Detoxplain: Explainable Toxicity Identification and Detoxification of Text

Cemre Biltekin
cemre.biltekin@tum.de

Onat Kaya
onat.kaya@tum.de

Selim Yagci
selim.yagci@tum.de

Abstract

Toxic language in online spheres is affecting individuals and communities in society. There has been an increasing demand to mitigate toxic content. In Ethical Artificial Intelligence (EAI) perspective, numerous research have been undertaken in detecting and alleviating the toxic content. In this paper, we aimed to combine several approaches together to implement a single pipeline for explainable toxicity identification and detoxification of text. We trained a transformer based language model for identification, adopted a local surrogate model for interpretability and leveraged the power of GPT3.5 model with zero-shot prompting for detoxification of text. Results showed that dataset size, annotator bias, bias in Large Language Models and evaluating content preservation are critical for this task. We concluded the paper presenting two possible future directions, reinforcement learning with human feedback for enhancing the quality of style transferring and re-ranking as a post-hoc debiasing method.

1 Introduction

Nowadays, the internet is used all around the world for many use cases, including but not limited to: news, entertainment, information sharing, and personal communication (Pew Research Center, 2021). In particular, social networks' impact on societies was better understood by people and communities after mass-scale political events like the anti-government protests in Egypt and Tunisia throughout the early 2010s (Kahn and Dennis, 2023) and the 2016 U.S. presidential elections (Pew Research Center, 2016).

Indeed, the mass adoption of the internet and its following platforms is proven by data: Statistical research indicates that in 2021 72% of the population in the United States were engaging at least with one social media platform, while this ratio had only been 5% in 2005 (Pew Research Center, 2021).

However, the increase in the employment these of relatively new technologies was observed to bring particular problems. In particular, the spread of toxic speech would be a crucial issue. This problem would come in many forms including sexual harassment, and threats of physical violence (Amnesty International, 2018; Anti-Defamation League, 2019). More specifically, it would be observed that toxic speech online would typically target the vulnerable individual's irremovable characteristics (e.g. race, ethnicity, disability status, sexual orientation) (Anti-Defamation League, 2019).

Furthermore, it would be noted that the toxic online spheres would lead to loss of confidence, anxiety, panic attacks, and sleep problems, among people who were subjected to it (Amnesty International, 2018; Anti-Defamation League, 2019).

As a result of these negative consequences stemming from toxicity online, a strong demand either from the government or tech companies to act upon this problem emerged (Anti-Defamation League, 2019). As a matter of fact, various governmental bodies have increased their attempts to further understand and mitigate toxic content online. For instance, the U.S. Senate (Gordon, 2017) and the European Union (Le Monde, 2022) are spending time and resources to better understand the dynamics of toxic speech and its impact on communities.

For the purpose of creating a safer online environment, our team has aimed to mitigate these malicious effects of toxic speech to some degree.

Consequently, we came up with a pipeline that would act like a "social media post texting tool assistant", and respectively detect the toxic piece of text, provide an explanation as to why it is deemed as toxic, and finally "detoxify" by neutralizing/debiasing the original input. For the implementation of the pipeline, rather than going with traditional machine learning methods, more recent encoder-decoder based transformer architectures and Large Language Models (LLM) were utilized.

Additionally, the proposed pipeline would be designed and applied in such a way that user feedback would be given in the process by explaining which parts of the texts are specifically deemed toxic.

The rest of the paper is outlined as follows: In Section 2, related literature in this field along with important keywords is going to be shared. Afterward, in Section 3, the methods and datasets which were used during this implementation of the pipeline will be explained. In Section 4, the outputs that were gained with this pipeline would be presented with a detailed evaluation detailing how many of the attempts at neutralizing the texts were successful, along with how much of the context was managed to be preserved. Subsequently, the takeaways that were gained are going to be shared in Section 5. In Section 6, the challenges that were faced during this process will be clarified. Finally, the possible future steps that could be taken to improve the project will be explained in Section 7.

2 Related Work

Toxicity It is an umbrella term including hate speech and abusive/harmful language. For the context of this project, we are referring to Jigsaw (Google), which defines toxicity as rude, disrespectful, or unreasonable language that is likely to make someone leave a discussion.

Toxicity Detection Supervised methods are very common especially fine-tuning transformer based language models in toxic speech detection tasks (Liu et al., 2019a; Safaya et al., 2020; Dai et al., 2020). Previous work on toxicity detection can be separated into two classes of identification approach: Span-level (Ghosh and Kumar, 2021; Pavlopoulos et al., 2021) and sentence-level identification. In this paper we are going to refer to the latter.

Interpretability of Toxicity Detection Despite the power of transformer-based models and transfer learning in detecting toxic speech, it can be difficult to understand the decision of the model. The feature of interpretability is an important part of the research in toxicity detection tasks, due to recent regulations maintaining the balance between free speech and safe communication (Xiang et al., 2021). Previous works in the literature show that deep learning interpretation techniques can also enhance content moderation (Gunturi et al., 2023).

Text Style Transfers The aim of Text Style Transferring (TST) is to control the style attributes of text while preserving the content (Jin et al., 2022). In this perspective, TST can serve as a helpful approach as it can be used to transfer toxic texts into normal speech. Setting the style attribute as toxicity, there are several approaches depending on the type of data that is available. Parallel data can be handled with supervised methods whereas nonparallel data requires self-supervision and larger models (Jin et al., 2022).

3 Methodology

The methodology is divided into three main tasks for our research purpose: (1) toxicity identification from textual input, (2) explainability of toxicity predictions, and (3) detoxification of the text.

3.1 Datasets

The Toxic Tweets Dataset (Iyer, 2021) was used as the non-parallel two-class textual dataset for the toxicity identification task. The tweets are short text posts that the users can create on Twitter, a social media platform. This dataset was chosen because it allows for supervised binary text classification, it was built with a mission to overcome the class imbalance in toxic text datasets, and it has sufficient data instances with about fifty-four thousand tweets (Iyer, 2021). The data is represented with its index, the textual content as a tweet, and its class label. The label "0" represents non-toxicity while label "1" represents toxicity. The tweets are in English.

3.2 Data Preprocessing

The Toxic Tweets Dataset (Iyer, 2021) was preprocessed appropriate to its domain attributes to be used for toxicity identification (Shah, 2020). First, the hashtags were extracted. The hashtags serve to indicate the topic of the tweet, however, there is a myriad of different hashtags generated and used every day. This makes it it harder to identify what the hashtags signal in an arbitrary tweets dataset where the topic is not shared across tweets, so it was disregarded. The remainder was cleaned from URLs, user mentions in the @user form, reserved words specific to Twitter, emojis, and smileys via *tweet-preprocessor* Python package (Özcan, 2020). Then, the cleaned-up tweets were further preprocessed with the removal of stop words, digits, and

punctuation. The preprocessed dataset was then used to train the toxicity identification model.

3.3 Models

3.3.1 Toxicity Identification

The toxicity identification task in our research corresponds to the classic natural language processing (NLP) task of text classification, where the task is to identify whether the given text is toxic or not (non-toxic). For this, we started by fine-tuning a fast and small transformer-based model DistilBERT (Sanh et al., 2020), a distilled version of BERT (Devlin et al., 2019) for this binary text classification task on the preprocessed Toxic Tweets Dataset (Iyer, 2021). The model's name is *distilbert-base-uncased*. This model was preferred over BERT to ease the computational load and the fine-tuning process. The dataset was split to 80% train set and %20 test set. The model was then trained for 2 epochs, batch size 2, learning rate of $3e^{-5}$.

The model was designed such way that would take the user-generated text as input and output the toxicity score, the probability of the input being classified as toxic.

3.3.2 Explainability

If the text is labeled as *toxic* by the DistilBERT model, Detoxplain moves on to show the user some explanations for the classification.

Explainability refers to the explanation to the user of how the model came up with the output or the result (Pluciński, 2022). Explainability is especially important in NLP models because it allows us to show possible social biases in the results (Pluciński, 2022) and assures trust in the created system (Bodria et al., 2020).

Since our system's goal is to make suggestions to the user, we needed to make sure the suggestions were informed, and the reasons behind the suggestions were indeed visible to the user. Thus, explainability in Detoxplain is divided into two layers: (1) feature importance, and (2) category assignment.

We acknowledged that the most intuitive approach for explainability is to show the users which words in their input have contributed to the toxicity classification. However, transformer-based models like DistilBERT are complex to explain. Thus, a *post hoc* explanation approach, Local Interpretable Agnostic-Model Explanations (LIME) (Ribeiro et al., 2016), was used for generating a heatmap for feature importance.

LIME treats the fine-tuned DistilBERT model as black-box as DistilBERT's predictions were passed to LIME. Then, LIME was trained as a local linear classifier with the user text input's variations where certain words were removed for certain samples as it treated the predictions of DistilBERT as ground truth (Pluciński, 2022). The number of features was chosen as 10 and the number of samples was chosen as 1000.

According to the weights of the trained model, the most important 10 features (words) were displayed and highlighted with different degrees of gradients. For our system, when the coefficient of a word is between -1 and 0, it means the word contributed to the non-toxic classification. On the other hand, if it shows a value between 0 and 1, it could be interpreted as the word contributing to the toxic classification.

This indication is important for the users to see which words have contributed to their text being classified as toxic, and which words are safe in this context. The users are allowed to try and rephrase the highlighted toxic words by LIME to decrease the toxicity score.

The second explainability step consists of identifying what type of toxicity the text carried within, and thus, the categorical analysis of toxicity. For this task, a RoBERTa model that had been pre-trained on Jigsaw Toxic Comments Dataset (Kaggle, 2018) was used (Liu et al., 2019b).

RoBERTa is the more developed version of another transformer-based model BERT (Devlin et al., 2019), where the necessary developments were applied such as pre-training the model for an increased amount of time with longer sequences and by employing larger batches (Liu et al., 2019b). This specific RoBERTa model trained on toxic corpus could be reached online (Hanu and Unitary team, 2020). This model online could additionally accomplish the toxicity identification task, however, only its toxicity categorization ability was in need since we had accomplished the task using fine-tuned DistilBERT for efficiency.

The toxicity categorization works by generating scores for each of the categories for the input text (except *All Good*, where the text is found to be non-toxic), and picking the category with the largest value. The available categories classifying each different type of toxicity could be listed as the following:

- All Good *(passed where there no toxicity was*

found by the fine-tuned DistilBERT model)

- Identity Attack
- Insult
- Sexual Explicit
- Threat
- Obscene

Thus, with these two explainability approaches, the users were given both the important word list that constitutes the reason for toxicity classification, and the toxicity category that highlights specifically the main reason and the subject behind toxicity.

3.3.3 Detoxification

Later on, depending on the toxic category given from the process "sentiment analysis" by our pre-trained RoBERTa model, a unique prompt via LLM was utilized to "detoxify" the original piece of text. It could be basically guessed that detoxification via prompting was applied to texts that were deemed as toxic, which corresponds to all the categories other than *All Good*.

The type of prompting could be identified as *"zero-shot prompting"*, where a prompt that is not part of the training data is requested to be processed, with no examples accompanying, unlike few-shot prompting (Tam, 2023).

With this approach, we have aimed to provide extra specificity to the prompt generated and the output given. This process could be named as "Categorical Prompting" since it summarizes the functionality of what had been accomplished.

Due to its API being accessible to developers and relatively improved nature, the series of models named GPT-3.5, developed and deployed by the Artificial Intelligence company OpenAI (OpenAI, a), was preferred as the LLM that would be employed.

More specifically, the model *text-davinci-003* was chosen, an improvement on other GPT-3.5 models *code-davinci-002* and *text-davinci-002* (OpenAI, c).

As for hyperparameters in the GPT-3.5 model, the following values were altered while the rest were kept as default:

- max_tokens = 1024 *(lowered from the default value of infinity to limit number of maximum tokens generated during prompting)* (OpenAI, b)
- temperature = 0.5 *(lowered from the default value of 1.0 to further decrease the randomness of the output generated)* (OpenAI, b)

The automatized nature of prompting was made possible by integrating the OpenAI API into our Google Colab Notebook. The necessary API key was provided by the TUM Social Computing Research group.

3.3.4 Additional Experiments

Besides the methods explained in the section, several other settings were investigated before deciding the final components of the pipeline.

Span-level Toxicity Identification Inspired by the Toxic Spans Detection task (Pavlopoulos et al., 2021), we implemented a span-level toxicity identification using the data from the SemEval task. The dataset was labeled with the locations of the toxic spans in the text. We had used the pre-trained RoBERTa model and fine-tuned it on predicting the index of the spans in the text. Empirical results showed that this approach is prone to swear word biases, and weakly performs on longer spans or toxicity in larger contexts.

Prompt-based Fine-tuning of LLMs For the detoxification task, we crafted fixed prompt with ParaDetox dataset (Logacheva et al., 2022), in the form of *'Instead of saying* [toxic text from the dataset], *i can express that in other way like* [nontoxic gold reference]', to train the GPT-NEO model (Black et al., 2022). We used the model version with 125m parameters and trained it for one epoch with over 20000 training data for predicting the next nontoxic sentences. The empirical results were not sufficient.

Few-shot Prompting We used the same model settings in Section 3.3.3 with few-shot in-context learning approach. Crafted prompts wrapping the randomly chosen parallel data from ParaDetox dataset are used. We set the number of few-shot parameters at 15. An example output could be seen in Table 2. Ultimately, we have decided on the zero-shot version with category information to make it generalize better by mitigating the need for parallel labeled data.

4 Results

In this section we provide example outputs of our system, our evaluation metrics and analysis of the performance with respect to the metrics.

4.1 Evaluation Metrics

For the evaluation, different metrics were applied for each part of the project.

In the identification model, we evaluated the model accuracy of predicting the binary classification task of the output label being toxic or not. We employed a test set with the same distribution as the training set. By virtue of the balanced nature of the dataset, in terms of the number of classes, accuracy captures a meaningful performance evaluation.

In the detoxification part, evaluation of the style-transferred output required more than correct target checking. We referred to commonly used three criteria that are generally used in the evaluation of text style transferring tasks: style strength, content preservation, and fluency (Jin et al., 2022). For our task, we mainly decided to focused on the first two. We used our classifier model from the first part of the project to infer the style of generated output, as toxic or non-toxic and compared them with the gold references.

Maintaining the semantics of the text was as much important as style strength in our overall objective. For the content preservation assessment, we were required to have parallel data. We used the ParaDetox dataset (Logacheva et al., 2022), which is a parallel corpus for the detoxification task of English texts annotated by a crowd-sourced platform. We calculated similarity scores between references in the dataset and generated text of our model using BLEURT (Sellam et al., 2020). BLEURT generated scores which are between 0 and 1, 0 indicating random output and 1 a perfect similarity. Similar to other automatic metrics such as BLEU (Papineni et al., 2002) and ROUGE (Lin, 2004), the scores are expected to be noisy.

For a better interpretation, we decided to calculate the distribution and the average across the texts in the test set, and we compared the distribution with WMT Metrics Shared Task (Freitag et al., 2022) results.

4.2 Performance Analysis

Table 1 shows the performance of the identification model. The results align with the other transformer-based models with the advantage of fewer parameters in the DistilBERT model. Inference time is also short. We get high precision and recall at the same time as opposed to similar tasks' performances, due to a balanced dataset setting. Since a low number of false positives (i.e. falsely flagged as toxic) was the main objective, performance was adequate.

Model	Acc	F1
BERT-FT (Xiang et al., 2021)	0.85	0.83
DistilBERT	0.92	0.91

Table 1: Evaluation of identification model in comparison to a baseline BERT model fine-tuned on toxic language classification task on OLID (Zampieri et al., 2019) dataset

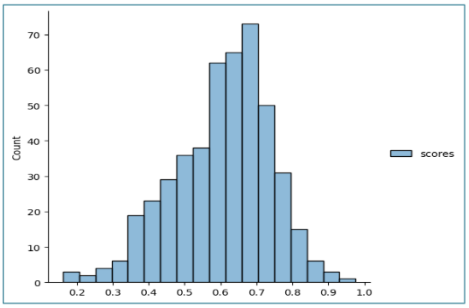

Figure 1: BLEURT scores distribution of detoxified outputs

The performance of the explainability features of the system relies on our identification model for the visual explanation with LIME and pre-trained RoBERTa model that is described in section 3.3.2. Example visual explanation output could be seen in Table 3.

For the performance of detoxification:

- We evaluated the style strength of the detoxified text samples with the accuracy of 0.61, i.e. the number of correctly labeled samples divided by the number of all generated non-toxic text for test.
- Figure 1 shows the distribution of BLEURT scores. The results aligned with our baseline in Figure 2 with respect to densely distributed around 0.7 score.

The results show that our system's identification model does not perform well in assessing our generated nontoxic outputs. Nevertheless, it also indicates LLMs can produce different toxic text while

Table 2: Example of generated outputs from different experiments

Original text	*millions of people paid money to watch this garbage*
Parallel reference	*millions of people paid money to watch this nonsense*
Zero-shot with category	*Millions of people voluntarily spent money to watch this*
Few-shot prompting	*Millions of people spent their resources to engage with this content*
Prompt fine-tuning	*Millions willingly paid to watch*

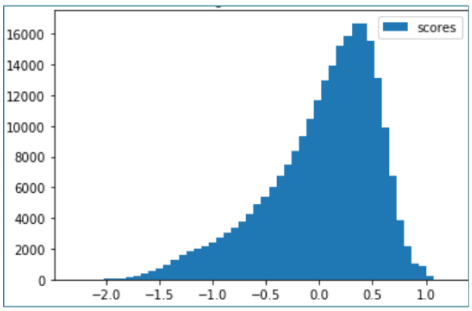

Figure 2: Disstribution of WMT Shared Task BLEURT scores

sh** whatever everyone ban cuz me talks

Table 3: Example LIME output for input text *i'm gonna ban everyone talks sh** about me, cuz i do whatever i want*

removing the toxic style. On the other hand, our system generated outputs that preserve the original content. The similarity between example-generated text and human reference can be seen in Table 2.

5 Conclusion

Overall our system promises a complete pipeline for dealing with toxic speech. Fine-tuning transformer-based models, even with smaller parameter sizes, is computation and resource-efficient for toxic speech detection tasks, which do not require leveraging the power of LLMs when there exists enough data.

For the end-user experience of such a system, LIME provides token-level explanations by highlighting important features affecting class decisions. Moreover, fine-grained text classification models can be utilized for enhancing interpretability even as hard-coded toxicity category information.

Using this information also in the detoxification phase, our system showed that zero-shot prompting with GPT3 and larger model as such gets improved similar to the Chain-of-Thought approach (Wei et al., 2023).

Nonetheless, we identified the reasons for performance issues and challenges in building such a system:

- Toxic speech classification is very ambiguous in real-world situations (Fortuna et al., 2020) and it heavily relies on the source, context, and target group of the speech.
- Identity, dialect and swear word biases in human annotated datasets are very challenging in supervised settings (Zhou et al., 2021).
- LLMs are prone to generate racist, sexist, or toxic speech inherited from large-scale training data (Gehman et al., 2020) and this stands still a factor for detoxification tasks despite the power of the model.
- The choice of the training dataset is challenging in terms of size as well. Larger and uncontrolled datasets have the tendency to inherit the bias, whereas smaller and controlled training data is prone to distribution shifts and is not generalizable.
- In text-style transferring tasks and specifically in our detoxification task one of the challenges is evaluating the generated text's quality. The content preservation criteria are an important part of such assessment and similarity scores are not yet expressive enough other than being a good approximation.

6 Limitations

There were certain constraints that have impacted design and methodology:

- Lack of high computational power and difficulties in steady testing in our collaborative development environment affected the design

choices. Fine-tuning LLMs with large-size parameters was either time-consuming or not possible, which was critical to reach better performances.

- There was also a time constraint preventing us from doing more advanced data preprocessing. Further analysis of special tokens like hashtags in the dataset could possibly improve toxicity identification.

7 Future Work

Although a considerable amount of steps were taken during this project, further improvements could be applied in the future to reach the objectives much closer.

One of the complementary works is utilizing a feedback interaction from the end-user and effectively integrating a Reinforcement Learning model into the system. Another way of optimizing our proposed system will be adding a constant data augmentation layer using the style-transferred generated text. A more general direction for research is to overcome bias issues. A post-hoc debiasing method to mitigate bias in generated text by re-ranking seems to be promising considering its efficiency.

References

Amnesty International. 2018. Toxic twitter - the psychological harms of violence and abuse against women online.

Anti-Defamation League. 2019. Online hate and harassment: The american experience.

Sidney Black, Stella Biderman, Eric Hallahan, Quentin Anthony, Leo Gao, Laurence Golding, Horace He, Connor Leahy, Kyle McDonell, Jason Phang, Michael Pieler, Usvsn Sai Prashanth, Shivanshu Purohit, Laria Reynolds, Jonathan Tow, Ben Wang, and Samuel Weinbach. 2022. GPT-NeoX-20B: An open-source autoregressive language model. In *Proceedings of BigScience Episode #5 – Workshop on Challenges & Perspectives in Creating Large Language Models*, pages 95–136, virtual+Dublin. Association for Computational Linguistics.

Francesco Bodria, André Panisson, Alan Perotti, and Simone Piaggesi. 2020. Explainability methods for natural language processing: Applications to sentiment analysis. In *Sistemi Evoluti per Basi di Dati*.

Wenliang Dai, Tiezheng Yu, Zihan Liu, and Pascale Fung. 2020. Kungfupanda at SemEval-2020 task 12: BERT-based multi-TaskLearning for offensive language detection. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, Barcelona (online). International Committee for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding.

Paula Fortuna, Juan Soler, and Leo Wanner. 2020. Toxic, hateful, offensive or abusive? what are we really classifying? an empirical analysis of hate speech datasets. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 6786–6794, Marseille, France. European Language Resources Association.

Markus Freitag, Ricardo Rei, Nitika Mathur, Chi-kiu Lo, Craig Stewart, Eleftherios Avramidis, Tom Kocmi, George Foster, Alon Lavie, and André F. T. Martins. 2022. Results of WMT22 metrics shared task: Stop using BLEU – neural metrics are better and more robust. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 46–68, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Samuel Gehman, Suchin Gururangan, Maarten Sap, Yejin Choi, and Noah A. Smith. 2020. Realtoxicityprompts: Evaluating neural toxic degeneration in language models.

Sreyan Ghosh and Sonal Kumar. 2021. Cisco at SemEval-2021 task 5: What's toxic?: Leveraging transformers for multiple toxic span extraction from online comments. In *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)*, Online. Association for Computational Linguistics.

Google. Google's safer internet unit.

Marcy Gordon. 2017. Ceos of 3 tech giants to testify at oct. 28 senate hearing. *The Associated Press*.

Uma Gunturi, Xiaohan Ding, and Eugenia H. Rho. 2023. Toxvis: Enabling interpretability of implicit vs. explicit toxicity detection models with interactive visualization.

Laura Hanu and Unitary team. 2020. Detoxify. Github. https://github.com/unitaryai/detoxify.

Ashwin U. Iyer. 2021. Toxic tweets dataset.

Di Jin, Zhijing Jin, Zhiting Hu, Olga Vechtomova, and Rada Mihalcea. 2022. Deep learning for text style transfer: A survey. *Computational Linguistics*, 48(1):155–205.

Kaggle. 2018. Toxic comment classification challenge.

Robert Kahn and Michael Aaron Dennis. 2023. Internet — Encyclopedia Britannica.

Le Monde. 2022. Eu law to protect internet users and their "fundamental rights online" against hate speech, disinformation.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Ping Liu, Wen Li, and Liang Zou. 2019a. NULI at SemEval-2019 task 6: Transfer learning for offensive language detection using bidirectional transformers. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019b. Roberta: A robustly optimized bert pretraining approach.

Varvara Logacheva, Daryna Dementieva, Sergey Ustyantsev, Daniil Moskovskiy, David Dale, Irina Krotova, Nikita Semenov, and Alexander Panchenko. 2022. ParaDetox: Detoxification with parallel data. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6804–6818, Dublin, Ireland. Association for Computational Linguistics.

OpenAI. a. About.

OpenAI. b. Create chat completion.

OpenAI. c. Model index for researchers.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

John Pavlopoulos, Jeffrey Sorensen, Léo Laugier, and Ion Androutsopoulos. 2021. SemEval-2021 task 5: Toxic spans detection. In *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)*, pages 59–69, Online. Association for Computational Linguistics.

Pew Research Center. 2016. Candidates differ in their use of social media to connect with the public.

Pew Research Center. 2021. Social media fact sheet.

Kamil Pluciński. 2022. Overview of explainable ai methods in nlp.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. "why should i trust you?": Explaining the predictions of any classifier.

Ali Safaya, Moutasem Abdullatif, and Deniz Yuret. 2020. KUISAIL at SemEval-2020 task 12: BERT-CNN for offensive speech identification in social media. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, Barcelona (online). International Committee for Computational Linguistics.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2020. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter.

Thibault Sellam, Dipanjan Das, and Ankur P. Parikh. 2020. Bleurt: Learning robust metrics for text generation.

Parthvi Shah. 2020. Basic tweet preprocessing in python.

Adrian Tam. 2023. What are zero-shot prompting and few-shot prompting.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. 2023. Chain-of-thought prompting elicits reasoning in large language models.

Tong Xiang, Sean MacAvaney, Eugene Yang, and Nazli Goharian. 2021. ToxCCIn: Toxic content classification with interpretability. In *Proceedings of the Eleventh Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 1–12, Online. Association for Computational Linguistics.

Marcos Zampieri, Shervin Malmasi, Preslav Nakov, Sara Rosenthal, Noura Farra, and Ritesh Kumar. 2019. Predicting the type and target of offensive posts in social media.

Xuhui Zhou, Maarten Sap, Swabha Swayamdipta, Yejin Choi, and Noah Smith. 2021. Challenges in automated debiasing for toxic language detection. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 3143–3155, Online. Association for Computational Linguistics.

Said Özcan. 2020. Tweet-preprocessor.