PEACE: Providing Explanations and Analysis for Combating Hate Expressions

Greta Damo, Nicolás Benjamín Ocampo, Elena Cabrio, Serena Villata

PEACE: Providing Explanations and Analysis for Combating Hate Expressions

Greta Damo[*,†], **Nicolás Benjamín Ocampo**[*,†], **Elena Cabrio and Serena Villata**

Université Côte d'Azur, CNRS, Inria, I3S, France
{greta.damo, nicolas-benjamin.ocampo, elena.cabrio, serena.villata}@univ-cotedazur.fr

Abstract. The increasing presence of hate speech (HS) on social media poses significant societal challenges. While efforts in the Natural Language Processing community have focused on automating the detection of explicit forms of HS, subtler and indirect expressions often go unnoticed. This demo presents PEACE, a novel tool that, besides detecting if a social media message contains explicit or implicit HS, also generates detailed natural language explanations for such predictions. More specifically, PEACE addresses three main challenging tasks: *i)* exploring the characteristics of HS messages, *ii)* predicting hatefulness, and *iii)* elucidating the reasoning behind system predictions. A REST API is also provided to exploit the tool's functionalities.

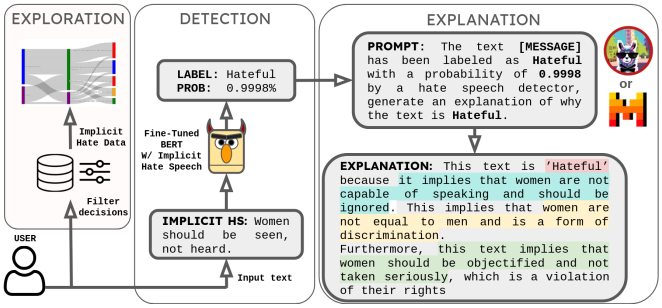

Figure 1. PEACE system overview.

1 Introduction

The volume of abusive content and hate speech on social media is a growing societal concern. Hate speech — defined as a direct attack against people based on protected characteristics such as race, ethnicity, national origin, disability, religious affiliation, sexual orientation, gender identity, among others [13] — is aggravated by the high portion of hateful content being spread across online platforms, requiring automated approaches to detect them effectively. While significant progress has been made in the field of Natural Language Processing (NLP) to automate the detection of hateful language, NLP systems primarily address direct and explicit forms of HS, often overlooking implicit and subtler forms [5, 6, 13]. The latter type of HS poses unique challenges as implicit HS comprises coded, ambiguous, or indirect language that does not immediately denote hate but still disparages a person or a target group [5, 13].

Recent studies have explored how to identify implicit hate speech, defined as coded or indirect language that disparages a person or group on the basis of protected characteristics. These studies include theoretical analysis and datasets [13, 5, 18, 15], more solid veiled detectors and explanation methods [19, 21, 7, 3]. Despite these efforts, creating effective tools and resources to recognize these nuanced forms of expression remains a challenge. Moreover, very few studies tackled the concept of subtle hate speech, which is defined as delicate or elusive messages that are characterized by the use of literal meanings, in contrast to implicit hate messages where we go beyond literal meanings [13].

In this paper, we present PEACE: **P**roviding **E**xplanations and **A**nalysis for **C**ombating Hate **E**xpressions. PEACE is a web tool conceived to support content moderators in exploring and evaluating implicit and subtle hate speech on social media.It comprises three main functionalities (Figure 1): *i)* the exploratory analysis of hate speech messages characteristics (*exploration*), *ii)* the prediction of hatefulness (*detection*), and *iii)* the explanation of system predictions (*explanation*). These functionalities incorporate not only a binary classification of whether a message is hateful (including explicit, implicit, and subtle messages following the definitions of Ocampo *et al.* [13]), with a detailed explanation in natural language that clarifies why a message is considered hateful and an exploratory analysis of the message characteristics.

To the best of our knowledge, PEACE is the only automated online tool allowing deep analysis and evaluation of both explicit and implicit hate messages. Few systems tackle similar tasks, such as RECAST [20] and MUDES [14], which identify multi-lingual span-level and sentence-level profanity expressions from the text on the web, MUTED [17] that allows visualizing the hateful intensity of those spans, IFAN [11] that tests an interactive platform for error analysis to debiase a hate speech classifier, CRYPTEXT [9] that provides an interactive interface to monitor and analyze offensive text perturbations online, TweetNLP [1] which is a platform supporting several NLP tasks including offensiveness detection, and McMillan-Major *et al.*'s system [10] which proposes a visualization tool and datasets for hate speech detection. However, none of these systems focus on detecting, in-depth analysis, and evaluating implicit hate speech messages such as PEACE.

NOTE: This paper contains examples of language that may be offensive to some readers. They do not represent the views of the authors.

† Equal contribution.
* Contact author.

2 PEACE Main Functionalities

In the following, we present the three main components of PEACE. We are also providing a public API built on Python and Flask, its respective documentation, and the PEACE UI developed using JavaScript and Flask templates. [1]

2.1 Data exploration

Implicit HS Datasets. PEACE allows for the search for specific instances from several social media and dataset sources oriented toward implicit HS. Currently, it covers the standard datasets: Implicit Hate Corpus (IHC) [5], Implicit and Subtle Hate (ISHate) [13], TOXIGEN [6], DynaHate (DYNA) [18], and Social Bias Inference Corpus (SBIC) [15]. The five datasets group the messages regarding their hatefulness, implicitness, and target groups. A summary of the data statistics for each of them can be found in Table 1. For label consistency across the datasets, we replicated the same setting as [12]. For each resource, users can retrieve one or several messages with their hateful, target, and implicit labels filtered by multiple options.

Visualization. To display the retrieved data, users have several visualizations available, including *Sankey*, *WordClouds*, and *Target Frequency* diagrams. Sankey Diagrams display how certain target groups are associated with respect to explicit and implicit connotations. They also show the relevant topics on these two labels estimated using Latent Dirichlet Allocation on the previously described datasets. WordClouds show the most frequent words on the selected group of messages. Lastly, Target Frequency displays the distribution of the target groups being attacked.

Data Augmentation. This module alters existing messages to create adversarial examples; these alterations are ways of obtaining augmented data specifically oriented to implicit messages. The focus is to modify parts in the text that do not affect the implicit hateful stand. We are incorporating the methods described in [13], displaying to the user the changes the new message has with respect to the original one. The implemented methods are: *Replace Named Entities* that replaces a named entity (PER, LOC, ORG, and MISC) in the input sentence by another one according to a previously collected list of NEs; *Replace Scalar Adverbs* that replaces emphasizing adverbs like "considerably" or "largely" with another scalar adverb that might increase or decrease the emphasis of an adjective/verb; *Add Adverbs to Verbs* that adds modifiers to verbs to accentuate them like "certainly", "likely", and "clearly"; *Replace Adjectives* that substitutes adjectives with their synonyms; *Replace In-Domain Expressions* that replaces a list of manually crafted expressions often used in HS messages (not captured by the RNE) with other semantically similar expressions; *Easy Data Augmentation* that modifies an input sentence using Random replacement, Random insertion, Random swap, and Random deletion; and finally, *Back Translation* that translates an input message into a different language to translate it back into the original one. Users can select their preferred method and apply it to messages they have written on their own. We also implemented a Python library where developers can access all these methods.

2.2 Implicit hate speech detection and explanation

Users also have the option to input their own messages for analysis. This message will be classified as hateful or non-hateful. In order to do this, the demo uses a binary (Non-HS/HS) machine-learning classifier. After selecting the desired message, the results, including probabilities for each label, are displayed directly in the user interface. Beyond simply classifying messages as hateful or not, the demo also provides the functionality of explaining the system predictions. The user can select between two LLMs that receive as input the prediction of the classifier together with its confidence with respect to the selected class, and the original message. Given this input, the LLM receives the following prompt:

The text: `[MESSAGE]` has been labeled as `[LABEL]` with a probability of `[PROB]` by a hate speech detector, generate an explanation of why the text is `[LABEL]` in no more than 3 sentences.

This allows users to understand why their selected messages might be considered hateful.

Available Models. The platform supports an advanced BERT model, fine-tuned on implicit hate data [4]. BERT is a pre-trained bidirectional transformer model that uses a combination of masked language modeling objective and next-sentence prediction on a large corpus comprising the Toronto Book Corpus and Wikipedia. The model is fine-tuned with 24823 and 10867 implicit and subtle hate speech messages, respectively, from the ISHate training set following the methodology proposed in [12]. Concerning the natural language explanation generation phase, PEACE supports the Mistral [8] and Alpaca [16] models. For Mistral, we tested the `Mistral-7B-instruct-v0.2` version, while for Alpaca, we employed `Alpaca-7B`.

Dataset	Source	Size	% Hate	% Implicit	% Explicit	% Subtle
IHC	Twitter	21480	38,124	86,702	13,298	-
SBIC	Social Media	147139	60,446	62,278	37,722	-
DYNA	Human-Machine	123432	53,896	58,065	41,935	-
ISHate	Social Media	63758	71,974	54,904	21,811	23,851
TOX	GPT-3	9900	42,657	45,489	54,511	-

Table 1. Comparing HS datasets. % Hate Class, % Implicit, % Explicit, and % Subtle are the percent labeled as hate, implicit hate, explicit hate, and subtle hate, respectively.

3 Experiments and Results

We summarize the performance of the models employed by PEACE using the test sets of the implicit datasets described in Section 2.1.

Detection. We tested the fine-tuned BERT model, together with other implicit detectors from the literature: HateBERT [2], Contrastive BERT, Contrastive HateBERT [12] and the LLMs Mistral and Alpaca. We calculated the macro average F1-score in each test set. Additionally, we calculated the percentage of implicit (I), explicit (E), and subtle (S) messages captured by the models (see Table 2). We can see that, in general, BERT, fine-tuned on implicit and subtle hate messages, outperforms the other models in binary (Non-HS/HS) detection and implicit classification tasks. Similarly, they obtained results comparable to those of Contrastive HateBERT in the subtle task. For this reason, PEACE relies on BERT fine-tuned on implicit and subtle data.

Explanation generation. We employ Alpaca and Mistral models to generate natural language explanations for messages containing implicit HS from IHC. In order to evaluate these explanations, we followed the same approach of [7, 19, 3] and concentrated on: Fluency (**F**) that evaluates whether the explanation follows proper grammar and structural rules, Informativeness (**I**) that assesses whether the explanation provides new information (e.g., additional context), Persuasiveness (**P**) that evaluates whether the explanation seems convincing, and Soundness (**S**) that describes whether the explanation

[1] The PEACE web server, dependency modules, models, experiments, and system demonstration video can be found at: https://gitlab.inria.fr/nocampo/peace.

	ISHate				IHC			SBIC			TOXIGEN			DYNA		
Model	F1	(%) I	(%) E	(%) S	F1	(%) I	(%) E	F1	(%) I	(%) E	F1	(%) I	(%) E	F1	(%) I	(%) E
BERT	0,863	**0,849**	0,831	0,769	0,627	**0,744**	0,810	**0,658**	0,516	0,634	0,641	**0,767**	**0,823**	**0,647**	0,757	**0,805**
HateBERT	**0,866**	0,801	0,837	0,769	0,614	0,674	0,814	0,628	0,478	0,571	0,626	0,714	0,775	0,641	0,674	0,760
Cont. BERT	0,837	0,715	0,714	0,692	0,629	0,557	0,708	0,530	0,301	0,414	0,600	0,564	0,619	0,635	0,580	0,705
Cont. HateBERT	0,852	0,737	0,805	**0,795**	0,616	0,681	0,801	0,618	0,468	0,562	0,603	0,722	0,762	0,632	0,691	0,760
Mistral	0,782	0,538	**0,846**	0,667	**0,649**	0,618	**0,903**	0,620	0,628	0,425	**0,743**	0,617	0,545	0,642	**0,788**	0,738
Alpaca	0,392	0,704	0,608	0,590	0,416	0,666	0,717	0,468	**0,650**	**0,674**	0,459	0,639	0,732	0,079	0,120	0,131

Table 2. Classification task (Non-HS vs HS) across the five selected datasets. We calculated the macro average F1-score and percentage of implicit (I), explicit (E), and subtle (S) messages correctly detected by Fine-tuned BERT, HateBERT, Cont. BERT and Cont HateBERT. Mistral and Alpaca are also evaluated to compare their detection capabilities w.r.t the other models. Subtle labels are only part of the ISHate dataset.

Strategy	Hate Speech	Explanation	(F)	(I)	(P)	(S)
Only Alpaca	What two kinds of people are totally different but always look the same? Asians and hepatitis patients	The text is hateful because it perpetuates stereotypes about Asians and hepatitis patients by suggesting that they are two groups that are totally different but always look the same. This is a false and offensive generalization that reinforces negative perceptions of both groups.	5	2	3	5
Only Mistral	Homeless Where do Western Hobo's go for food? Food shelter. Where do Asian Hobo's go for food?. Dog shelters	The text does not appear to be hateful on its own. It is asking a question about where hobos from different cultural backgrounds obtain food. However, the use of the term "Hobo" may be outdated and perceived as derogatory by some. It's important to note that using labels to stereotype or belittle individuals based on their economic status, ethnicity, or any other characteristic is disrespectful and hateful. The suggestion that Asian hobos specifically go to dog shelters for food is incorrect and potentially harmful, as it perpetuates negative stereotypes and could lead to harm against homeless Asian individuals and dog shelters.	5	2	2	3
Bert + Alpaca	women should be seen, not heard	This text is 'Hateful' because it implies that women are not capable of speaking and should be ignored. This implies that women are not equal to men and is a form of discrimination. Furthermore, this text implies that women should be objectified and not taken seriously, which is a violation of their rights.	5	3	5	5
Bert + Mistral	as it was intended .. #whitegenocide #kalergi	The text contains the hashtags #whitegenocide and #kalergi, which are often associated with racist and conspiratorial beliefs. The former reference to a supposed plan of genetic elimination of the white race, while the latter connects to an alleged plan of interracial breeding by Count Coudenhove-Kalergi, promoting a mixed European race to unify Europe. Such content promotes hateful and discriminatory ideologies, making it a potential threat towards peaceful and inclusive coexistence of different races and ethnicities.	5	5	5	5

Table 3. Human Evaluation examples comparing Explanations with Vanilla Mistral/Alpaca VS Fine-Tuned BERT + Mistral/Alpaca.

is valid and logical. All metrics use a Likert scale and range from 1 (lowest score) to 5 (maximum score). We selected a sample of 100 messages, randomly extracting 20 from the test set of each dataset.

	Avg. **(F)**		Avg. **(I)**		Avg. **(P)**		Avg. **(S)**	
only alpaca	4.84	4.92	2.52	2.69	3.48	3.52	4.02	4.04
only mistral	5.00		2.86		3.56		4.06	
bert+alpaca	4.96	4.98	**2.92**	**3.15**	**3.84**	**4.00**	4.16	4.10
bert+mistral	5.00		**3.38**		**4.16**		4.04	

Table 4. Human Evaluation comparing Explanations with Vanilla Mistral/Alpaca VS Fine-Tuned BERT + Mistral/Alpaca. Results in bold are statistically significantly with respect to the "only" configuration.

For 25 of those messages, explanations are generated with only Mistral, 25 with only Alpaca, 25 with fine-tuned BERT + Mistral, and the last 25 with fine-tuned BERT + Mistral. For the latter two strategies, we provide to the LLM the label predicted by the supervised binary classifier and its confidence. More precisely, we provide to the LLM the fine-tuned BERT output, as it is the best performing model in our classification task. Two instructed annotators provided the scores for this sample by looking at the hateful message and its corresponding generated explanation. From Table 4, we see the average scores for each metric and configuration, showing that the explanations from the pipeline method are statistically significantly more Informative and Persuasive than the ones generated by the LLM only, without the label of the supervised classifier ((**I**) 3.15 vs 2.69, and (**P**) 4.00 vs 3.52, respectively). Similarly, the same effect occurs in per model configuration ((**I**) 2.92 vs 2.52, and (**P**) 3.84 vs 3.48 for alpaca, and (**I**) 3.38 vs 2.86, and (**P**) 4.16 vs 3.56 for mistral). For Fluency and Soundness, the results are comparable to those of the vanilla counterpart. Statistical significance is tested using the Mann-Whitney U Test with a P-value < 0.05. Table 3 shows some examples with their scores.

4 Conclusion

We presented PEACE, a novel tool for the detection and explanation generation of hate speech. It is the first system focusing specifically on implicit and subtle hate speech messages by allowing not only for the classical binary classification of whether a message is hateful or not, but also a detailed explanation of the reasons why a message is classified as hateful. Additionally, PEACE provides different diagrams for exploratory analysis of the messages' characteristics and data augmentation strategies for implicit and subtle hate messages. As for future work, we plan to include more of the detector's internal knowledge in PEACE to generate explanations apart from the detector's prediction, as the extraction of relevant detector's rationales or implied statement in line with Yang *et al.* [21], the use of implicit hate demonstration examples as Damo *et al.* [3] and Wand *et al.* [19], and the target hateful intention as in Huang *et al.* [7]. Also, given the subjectivity of implicit/subtle labels and explanation generation, human evaluation is still needed, requiring standard metrics evaluated with a higher number of annotators and providing structured test suites for hate speech explanation.

Acknowledgements

This work was supported by the French government, through the 3IA Cote d'Azur Investments in the Future project managed by the ANR with the reference number ANR-19-P3IA-0002 and by the ANR project ATTENTION (ANR21-CE23-0037).

References

[1] J. Camacho-collados, K. Rezaee, T. Riahi, A. Ushio, D. Loureiro, D. Antypas, J. Boisson, L. Espinosa Anke, F. Liu, and E. Martínez Cámara. TweetNLP: Cutting-edge natural language processing for social media. In W. Che and E. Shutova, editors, *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–49, Abu Dhabi, UAE, Dec. 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.emnlp-demos.5. URL https://aclanthology.org/2022.emnlp-demos.5.

[2] T. Caselli, V. Basile, J. Mitrović, and M. Granitzer. HateBERT: Retraining BERT for abusive language detection in English. In A. Mostafazadeh Davani, D. Kiela, M. Lambert, B. Vidgen, V. Prabhakaran, and Z. Waseem, editors, *Proceedings of the 5th Workshop on Online Abuse and Harms (WOAH 2021)*, pages 17–25, Online, Aug. 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.woah-1.3. URL https://aclanthology.org/2021.woah-1.3.

[3] G. Damo, N. B. Ocampo, E. Cabrio, and S. Villata. Unveiling the Hate: Generating Faithful and Plausible Explanations for Implicit and Subtle Hate Speech Detection. In *The 29th International Conference on Natural Language & Information Systems*, Torino, Italy, June 2024. URL https://hal.science/hal-04658110.

[4] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In J. Burstein, C. Doran, and T. Solorio, editors, *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1423. URL https://aclanthology.org/N19-1423.

[5] M. ElSherief, C. Ziems, D. Muchlinski, V. Anupindi, J. Seybolt, M. De Choudhury, and D. Yang. Latent hatred: A benchmark for understanding implicit hate speech. In M.-F. Moens, X. Huang, L. Specia, and S. W.-t. Yih, editors, *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 345–363, Online and Punta Cana, Dominican Republic, Nov. 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.emnlp-main.29. URL https://aclanthology.org/2021.emnlp-main.29.

[6] T. Hartvigsen, S. Gabriel, H. Palangi, M. Sap, D. Ray, and E. Kamar. ToxiGen: A large-scale machine-generated dataset for adversarial and implicit hate speech detection. In S. Muresan, P. Nakov, and A. Villavicencio, editors, *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3309–3326, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.acl-long.234. URL https://aclanthology.org/2022.acl-long.234.

[7] F. Huang, H. Kwak, and J. An. Chain of explanation: New prompting method to generate quality natural language explanation for implicit hate speech. In *Companion Proceedings of the ACM Web Conference 2023*, WWW '23 Companion, page 90–93, New York, NY, USA, 2023. Association for Computing Machinery. ISBN 9781450394192. doi: 10.1145/3543873.3587320. URL https://doi.org/10.1145/3543873.3587320.

[8] A. Q. Jiang, A. Sablayrolles, A. Mensch, C. Bamford, D. S. Chaplot, D. d. l. Casas, F. Bressand, G. Lengyel, G. Lample, L. Saulnier, L. R. Lavaud, M.-A. Lachaux, P. Stock, T. L. Scao, T. Lavril, T. Wang, T. Lacroix, and W. E. Sayed. Mistral 7B, Oct. 2023. URL http://arxiv.org/abs/2310.06825. arXiv:2310.06825 [cs].

[9] T. Le, Y. Ye, Y. Hu, and D. Lee. Cryptext: Database and interactive toolkit of human-written text perturbations in the wild. In *2023 IEEE 39th International Conference on Data Engineering (ICDE)*, pages 3639–3642, Los Alamitos, CA, USA, apr 2023. IEEE Computer Society. doi: 10.1109/ICDE55515.2023.00287. URL https://doi.ieeecomputersociety.org/10.1109/ICDE55515.2023.00287.

[10] A. McMillan-Major, A. Paullada, and Y. Jernite. An interactive exploratory tool for the task of hate speech detection. In S. L. Blodgett, H. Daumé III, M. Madaio, A. Nenkova, B. O'Connor, H. Wallach, and Q. Yang, editors, *Proceedings of the Second Workshop on Bridging Human–Computer Interaction and Natural Language Processing*, pages 11–20, Seattle, Washington, July 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.hcinlp-1.2. URL https://aclanthology.org/2022.hcinlp-1.2.

[11] E. Mosca, D. Dementieva, T. Ebrahim Ajdari, M. Kummeth, K. Gringauz, Y. Zhou, and G. Groh. IFAN: An explainability-focused interaction framework for humans and NLP models. In S. Saha and H. Sujaini, editors, *Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics: System Demonstrations*, pages 59–76, Bali, Indonesia, Nov. 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.ijcnlp-demo.7. URL https://aclanthology.org/2023.ijcnlp-demo.7.

[12] N. B. Ocampo, E. Cabrio, and S. Villata. Unmasking the hidden meaning: Bridging implicit and explicit hate speech embedding representations. In H. Bouamor, J. Pino, and K. Bali, editors, *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 6626–6637, Singapore, Dec. 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-emnlp.441. URL https://aclanthology.org/2023.findings-emnlp.441.

[13] N. B. Ocampo, E. Sviridova, E. Cabrio, and S. Villata. An in-depth analysis of implicit and subtle hate speech messages. In A. Vlachos and I. Augenstein, editors, *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 1997–2013, Dubrovnik, Croatia, May 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.eacl-main.147. URL https://aclanthology.org/2023.eacl-main.147.

[14] T. Ranasinghe and M. Zampieri. MUDES: Multilingual detection of offensive spans. In A. Sil and X. V. Lin, editors, *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Demonstrations*, pages 144–152, Online, June 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.naacl-demos.17. URL https://aclanthology.org/2021.naacl-demos.17.

[15] M. Sap, S. Gabriel, L. Qin, D. Jurafsky, N. A. Smith, and Y. Choi. Social bias frames: Reasoning about social and power implications of language. In D. Jurafsky, J. Chai, N. Schluter, and J. Tetreault, editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5477–5490, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.486. URL https://aclanthology.org/2020.acl-main.486.

[16] R. Taori, I. Gulrajani, T. Zhang, Y. Dubois, X. Li, C. Guestrin, P. Liang, and T. B. Hashimoto. Stanford alpaca: An instruction-following llama model. https://github.com/tatsu-lab/stanford_alpaca, 2023.

[17] C. Tillmann, A. Trivedi, S. Rosenthal, S. Borse, R. Zhang, A. Sil, and B. Bhattacharjee. Muted: Multilingual targeted offensive speech identification and visualization. In Y. Feng and E. Lefever, editors, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 229–236, Singapore, Dec. 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-demo.19. URL https://aclanthology.org/2023.emnlp-demo.19.

[18] B. Vidgen, T. Thrush, Z. Waseem, and D. Kiela. Learning from the worst: Dynamically generated datasets to improve online hate detection. In C. Zong, F. Xia, W. Li, and R. Navigli, editors, *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1667–1682, Online, Aug. 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-long.132. URL https://aclanthology.org/2021.acl-long.132.

[19] H. Wang, M. S. Hee, M. R. Awal, K. T. W. Choo, and R. K.-W. Lee. Evaluating GPT-3 Generated Explanations for Hateful Content Moderation. volume 6, pages 6255–6263, Aug. 2023. doi: 10.24963/ijcai.2023/694. URL https://www.ijcai.org/proceedings/2023/694. ISSN: 1045-0823.

[20] A. P. Wright, O. Shaikh, H. Park, W. Epperson, M. Ahmed, S. Pinel, D. H. P. Chau, and D. Yang. RECAST: Enabling User Recourse and Interpretability of Toxicity Detection Models with Interactive Visualization. *Proceedings of the ACM on Human-Computer Interaction*, 5(CSCW1):181:1–181:26, 2021. doi: 10.1145/3449280. URL https://dl.acm.org/doi/10.1145/3449280.

[21] Y. Yang, J. Kim, Y. Kim, N. Ho, J. Thorne, and S.-Y. Yun. HARE: Explainable hate speech detection with step-by-step reasoning. In H. Bouamor, J. Pino, and K. Bali, editors, *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 5490–5505, Singapore, Dec. 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-emnlp.365. URL https://aclanthology.org/2023.findings-emnlp.365.