Biases in Toxicity Detection Models

Gianluca Nogara[1,†], Francesco Pierri[2,†], Stefano Cresci[3], Luca Luceri[4], Petter Törnberg[5] and Silvia Giordano[1]

University of Applied Sciences and Arts of Southern Switzerland, Switzerland
Politecnico di Milano, Italy
IIT-CNR, Italy
USC Information Sciences Institute, USA
University of Amsterdam, Netherlands

Keywords
API, bias, language models, toxicity,

Introduction

Online abuse has become increasingly prevalent in recent years, affecting approximately 40% of U.S. adults according to self-reported data [1]. This rise in harmful interactions – commonly labeled as online "toxicity" – particularly on social media platforms, has raised concerns among researchers and the general public. A primary tool for detecting such toxicity is the Perspective API, developed by the Jigsaw unit of Google [2], specifically designed to mitigate toxicity and promote healthy online dialogue. Widely utilized within both academic research and content moderation efforts, the Perspective API boasts over 1,400 mentions on Google Scholar as of January 2024.

Perspective API employs supervised learning, leveraging a vast dataset of millions of comments sourced from diverse online platforms spanning over 20 languages, including forums like Wikipedia and The New York Times. It defines "toxic" messages as those containing "rude, disrespectful, or unreasonable language likely to disrupt discussions" [3, 4]. In a 2019 SemEval Task, the Perspective API demonstrated superior performance compared to other transformer-based models in hate speech detection [5]. However, recent studies have noted disparities between its scores and human labels [4]. The toxicity score, ranging from 0 to 1, lacks absolute meaning, and typically a threshold (usually 0.5-0.7) is set on the toxicity score, above which content is deemed toxic [6].

Technologies like the Perspective API can foster a safer and more respectful online environment, but this is highly dependent on their ability to maintain accuracy and impartiality across a spectrum of languages and cultural settings. This extended abstract discusses the results of our paper [7], where we aim to address the following research question: **Is Perspective API biased towards the German language?**

To achieve this goal, we utilize two extensive datasets comprising millions of multilingual Twitter conversations and thousands of random Wikipedia summaries. In each dataset, we apply the Perspective API to all available texts, focusing on the attribute "Toxicity", which is widely utilized in the literature due to its general applicability. Specifically, in September 2022,

SEBD 2024: 32nd Symposium on Advanced Database Systems, June 23-26, 2024, Villasimius, Sardinia, Italy
[†]These authors contributed equally.

we utilized the Perspective API to obtain toxicity scores for Dataset 1, consisting of random tweets shared in European countries. Subsequently, between September and December 2023, we employed the Perspective API to label Datasets 2 and 3. Dataset 2 comprises COVID-19-related tweets in German and Italian languages, while Dataset 3 encompasses Wikipedia summaries in German, English, and Italian languages.

Results

Our initial analysis focuses on examining the distribution of toxicity scores for tweets originating from German-speaking countries, namely Austria, Germany, and Switzerland, in comparison to those from other countries (Dataset 1). As illustrated in Figure 1, tweets originating from German-speaking countries, predominantly in the German language, exhibit significantly higher toxicity levels (Kruskal-Wallis $P < .001$) compared to tweets from other regions, with a median toxicity score of 0.075 versus 0.023, respectively. The toxicity distribution of tweets from other countries closely resembles a smooth exponential distribution, a characteristic often observed in various social media phenomena. Conversely, the distribution of toxicity scores for tweets from German-speaking countries appears spiky, indicating the potential presence of classification errors or artifacts.

To assess the practical implications of the identified biases in toxicity scores, we turn to Dataset 2, consisting of tweets related to COVID-19 vaccines. Given the critical importance of accurate content moderation in disseminating reliable COVID-19 information [8], this dataset provides a relevant context for our investigation. To investigate whether the higher toxicity scores assigned to German texts stem from inherently more toxic content, we translate German tweets into English. Subsequently, we utilize the Perspective API to compute toxicity scores for the English-translated texts. We employ Argos Translate to obtain English translations of our tweets. Figure 2 depicts the distribution of toxicity scores for all German tweets (solid line) and their English translation equivalents (dashed line). Notably, we observe that the tweets translated into English exhibit significantly lower toxicity, with their distribution resembling that of generic tweets in non-German languages, as shown in Figure 1. This finding suggests that the observed higher toxicity scores in German texts may not necessarily reflect the content's inherent toxicity but could be influenced by language-specific biases in the Perspective API's classification.

To evaluate the impact of this bias, we examine how moderation strategies aimed at reducing online toxicity may influence decision-making processes, particularly when moderators or automated systems rely on toxicity scores from the Perspective API to supervise and regulate discussions on social media platforms [9]. Specifically, we investigate how different toxicity thresholds used for removing tweets or users could affect conversations regarding COVID-19 vaccines in German compared to their translation in English. In the top panel of Figure 4, we simulate the removal of all tweets exceeding a given toxicity threshold and depict the percentage of additional tweets removed when considering German toxicity scores compared to English scores. For instance, adopting the threshold value of 0.7 recommended by the Perspective API, ten times more German tweets would be removed compared to English tweets. Similarly, this trend extends to the removal of German users, as illustrated in the bottom panel of the figure. The median increase in percentage points is +429.94 for tweets and +409.84 for users, indicating that, on average, more than four times the number of German tweets and users would be removed compared to English.

To confirm that the identified biases are not limited to Twitter messages, we extend our analysis to include the distribution of toxicity scores when classifying a random sample of summaries from Wikipedia (Dataset 3). In this dataset, each summary is provided in both English and German languages. As depicted in Figure 3, we find that German summaries demonstrate significantly higher toxicity scores compared to those in English (Kruskal-Wallis $P < .001$). Additionally, similar to previous distributions observed, such as Figure 1, we observe peaks in the distribution of German scores. These findings suggest that the observed biases persist across different types of text sources, reinforcing the need for careful consideration and mitigation of language-specific biases in toxicity classification algorithms.

Discussion

The results presented in this paper underscore the inherent risks associated with the growing reliance of researchers on proprietary models for data analysis and artificial intelligence. Closed and proprietary APIs provided by major technology companies offer researchers convenient access to vast datasets and sophisticated analytical methods that might otherwise be difficult to obtain. This is invaluable for studying complex social phenomena, trends, and behaviors in the digital age. However, this increasing dependence on closed models also raises significant concerns. The opaque nature of these proprietary systems means that researchers are unable to scrutinize or comprehend the inner workings of the algorithms, making it difficult to identify potential biases and limitations. Consequently, there is a substantial risk that systemic biases may become embedded and perpetuated through the utilization of these tools. Moreover, the lack of transparency can inadvertently influence the direction and conclusions of research studies, as researchers are constrained by the methodologies predefined by these tools. While proprietary APIs have undoubtedly opened up new avenues for exploration in the social sciences, it is imperative to adopt a cautious and critical approach to ensure the integrity and reproducibility of research. This entails conducting thorough assessments and validations of the outputs generated by these models, rather than relying solely on the assurances provided by technology firms [10].

Several limitations are present in our study. Firstly, our investigation relies on data from Twitter, which may not be the most widely utilized social platform in many countries, thus potentially limiting the representativeness of our findings to the broader population. However, our robustness analyses using Wikipedia data indicate that the identified issues are not contingent upon user demographics or the specific context or topic of conversation. Secondly, our comparison of toxicity scores primarily focuses on German in contrast to other Western-based languages. We acknowledge that biases in toxicity scores may also be present in non-Western languages; however, this aspect remains unexplored in our study. Lastly, we did not assess whether similar biases manifest in other language model-based (LLM-based) models for toxicity detection. Exploring the extent of these biases across different models could provide valuable insights into the generalizability of our findings and the broader implications for toxicity detection methodologies.

Our work carries significant ethical implications, particularly for future research on online toxicity and the design of policies for online platforms. The identification of intrinsic biases within the multilingual Perspective API highlights the risk of inadvertently amplifying or suppressing certain voices in online spaces. Specifically, the higher levels of toxicity assigned to German content compared to other languages raise concerns about the fairness and equity

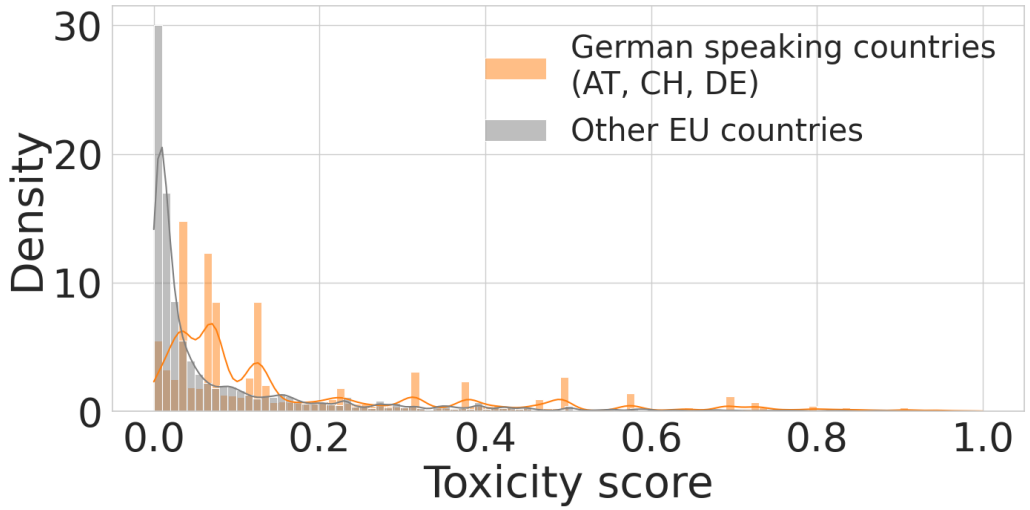

Figure 1: Distribution of toxicity scores for tweets shared in German-speaking countries (Austria, Switzerland, and Germany) versus those in other EU countries, from Dataset 1.

of current content moderation processes. We emphasize the importance of transparency, accountability, and continuous refinement to ensure that these technologies align with principles of impartiality and cultural sensitivity. By addressing these ethical concerns, we can mitigate the inadvertent amplification of biases in online discourse and promote a more equitable and inclusive online environment for all users.

References

[1] E. A. Vogels, The state of online harassment, Pew Research Center 13 (2021) 625.

[2] A. Lees, V. Q. Tran, Y. Tay, J. Sorensen, J. Gupta, D. Metzler, L. Vasserman, A new generation of Perspective API: Efficient multilingual character-level transformers, in: The 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD'22), 2022, pp. 3197–3207.

[3] S. Gehman, S. Gururangan, M. Sap, Y. Choi, N. A. Smith, RealToxicityPrompts: Evaluating neural toxic degeneration in language models, in: The 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP'20), 2020, pp. 3356–3369.

[4] J. Welbl, A. Glaese, J. Uesato, S. Dathathri, J. Mellor, L. A. Hendricks, K. Anderson, P. Kohli, B. Coppin, P.-S. Huang, Challenges in detoxifying language models, in: The 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP'21), 2021, pp. 2447–2469.

[5] J. Pavlopoulos, N. Thain, L. Dixon, I. Androutsopoulos, ConvAI at SemEval-2019 task 6: Offensive language identification and categorization with Perspective and BERT, in:

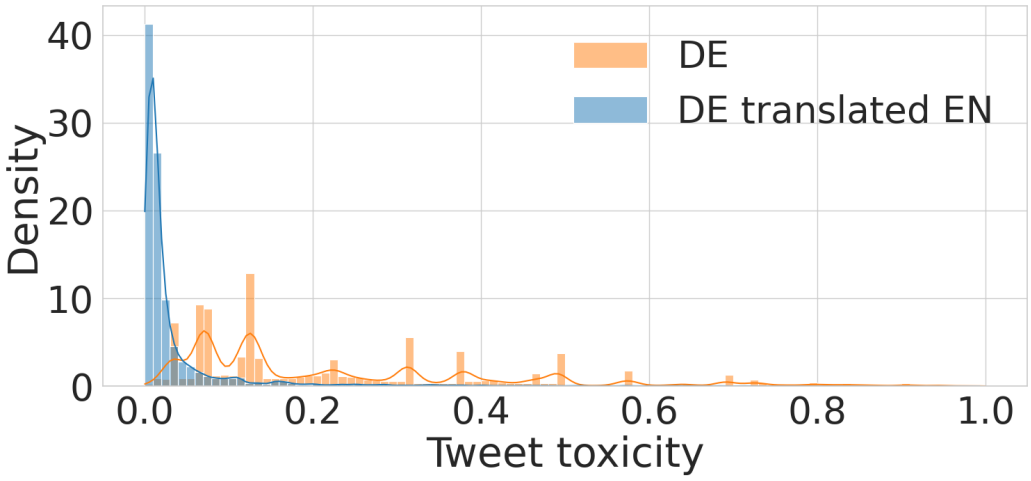

Figure 2: Distribution of toxicity scores for German COVID-19 vaccine-related tweets and for their English translation, from Dataset 2.

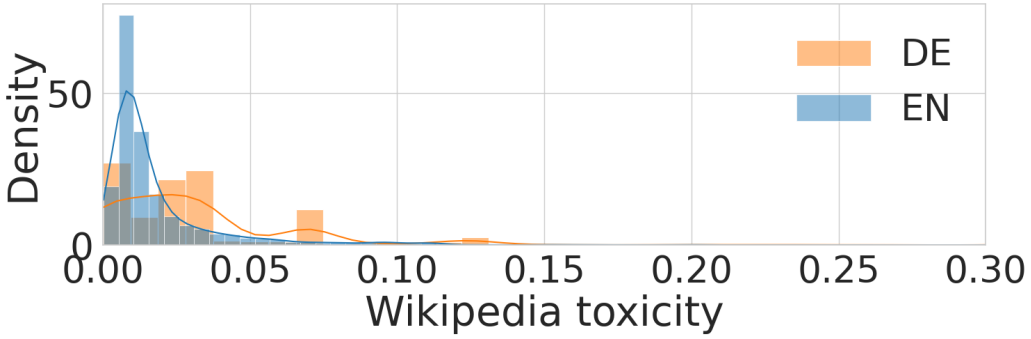

Figure 3: Distribution of toxicity for 12000 random Wikipedia page summaries in English and German language from Dataset 3.

Proceedings of the 13th International Workshop on Semantic Evaluation (SemEval'19), 2019, pp. 571–576.

[6] Y. Hua, M. Naaman, T. Ristenpart, Characterizing Twitter users who engage in adversarial interactions against political candidates, in: The 2020 CHI Conference on Human Factors in Computing Systems (CHI'20), 2020, pp. 1–13.

[7] G. Nogara, F. Pierri, S. Cresci, L. Luceri, P. Törnberg, S. Giordano, Toxic bias: Perspective api misreads german as more toxic, arXiv preprint arXiv:2312.12651 (2023).

[8] E. Ferrara, S. Cresci, L. Luceri, Misinformation, manipulation, and abuse on social media in the era of COVID-19, Journal of Computational Social Science 3 (2020) 271–277.

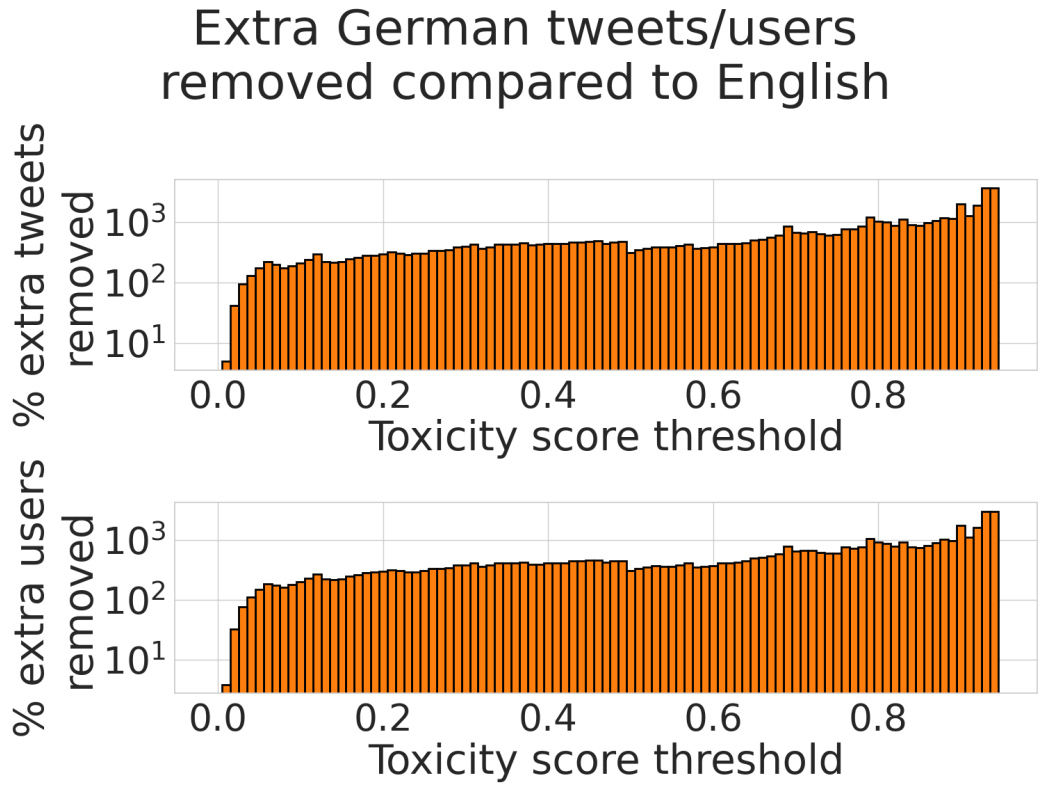

Figure 4: Proportion of extra tweets (top) and users (bottom) removed in German compared to English for different choices of toxicity score as a threshold, i.e., tweets/users above a given score would be removed.

[9] B. Rieder, Y. Skop, The fabrics of machine moderation: Studying the technical, normative, and organizational structure of Perspective API, Big Data & Society 8 (2021).

[10] L. A. Pozzobon, B. Ermis, P. Lewis, S. Hooker, On the challenges of using black-box apis for toxicity evaluation in research, in: ICLR 2023 Workshop on Trustworthy and Reliable Large-Scale Machine Learning Models, 2023.