Evaluating Hate Speech Detection to Unseen Target Groups

Alexandre Negretti[1] and Marcos M. Raimundo[1]

[1]Instituto de Computação, Universidade de Campinas (UNICAMP), Campinas – SP – Brazil
a233609@dac.unicamp.br, mrai@unicamp.br

November 26, 2024

Abstract. *LLMs trained to detect hate speech have a significant challenge on identifying hate speech directed toward new or less common target groups. This happens because the models are primarily trained on data focused on more prevalent forms of hate, targeting groups that have historically been subjected to hate speech. Not only the way of defamation evolves through time, but new targets may emerge, presenting forms of hate that were previously non-existent in datasets. This work presents analyses of the influence of targeted groups on model prediction. We evaluate training strategies that address target group bias in hate speech detectors. Lastly, we present a novel dataset composed of text posts from Twitter regarding the 2022 Russia-Ukraine war.*

1. Introduction

The ways of directing hate and creating harmful content change constantly; depending on the social network, discourse characteristics such as the intensity of the harassment, the usage of words, and the groups targeted may vary immensely [Maarouf et al. 2022]. Regarding target groups, a leading concern is fairness and how models can equally classify hate speech varying the target groups without perpetuating stereotypes related to them [Kovatchev et al. 2022]. That is a great challenge since data is not equally distributed between groups, mainly because hateful behavior has a temporal attribute, and it follows trends of internet nature [Ayo F 2020, Velankar et al. 2022]. Another source of bias is regarding annotation. Hate speech sounds differently from person to person. It may be seen as more or less harmful depending on how close the individual is to the targeted audience of the hate, which means that datasets carry annotator bias related to their world perception [Cai et al. 2022].

2. Slavic Hate Dataset

Anti-Slavism is a racist and xenophobic movement that was inflamed and reached its peak during World War II (1939-1945), after the invasion of Ukraine and Belarus, with the narrative that natives from these countries were subhumans and must be exterminated. Even after the end of WWII, Slavic hatred continued latent. During the Cold War (1947-1991), most North American narratives were against Russia and the Soviet Union, reinforcing the discourse of 'us' against 'them'. Anti-Slavic sentiment had never ended, and it existed at various moments of high tension. Our proposition with this dataset was to use the 2022 Russia-Ukraine war as a thermometer and examine Twitter posts to extract potentially hateful examples of anti-Slavic sentiment and Slavic Hate.

The raw data from the dataset was collected using a Python library called snscrape[1]. This library allows to query tweets based on keywords in a specified time range. For this dataset we used the keywords: *russia*, *ukraine*, *putin*, *zelensky*, and *war*. The collection period spanned 2022/01/01 to 2023/04/30 to capture sentiment before and after the conflict's start date (2022/02/24). Since the war had not been finished until the publication of this work, an end date close to the day of the collection was chosen.

[1] Available at github.com/JustAnotherArchivist/snscrape

Date	Raw Text	Text Treated	Tweet URL	Tweet ID
2022-06-04 23:59:11	Fuck humiliation! I just wonder why the west hasn't taken out all of russia's nukes, yet!	fuck humiliation! i just wonder why the west hasn't taken out all of russia's nukes, yet!	https://twitter.com/ IAmKarlCensored/status/ 15332369***	15332369***
2023-02-10 23:58:03	Zelensky is a nazi and Rusia already know it. Rusia is winning and Orban is right	zelensky is a nazi and rusia already know it. rusia is winning and orban is right	https://twitter.com/ FEscman/status/ 16241960***	16241960***

Table 1. Examples of data from Slavic Hate dataset

For example, Table 1 shows tweets that compose the dataset from different dates. It was chosen to show instances with low insulting and slurs since it is sensitive content and may offend potential readers. Snscrape has a search limit that stops retrieving data when it cannot refresh the Twitter search page. Since it can take too long to stop, and for better balancing the number of examples for each keyword, we set an upper threshold of 500 tweets per keyword per day in the date range. After the collection has finished, the entire dataset has a total of 1,021,221 tweets.

3. Preliminary results

3.1. Slavic Hate Dataset Exploration

After collecting data from our dataset, we evaluated each tweet using a pre-trained classifier model. We choose the model shared in the work of [Vidgen et al. 2021]. Even with the premise proposed by our work that models may not perform well on unseen target groups, we intended to know how many tweets from our dataset may have hateful content, even without fine-tuning. The model was used on treated tweets without user citations, hashtags, and all lowercase. By processing the entire dataset, we got 963,026 (94%) non-hate tweets and 58,195 (6%) hate tweets from 1,021,221 tweets. The number of hate tweets may seem low in percentage, but compared to relevant datasets such as HateXplain [Mathew et al. 2020] (19,229 rows) and ToxiGen [Hartvigsen et al. 2022] (260,851 rows), we believe that we have a relevant amount of data for our study.

Since ToxiGen [Hartvigsen et al. 2022] has a great standard of dataset with target group segmentation, we decided to create a subset similar to theirs from our dataset. For each target group, ToxiGen data has around 20,000 rows, of which half are hate and the others non-hate. We created a subset from our dataset of 10,000 non-hate rows and another 10,000 hate rows, which were predicted as hate in the process described above.

3.2. Masked Data

One of our first investigations in this research was to verify the effect of words related to target groups in models' inference. We decided to use ToxiGen data since it has the biggest diversity of target groups, and to choose which word to analyse for each sentence, we decided to look into the actor and the object of the phrase, which usually carries the identity terms. Since the data set is too large, we performed a programmatic approach to tag the words, using spaCy [Montani et al. 2023], a library used to extract Linguistic Features from a text, such as the subject and object of the phrase. It is a handful library since tagging by searching for common identity terms could lack coverage of rare words or even word variations.

3.3. Model Behavior on Masked Data

We ran the ToxiGen model on both Subject and Object masked data and evaluated the shift on the classification; if before the mask the prediction was hate and after the mask it became non-hate, if before the mask the prediction was non-hate and after the mask it became hate, or if it stayed the same. Of the 13 target groups of ToxiGen, Table 2 presents the groups with the highest shifts. For the Latino target group was identified the highest shift from hate to non-hate; for mental disorders, it was detected the highest shift from non-hate to hate; and for the Middle East target group, the smallest shift was observed.

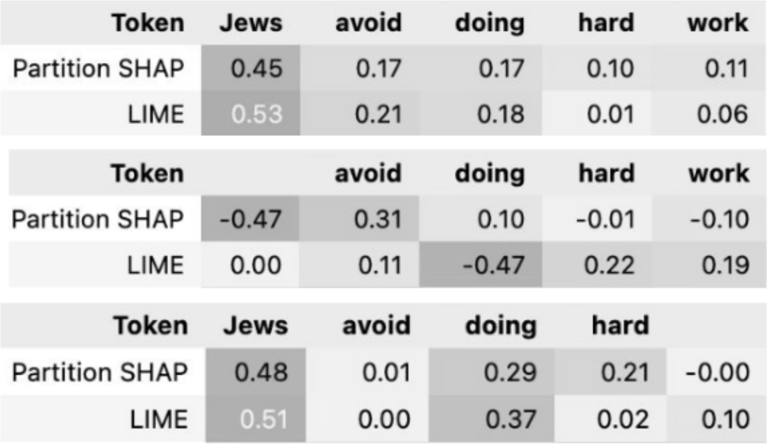
Group	Total	eval shift	percentage eval shift
latino	18545	same 12426	67.00%
		became hate 5259	28.36%
		became non-hate 860	**4.64%**
mental_dis	18659	same 12085	64.77%
		became hate 6429	**34.46%**
		became non-hate 145	0.78%
Slavic Hate Dataset	20000	same 18545	92.72%
		became hate 1201	6.01%
		became non-hate 254	1.27%

Table 2. Evaluation Shift on Subject Masked Data

We also ran the masking process on our 20k sampled Slavic Hate dataset and the prediction of the Toxigen model, whose results are also shown in Table 2. We found the smallest shifts from our dataset; the model's lack of knowledge about this target group likely explains why Slavic words rarely trigger a hate speech classification.

3.4. Fairness and Interpretability on Masked Data

To better understand how masked Subjects and Objects interfere in model classification, we used the ferret library proposed by [Attanasio et al. 2023], which is capable of using common explainer methods, such as SHAP [Lundberg and Lee 2017] and LIME [Ribeiro et al. 2016], to show the weight of each word in model's inference. This library uses a given classification model, and we choose to use the model proposed by ToxiGen. Figure 1 shows a tweet and exemplifies how the weights of classification change based on which word is masked.



Token	Jews	avoid	doing	hard	work
Partition SHAP	0.45	0.17	0.17	0.10	0.11
LIME	0.53	0.21	0.18	0.01	0.06

Token		avoid	doing	hard	work
Partition SHAP	-0.47	0.31	0.10	-0.01	-0.10
LIME	0.00	0.11	-0.47	0.22	0.19

Token	Jews	avoid	doing	hard	
Partition SHAP	0.48	0.01	0.29	0.21	-0.00
LIME	0.51	0.00	0.37	0.02	0.10

Figure 1. Evaluation explanation for masked sentence

4. Conclusion

This study aims to develop an analysis on how LLMs behave when evaluating hate speech targeting different and under-represented groups. During our research, we identified some limitations in the field: (1) Proposing a new dataset regarding an uncovered target group is challenging; the first challenge has already been surpassed, which is data collection. Now, another key factor that can impact the quality of our data is annotation. Achieving a completely annotated dataset is unfeasible, so methods to partially annotate it or use it without annotation are considerations that may impact our results and must be addressed carefully. (2) We aim to use Semi-supervised and Unsupervised methods to deal with the lack of annotation, this will be a challenge, since those methods are not easy to converge or may demand too much data to achieve good results. (3) LLMs are a field of rapid development, where new models, including open-source models, are continuously being released. We have to make sure that our work is robust enough to be comparable to current and future models that use the same architecture. Successful handling of these challenges will be key to ensuring the robustness of the work and its ability to deliver reliable results in the field of hate speech detection.

References

Attanasio, G., Pastor, E., Di Bonaventura, C., and Nozza, D. (2023). ferret: a framework for benchmarking explainers on transformers. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*. Association for Computational Linguistics.

Ayo F, Folorunso O, I. F. (2020). Machine learning techniques for hate speech classification of twitter data: State-of-the-art, future challenges and research directions. *Computer Science Review*, 38.

Cai, Y., Zimek, A., Wunder, G., and Ntoutsi, E. (2022). Power of explanations: Towards automatic debiasing in hate speech detection. *2022 IEEE 9th International Conference on Data Science and Advanced Analytics (DSAA)*, pages 1–10.

Hartvigsen, T., Gabriel, S., Palangi, H., Sap, M., Ray, D., and Kamar, E. (2022). Toxigen: A large-scale machine-generated dataset for adversarial and implicit hate speech detection. In *Annual Meeting of the Association for Computational Linguistics*.

Kovatchev, V., Gupta, S., and Lease, M. (2022). Fairly accurate: Learning optimal accuracy vs. fairness tradeoffs for hate speech detection. *ArXiv*, abs/2204.07661.

Lundberg, S. M. and Lee, S.-I. (2017). A unified approach to interpreting model predictions. In Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R., editors, *Advances in Neural Information Processing Systems 30*, pages 4765–4774. Curran Associates, Inc.

Maarouf, A., Pröllochs, N., and Feuerriegel, S. (2022). The virality of hate speech on social media.

Mathew, B., Saha, P., Yimam, S. M., Biemann, C., Goyal, P., and Mukherjee, A. (2020). Hatexplain: A benchmark dataset for explainable hate speech detection. In *AAAI Conference on Artificial Intelligence*.

Montani, I., Honnibal, M., Honnibal, M., Boyd, A., Landeghem, S. V., and Peters, H. (2023). spacy: Industrial-strength nlp.

Ribeiro, M. T., Singh, S., and Guestrin, C. (2016). ”why should I trust you?”: Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, August 13-17, 2016*, pages 1135–1144.

Velankar, A., Patil, H., and Joshi, R. (2022). A review of challenges in machine learning based automated hate speech detection.

Vidgen, B., Thrush, T., Waseem, Z., and Kiela, D. (2021). Learning from the worst. In *ACL*.