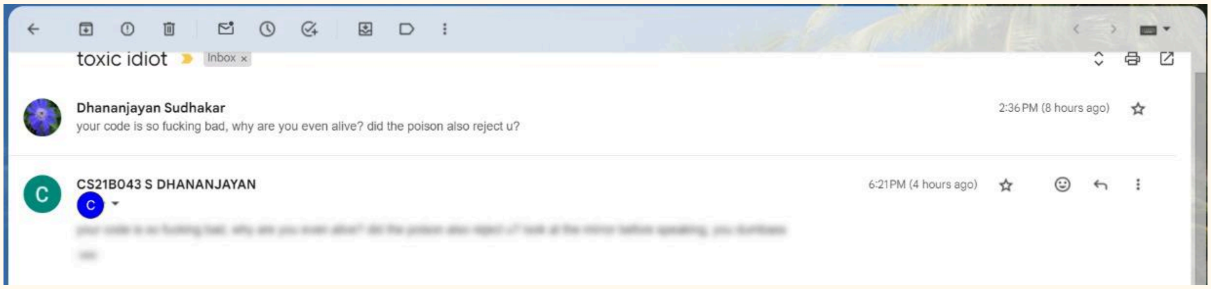
Error Docs: Click Here

Datasets Doc: Click Here

FEATURES

Blurring Toxic Comments - SE Specific

Takes comments from Gmail/GitHub and checks for toxicity level.The BERT model returns a toxicity score which is displayed in the toxicity bar . If the toxicity score is greater than a particular threshold, then the comment gets blurred.

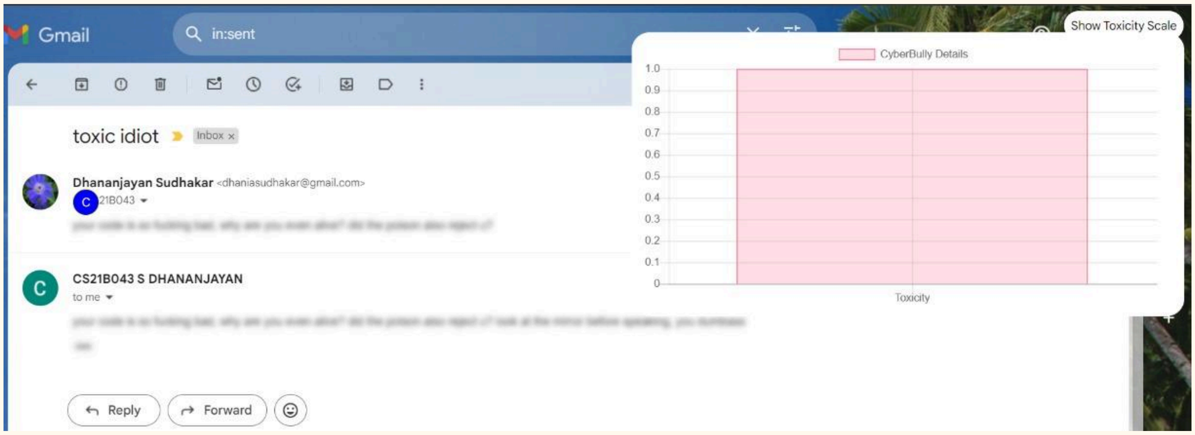


Dhananjayan Sudhakar 2:36 PM (8 hours ago)
your code is so fucking bad, why are you even alive? did the poison also reject u?

CS21B043 S DHANANJAYAN 6:21 PM (4 hours ago)

Toxicity Bar

On clicking the comment/reply, a highlight button pops up and on clicking the button, the toxicity scale is displayed.

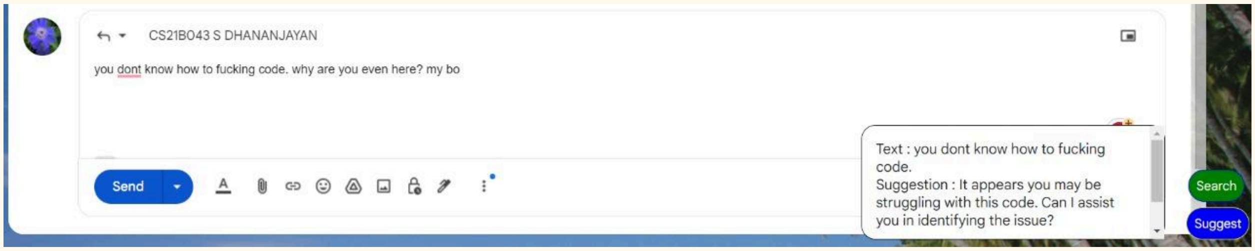


Dhananjayan Sudhakar <dhaniasudhakar@gmail.com>
21B043

CS21B043 S DHANANJAYAN
to me

SE-Specific Detoxification Feature

During the user interaction, if the comment/reply that is being typed is toxic, a suggestion panel pops up which contains search and suggest buttons. On clicking the suggest button, a better non-toxic alternative provided by the Gemini API will be displayed. The user can click on the replace button to replace the toxic comment with the suggested non-toxic comment.

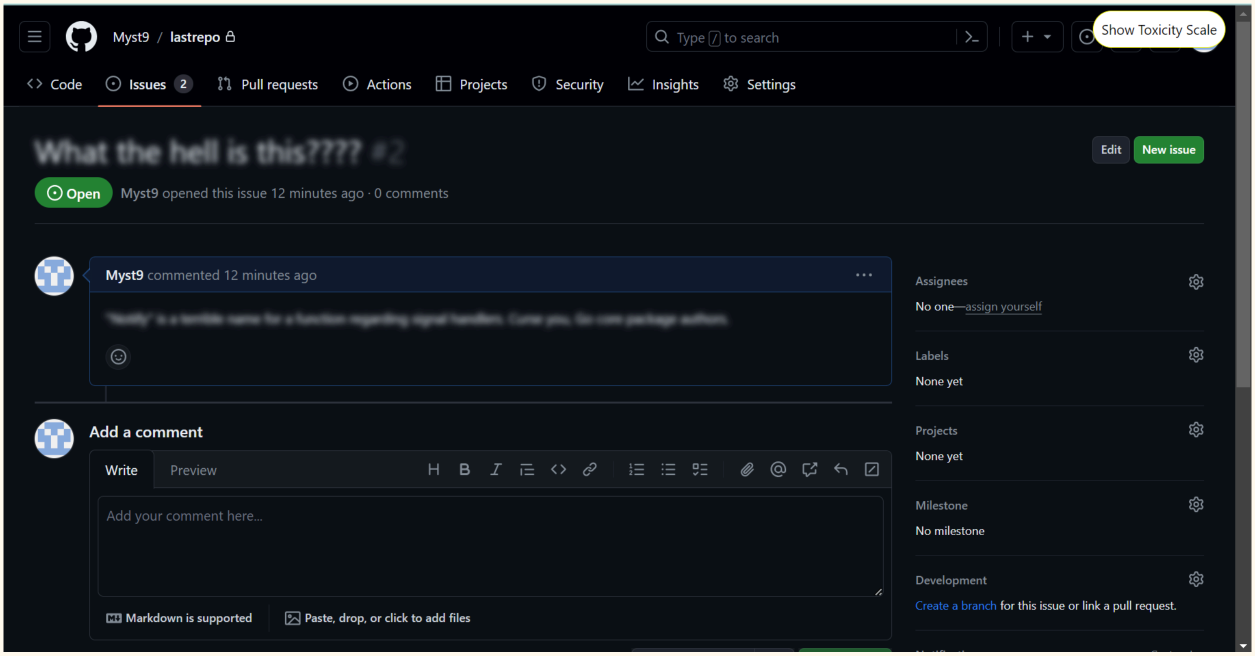

CS21B043 S DHANANJAYAN

you dont know how to fucking code. why are you even here? my bo

Text : you dont know how to fucking code.
Suggestion : It appears you may be struggling with this code. Can I assist you in identifying the issue?

Github

Blurs toxic comments


Myst9 / lastrepo

What the hell is this???? #2

Open Myst9 opened this issue 12 minutes ago · 0 comments

Myst9 commented 12 minutes ago

"Notify" is a terrible name for a function regarding signal handlers. Curse you, Go core package authors.

Add a comment

Add your comment here...

Assignees
No one—assign yourself

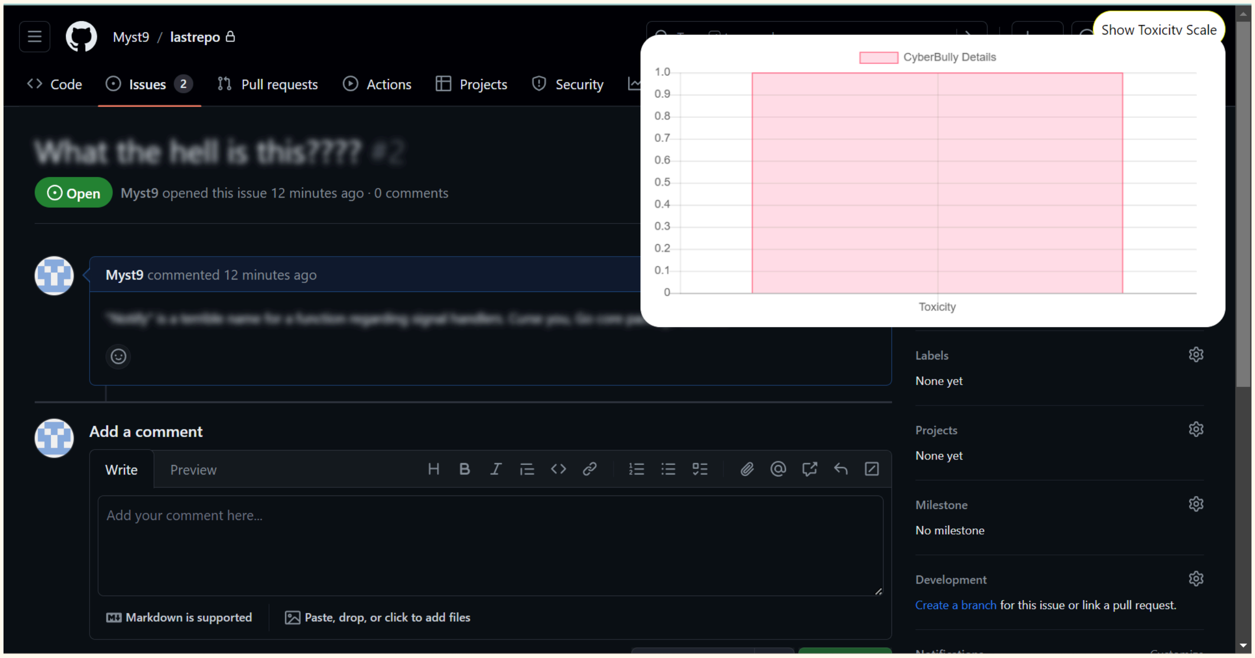
Labels
None yet

Projects
None yet

Milestone
No milestone

Development
Create a branch for this issue or link a pull request.

Toxicity scale for each comment



What the hell is this???? #2

Open Myst9 opened this issue 12 minutes ago · 0 comments

Myst9 commented 12 minutes ago

"Notify" is a terrible name for a function regarding signal handlers. Curse you, Go core pa

Labels: None yet

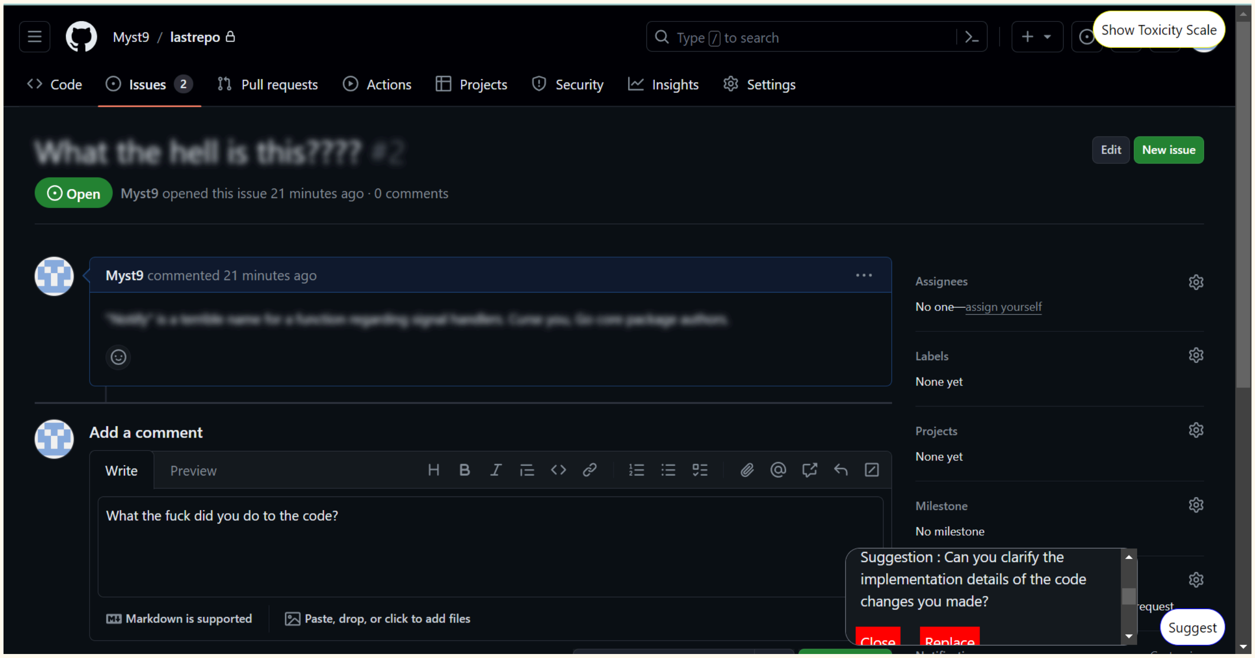
Projects: None yet

Milestone: No milestone

Development: Create a branch for this issue or link a pull request.

Add your comment here...

Suggest Feature



What the hell is this???? #2

Open Myst9 opened this issue 21 minutes ago · 0 comments

Myst9 commented 21 minutes ago

"Notify" is a terrible name for a function regarding signal handlers. Curse you, Go core package authors.

Assignees: No one—assign yourself

Labels: None yet

Projects: None yet

Milestone: No milestone

What the fuck did you do to the code?

Suggestion : Can you clarify the implementation details of the code changes you made?

METHODOLOGY AND TECHNIQUES

We took one full week to search thoroughly across the Internet, especially Google Scholar, for pre-existing work in this area. We read several articles regarding the negative effects of harsh comments on sites like StackOverflow and how it mentally affected even some of the topmost Software Engineers on the planet.

We also read several research papers published in this area, especially by experts like **Jaydeb Sarker**. The comprehensive list of papers has been attached at the beginning of this document.

We concluded that although **they identify Toxic comments**, they ***don't actually attempt to do anything about it***. We intended to change that by :

A) **Preventing/reducing the sending of Toxic Comments** at the user level by showing the users a positive alternative version of every toxic comment they're about to post.

B) **Automating the reporting** of Individuals who post hateful content in these SE spaces. This remains in the ideation phase and may be fruitful in Release 2.

1. TOXICITY DETECTION MODEL

- **Dataset: ToxiCR dataset**(refer to Datasets Doc for more information)
- This is because we needed our detection model to be SE Specific.
- **19657 Comments from four popular FOSS communities (i.e., Android, Chromium OS, OpenStack, and LibreOffice). About 3757 were Toxic(19%)**
- We **preprocessed this data** and oversampled the data, so that the number of **Toxic and Non-Toxic comments is roughly equal**(so that the model doesn't just predict everything it sees as non-toxic).
- We fine-tuned the standard BERT model using this data.
- We used Keras for preprocessing comments before sending them to the model.
- We used the BERT Tokenizer from Transformers to convert Tokens into their index numbers in the BERT Vocabulary and PyTorch to run our model and get the output.

2. REPHRASING TOXIC COMMENTS

- We chose not to use existing models like Detox-BERT because they were too generic and were not SE Specific.
- Instead, we used Gemini to rephrase the comments. The advantage of this design decision is that tomorrow if we want to extend to something else, it's much easier with the Gemini Architecture to adapt our Backend accordingly.
- We passed on the requirement that the rephrasing must be SE-specific through the prompt.

TECH STACK AND PIPELINE

FrontEnd:

- Vanilla JS
- Chart JS

BackEnd:

- Flask (Python)
- WSGI Server
- Google.GenerativeAI
- Keras
- Transformers
- PyTorch

Pipeline:

GitHub DATA FETCH -> EXTENSION API REQUEST -> LOCAL BACKEND SERVER PRE PROCESSING, RUN MODEL, API RESPONSE -> EXTENSION MOUSE CLICK -> DISPLAY TOXICITY SCALE ABOVE THRESHOLD -> BLURS THE TEXT

PLANS FOR RELEASE 2:

According to Sridhar Chimalakonda Sir's advice, we are planning to do the following thing for Release 2:

Find out the Toxicity Score of a GitHub Repository using some metrics and data, possibly the comments, the issues and the way the contributors have responded to pull requests and then display it as a neat graphic at the root page of every repository.

We are also looking at the following possibilities for improvement:

- Change the model we fine-tuned on(We want to try Toxic-BERT instead of plain BERT)
- Change the visualization of the Toxicity Scale to circles or some other more appealing form.
- Improve the FrontEnd of the extension. The UI has a lot of room for improvement.
- Make the alternate suggestion automatic instead of after a button click.
- Try to blur the toxic comments automatically rather than waiting for the user to click on that particular mail and hover over that text.
- Maybe display the 6 toxicity features that were previously displayed in ToxiCheck as well in addition to our model's toxicity scale.
- Extend it to other websites and perhaps to other browsers.

INDIVIDUAL CONTRIBUTIONS:

- **CS21B013 BODALA SRI VARSHITHA:**
    - ***Preprocess*** the ***ToxiCR dataset*** to make it unbiased(roughly equal number of toxic to non-toxic comments)
    - ***Train/Fine-Tune*** the BERT model on the SE-specific Dataset and store it for further use.
    - ***Preprocess*** any ***input string***/text so that we can feed it directly to the model.
    - Fixed bug in server call from front end.
    - Created repository and added us as collaborators
    - ***Error Docs, WSR updation*** regarding her contributions.

- **CS21B043 S DHANANJAYAN:**
  - ***Searched Google Scholar for*** SE-Specific toxic comment ***datasets*** and tabularized my results in a doc with links to papers, repos and datasets.
  - ***Backend:*** Wrote the code for ***two API endpoints***(/predict and /suggest) and integrated Varshitha's model loading/preprocessing code here.
  - ***Integrated*** our code ***with ToxiCheck***: Ishaan Kulkarni bhaiya's tool.
  - ***Report*** making
  - ***Error Docs, WSR***: mentioned my contributions and summarized the discussion with Ishaan Kulkarni Bhaiya there.
- **CS21B045 SANTHOSI RM:**
  - SE-Specific ***detoxification*** of comments ***using Gemini API***: she wrote the ***code*** and the ***prompt***, generated the ***API key*** and ***integrated*** it with the ***Backend*** Code. Later, she modified the prompt to get better results.
  - ***Error Docs, WSR Doc***: she created the WSR doc and added her contributions
  - ***Slides preparation*** for the presentation
  - Helped Varshitha in the model training and Dhananjayan in the process of writing the Backend(by ***referring*** to the internet/***documentation*** for proper ***syntax***)
  - Helped Dhananjayan in the report making.
- **CS21B047 TADISETTI HEMA SRI:**
  - ***Error Docs, WSR Doc***: Added her contribution and summarized our discussions with Sridhar Chimalakonda sir.
  - ***Making the Readme*** for our tool with proper ***installation*** tips, ***visualizations*** of our ***architecture*** and breaking down our tool into simple and elegant sentences.
  - Helped Dhananjayan and Santhosi in making the Backend code.
  - ***Tested*** Dhananjayan's integration by downloading and running it. ***Reported errors*** which were ***subsequently fixed***.
  - Explored the possibility of ***automatic reporting*** to sites like GitHub and concluded that it was ***not possible to automate*** that task as of now.
- **CS21B053 VELAGALA SWETHA REDDY:**
  - ***Study of ToxiCheck***: During the initial days, she ***downloaded*** Ishaan Kulkarni's ToxiCheck, ***ran it, read their code*** and readme, ***understood*** it and explained it to the rest of us.

- ***Finding weak points*** in ToxiCheck for ***further improvements***: She noted a ***list of issues with ToxiCheck*** and came up with suggestions for improving it in Harmonize.

- ***Error Docs, WSR Doc***: Coordinated the meeting with Jahnavi mam and summarized the key points discussed in the meeting.

- ***Testing of Integration***: She also downloaded the integrated code, but ran into Pytorch installation issues, which, unfortunately, till date, haven't been resolved. In the process, her VM itself has gotten corrupted and isn't working anymore.
- ***Slides preparation*** for the presentation

*********THE END*********