Hatespeech on software engineering forums

Why Software Engineer wants to kill children

Paweł Walkowiak, Bartosz Walkowiak, Bartłomiej Góral
Wroclaw University of Science and Technology, Poland

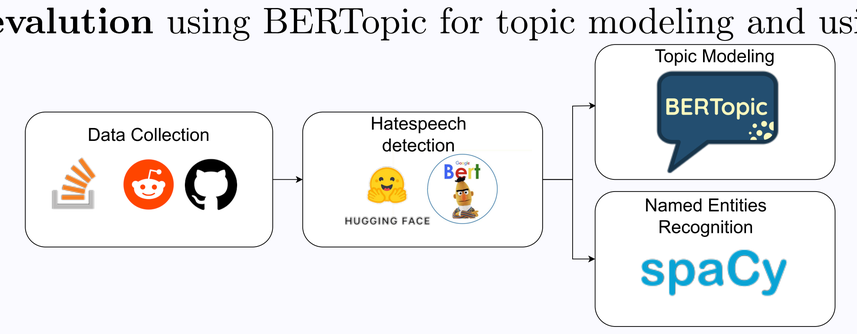
Department of Artificial Intelligence

The aim of this project

There is a belief that online forums for software engineering, like Stack Overflow and Reddit, are filled with individuals who spread negativity and that young, inexperienced engineers often face hateful comments when seeking help. Our goal is to investigate this belief and determine its validity, as well as identify any specific topics that tend to attract hate and offensive behavior.

Experiments setup and tools

Experiment pipeline consist of steps: **Datacollection**, which included using official Github api and Reddit api, as well as downloading public StackOverflow database dump. **Hatespeech detection** on datasets, using Large-Scale Hate Speech Detection model trained on data from digital media[1]. **Model results evalution** using BERTopic for topic modeling and using spaCy NER.



Datasets

Data for experiments comes from ther different SE forums; Stack Overflow, Reddit and Github issues.

Dataset	Lang	Size	Source
Stackoverflow	EN	673 836	SE forums db dump
Reddit	EN	138 379	Reddit API, praw library
Github	EN	13 180	Github API, PyGithub library

Examples

Correctly classified

Theoretically if you figured out cold fusion and you came out with it, would the government take it and/or kill you.

Incorrectly classified

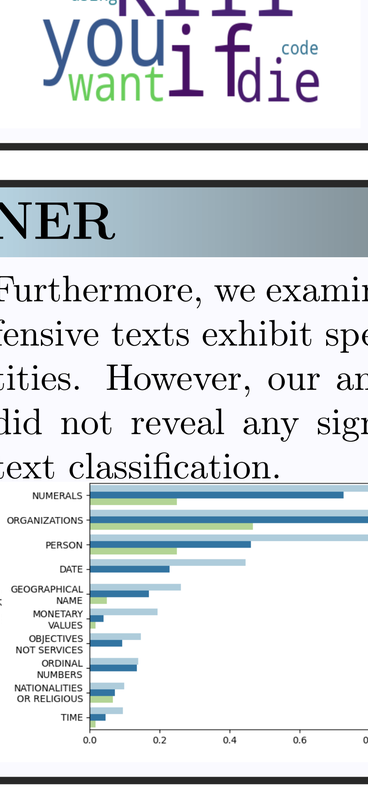
I wrote a program that forks some processes with fork(). I want to kill all child and the mother process if there is an error. If I use exit(EXIT_FAILURE) only the child process is killed.

On UNIX, you need to fork twice in a row and let the parent die.

Topic Modeling

To detect hate speech keywords, we utilized the BERTopic model. We ran topic modeling separately on hate and offensive texts. **Topic word-clouds from offensive posts:**

Topic word-clouds from hate posts:



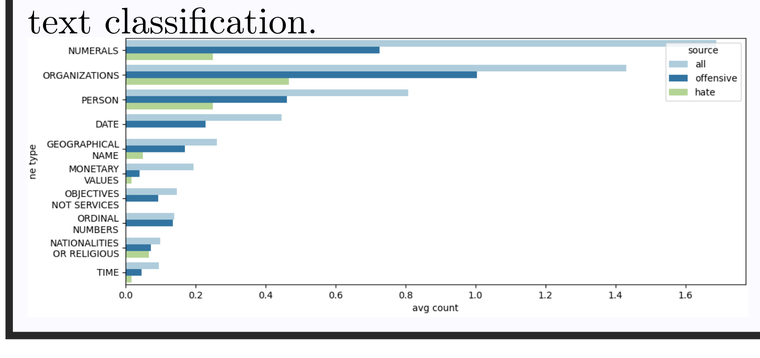
Results

As can be seen in tables above hatespeech models sometimes misclassifies examples that are using specific SE collocations. Overall percentage of hate and offensive speech is presented in table below.

Dataset	Hate[%]	Offensive[%]	Neutral[%]
Stackoverflow	0.006	0.389	99.605
Reddit	0.010	0.514	99.476
Reddit Comments	**0.019**	**2.302**	97.680
Github	0.000	0.046	**99.954**
Github Comments	0.000	0.273	99.727
Average	0.007	0.616	99.377

NER

Furthermore, we examined whether hate and offensive texts exhibit specific types of named entities. However, our analysis of named entities did not reveal any significant correlation with text classification.



Conclusions

- Belief that SE forums are full of haters is wrong, such forums indicate a low percentage of hate and offensive speech.
- The frequency of hate and offensive remarks is more prevalent in comments compared to the original posts, and ranges between SE formus, with 10 times more offensive statements on Reddit compared to Github.
- Experiments (model evaluation based on categorised data from SO administrators) showed that model trained on tweets dataset [1] can be used for hatespeech detection on SE forums.
- The hatespeech detection in utterances from software forums is non-trivial, due to the contextualisation of semantics evaporates into that used in specialist speech, e.g 'kill child [process]'. This problem occurs in approximately 20% of the samples classified as hateful/offensive.

Future Works

In future studies, we aim to explore the connection between specialist speech and detecting hate speech. Our experiments have shown that identifying hate and offensive content on SE forums is difficult. Creating an accurate model for this task would greatly benefit forum administrators.

Take away message

The majority of software engineers do not have any ill intentions towards you, even though they may occasionally terminate certain child and parent processes.

References

[1] Cagri Toraman, Furkan Şahinuç, and Eyup Halit Yilmaz. Large-scale hate speech detection with cross-domain transfer. In *Proceedings of the Language Resources and Evaluation Conference*, pages 2215–2225, Marseille, France, June 2022. European Language Resources Association.

Project done as part of the "Social Media Analysis" course for the Artificial Intelligence in 2023 under supervision of professor Tomasz Kajdanowicz.