Proceedings of the Thirty-Fifth International Joint Conference on Artificial Intelligence (IJCAI-26)
Demonstrations Track

PEACE 2.0: Grounded Explanations and Counter-Speech for Combating Hate Expressions

Greta Damo[1], **Stéphane Petiot**[2], **Elena Cabrio**[1] and **Serena Villata**[1]
[1]Université Côte d'Azur, CNRS, Inria, I3S, France
[2]Université Côte d'Azur, Institut 3IA Côte d'Azur, Techpool, France
{firstname.surname}@univ-cotedazur.fr, stephane.petiot@inria.fr

Abstract

The increasing volume of hate speech on online platforms poses significant societal challenges. While the Natural Language Processing community has developed effective methods to automatically detect the presence of hate speech, responses to it, called counter-speech, are still an open challenge. We present PEACE 2.0, a novel tool that, besides analysing and explaining why a message is considered hateful or not, also generates a response to it. More specifically, PEACE 2.0 has three main new functionalities: leveraging a Retrieval-Augmented Generation (RAG) pipeline *i)* to ground HS explanations into evidence and facts, *ii)* to automatically generate evidence-grounded counter-speech, and *iii)* exploring the characteristics of counter-speech replies. By integrating these capabilities, PEACE 2.0 enables in-depth analysis and response generation for both explicit and implicit hateful messages.

1 Introduction

The growing volume of hate speech (HS) on social media represents a major societal challenge, requiring automated tools that go beyond surface-level detection. Hate speech can appear in explicit, implicit, or subtle forms, with the latter relying on coded, indirect, or context-dependent language that is particularly difficult to identify and interpret [Ocampo *et al.*, 2023]. While advances in Natural Language Processing (NLP) and Generation (NLG) have improved detection of explicit content, understanding and addressing nuanced forms of hate remains an open problem. To address this challenge, the tool PEACE was introduced (Providing Explanations and Analysis for Combating Hate Expressions) [Damo *et al.*, 2024], a web-based system designed to support exploration, detection, and explanation of explicit, and implicit HS. PEACE enables content moderators and researchers to analyze hateful messages, inspect model predictions, and obtain natural language explanations clarifying why a message is considered hateful. In this paper, we present PEACE 2.0, an extended version that moves beyond analysis and explanation to support actionable responses through automatic counter-speech (CS) generation. Counter-speech challenges hateful messages with factual information or alternative perspectives, offering a constructive alternative to content removal or censorship [Benesch, 2014]. However, generating effective CS is particularly challenging for implicit hate, where responses must be both informative and carefully framed. PEACE 2.0 introduces three key new functionalities to address these challenges: **1. Knowledge-grounded counter-speech generation.** A Retrieval-Augmented Generation (RAG) pipeline, as in [Damo *et al.*, 2025a], retrieves evidence from authoritative human rights sources and conditions counter-speech on this information. Users can explicitly compare responses generated with and without retrieval, highlighting the impact of knowledge grounding on factuality, relevance, and effectiveness. **2. Evidence-grounded explanations for hate speech classification.** Using the same RAG mechanism, PEACE 2.0 generates explanations that justify the predictions of a fine-tuned BERT classifier, grounding label decisions in retrieved evidence to improve transparency and interpretability. **3. Visual analytics for counter-speech exploration.** Interactive tools allow analysis of numerous counter-speech datasets supporting experimental evaluation of counter-speech performance on explicit versus implicit hate speech and offering insights into the role of retrieval-based grounding.

Related Work. Existing systems address HS detection, visualization, monitoring, or CS generation separately. Tools such as RECAST [Wright *et al.*, 2021], MUDES [Ranasinghe and Zampieri, 2021], MUTED [Tillmann *et al.*, 2023], IFAN [Mosca *et al.*, 2023], CRYPTEXT [Le *et al.*, 2023], TweetNLP [Camacho-collados *et al.*, 2022], and McMillan-Major et al.'s framework [McMillan-Major *et al.*, 2022] provide detection and analysis interfaces, while dashboards like the Indonesian HS monitoring system [Wijanarko *et al.*, 2024] support large-scale tracking. In the CS domain, work is limited; CounterHelp [Babic *et al.*, 2025] leverages large language models for context-sensitive responses. However, prior systems treat these components in isolation, focus mainly on explicit hate, and do not systematically compare knowledge-grounded and non-grounded generation.

Overall, PEACE 2.0 offers a unified, interactive platform for hate speech exploration, detection, explanation, and response generation, bridging analytical insights and practical mitigation. To our knowledge, it is the only online tool providing in-depth analysis of both explicit and implicit hate speech together with knowledge-grounded CS generation.

2 PEACE 2.0 Main Functionalities

In this section, we report the main features of the original demo, and the new functionalities of PEACE 2.0[1]

2.1 Data Exploration and Visualization

This module offers interactive visualizations for exploring HS and CS datasets. Users can switch between the two views.

Hate Speech. The hate speech view covers implicit hate datasets used in PEACE: the Implicit Hate Corpus (IHC) [ElSherief *et al.*, 2021], Implicit and Subtle Hate (ISHate) [Ocampo *et al.*, 2023], TOXIGEN [Hartvigsen *et al.*, 2022], DynaHate (DYNA) [Vidgen *et al.*, 2021], and Social Bias Inference Corpus (SBIC) [Sap *et al.*, 2020]. Messages are organized by hatefulness, implicitness, and target group, with sanitized labels for consistency, that users can filter. Visualizations include: *Sankey diagrams* linking target groups and HS categories (explicit vs. implicit) with topic distributions derived via Latent Dirichlet Allocation (LDA); *Word Clouds* displaying the most frequent lexical items within selected attributes; and *Target Frequency* charts displaying attacked groups across datasets and implicitness levels.

Counter-speech. In PEACE 2.0, the same visual framework is extended to CS datasets, including CONAN [Chung *et al.*, 2019], Multitarget-CONAN [Fanton *et al.*, 2021], Twitter and YouTube datasets [Mathew *et al.*, 2020; Albanyan and Blanco, 2022; Mathew *et al.*, 2019], knowledge-grounded human expert datasets [Bonaldi *et al.*, 2025; Chung *et al.*, 2021], and RAG-generated responses from multiple LLMs and retrieval strategies [Damo *et al.*, 2025a]. Labels follow the same sanitization scheme as PEACE. *Sankey diagrams* connect targets, counter-speech sources (expert, user, RAG, No-RAG), and LDA topics; *Word Clouds* highlight frequent terms within filtered messages, and *Frequency* charts display CS volume per target and source, and the distribution of different CS strategies across targets.

2.2 Data Augmentation

This module generates adversarial examples by modifying messages while preserving their implicit hateful meaning. The aim is to augment data for implicit HS by altering surface-level elements without changing the underlying stance. Following [Ocampo *et al.*, 2023], the strategies are: named entity replacement; adjustment of scalar adverbs; addition of adverbial modifiers; adjective synonym substitution; replacement of domain-specific expressions with semantically similar variants; Easy Data Augmentation (random replacement, insertion, swap, deletion); and back-translation. Users can apply these methods to custom inputs.

2.3 Hate Speech Detection and Explanation

This module allows users to input custom messages for binary hate speech classification. The system outputs the predicted label with its confidence score. In addition, PEACE 2.0 generates concise, human-readable explanations conditioned on the message, predicted label, and probability, helping users understand the rationale behind the decision. PEACE 2.0 further introduces an optional RAG pipeline from [Damo *et al.*, 2025a], to produce evidence-grounded explanations. When enabled, the system retrieves relevant passages from an authoritative knowledge base, summarizes them into factual context, and integrates this information into the LLM prompt. Users can choose whether to use RAG and select the underlying LLM. For transparency, the system also shows the retrieved paragraphs together with their similarity score.

2.4 Knowledge-Grounded CS Generation

Building on the RAG pipeline, a central innovation in PEACE 2.0 is evidence-grounded CS generation based on a curated human rights knowledge base. Users can input a hateful message and generate counter-speech responses with or without knowledge grounding, enabling direct comparison of evidence on outputs in terms of relevance, factuality, and persuasiveness. Multiple LLMs are supported, offering flexibility in style and quality. The CS generation pipeline consists of three steps. First, the input message is encoded using the BGE-M3 sentence transformer, and FAISS performs inner-product similarity search over precomputed paragraph embeddings to retrieve the top-3 most relevant evidence passages, with deduplication applied. Second, the retrieved passages are concatenated and summarized into a concise summary using the selected LLM. Finally, the original message and the evidence summary are provided to the same LLM to generate a respectful and persuasive CS response suitable for social media. Users may also disable retrieval to compare RAG and non-RAG outputs. For transparency, the system also shows the retrieved paragraphs together with their similarity score.

Available models. For detection, the demo uses a BERT classifier fine-tuned on the ISHate training set. For explanation and CS generation, PEACE 2.0 supports open-source LLMs: Mistral (`Mistral-7B-Instruct-v0.3`), LLaMa (`Llama-3.1-8B-Instruct`), and CommandR (`c4ai-command-r7b-12-2024`). The knowledge base comprises 32,792 documents from the United Nations Digital Library, Eur-Lex, and the European Agency for Fundamental Rights (from 2000 to 2025), totaling 3,173,630 tokenized paragraphs.

3 Experiments and Results

3.1 Experimental Setting

We hypothesize that RAG-grounded explanations and CS will outperform non-RAG generations for both explicit and implicit HS, demonstrating the value of evidence-grounded generation. Specifically, we test: **H1**: RAG outputs are better overall and more informative for both explanations and CS. **H2**: RAG improves persuasiveness; **H3**: RAG improves outputs for both implicit and explicit cases. To evaluate this, from the implicit HS datasets (IHC, ISHate, TOXIGEN, DYNA, SBIC), we randomly sample 20 messages per dataset, evenly split between explicit and implicit, resulting in 100 HS examples. For each message, we generate explanations with and without RAG (100 RAG, 100 non-RAG). The same procedure is applied to CS, producing 200 responses. We assess all generations through both automatic and human evaluation.

[1]The demo is available at https://3ia-demos.inria.fr/peace/ and the video by clicking here. We also provide a public API at https://gitlab.inria.fr/nocampo/peace.

Human Evaluation. We evaluate a subset of 100 explanations and 100 CS responses. For each task, 50 explicit and 50 implicit (25 RAG, 25 non-RAG) cases are sampled, resulting in 200 human-evaluated instances overall. Each example is assessed by three trained annotators. Evaluation criteria for explanations follow PEACE, while CS metrics are adapted from [Zheng *et al.*, 2023; Bonaldi *et al.*, 2024; Damo *et al.*, 2025b]. Both explanations and CS are rated on: **Fluency (F)** (grammatical correctness), **Informativeness (I)** (relevant contextual or factual content), **Persuasiveness (P)** (ability to convince the hater and foster empathy in bystanders), **Soundness (SO)** (logical coherence), and **Specificity (SP)** (direct engagement with the HS and target). All dimensions are rated on a 1-5 Likert scale (5 = highest).

Automatic metrics. We compute automatic measures of linguistic quality, diversity, and faithfulness. These include *Distinct-3* for lexical diversity; *Semantic Similarity* (Sentence-BERT) between HS and generated explanations or counter-speech; and *Perplexity* as a proxy for fluency. For RAG outputs, we also measure *faithfulness* via similarity between generations and retrieved evidence. Finally, *NLI-based metrics* (entailment and contradiction with `roberta-large-nli`) assess whether outputs appropriately address the original HS and remain consistent with the supporting evidence.

Metric	Exp_{RAG}	Exp_{NoRAG}	Imp_{RAG}	Imp_{NoRAG}
Explanations				
F	5.00	5.00	5.00	5.00
SO	4.88	4.56	4.80	4.58
I	4.38	2.84	4.64	2.72
SP	4.86	3.78	4.88	4.40
P	4.68	3.52	4.72	3.86
Overall	**4.76**	**3.94**	**4.81**	**4.11**
Counter-speech				
F	5.00	5.00	5.00	5.00
SO	4.82	3.92	4.88	4.52
I	4.66	2.52	4.80	2.86
SP	4.90	2.98	4.90	3.32
P	4.68	2.64	4.94	3.22
Overall	**4.81**	**3.41**	**4.90**	**3.78**

Table 1: Mean human ratings for RAG vs. No-RAG outputs: Fluency, SOundness, Informativeness, SPecificity, Persuasiveness.

3.2 Results

Human evaluation. From Table 1, RAG-generated outputs consistently outperform non-RAG outputs across all metrics. Supporting H1, *RAG outputs are significantly more informative*, particularly for implicit content (Explanation-Imp.: 4.64 vs. 2.72; Counter-speech-Imp.: 4.80 vs. 2.86), *and of better quality* (see highlighted results). In line with H2, *RAG also improves persuasiveness* for both explicit and implicit hate (e.g., Explanation-Exp.: 4.68 vs. 3.52; Counter-speech-Exp.: 4.68 vs. 2.64). Consistent with H3, *gains are larger for both implicit and explicit cases*, highlighting the benefit of retrieval in subtle contexts. Fluency and Soundness remain comparable across RAG and non-RAG outputs, indicating that these improvements do not compromise readability or coherence. Wilcoxon signed-rank tests confirm these differences are statistically significant ($p < 0.05$). Inter-annotator agreement, measured using Krippendorff's α, is substantial to perfect across dimensions (k = 0.57-1).

Metric	Exp_{RAG}	Exp_{NoRAG}	Imp_{RAG}	Imp_{NoRAG}
Explanations				
Sem. Sim.	**0.57**	0.49	**0.56**	0.47
Faithfulness	0.57	-	0.56	-
Perplexity	**25.66**	37.38	**25.43**	37.21
Distinct-3	0.98	**0.99**	0.97	**0.99**
Hate-Ent.	**0.18**	0.08	**0.12**	0.08
Ev.-Contr.	0.05	-	0.05	-
Ev.-Ent.	0.26	-	0.21	-
Counter-speech				
Sem. Sim.	**0.51**	0.45	**0.50**	0.41
Faithfulness	0.65	-	0.61	-
Perplexity	**14.47**	21.74	**14.36**	22.91
Distinct-3	0.99	**1.00**	0.99	**1.00**
Hate Ent.	**0.04**	0.03	**0.09**	0.05
Ev.Contr.	0.05	-	0.03	-
Ev. Ent.	0.18	-	0.15	-

Table 2: Automatic metrics results. Abbreviations: Explicit (Exp), Implicit (Imp), Semantic Similarity (Sem. Sim.), Entailment (Ent.), Contradiction (Contr.), Evidence (Ev.). Best results are in bold.

Automatic metrics. From Table 2, we see that RAG-generated outputs consistently outperform No-RAG in both Explanation and CS tasks. They achieve higher semantic similarity, are faithful to retrieved evidence, have lower perplexity, indicating more informative and fluent outputs, while maintaining diversity (Distinct-3). NLI metrics further show that RAG outputs have higher Hate-Entailment and low Evidence-Contradiction, indicating that they appropriately address the content of the original HS without supporting it, and are aligned with retrieved content. Gains are for both implicit and explicit content, confirming that RAG improves informativeness, persuasiveness, and faithfulness across both tasks. Differences are statistically significant with Wilcoxon signed-rank tests.

4 Conclusion

By integrating retrieval, summarization, and generation within a single pipeline, the PEACE 2.0 tool bridges analytical understanding and practical mitigation of both implicit and explicit HS messages, producing evidence-backed CS that improves trustworthiness and effectiveness. The explanation and generation modules enhance the transparency of the abusive language classification and the CS generation, making PEACE 2.0 a suitable tool for e-democracy applications with the aim to enhance inclusiveness and fairness. Future work includes the integration of adaptive retrieval strategies, dynamic knowledge base updates, and the inclusion of evaluation metrics to assess CS quality.

Acknowledgments

This work has been partially supported by the French government, through the 3IA Côte d'Azur investments in the project managed by the National Research Agency (ANR) with the reference number ANR-23-IACL-0001.

References

[Albanyan and Blanco, 2022] Abdullah Albanyan and Eduardo Blanco. Pinpointing fine-grained relationships between hateful tweets and replies. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2022.

[Babic *et al.*, 2025] Andreas Babic, Xihui Chen, Djordje Slijepčević, Adrian J. Böck, and Matthias Zeppelzauer. Counterhelp: Promoting online civil courage among young people through ai-generated counterspeech. In *Proceedings of the 33rd ACM International Conference on Multimedia*. Association for Computing Machinery, 2025.

[Benesch, 2014] Susan Benesch. Countering dangerous speech: New ideas for genocide prevention. *SSRN Electronic Journal*, 2014.

[Bonaldi *et al.*, 2024] Helena Bonaldi, Greta Damo, Nicolás Benjamín Ocampo, Elena Cabrio, Serena Villata, and Marco Guerini. Is safer better? the impact of guardrails on the argumentative strength of LLMs in hate speech countering. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, Miami, Florida, USA, 2024.

[Bonaldi *et al.*, 2025] Helena Bonaldi, María Estrella Vallecillo-Rodríguez, Irune Zubiaga, Arturo Montejo-Ráez, Aitor Soroa, María-Teresa Martín-Valdivia, Marco Guerini, and Rodrigo Agerri. The first workshop on multilingual counterspeech generation at coling 2025: Overview of the shared task. In *Proceedings of the First Workshop on Multilingual Counterspeech Generation*, 2025.

[Camacho-collados *et al.*, 2022] Jose Camacho-collados, Kiamehr Rezaee, Talayeh Riahi, Asahi Ushio, Daniel Loureiro, Dimosthenis Antypas, Joanne Boisson, Luis Espinosa Anke, Fangyu Liu, and Eugenio Martínez Cámara. TweetNLP: Cutting-edge natural language processing for social media. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, Abu Dhabi, UAE, 2022.

[Chung *et al.*, 2019] Yi-Ling Chung, Elizaveta Kuzmenko, Serra Sinem Tekiroğlu, and Marco Guerini. Conan-counter narratives through nichesourcing: a multilingual dataset of responses to fight online hate speech. In *Proceedings of the 57th annual meeting of the association for computational linguistics*, 2019.

[Chung *et al.*, 2021] Yi-Ling Chung, Serra Sinem Tekiroğlu, and Marco Guerini. Towards knowledge-grounded counter narrative generation for hate speech. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, Online, 2021.

[Damo *et al.*, 2024] Greta Damo, Nicolás Benjamín Ocampo, Elena Cabrio, and Serena Villata. Peace: Providing explanations and analysis for combating hate expressions. In *ECAI 2024-27th European Conference on Artificial Intelligence*, 2024.

[Damo *et al.*, 2025a] Greta Damo, Elena Cabrio, and Serena Villata. Beating harmful stereotypes through facts: Rag-based counter-speech generation. *arXiv preprint arXiv:2510.12316*, 2025.

[Damo *et al.*, 2025b] Greta Damo, Elena Cabrio, and Serena Villata. Effectiveness of Counter-Speech against Abusive Content: A Multidimensional Annotation and Classification Study. In *IEEE Xplore*, London, United Kingdom, 2025.

[ElSherief *et al.*, 2021] Mai ElSherief, Caleb Ziems, David Muchlinski, Vaishnavi Anupindi, Jordyn Seybolt, Munmun De Choudhury, and Diyi Yang. Latent hatred: A benchmark for understanding implicit hate speech. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, Online and Punta Cana, Dominican Republic, 2021.

[Fanton *et al.*, 2021] Margherita Fanton, Helena Bonaldi, Serra Sinem Tekiroğlu, and Marco Guerini. Human-in-the-loop for data collection: a multi-target counter narrative dataset to fight online hate speech. In *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)*, Online, 2021.

[Hartvigsen *et al.*, 2022] Thomas Hartvigsen, Saadia Gabriel, Hamid Palangi, Maarten Sap, Dipankar Ray, and Ece Kamar. ToxiGen: A large-scale machine-generated dataset for adversarial and implicit hate speech detection. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Dublin, Ireland, 2022.

[Le *et al.*, 2023] Thai Le, Yiran Ye, Yifan Hu, and Dongwon 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. IEEE, 2023.

[Mathew *et al.*, 2019] Binny Mathew, Punyajoy Saha, Hardik Tharad, Subham Rajgaria, Prajwal Singhania, Suman Kalyan Maity, Pawan Goyal, and Animesh Mukherjee. Thou shalt not hate: Countering online hate speech. In *Proceedings of the international AAAI conference on web and social media*, 2019.

[Mathew *et al.*, 2020] Binny Mathew, Navish Kumar, Pawan Goyal, and Animesh Mukherjee. Interaction dynamics between hate and counter users on twitter. In *Proceedings of the 7th ACM IKDD CoDS and 25th COMAD*. 2020.

[McMillan-Major *et al.*, 2022] Angelina McMillan-Major, Amandalynne Paullada, and Yacine Jernite. An interactive exploratory tool for the task of hate speech detection. In *Proceedings of the Second Workshop on Bridging*

Human–Computer Interaction and Natural Language Processing, Seattle, Washington, 2022.

[Mosca *et al.*, 2023] Edoardo Mosca, Daryna Dementieva, Tohid Ebrahim Ajdari, Maximilian Kummeth, Kirill Gringauz, Yutong Zhou, and Georg Groh. IFAN: An explainability-focused interaction framework for humans and NLP models. In *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*, Bali, Indonesia, 2023.

[Ocampo *et al.*, 2023] Nicolás Benjamín Ocampo, Ekaterina Sviridova, Elena Cabrio, and Serena Villata. An in-depth analysis of implicit and subtle hate speech messages. In *EACL 2023-17th Conference of the European Chapter of the Association for Computational Linguistics*, volume 2023, pages 1997–2013. Association for Computational Linguistics, 2023.

[Ranasinghe and Zampieri, 2021] Tharindu Ranasinghe and Marcos Zampieri. MUDES: Multilingual detection of offensive spans. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Demonstrations*, Online, 2021.

[Sap *et al.*, 2020] Maarten Sap, Saadia Gabriel, Lianhui Qin, Dan Jurafsky, Noah A. Smith, and Yejin Choi. Social bias frames: Reasoning about social and power implications of language. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, Online, 2020.

[Tillmann *et al.*, 2023] Christoph Tillmann, Aashka Trivedi, Sara Rosenthal, Santosh Borse, Rong Zhang, Avirup Sil, and Bishwaranjan Bhattacharjee. Muted: Multilingual targeted offensive speech identification and visualization. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, Singapore, 2023.

[Vidgen *et al.*, 2021] Bertie Vidgen, Tristan Thrush, Zeerak Waseem, and Douwe Kiela. Learning from the worst: Dynamically generated datasets to improve online hate detection. In *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)*, Online, 2021.

[Wijanarko *et al.*, 2024] Musa Izzanardi Wijanarko, Lucky Susanto, Prasetia Anugrah Pratama, Ika Karlina Idris, Traci Hong, and Derry Tanti Wijaya. Monitoring hate speech in Indonesia: An NLP-based classification of social media texts. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, Miami, Florida, USA, 2024.

[Wright *et al.*, 2021] Austin P. Wright, Omar Shaikh, Haekyu Park, Will Epperson, Muhammed Ahmed, Stephane Pinel, Duen Horng (Polo) Chau, and Diyi Yang. Recast: Enabling user recourse and interpretability of toxicity detection models with interactive visualization. 2021.

[Zheng *et al.*, 2023] Yi Zheng, Björn Ross, and Walid Magdy. What makes good counterspeech? a comparison of generation approaches and evaluation metrics. In *Proceedings of the 1st Workshop on CounterSpeech for Online Abuse (CS4OA)*, Prague, Czechia, 2023.