Hate Explained: Evaluating NER-Enriched Text in Human and Machine Moderation of Hate Speech

Andrés Carvallo[1*], **Marcelo Mendoza**[1,2], **Miguel Fernández**[1,2], **Maximiliano Ojeda**[2], **Lilly Guevara**[3], **Diego Varela**[3], **Martín Bórquez**[2], **Nicolás Buzeta**[2], **Felipe Ayala**[3]
[1]National Center for Artificial Intelligence (CENIA)
[2]Pontificia Universidad Católica de Chile
[3]Universidad Técnica Federico Santa María

Abstract

Hate speech detection is vital for creating safe online environments, as harmful content can drive social polarization. This study explores the impact of enriching text with intent and group tags on machine performance and human moderation workflows. For machine performance, we enriched text with intent and group tags to train hate speech classifiers. Intent tags were the most effective, achieving state-of-the-art F1-score improvements on the IHC, SBIC, and DH datasets, respectively. Cross-dataset evaluations further demonstrated the superior generalization of intent-tagged models compared to other pre-trained approaches. Then, through a user study (N=100), we evaluated seven moderation settings, including intent tags, group tags, model probabilities, and randomized counterparts. Intent annotations significantly improved the accuracy of the moderators, allowing them to outperform machine classifiers by 12.9%. Moderators also rated intent tags as the most useful explanation tool, with a 41% increase in perceived helpfulness over the control group. Our findings demonstrate that intent-based annotations enhance both machine classification performance and human moderation workflows.

1 Introduction

Warning: *This paper contains content that may be offensive or upsetting.*

Social media platforms face persistent challenges in moderating harmful content, such as hate speech, which violates community guidelines and poses significant risks to user safety. Although automated systems are critical for detecting policy violations, ambiguous cases and flagged content often require human moderators to make final decisions (Leo et al., 2023). Given the overwhelming volume of content that requires review, improving the efficacy of both human and machine moderation processes is essential (Halevy et al., 2022).

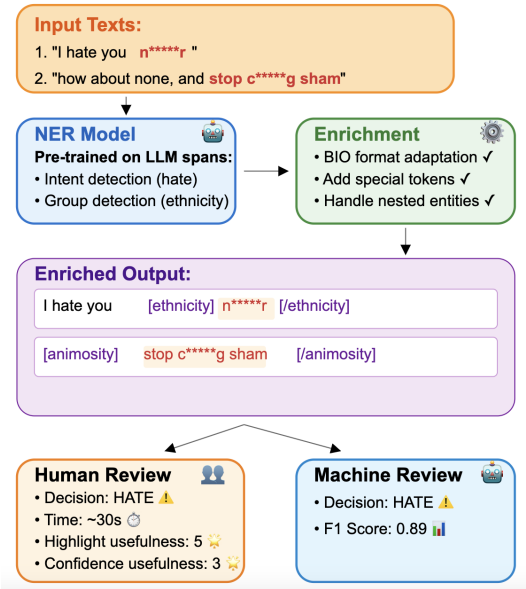


Figure 1: Input texts are processed by a named entity recognition model (NER), pre-trained on LLM-generated spans for intent and group detection. The outputs are enriched with format adaptations, special tokens, and nested entity handling. Enriched texts are reviewed by both human moderators and classifiers.

Recent advances in hate speech detection have highlighted the potential of annotations to improve both performance and model explainability (Calabrese et al., 2022; Mosca et al., 2021; MacAvaney et al., 2019). Annotations highlighting key elements of hate speech, such as intent and group mentions, offer a promising direction for improving detection systems. However, **their specific impact on machine classifiers and human moderation workflows remains underexplored**. Addressing this gap is essential to developing tools that are not only effective but also interpretable for end-users.

To address this gap, **we semantically enrich the text with intent and group annotations to evaluate their impact on hate speech detection** (see

* Corresponding author: andres.carvallo@cenia.cl

Proceedings of the The 9th Workshop on Online Abuse and Harms (WOAH), pages 458–467

Figure 1). Using a cross-dataset evaluation, our findings show that intent tags produce the most significant improvement, achieving state-of-the-art F1 score gains on the IHC (ElSherief et al., 2021), SBIC (Vidgen et al., 2021) and DH (Sap et al., 2020) datasets, respectively. These datasets, recognized benchmarks in the field, address implicit hate, social bias, and power implications in language and dynamically generated hate speech content. Cross-dataset evaluations further demonstrated the superior generalization of intent-tagged models compared to other pre-trained approaches.

To further validate these findings, we conducted a user study comparing seven experimental settings, including configurations with intent tags, group tags, model uncertainty, and randomized counterparts. The results show that intent annotations significantly improve the accuracy of human moderators, allowing them to outperform machine classifiers by 12.9%. Moderators also perceived intent tags as the most helpful source of explanation, with a 41% increase in perceived help compared to the control group. This work makes the following contributions.

- We demonstrate that intent tags improve the performance of machine classifiers, achieving state-of-the-art results on benchmark datasets.
- We conducted a user study (N=100) showing that intent annotations significantly enhance human moderation accuracy and are perceived as the most helpful explanation.
- We directly compare human and machine performance, highlighting scenarios where moderators augmented with enriched spans outperform automated systems.
- To support reproducibility and further research, enriched datasets, code, and trained models will be publicly available [1].

The remainder of this paper is organized as follows. Section 2 reviews related work. Section 3 describes the study design. Section 4 describes the methodology of the study. Section 5 presents the results. Section 6 concludes with a summary of contributions and future research directions.

[1]The model weights and the enriched datasets are available; however, the links have not been included to comply with the double-blind review process. We release the codes and prompts in https://anonymous.4open.science/r/hate_speech_enrichment-6A87/README.md

2 Related Work

The explainability to detect hate speech has been studied (MacAvaney et al., 2019; Mosca et al., 2021; Siddiqui et al., 2024; Sridhar and Yang, 2022; Zhou et al., 2023; Yadav et al., 2024). Kim et al. (2022a) introduced Masked Rationale Prediction (MRP) to improve detection by predicting human rationales, improving bias mitigation. Mittal and Singh (2023) proposed explainable models such as KTrain to improve the interpretability of classifiers. Calabrese et al. (2024) investigated the usage of tags for eXplainable AI (XAI), focusing primarily on human-side evaluation with user validation (N=25). Our study expands on this by incorporating a larger-scale assessment of human and machine moderators.

Open-source dataset initiatives such as PLEAD (Calabrese et al., 2022) have provided intent annotations. In contrast, ToxyGen (Hartvigsen et al., 2022) addresses adversarial hate speech but lacks the intent and group-span annotations essential for subtle moderation. Wang et al. (2023) highlighted the risks of misinterpreting explanations in moderation, which we address by integrating specific NER tags for intents and groups. Recent methods such as ConPrompt's contrastive learning approach (Kim et al., 2023) and HARE's LLM-based reasoning (Yang et al., 2023) achieve competitive results in automated hate speech detection. Our work extends these approaches by enriching text with intent and group tags and further differentiates them by directly evaluating their effectiveness in improving human moderation.

Beyond hate speech, explainability has been explored in other domains. In healthcare, prior work applied explainable and active learning approaches for document screening and evidence-based text classification (Carvallo et al., 2020a,b, 2023b). In education, explainable NLP techniques have supported moral discourse analysis and peer influence modeling (Alvarez et al., 2021; Álvarez et al., 2023). In low-resource machine translation, explainability and data curation strategies were used to support indigenous language processing (Pendas et al., 2023; Carvallo et al., 2023a). Additionally, recent work has proposed enriching hate speech classification using named entity tags for identity groups (Carvallo et al., 2024), and visualization strategies leveraging attention weights in transformers have also been explored for text classification transparency (Parra et al., 2019).

3 Study Design

This work aims to evaluate two key objectives. First, we examine whether enriching text with tags, such as intents or mentioned groups, improves the performance of hate speech classifiers. Following Röttger et al. (2021), we work with intent tags that capture the motivations behind the statements, including derogation, threats, hate crimes, comparisons, and animosity. In contrast, group tags identify mentions of demographic or social groups, ranging from neutral descriptors to pejorative terms that denigrate or dehumanize [2]. Second, we investigate whether these tags help human moderators identify hate speech. This involves assessing the usefulness of enriched text for distinguishing hate from non-hate and determining which types of tags most effectively support moderation along with other variables, such as the model's certainty.

4 Methodology

4.1 Model Implementation

Our study follows a two-stage pipeline for hate speech detection, as illustrated in Figure 1. The process integrates a text enrichment phase in which the input text is annotated with tags.

In the first stage, GPT-4o generates intent and group tags for the datasets (train partitions) based on prompts that include annotation guidelines [3]. Then, these tags were used to train RoBERTa-large-based NER models (Liu et al., 2019). NER models identify tags in text to create an enriched version of each dataset. We define functions to adapt the format and handle nested entities.

In the second stage, we fine-tune HateBERT models (Caselli et al., 2021) using each enriched train dataset as input. The complete process can be formalized as follows.

$$\hat{y}i = h_\theta(g(x_i, f_{NER}(x_i; \theta_{NER})); \theta_h) \quad (1)$$

where x_i is the original text input, f_{NER} is the NER model with parameters θ_{NER} trained on LLM-generated tags, g is the enrichment function that combines the original text with the identified tags, and h_θ is the HateBERT classifier with parameters θ_h that produces the final classification $\hat{y}_i$.

The NER model identifies and inserts intent and group tags into the text, enabling structured enrichment for improved classifier training.

[2] A detailed list of the intents and groups considered in this study can be reviewed in the appendices (see Table 4).

[3] We include the prompts and the annotation guideline used with GPT-4o in the appendices (see Sections A.1 and B.1)

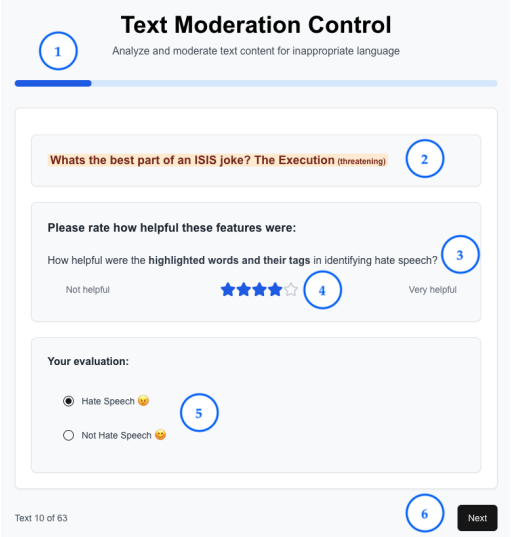


Figure 2: Interface of the text moderation control platform used in the user study: (1) a task progress bar, (2) a display of the text to be moderated, (3) a query regarding the user's perception of the utility of explanations, (4) a star rating system to evaluate the helpfulness of features, (5) options for classifying content as hate speech or not, and (6) a *Next* task button.

We evaluate the performance of our proposal against several state-of-the-art models: ConPrompt (Kim et al., 2023), fBERT (Sarkar et al., 2021), BERT (Devlin et al., 2019), and text enriched with offensive words of MUDES (Ranasinghe and Zampieri, 2021) and PLEAD intents (Calabrese et al., 2022). Our experiments were carried out on three well-known hate speech benchmark datasets considering the original partitions for train, validation, and testing: the Implicit Hate Corpus (IHC) (ElSherief et al., 2021); DynaHate (DH) (Vidgen et al., 2021); and SBIC (Social Bias Inference Corpus-Hate) (Kim et al., 2022b) [4].

4.2 User Study

This user study was designed to assess whether the inclusion of NER-based explanations (e.g., intent and group tags) improves moderation decisions made by humans, complementing our evaluation of their effect on machine learning models. In other words, we aimed to test if these explanations serve as helpful signals for both humans and machines, enhancing accuracy, efficiency, and perceived utility.

In the first stage of our user study, five senior moderators independently selected and curated a representative subset of the enriched IHC dataset.

After conducting curation without interference

[4] Data statistics are provided in the appendices (Table 3)

among moderators, the samples that achieved agreement from at least four out of the five moderators were included in the study.

Subsequently, 100 moderators were recruited via Clickworker, ensuring gender parity (male, female, non-binary), age diversity (18–60), and representation from major English-speaking countries (e.g., USA, UK, Australia). These participants, all native English speakers, were asked to assess the utility of tags spanning different settings. Using a factorial experimental design, each factor was isolated and compared with a randomly generated baseline in unifactorial experiments, with participants moderating around 63 examples each. The study included a highlighted group, intent, no highlights, random intent highlights, random group highlights, model-generated probabilities, and random probabilities, all implemented through a web moderation tool [5]. Statistical differences in performance, time spent and perceived usefulness between factors were analyzed. As shown in Figure 2, the moderation tool includes components that allow data collection for the study.

5 Results

5.1 Model Results

We evaluated the performance of hate speech classification models using F1 scores (mean and deviations) across five independent trials for each setting. To assess the generalization of these models, we performed cross-dataset evaluations. In each case, the model trained on a training partition (antecedent of a transfer setting) was evaluated on the corresponding testing partition (consequent of a transfer setting), following a structured approach to measure transfer learning.

As shown in Table 1, the results demonstrate that intent tag enrichment consistently outperformed other approaches. Intent-Tag surpassed the second best performing method in 8 of 9 evaluation settings, significantly improving cross-dataset tasks such as IHC → DH (10.5%), DH → IHC (7.5%), and DH → SBIC (8.2%). These results highlight the effectiveness of providing explicit semantic context through intent tags, enabling better generalization to unseen hate speech contexts. An exception was observed within SBIC → SBIC, where offensive word tags (MUDES-NER) slightly outperformed intent tags, reflecting the dataset's focus on explicit hate speech. This difference was not statistically significant.

[5]Details of the web moderation tool are included in the appendices (see Section B.2)

Model	**IHC → SBIC**	**IHC → DH**	**IHC → IHC**
BERT	62.0±0.01	54.3±0.01	75.3±0.003
HateBERT	56.6±0.01	51.7±0.003	72.9±0.004
fBERT	56.0±0.01	50.3±0.01	72.6±0.004
ConPrompt	67.0±0.02	60.0±0.01	76.0±0.011
PLEAD-NER	76.3±0.01	56.5±0.01	76.4±0.003
MUDES-NER	58.8±0.01	53.2±0.003	76.8±0.004
Group-Tag	57.0±0.01	53.7±0.01	61.1±0.004
Intent-Tag	**83.0±0.02***	**73.0±0.01***	**79.1±0.01***
Full-Tag	67.8±0.01	62.5±0.01	70.3±0.003
Gain	6.7	10.5	2.3
Model	**SBIC → IHC**	**SBIC → DH**	**SBIC → SBIC**
BERT	60.6±0.60	63.7±0.63	88.4±0.88
HateBERT	60.8±0.60	65.5±0.64	88.0±0.87
fBERT	58.2±0.58	64.2±0.63	88.0±0.88
ConPrompt	63.0±0.62	67.0±0.65	89.0±0.88
PLEAD-NER	65.8±0.60	65.0±0.63	88.6±0.88
MUDES-NER	63.5±0.60	65.0±0.64	**89.3±0.87**
Group-Tag	60.6±0.58	60.6±0.63	83.1±0.88
Intent-Tag	**70.3±0.62***	**70.5±0.65***	88.2±0.88
Full-Tag	65.9±0.60	65.3±0.63	83.2±0.88
Gain	4.4	3.5	-
Model	**DH → IHC**	**DH → SBIC**	**DH → DH**
BERT	65.2±0.003	75.6±0.01	74.4±0.003
HateBERT	64.4±0.002	74.0±0.003	75.8±0.004
fBERT	64.6±0.004	75.2±0.004	76.0±0.002
ConPrompt	66.0±0.002	76.0±0.003	77.0±0.002
PLEAD-NER	65.8±0.003	74.2±0.01	76.6±0.003
MUDES-NER	65.5±0.003	75.7±0.004	78.1±0.004
Group-Tag	65.3±0.004	65.0±0.004	70.9±0.002
Intent-Tag	**73.5±0.002***	**84.2±0.003***	**78.6±0.002***
Full-Tag	64.3±0.002	66.9±0.004	70.0±0.003
Gain	7.5	8.2	0.7

Table 1: Performance comparison of different models and text enrichment methods across datasets and evaluation settings using F1 scores in testing partitions. The symbol * indicates the statistically significant best result compared to the former for each setting using a t-test.

5.2 User Study Results

Setting	Human Accuracy†	Model Accuracy	Avg. Time† (s)	Confidence Rating	Highlight Rating†
AI Confidence	**0.71**	0.70	31.78	3.08	-
Random AI Conf.	0.67	0.70	33.67	**3.12**	-
Highlight Group	**0.68**	0.64	34.12	-	**3.46**
Random Group	0.64	0.70	39.52	-	2.54
Highlight Intent	**0.79**	0.70	33.33	-	**3.81**
Random Intent	0.66	0.70	38.28	-	2.69
No Highlights	0.67	0.70	**27.04**	-	-
p-value	$< .001$	-	$< .001$	0.99	$< .001$

Table 2: Performance across different visualization settings for hate speech detection. **Bold** values indicate the best performance compared to corresponding random counterparts. † indicates significant differences between settings ($p < .001$) using Welch's ANOVA.

Table 2 presents human moderator performance across settings. Model accuracy reflects classifier correctness, while human accuracy measures user performance. Highlight Rating (1–5) captures perceived usefulness of highlighted words, Confidence Rating (1–5) indicates perceived model certainty, and Average Time is the time (in seconds) spent per decision. **Highlight Intent** yielded the highest

human accuracy (79%), outperforming Random Intent (66%), No Highlights (67%), and the classifier (70%). **Highlight Group** also improved accuracy (68%) over Random Group (64%) and matched the classifier (64%). Perceived usefulness aligned with performance: Highlight Intent received the highest rating (3.81), followed by Highlight Group (3.46), both significantly above their random baselines (2.69 and 2.54, respectively). Confidence ratings were similar between AI Confidence (3.08) and Random AI Confidence (3.12), suggesting users could not reliably distinguish meaningful confidence signals. In terms of speed, No Highlights was fastest (27s), followed by Highlight Group (30s) and Highlight Intent (33s), which had the best performance despite the slight time increase. Overall, random highlights underperformed, emphasizing the importance of meaningful, span-based explanations in supporting effective moderation.

6 Conclusions

This study demonstrated that the use of intent tags significantly improves the detection of hate speech. The evaluation in benchmark data confirmed that intent tags boost classifier performance to the state of the art. A user study showed that these tags improve human accuracy and serve as valuable explanation tools for the moderation of hate speech. Future research could explore the extension of these annotations to multilingual contexts to broaden their effectiveness. In addition, efforts to incorporate human curation of intent tags provided by LLMs are essential to clarify the differences between human and AI data annotations. Evaluation benchmarks for Spanish sentence representations (Araujo et al., 2022), as well as lightweight Spanish language models such as ALBETO and DistilBETO (Cañete et al., 2022), provide a promising foundation for expanding this work beyond English-centric approaches. Furthermore, integrating entity-enriched hate speech detection into social network analysis pipelines—such as those enabled by toolkits like Tsundoku (Graells-Garrido et al., 2025)—could support broader applications in content moderation and network-level intervention strategies.

7 Limitations

Although our approach demonstrates promising results, three main limitations should be declared. First, our multistage approach (LLM → NER → Enrichment → Classification) means that errors can cascade through the system, with each stage potentially introducing its own uncertainties that propagate to subsequent steps. Second, while we used the most recent version of GPT-4o [6] to generate text spans, the rapid evolution of LLM means that the quality and nature of the generated intent and group tags could improve or change over time, potentially requiring periodic revalidation and updates to maintain optimal performance. In this regard, this study relies on the automatic annotations generated by a single LLM. This model was chosen because it has a prominent position on the LLM leaderboards for language understanding tasks [7]. However, it is important to examine the impact of using other LLMs on these results, determining whether the study's conclusions depend specifically on GPT-4o or are generalisable to other language models. Finally, another limitation lies in the fact that the user study was conducted solely on a dataset curated by moderators with annotations from IHC. This means that the conclusions of the study are not necessarily generalisable to other datasets such as DH and SBIC. However, we chose to focus on a human evaluation based on IHC because this dataset includes examples of both implicit and explicit hate speech, which we considered to provide a broader variety of hate speech examples compared to those included in DH and SBIC.

Acknowledgments

This work was supported by National Center for Artificial Intelligence CENIA FB210017, Basal ANID, Postdoctoral FONDECYT 3240001 and FONDECYT 1241462.

References

Claudio Álvarez, Gustavo Zurita, and Andrés Carvallo. 2023. Analyzing peer influence in ethical judgment: collaborative ranking in a case-based scenario. In *International Conference on Collaboration Technologies and Social Computing*, pages 19–35. Springer.

Claudio Alvarez, Gustavo Zurita, Andrés Carvallo, Pablo Ramírez, Eugenio Bravo, and Nelson Baloian. 2021. Automatic content analysis of student moral discourse in a collaborative learning activity. In *Collaboration Technologies and Social Computing: 27th International Conference, CollabTech 2021, Virtual Event, August 31–September 3, 2021, Proceedings 27*, pages 3–19. Springer.

[6] ChatGPT-4o-latest (2024-11-20)

[7] https://lmarena.ai/?leaderboard

Anastasios N. Angelopoulos and Stephen Bates. 2023. Conformal prediction: A gentle introduction. *Found. Trends Mach. Learn.*, 16(4):494–591.

Vladimir Araujo, Andrés Carvallo, Souvik Kundu, José Cañete, Marcelo Mendoza, Robert E Mercer, Felipe Bravo-Marquez, Marie-Francine Moens, and Alvaro Soto. 2022. Evaluation benchmarks for spanish sentence representations. *arXiv preprint arXiv:2204.07571*.

Agostina Calabrese, Leonardo Neves, Neil Shah, Maarten Bos, Björn Ross, Mirella Lapata, and Francesco Barbieri. 2024. Explainability and hate speech: Structured explanations make social media moderators faster. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 398–408, Bangkok, Thailand. Association for Computational Linguistics.

Agostina Calabrese, Björn Ross, and Mirella Lapata. 2022. Explainable abuse detection as intent classification and slot filling. *Transactions of the Association for Computational Linguistics*, 10:1440–1454.

José Cañete, Sebastian Donoso, Felipe Bravo-Marquez, Andrés Carvallo, and Vladimir Araujo. 2022. Albeto and distilbeto: Lightweight spanish language models. *arXiv preprint arXiv:2204.09145*.

Andrés Carvallo, Ignacio Jorquera, and Carlos Aspillaga. 2023a. Cotranslate: A web-based tool for crowdsourcing high-quality sentence pair corpora. *SoftwareX*, 23:101508.

Andres Carvallo, Denis Parra, Hans Lobel, and Alvaro Soto. 2020a. Automatic document screening of medical literature using word and text embeddings in an active learning setting. *Scientometrics*, 125(3):3047–3084.

Andres Carvallo, Denis Parra, Gabriel Rada, Daniel Pérez, Juan Ignacio Vasquez, and Camilo Vergara. 2020b. Neural language models for text classification in evidence-based medicine. *arXiv preprint arXiv:2012.00584*.

Andrés Carvallo, Tamara Quiroga, Carlos Aspillaga, and Marcelo Mendoza. 2024. Unveiling social media comments with a novel named entity recognition system for identity groups. *arXiv preprint arXiv:2405.13011*.

Andrés Carvallo, Matías Rojas, Carlos Muñoz-Castro, Claudio Aracena, Rodrigo Guerra, Benjamín Pizarro, and Jocelyn Dunstan. 2023b. Automatic section classification in spanish clinical narratives using chunked named entity recognition. In *IberLEF@ SEPLN*.

Tommaso Caselli, Valerio Basile, Jelena Mitrović, and Michael Granitzer. 2021. HateBERT: Retraining BERT for abusive language detection in English. In *Proceedings of the 5th Workshop on Online Abuse and Harms (WOAH 2021)*, pages 17–25, Online. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 4171–4186. Association for Computational Linguistics.

Mai ElSherief, Caleb Ziems, David Muchlinski, Vaishnavi Anupindi, Jordyn Seybolt, Munmun De Choudhury, and Diyi Yang. 2021. Latent hatred: A benchmark for understanding implicit hate speech. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 345–363, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Eduardo Graells-Garrido, Nicolás García, and Andrés Carvallo. 2025. Tsundoku: A python toolkit for social network analysis. *SoftwareX*, 29:102008.

Alon Halevy, Cristian Canton-Ferrer, Hao Ma, Umut Ozertem, Patrick Pantel, Marzieh Saeidi, Fabrizio Silvestri, and Ves Stoyanov. 2022. Preserving integrity in online social networks. *Commun. ACM*, 65(2):92–98.

Thomas Hartvigsen, Saadia Gabriel, Hamid Palangi, Maarten Sap, Dipankar Ray, and Ece Kamar. 2022. 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)*, pages 3309–3326, Dublin, Ireland. Association for Computational Linguistics.

Jiyun Kim, Byounghan Lee, and Kyung-Ah Sohn. 2022a. Why is it hate speech? masked rationale prediction for explainable hate speech detection. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 6644–6655, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Youngwook Kim, Shinwoo Park, and Yo-Sub Han. 2022b. Generalizable implicit hate speech detection using contrastive learning. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 6667–6679, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Youngwook Kim, Shinwoo Park, Youngsoo Namgoong, and Yo-Sub Han. 2023. ConPrompt: Pre-training a language model with machine-generated data for implicit hate speech detection. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 10964–10980, Singapore. Association for Computational Linguistics.

Chelsea Olivia Leo, B. J. Santoso, and B. Pratomo. 2023. Enhancing hate speech detection for social media moderation: A comparative analysis of machine learning algorithms. *2023 International Conference on Advanced Mechatronics, Intelligent Manufacture and Industrial Automation (ICAMIMIA)*, pages 960–964.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692.

Sean MacAvaney, Hao-Ren Yao, Eugene Yang, Katina Russell, Nazli Goharian, and O. Frieder. 2019. Hate speech detection: Challenges and solutions. *PLoS ONE*, 14.

D. Mittal and Harmeet Singh. 2023. Enhancing hate speech detection through explainable ai. *2023 3rd International Conference on Smart Data Intelligence (ICSMDI)*, pages 118–123.

Edoardo Mosca, Maximilian Wich, and Georg Groh. 2021. Understanding and interpreting the impact of user context in hate speech detection. In *Proceedings of the Ninth International Workshop on Natural Language Processing for Social Media*, pages 91–102, Online. Association for Computational Linguistics.

D Parra, H Valdivieso, A Carvallo, G Rada, K Verbert, and T Schreck. 2019. Analyzing the design space for visualizing neural attention in text classification. In *Proc. ieee vis workshop on vis x ai: 2nd workshop on visualization for ai explainability (visxai)*.

Begoña Pendas, Andrés Carvallo, and Carlos Aspillaga. 2023. Neural machine translation through active learning on low-resource languages: The case of spanish to mapudungun. In *Proceedings of the Workshop on Natural Language Processing for Indigenous Languages of the Americas (AmericasNLP)*, pages 6–11.

Tharindu Ranasinghe and Marcos Zampieri. 2021. 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*, pages 144–152, Online. Association for Computational Linguistics.

Paul Röttger, Bertie Vidgen, Dong Nguyen, Zeerak Waseem, Helen Margetts, and Janet Pierrehumbert. 2021. HateCheck: Functional tests for hate speech detection models. 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)*, pages 41–58, Online. Association for Computational Linguistics.

Maarten Sap, Saadia Gabriel, Lianhui Qin, Dan Jurafsky, Noah A. Smith, and Yejin Choi. 2020. Social bias frames: Reasoning about social and power implications of language. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5477–5490, Online. Association for Computational Linguistics.

Diptanu Sarkar, Marcos Zampieri, Tharindu Ranasinghe, and Alexander Ororbia. 2021. fBERT: A neural transformer for identifying offensive content. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 1792–1798, Punta Cana, Dominican Republic. Association for Computational Linguistics.

J. A. Siddiqui, S. Yuhaniz, Ghulam Mujtaba, Safdar Ali Soomro, and Zafar Ali Mahar. 2024. Fine-grained multilingual hate speech detection using explainable ai and transformers. *IEEE Access*, 12:143177–143192.

Rohit Sridhar and Diyi Yang. 2022. Explaining toxic text via knowledge enhanced text generation. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 811–826.

Bertie Vidgen, Tristan Thrush, Zeerak Waseem, and Douwe Kiela. 2021. 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)*, pages 1667–1682, Online. Association for Computational Linguistics.

Han Wang, Ming Shan Hee, Md Rabiul Awal, Kenny Tsu Wei Choo, and Roy Ka-Wei Lee. 2023. Evaluating GPT-3 generated explanations for hateful content moderation. In *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence*, IJCAI '23.

Neemesh Yadav, Sarah Masud, Vikram Goyal, Md Shad Akhtar, and Tanmoy Chakraborty. 2024. Tox-bart: Leveraging toxicity attributes for explanation generation of implicit hate speech. *arXiv preprint arXiv:2406.03953*.

Yongjin Yang, Joonkee Kim, Yujin Kim, Namgyu Ho, James Thorne, and Se-Young Yun. 2023. HARE: Explainable hate speech detection with step-by-step reasoning. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 5490–5505, Singapore. Association for Computational Linguistics.

Xuhui Zhou, Hao Zhu, Akhila Yerukola, Thomas Davidson, Jena D Hwang, Swabha Swayamdipta, and Maarten Sap. 2023. Cobra frames: Contextual reasoning about effects and harms of offensive statements. *arXiv preprint arXiv:2306.01985*.

SUPPLEMENTARY MATERIALS

APPENDICES

A Ethical considerations

This study involved professional moderators trained in the handling of hate speech, using posts from public datasets. No personal or demographic data were collected from the moderators to protect their privacy. All annotations were anonymized and no private user data was used. This ensures confidentiality and aligns with ethical standards while promoting transparency and reproducibility.

All participants in the study voluntarily agreed to participate, having been informed of the nature of the study and the use of the data for research purposes. Participants gave their informed consent. A data collection protocol was approved by the institutional ethics review board.

B Experimental Design

B.1 User Study Settings

This appendix provides a visual representation of all the configurations employed in the user study, illustrating the impact of various text annotations on moderator decision-making. Each setting is shown and described below to demonstrate how it affects the hate speech moderation process:

- **No Highlights**: Serves as the baseline setting where the text is presented without annotation. This configuration is visualized in Figure 3a and is used to estimate moderator performance without influence from additional environments.
- **Model Predicted Probability**: Displays texts with a model-generated probability score that indicates the likelihood of the text being hate speech. This setting, shown in Figure 3b, tests the utility of an automated model's certainty to the moderator to decide.
- **Intent Tags Highlight**: Involves texts annotated with tags that clarify the underlying intents of statements, such as threats or derogation. Figure 3c shows this setting to determine if identifying and visualizing intents can enhance moderation performance.
- **Group Tags Highlight**: Features texts highlighted with tags identifying group mentions or offensive denominations. Figure 3d shows an example of a highlighted group.

Additionally, control group settings utilize random probability values between 0 and 1 and random highlighted words, providing a comparative baseline to assess the effectiveness of each setting. Furthermore, the model utilized in the user study was the hate speech classifier enriched with NER intent tags on the IHC dataset along with the NER to tag groups and intents, which showed the best performance in the previous experiments (Table 1) and was calibrated using a conformal prediction method (Angelopoulos and Bates, 2023), ensuring that the probabilities provided to users were well-estimated.

Annotation Guidelines

The annotation guide used in this study provides detailed instructions and criteria essential for ensuring the consistency and reliability of hate speech annotations. Following Röttger et al. (2021), the annotation guidelines define intent tags that capture the motivations behind hate speech examples, including derogation, threats, hate crimes, comparisons, and animosity. They also include group tags that identify mentions of demographic or social groups, ranging from neutral descriptors to pejorative terms that denigrate or dehumanise. The guidelines establish general annotation rules, as well as specific rules for annotating groups, offensive denominations, hate speech intents, and offensive words. The general rules include inclusion and exclusion criteria, with examples illustrating their application.
The annotation guidelines were used to guide human moderators during the user study. Additionally, the same guidelines were employed to define the prompts used with GPT-4o, which served as the basis for annotating the datasets. The guide is available at:

`https://anonymous.4open.science/w/hate-speech-annotation-F8DE/`

B.2 Model settings

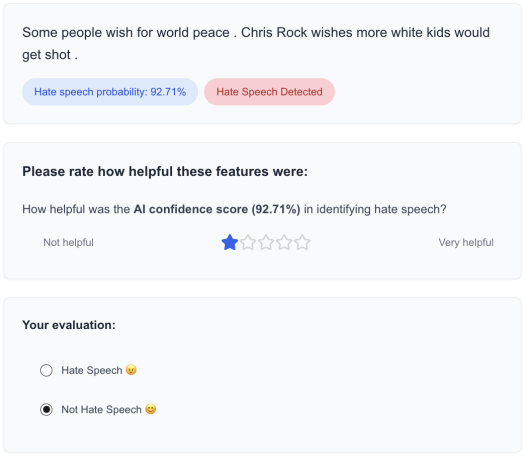
The training protocol adheres to established practices for reliability and reproducibility. Models are fine-tuned for six epochs with a batch size of 8, and a learning rate search is conducted over 5e-6, 1e-5, 2e-5, 3e-5, 5e-5, following Kim et al. (Kim et al., 2022b). Macro F1 scores on test sets are reported, averaging results across five fine-tuned models initialized with different random seeds (0, 1, 2, 3, 4). All experiments were performed on an NVIDIA A40 GPU, ensuring sufficient computational resources for training and evaluation. Concerning LLM used to generate named entity recognition tags to enrich datasets for classification, we used GPT-4o-latest (2024-11-20).

C SOFTWARE

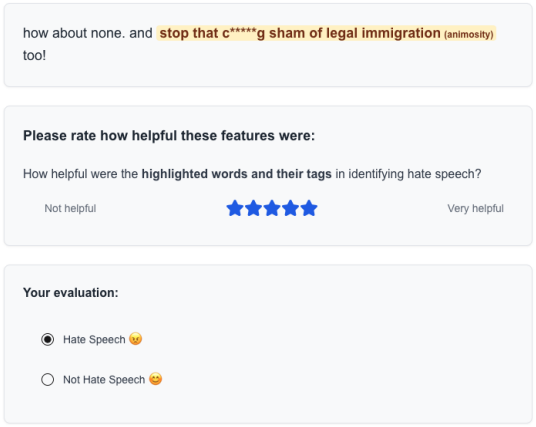
C.1 Source code

The source code related to Named Entity Recognition (NER) training and Hate speech

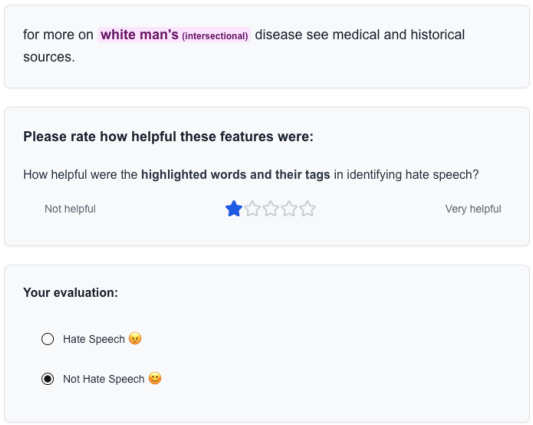
(a) No highlights



(b) Model confidence without highlights



(c) Highlighted Intent tags



(d) Highlighted group tags

Figure 3: Comparison of different text annotation settings in the text moderation platform: (a) No highlights, (b) Model confidence without highlights, (c) Highlighted intent tags, and (d) Highlighted group tags.

classifier, including custom prompts designed for identifying specific intents and groups, is available in our project repository: https://anonymous.4open.science/r/hate_speech_enrichment-6A87/README.md

C.2 Web application

The user study was conducted using a web application developed with ReactJS for a user-friendly interface. The application was containerized with Docker, ensuring consistent deployment and encapsulating all dependencies. It was hosted on an SSH server, providing secure remote access to participants. All responses were securely stored on the server hosting the Docker container, ensuring data privacy and efficient collection.

D DATA

This section shows the datasets used for training and evaluating the hate speech classification models. Table 3 presents the distribution of train, validation, and test sets across the three benchmark datasets utilized in this work.

Dataset	Train Set	Validation Set	Test Set
IHC	11,199	3,733	3,734
SBIC	29,422	3,948	3,978
DH	33,006	4,125	4,124

Table 3: Statistics of the datasets used for model fine-tuning and evaluation.

Table 4 presents the classification of hate speech intents, group mentions, and offensive denominations, along with their respective definitions.

Type	Group	Definition
Hate Speech Intent	Derogation	Statements intended to belittle or demean.
	Threat	Expressions of intent to cause harm.
	Hate Crime	Incitement to criminal acts motivated by hate.
	Comparison	Negative comparisons between groups.
	Animosity	General hostility toward a group.
Offensive Denomination	Ethnicity	Slurs or derogatory terms based on ethnicity.
	Religion	Offensive remarks targeting religious beliefs.
	Gender	Derogatory terms aimed at gender identity.
	Sexual Orientation	Terms attacking sexual identity.
	Disability	Mocking or demeaning disabilities.
	Working Class	Insults based on socioeconomic status.
	Ideological Group	Attacks on political or social ideologies.
	Intersectional	Combines multiple identity aspects.
Group Mention	Ethnicity	Neutral or factual mentions of ethnicity.
	Religion	Neutral mentions of religious groups.
	Gender	Neutral mentions of gender groups.
	Sexual Orientation	Neutral mentions of sexual identity.
	Disability	References to disabilities.
	Working Class	Mentions of socioeconomic groups.
	Ideological Group	Factual mentions of ideologies.
	Intersectional	Combines multiple identity aspects.

Table 4: Categories of hate speech intents and groups.