Explainable AI for Ethical Counter Speech Generation in Hate Speech Mitigation

Ashiful Islam Ridoy, Mohammed Faisal, Yogesh Kumar, Md Mamun-Ur Rashid, Marina Ernst, Frank Hopfgartner
University of Koblenz, Germany

{ashifulridoy, mohammed_faisal, yogesh-kumar, rashid, marinaernst, hopfgartner}@uni-koblenz.de

Abstract

The proliferation of hate speech in digital communication platforms poses significant challenges to online safety and social cohesion. While automated hate speech detection systems have shown promise, their black-box nature limits user trust and understanding of AI-driven content moderation decisions. This paper presents a framework that integrates explainable AI (XAI) techniques with counter-speech generation to create transparent, ethical solutions for hate speech mitigation. Our approach combines a fine-tuned HateBERT model, with a specialized Llama 3.1-8B-Instruct model for generating empathetic counter-narratives. The system employs five distinct XAI methods: Integrated Gradients, Attention Visualization, LIME, Counterfactual Analysis, and Natural Language Explanations to provide interpretable reasoning behind both detection and response generation decisions. The integration of explainability mechanisms with counter-speech generation represents a novel contribution to ethical AI systems, fostering transparency and trust in automated hate speech mitigation while maintaining high performance standards for real-world deployment.

Keywords: Explainable AI, Hate Speech Detection, Counter Speech Generation, Ethical AI, Natural Language Processing

1. Introduction

The rapid growth of online communication platforms has transformed the way people interact and exchange information on a global scale. However, this digital revolution has also fuelled the spread of hate speech, creating major challenges in maintaining inclusive and safe online spaces. Hate speech is defined as threatening or abusive language directed at people or groups because of protected traits like sexual orientation, gender, race, or religion (United Nations). Maintaining inclusive and safe online spaces under this immense pressure has become one of the most pressing Human-Computer Interaction (HCI) challenges of the digital age (Castaño-Pulgarín et al., 2021).

Traditional approaches to hate speech mitigation rely heavily on content removal and user suspension, strategies that often prove inadequate for addressing the underlying attitudes and beliefs that fuel hateful discourse (Hangartner et al., 2021). These reactive actions miss chances to turn hostile encounters into meaningful discussions that could lessen prejudice and advance understanding because they ignore the transformative and educational potential of constructive dialogue.

Recent advances in natural language processing have enabled the development of sophisticated automated systems for detecting hate speech with increasing accuracy (Jahan and Oussalah, 2023; Zhao et al., 2022). However, the majority of current solutions function as black-boxes making classification decisions without offering clear justification for why particular content is marked as problematic (Nirmal et al., 2024). This opacity reduces the educational value of moderation, discourages users from accepting it, and raises concerns about algorithmic bias and the fairness of content decisions.

Instead of using censorship, counter-speech generation is a promising alternative strategy that aims to counter hate speech through constructive engagement. By responding to hateful content with empathetic, fact-based counter-narratives, these systems can potentially de-escalate conflicts, correct misinformation, and model respectful discourse (Bonaldi et al., 2024). However, generating effective counter-speech requires a nuanced understanding of context, audience, and persuasive communication strategies that current automated systems struggle to achieve consistently.

Integrating explainable AI (XAI) with counter-speech offers a novel solution by providing transparent explanations for both hate speech detection and the reasoning behind generated responses. This approach enhances user trust, supports learning about harmful discourse, and enables more targeted counter-narratives that address specific problematic elements identified in the analysis.

This paper presents a framework for explainable AI-driven counter-speech generation that combines state-of-the-art hate speech detection with interpretable response generation. Our approach leverages multiple complementary XAI methods to

Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026), pages 2104–2114
11–16 May 2026.

provide diverse perspectives on model decisions, enabling both human understanding and informed counter-speech synthesis. Through detailed evaluation across multiple dimensions we demonstrate the effectiveness of this integrated approach for ethical hate speech mitigation.

The remainder of this paper is structured as follows. Section 2 reviews the related work on hate speech mitigation approaches, including the state-of-the-art methods of detection, as well as explainable AI techniques. Section 3 describes the proposed methodology, including details on the datasets used, model configuration, and evaluation procedures. Section 4 presents and discusses the experimental results. Section 5 outlines future research directions and concludes the paper.

2. Related Works

2.1. Hate Speech Detection and Classification

The automated detection of hate speech has evolved significantly over the past decade, progressing from simple keyword-based filtering to complex deep learning approaches capable of understanding contextual nuances and implicit meanings. Early work by Davidson et al. (2017) established important benchmarks for hate speech classification using traditional machine learning approaches combined with linguistic features and lexical resources. Some RRN methods, such as LSTM (Long Short-Term Memory) model (Mehta and Passi, 2022), have archived promising results in detecting hate speech across various benchmarks. However, these methods struggled with the contextual complexity and evolving nature of hateful discourse.

The introduction of transformer-based architectures revolutionized hate speech detection capabilities. Founta et al. (2018) demonstrated that BERT-based models could achieve improved performance by leveraging pre-trained contextual representations. For instance, BERT along with MPNet model have demonstrated high detection accuracy in study by Clarke et al. (2023). Caselli et al. (2021) developed HateBERT, a specialized variant pre-trained on offensive language data that showed outstanding performance on hate speech detection tasks compared to general-purpose language models. Rise of LLMs contributed to the hate speech detection and explanation. HARE, a detection framework introduced by Yang et al. (2023), uses the reasoning capabilities of LLMs to address the lack of explanation in hate speech detection, thus enabling effective supervision of detection models.

Recent advances have focused on addressing specific challenges in hate speech detection, such as dataset bias, cross-platform generalization, and multilingual capabilities. Hartvigsen et al. (2022) introduced the ToxiGen dataset to address synthetic data generation for hate speech detection, while Vidgen et al. (2021) proposed evaluation frameworks that assess model performance across different demographic groups and hate speech categories.

2.2. Counter-Speech Generation and Dialogue Systems

Counter-speech generation emerged as a proactive approach to hate speech mitigation, drawing inspiration from research in persuasive communication and conflict resolution. The CONAN dataset (Chung et al., 2019) provided the first large-scale collection of human-generated counter-narratives paired with hate speech examples, establishing benchmarks for automated counter-speech generation systems.

Early computational approaches to counter-speech generation relied on template-based systems and rule-driven response selection (Qian et al., 2019). While these methods could produce contextually relevant responses, they lacked the flexibility and authenticity required for engaging with diverse forms of hate speech. The advent of large language models enabled more sophisticated generation capabilities, with researchers exploring fine-tuning approaches for creating empathetic, fact-based counter-responses (Tekiroğlu et al., 2020).

Recent work has emphasized the importance of response quality beyond relevance, incorporating measures of empathy, factual accuracy, and persuasive effectiveness (Zheng et al., 2023). Mathew et al. (2018) developed evaluation frameworks for counter-speech that consider both automatic metrics and human assessments of response quality. Their work highlighted the challenges of balancing non-confrontational approaches with effective persuasion in counter-narrative generation.

2.3. Explainable AI in Natural Language Processing

The rapid adoption of deep learning models in natural language processing applications has sparked significant interest in explainability techniques that can illuminate the reasoning behind model decisions (Hassan et al., 2025). Attention visualization, one of the earliest XAI approaches for NLP, leverages the attention mechanisms inherent in transformer architectures to identify important input tokens (Clark et al., 2019). However, research has shown that attention weights do not always correspond to model reasoning and may provide misleading explanations (Jain and Wallace, 2019).

Gradient-based attribution methods offer alternative approaches to understanding model behavior. Integrated Gradients (Sundararajan et al., 2017) computes feature importance by integrating gradients along paths from baseline inputs to actual inputs, providing theoretically grounded attributions that satisfy important axioms, including sensitivity and implementation invariance. LIME (Ribeiro et al., 2016) takes a different approach by training local surrogate models to approximate behaviour in the neighbourhood of specific instances, offering model-agnostic explanations that can be applied to any black-box system.

The application of XAI techniques to hate speech detection has received limited attention in the literature. Mollas et al. (2020) conducted preliminary investigations of attention-based explanations for abusive language detection, while Gencoglu (2021) explored the use of LIME for understanding toxic comment classification. However, frameworks that integrate multiple XAI approaches specifically for hate speech analysis remain largely unexplored.

3. Methodology

Using a three-stage pipeline, our explainable AI system for ethical counter-speech generation generates suitable counter-responses while offering transparent, thorough analysis of hate speech content. Hate speech recognition, multi-modal explanation generation, and XAI-informed counter-speech synthesis are all integrated into the system architecture to create a coherent framework that puts efficacy and interpretability first. An overview of the proposed methodology is presented in Figure 1. To ensure full transparency and reproducibility, we make the codebase of our experiments available[1].

A fine-tuned HateBERT model is used in the first stage to classify input text as either toxic or non-toxic. This model, specialized for hate speech detection through domain-specific pre-training and fine-tuning on the ToxiGen (Hartvigsen et al., 2022) dataset, serves as the foundation for subsequent analysis stages. The second stage generates explanations using five complementary XAI methods: Integrated Gradients for gradient-based attribution, Attention Visualization for transformer-specific insights, LIME for local interpretability, Counterfactual Analysis for decision boundary exploration, and Natural Language Explanations for human-readable synthesis. The final stage leverages these explanations to inform counter-speech generation using a fine-tuned Llama 3.1-8B-Instruct model optimized for producing empathetic, constructive responses.

[1]The code is available at https://github.com/AshifulRidoy/XAI-on-HateBERT/tree/main

3.1. Datasets

In this study, two datasets are used: ToxiGen (Hartvigsen et al., 2022) is utilised for the first stage of pipeline to fine-tune HateBERT, and the CONAN (Chung et al., 2019) dataset to enhance counter speech generation.

ToxiGen contains over 274,000 machine-generated and human-annotated samples covering diverse demographic groups and hate speech categories. The dataset was created using a large pretrained language model (GPT-3) via two main techniques: Demonstration-based prompting and ALICE (Adversarial Language Imitation with Constrained Exemplars). Those techniques enabled human-like quality in synthetic data. The dataset is balanced, containing approximately equal numbers of toxic and benign statements. The ToxiGen dataset offers diversity, encompassing statements related to 13 minority identity groups (e.g., African Americans, women, LGBTQ+ individuals). This diversity helps mitigate bias in hate speech detection systems, which often produce false positives when texts merely mention minority groups, that are frequently targeted in online hate.

The CONAN (COunter NArratives through Nichesourcing) provides a reliable resource for research on automatically generating effective counter-narratives to fight online hate speech. Dataset contains of 4,078 original hate speech/counter-narrative pairs on three languages: 1,288 for English, 1,719 for French, 1,071 for Italian. The counter-narratives were written by over 100 trained experts in opposing online hate speech and creating non-offensive, fact-based responses to de-escalate conversation.

3.2. Hate Speech Detection Model

HateBERT (Caselli et al., 2021) is chosen as a foundation model due to its specialized pre-training on offensive language data and demonstrates high performance on hate speech detection tasks. The model employs a BERT-based transformer architecture with 110 million parameters, optimized for processing sequences up to 96 tokens in length. This sequence length constraint balances computational efficiency with adequate context coverage for most hate speech instances.

Fine-tuning was performed on the ToxiGen dataset, describe in details in Subsection 3.1. To guarantee the integrity of the training data, the dataset underwent extensive preparation: text normalisation, language detection filtering, and quality validation. In order to preserve data quality, our preprocessing pipeline processes emoji using text descriptions, transforms URLs and user mentions to standardised tokens, and implements stringent

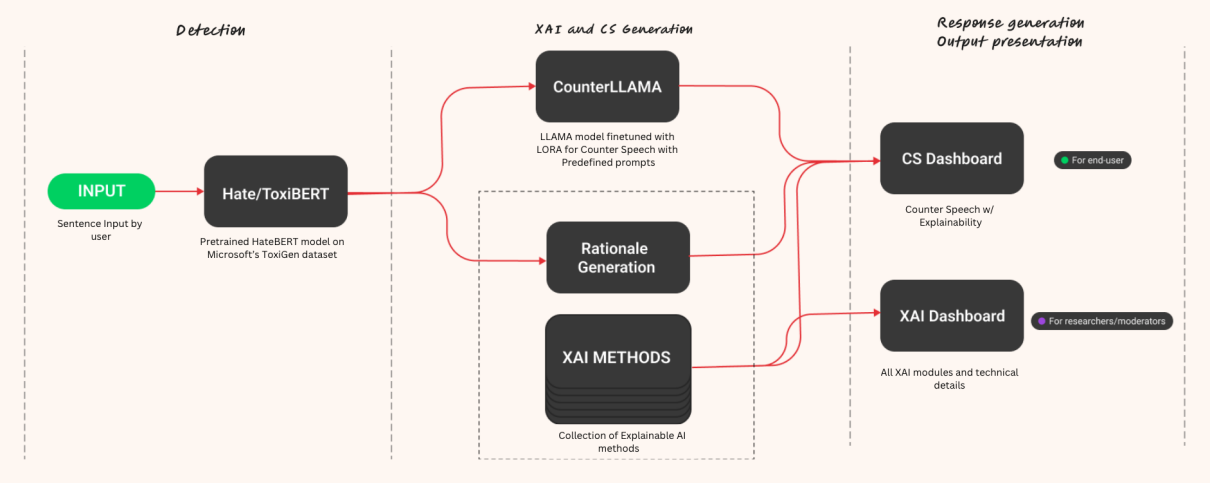

Figure 1: System architecture overview showing the three-stage pipeline: (1) HateBERT-based hate speech detection, (2) Multi-modal XAI explanation generation using five complementary methods, and (3) Llama 3.1-8B-based counter-speech generation informed by XAI insights.

length and content screening.

Advanced optimisation approaches, such as gradient clipping for stability, weighted sampling for balanced representation, and focal loss are used during training. Using AdamW optimisation (Loshchilov and Hutter, 2017), the model is trained for three epochs at a learning rate of 2e-5. on NVIDIA H100 hardware, it achieved convergence in 17.3 minutes while retaining computational efficiency appropriate for production deployment.

3.3. Multi-Modal XAI Framework

The framework implements the three complementary gradient-based attribution methods identified in Section 2.3 to capture different aspects of token importance in hate speech detection decisions.

Integrated Gradients computes attribution scores by integrating gradients along paths from baseline inputs (zero embeddings) to actual inputs, providing theoretically grounded explanations that satisfy sensitivity and implementation invariance axioms. The implementation uses 50 integration steps with Riemann approximation for computational efficiency while maintaining attribution quality.

Attention Visualization leverages the inherent attention mechanisms in the transformer architecture to identify tokens that receive high attention weights during classification. While attention weights do not always correspond directly to importance, they provide valuable insights into model behaviour and token interactions that complement gradient-based explanations.

LIME generates local linear approximations of model behaviour in the neighbourhood of specific instances. Our implementation uses perturbation strategies, including synonym replacement and token masking, to create local datasets for training interpretable surrogate models that approximate the complex decision boundaries of the HateBERT model.

Additionally, *counterfactual analysis* employs seven transformation strategies designed to generate minimal modifications that flip model predictions while preserving semantic coherence. Lexical transformations include hate word removal, neutral replacement, and language softening. Syntactic transformations encompass pronoun neutralization, uncertainty qualification, and statement negation. Pragmatic transformations involve context softening to add empathetic framing.

Counterfactual validation ensures that generated examples meet quality criteria, including minimal edit distance, semantic coherence, significant probability changes (minimum 0.1 threshold), and causal validity. Generated counterfactuals are ranked by impact magnitude weighted by semantic preservation scores to identify the most informative examples for explanation purposes.

3.4. Counter-Speech Generation System

Counter-speech generation employs a fine-tuned Llama 3.1-8B-Instruct model[2] optimized through Low-Rank Adaptation (LoRA) for parameter-efficient training. The LoRA configuration uses rank 64 with alpha 16, targeting attention projection layers (q_proj, k_proj, v_proj, o_proj) while applying 4-bit NF4 quantization for memory efficiency.

Training data preparation utilizes the CONAN dataset (Chung et al., 2019), describe in details Subsection 3.1. System prompt engineering in-

[2] https://huggingface.co/meta-llama/Llama-3.1-8B-Instruct

corporated specific guidelines for response quality, including de-escalation strategies, empathy demonstration, factual grounding, and length constraints (150-300 characters) suitable for social media deployment.

The supervised fine-tuning process employed gradient accumulation (effective batch size 16), mixed precision training (bfloat16), and careful hyperparameter tuning, including learning rate 1e-4, weight decay 0.01, and gradient clipping at 1.0 for training stability. This configuration achieved an optimal balance between response quality and computational efficiency.

A key contribution of this work is leveraging XAI insights to inform counter-speech generation strategies. Token attribution results guide attention to specific problematic elements that require addressing, while concept activation scores inform thematic response strategies. Targeted educational content is made possible by rule-based pattern recognition, while counterfactual examples offer templates for illustrating alternative phrasings.

In order to enable the model to generate responses that directly address identified hate speech elements, the integration process uses dynamic prompt engineering, which integrates explanation insights into generation prompts. Compared to generic response generation methods, this XAI-informed approach produces more relevant and targeted counter-speech.

Using a multi-layer framework that ensures ethical standards, generated counter-speech is thoroughly validated for safety. Using the same HateBERT model, primary validation reassesses generated responses with a lower threshold (0.3) for increased sensitivity. Content filtering identifies offensive language or aggressive tendencies, while sentiment analysis guarantees a neutral or positive tone (minimum threshold 0.1).

3.5. Evaluation Framework

Hate speech detector based on HateBERT is evaluated with standard metrics: Precision, Recall and F1-score. We further compare our fine-tuned HateBERT with the state-of-the-art models used for hate speech detection: LSTM (Mehta and Passi, 2022), HARE (Yang et al., 2023), BERT and MPNet (Clarke et al., 2023)

The XAI methods are evaluated using metrics that capture multiple aspects of explanation quality. Measures of faithfulness include infidelity scores, which use perturbation analysis to assess the consistency between attributions and model behaviour, and deletion/insertion AUC (Opitz, 2024), which quantifies how model confidence changes when significant tokens are added or removed.

Localization accuracy is assessed through span-based evaluation comparing explanation highlights with human-annotated toxic content regions, using precision, recall, and F1-metrics to quantify alignment between computational and human understanding of hate speech indicators.

Method-specific metrics include LIME stability and fidelity scores, attention visualization coherence measures, and counterfactual validity assessments, ensuring generated examples meet quality criteria for meaningful interpretation.

Counter-speech evaluation combines automatic metrics with quality assessment. Semantic quality is measured using BERTScore (Zhang et al., 2019) for semantic similarity, BLEU scores (Papineni et al., 2002) for lexical overlap, and ROUGE metrics (Lin, 2004) for content preservation compared to reference counter-narratives from the CONAN dataset.

Diversity and creativity metrics include distinct-n scores measuring unique n-gram ratios and overall diversity calculations to ensure generated responses avoid repetitive patterns (Tevet and Berant, 2021). Safety validation tracks pass rates through the multi-layer safety framework and analyses failure modes to identify potential risks.

4. Results & Discussion

4.1. Hate Speech Detection Performance

The fine-tuned HateBERT model demonstrated exceptional performance in hate speech classification. Detailed metrics are shown in Table 1. With an overall accuracy of 89.2%, the refined HateBERT model demonstrated remarkable performance, exhibiting balanced precision (89.6%) and recall (89.2%) across both toxic and non-toxic content categories. The model significantly outperformed baseline approaches while maintaining balanced precision-recall characteristics essential for practical content moderation applications (Figure 2).

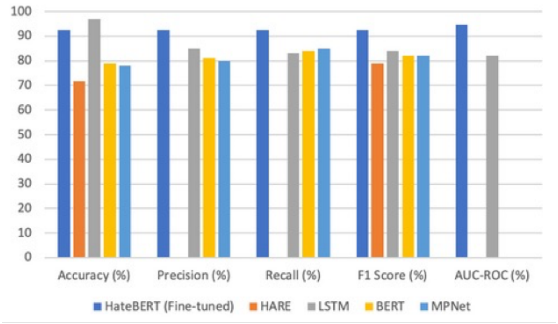

Figure 2: HateBERT results visualization showing performance comparisons with baseline methods HARE (Yang et al., 2023), LSTM (Mehta and Passi, 2022), BERT and MPNet (Clarke et al., 2023)

Advanced performance validation included cal-

ibration assessment with a Brier score of 0.081, indicating well-calibrated probability estimates crucial for confidence-based applications. ROC analysis, shown in Figure 3 with an AUC of 0.9625, indicating excellent discriminative capability. With 95% CIs verifying significant improvements over baseline methods exceeding 37 percentage points, bootstrap analysis was used to confirm statistical robustness.

Table 1: HateBERT Classification Performance

Class	Precision	Recall	F1
Toxic	0.936	0.848	0.890
Non-toxic	0.854	0.938	0.894
Macro Avg	0.895	0.893	0.892
Weigh. Avg	0.896	0.892	0.892

4.2. Explainable AI Method Performance

The evaluation of XAI methods revealed significant performance variations across different explanation approaches and content types. Table 2 presents detailed faithfulness metrics across all methods and content categories.

Attention Visualization proved most effective for toxic content analysis, with deletion AUCs of 0.664 and 0.417 on explicit hate speech samples. This suggests transformer attention aligns well with hate speech detection, offering faithful explanations when offensive language is present. Integrated Gradients showed stable, moderate performance (AUCs 0.596 and 0.520), indicating reliable gradient-based attribution, though slightly less faithful than attention for explicit toxicity. LIME displayed high variability with AUC fluctuating in range form 0.313 to 0.580), with poor results on one toxic sample (0.313), highlighting limitations in capturing complex or subtle toxic patterns due to its linear surrogate model. All methods performed strongly on benign content with AUC over 0.88, demonstrating consistent explanation quality where decision boundaries are clear.

For infidelity Integrated Gradients got the lowest Integrated Gradients across all content types, ranging from 0.0004 to 0.012. These exceptionally low values indicate high consistency between explanation attributions and actual model behaviour, confirming that gradient-based explanations most faithfully represent the model's decision-making process. Natural Language Explanations, on the other had, demonstrated competitive infidelity, validating the synthesis approach for maintaining explanation faithfulness while improving human interpretability.

Counterfactual Analysis showed notably higher infidelity scores on toxic samples (0.015-0.060), indicating potential limitations in decision boundary approximation for complex hate speech patterns. This performance degradation suggests that generating meaningful minimal modifications to toxic text that flip model predictions while maintaining semantic coherence presents significant challenges. Attention Visualization achieved the most promising span-based performance with F1-scores of 0.5 and 0.471 on toxic samples, demonstrating perfect precision (1.0) but limited recall (0.33). This precision-focused behaviour indicates conservative but accurate identification of hate speech elements, minimizing false positive flagging while ensuring that highlighted content is genuinely problematic.

Integrated Gradients showed zero span overlap with human-annotated toxic regions, revealing a critical limitation despite high faithfulness scores. This misalignment between gradient-based attributions and human intuition about hate speech localization highlights the distinction between model faithfulness and human interpretability.

Overall, the results highlight significant variability across XAI methods, metrics, and content types. Different techniques capture different aspects of model behaviour, and they may even yield conflicting interpretations. This variability is particularly important when explanations are used to inform counter-speech generation, emphasising the need for further robustness checks and validation.

4.3. Counter-Speech Generation Performance

The CounterLLAMA model demonstrated exceptional performance across all evaluation dimensions, significantly outperforming established baseline approaches.

We compare our proposed approach with the following baseline models: GPS by Zhu and Bhat (2021), Seq2Seg by Sutskever et al. (2014) and MMI by Li et al. (2016). Detailed metrics provided in Table 3.

BERTScore of 0.854 (85.4%) represents improvements over Seq2Seq (0.821), MMI (0.808), and GPS (0.839) models. ROUGE score equal to 0.117 confirmed CounterLLAMA's content preservation capabilities. This recall-oriented performance indicates more effective capture of essential semantic elements characterizing effective counter-speech.

The most dramatic performance difference was seen in linguistic diversity metrics, where CounterLLAMA outperformed baseline models with remarkable scores of 67.5%. The ability to produce diverse, non-repetitive counter-speech responses that steer clear of template-based patterns is demonstrated by this more than 90% improvement in linguistic variety.

The safety validation framework achieved decent

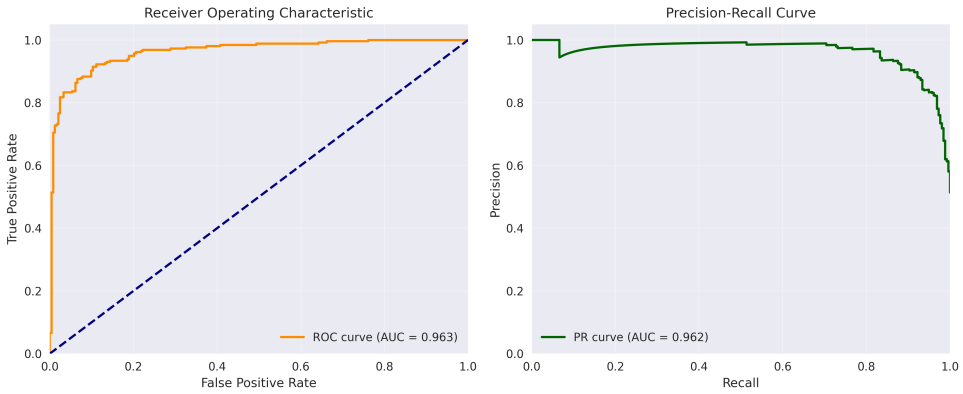

Figure 3: HateBERT ROC curve and calibration analysis demonstrating excellent discriminative capability (AUC=0.9625).

Table 2: XAI methods Faithfulness Evaluation

Method	Toxic-1	Toxic-2	Benign	Ambiguous	Avg Infidelity	Runtime (s)
Attention Viz	0.664	0.417	0.890	0.866	0.008	3.0
Integrated Grad	0.596	0.520	0.884	0.790	0.004	2.8
LIME	0.313	0.580	0.902	0.850	0.025	34.5
Counterfactual	0.450	0.535	0.886	0.820	0.038	2.9
Natural Language	0.520	0.480	0.875	0.800	0.012	2.8

pass rates across all generated counter-speech samples. Primary hate speech re-evaluation using HateBERT with enhanced sensitivity (threshold 0.3) confirmed no toxic content generation. Sentiment analysis validated neutral to positive tone orientation across all responses, exceeding the minimum threshold of 0.1 for ethical counter-speech generation.

Response length management achieved target constraints with an average generation of 147.4 characters, appropriate for social media platform requirements, while approaching the lower bound of the optimal 150-300 character range.

4.4. Computational Efficiency and Scalability

Computational efficiency analysis revealed a clear performance hierarchy with significant implications for practical deployment. Table 4 presents runtime comparisons across all XAI methods.

Integrated Gradients achieved optimal efficiency at 2.6-3.0 seconds, followed closely by Natural Language Explanations (2.7-3.0 seconds), demonstrating that gradient-based methods and synthesis approaches maintain computational tractability for real-time applications.

LIME exhibited higher computational costs at 32-37 seconds, representing 10-12x overhead compared to other methods. The dramatic efficiency difference for LIME poses significant constraints for real-time applications, particularly in interactive counter-speech generation scenarios where response latency affects user experience.

The integrated framework demonstrated efficient resource utilization across all components. HateBERT inference required 8.2 GB GPU memory during training and achieved approximately 1,200 samples/second throughput on H100 hardware, indicating scalability for high-volume content moderation applications.

Counter-speech generation with LoRA fine-tuning maintained memory efficiency through 4-bit quantization while preserving generation quality. Training efficiency metrics showed HateBERT convergence in 17.3 minutes for 238,987 samples, while CounterLLAMA fine-tuning completed within reasonable computational budgets suitable for both research and production deployment.

5. Conclusion

This work presents an explainable AI framework for counter-speech generation, bridging major gaps in current hate speech mitigation methods. By combining multiple XAI techniques with advanced counter-narrative generation, the system achieves strong transparency and performance across detection, explanation, and response quality.

The fine-tuned HateBERT model achieved exceptional performance with 89.2% accuracy in hate speech detection, significantly outperforming baseline approaches while maintaining balanced

Table 3: Counter-Speech Generation Baseline Comparison

Model	BERTScore	ROUGE	BLEU	Diversity
CounterLLAMA	**0.854**	**0.117**	0.039	**67.5%**
Seq2Seq	0.821	0.089	0.035	6.7%
MMI	0.808	0.072	0.042	7.0%
GPS	0.839	0.081	0.038	7.4%
Improvement over best baseline:				
	+3.3pp	+44.4%	-7.1%	+810%

Method	Avg Runtime(s)	Relative Cost
Integ. Grad.	2.8	1.0x
Natural Lang.	2.8	1.0x
Counterfact.	2.9	1.0x
Attent. Vis.	3.1	1.1x
LIME	34.5	12.3x

Table 4: XAI Method Computational Efficiency

precision-recall characteristics essential for practical deployment. The model's conservative classification approach, evidenced by high specificity (93.8%) and controlled false positive rates (6.2%), proves particularly valuable for content moderation applications.

Our multi-modal XAI framework revealed important insights about explanation method performance and complementarity. Attention Visualization emerged as superior for span localization with perfect precision, while Integrated Gradients provided the most faithful explanations with infidelity scores as low as 0.0004. The significant performance variations across different content types highlight the complexity of explaining hate speech detection decisions. The CounterLLAMA model demonstrated exceptional performance improvements over baseline approaches, achieving 85.4% semantic similarity (BERTScore) and remarkable linguistic diversity (67.5%). This breakthrough in response variety addresses a fundamental limitation of template-based counter-speech systems while maintaining semantic quality and safety standards.

The study's findings might be affected by data bias and representation issues. Reliance on datasets like ToxiGen and CONAN may introduce systematic demographic and cultural biases, limiting generalization across diverse user groups and communication styles. Additionally, the static nature of model training can result in temporal and cultural limitations, potentially causing the system to fail to capture the evolving dynamics of hate speech.

Another important limitation of this study is the lack of human evaluation of the generated counter-speech. Although automatic evaluation metrics offer a scalable and reproducible means of assessing system outputs, they may not adequately capture nuanced aspects such as contextual appropriateness, perceived sincerity, persuasive effectiveness or potential unintended harm. Incorporating human judgments was beyond the scope of the present study due to resource and design constraints, as the primary focus was on automated detection and generation performance. Therefore, systematic human-centered evaluation of generated counter-speech is the key direction for future research. Further future research should focus on developing bias detection and mitigation techniques to enhance fairness in diverse demographic and cultural contexts, and on conducting large-scale deployment studies to assess adaptability and effectiveness in evolving, real-world communication environments.

6. Limitations

The present study is subject to several limitations related to data sources, model architecture, and generalizability. The hate speech detection model architecture is affected by the HateBERT 96-token sequence length, thus poses significant constraints for analysing longer textual content where hate speech elements may be embedded within extensive context. It is also important to note, that binary classification approach reduces the complexity of Real-world hate speech to toxic/non-toxic categories and therefore may miss important gradations that could inform more sophisticated counter-speech strategies.

From XAI point of view, the significant misalignment between gradient-based attributions and human-annotated span locations indicates that computational faithfulness does not necessarily translate to human interpretability. This disconnect poses challenges for user-facing applications and requires extinctive work on user perception.

From cultural perspective it is also important to note that hate speech patterns evolve rapidly in response to social developments, platform policies, and adversarial attempts to evade detection systems. This temporal drift, when not addressed, might potentially leading to degraded performance over time as new forms of hate speech emerge.

Cultural and linguistic variations in hate speech expression present additional challenges for system generalization.

7. Ethics Statement

This work adheres to the ethical standards of research in natural language processing and computational social science. All datasets used in this study are publicly available and contain no personally identifiable information. The examples of hate speech included were handled responsibly and solely for the purpose of developing and evaluating automated mitigation systems. Since the research does not involve direct human participation or the collection of private data, no formal ethics approval was required.

We acknowledge the societal impact of automated hate speech detection and counter-speech generation systems. While these tools can encourage positive online discussions and assist moderators, there is a risk of misclassification or biased responses, representing specific cultural and ideological viewpoints that are ingrained in training data.

While increasing transparency, the XAI framework's thorough explanations may also help adversarial users better comprehend detection methods and create more complex evasion techniques. In deployment scenarios, this trade-off between security and transparency must be carefully considered.

8. Bibliographical References

Helena Bonaldi, Yi-Ling Chung, Gavin Abercrombie, and Marco Guerini. 2024. NLP for counterspeech against hate: A survey and how-to guide. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 3480–3499, Mexico City, Mexico. Association for Computational Linguistics.

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.

Sergio Andrés Castaño-Pulgarín, Natalia Suárez-Betancur, Luz Magnolia Tilano Vega, and Harvey Mauricio Herrera López. 2021. Internet, social media and online hate speech. systematic review. *Aggression and Violent Behavior*, 58:101608.

Yi-Ling Chung, Elizaveta Kuzmenko, Serra Sinem Tekiroglu, and Marco Guerini. 2019. 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*, pages 2819–2829, Florence, Italy. Association for Computational Linguistics.

Kevin Clark, Urvashi Khandelwal, Omer Levy, and Christopher D. Manning. 2019. What does BERT look at? an analysis of BERT's attention. In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 276–286, Florence, Italy. Association for Computational Linguistics.

Christopher Clarke, Matthew Hall, Gaurav Mittal, Ye Yu, Sandra Sajeev, Jason Mars, and Mei Chen. 2023. Rule by example: Harnessing logical rules for explainable hate speech detection. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 364–376, Toronto, Canada. Association for Computational Linguistics.

Thomas Davidson, Dana Warmsley, Michael Macy, and Ingmar Weber. 2017. Automated hate speech detection and the problem of offensive language. *Proceedings of the International AAAI Conference on Web and Social Media*, 11(1):512–515.

Antigoni Founta, Constantinos Djouvas, Despoina Chatzakou, Ilias Leontiadis, Jeremy Blackburn, Gianluca Stringhini, Athena Vakali, Michael Sirivianos, and Nicolas Kourtellis. 2018. Large scale crowdsourcing and characterization of twitter abusive behavior. *Proceedings of the International AAAI Conference on Web and Social Media*, 12(1).

Oguzhan Gencoglu. 2021. Cyberbullying Detection With Fairness Constraints . *IEEE Internet Computing*, 25(01):20–29.

Dominik Hangartner, Gloria Gennaro, Sary Alasiri, Nicholas Bahrich, Alexandra Bornhoft, Joseph Boucher, Buket Buse Demirci, Laurenz Derksen, Aldo Hall, Matthias Jochum, Maria Murias Munoz, Marc Richter, Franziska Vogel, Salomé Wittwer, Felix Wüthrich, Fabrizio Gilardi, and Karsten Donnay. 2021. Empathy-based counterspeech can reduce racist hate speech in a social media field experiment. *Proceedings of the National Academy of Sciences*, 118(50):e2116310118.

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.

Md. Mehedi Hassan, Anindya Nag, Riya Biswas, Md Shahin Ali, Sadika Zaman, Anupam Kumar Bairagi, and Chetna Kaushal. 2025. Explainable artificial intelligence for natural language processing: A survey. *Data Knowledge Engineering*, 160:102470.

Md Saroar Jahan and Mourad Oussalah. 2023. A systematic review of hate speech automatic detection using natural language processing. *Neurocomputing*, 546:126232.

Sarthak Jain and Byron C. Wallace. 2019. Attention is not Explanation. In *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 3543–3556, Minneapolis, Minnesota. Association for Computational Linguistics.

Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. 2016. A diversity-promoting objective function for neural conversation models. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 110–119, San Diego, California. Association for Computational Linguistics.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Ilya Loshchilov and Frank Hutter. 2017. Decoupled weight decay regularization. In *International Conference on Learning Representations*.

Binny Mathew, Navish Kumar, Ravina, Pawan Goyal, and Animesh Mukherjee. 2018. Analyzing the hate and counter speech accounts on twitter. *ArXiv*, abs/1812.02712.

Harshkumar Mehta and Kalpdrum Passi. 2022. Social media hate speech detection using explainable artificial intelligence (xai). *Algorithms*, 15(8).

Ioannis Mollas, Zoe Chrysopoulou, Stamatis Karlos, and Grigorios Tsoumakas. 2020. Ethos: a multi-label hate speech detection dataset. *Complex & Intelligent Systems*, pages 1–16.

Ayushi Nirmal, Amrita Bhattacharjee, Paras Sheth, and Huan Liu. 2024. Towards interpretable hate speech detection using large language model-extracted rationales. In *Proceedings of the 8th Workshop on Online Abuse and Harms (WOAH 2024)*, pages 223–233, Mexico City, Mexico. Association for Computational Linguistics.

Juri Opitz. 2024. Schroedinger's threshold: When the AUC doesn't predict accuracy. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 14400–14406, Torino, Italia. ELRA and ICCL.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Jing Qian, Anna Bethke, Yinyin Liu, Elizabeth Belding, and William Yang Wang. 2019. A benchmark dataset for learning to intervene in online hate speech. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4755–4764, Hong Kong, China. Association for Computational Linguistics.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. "why should i trust you?": Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, page 1135–1144, New York, NY, USA. Association for Computing Machinery.

Mukund Sundararajan, Ankur Taly, and Qiqi Yan. 2017. Axiomatic attribution for deep networks. In *International Conference on Machine Learning*.

Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. 2014. Sequence to sequence learning with neural networks. In *Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 2*, NIPS'14, page 3104–3112, Cambridge, MA, USA. MIT Press.

Serra Sinem Tekiroğlu, Yi-Ling Chung, and Marco Guerini. 2020. Generating counter narratives against online hate speech: Data and strategies. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1177–1190, Online. Association for Computational Linguistics.

Guy Tevet and Jonathan Berant. 2021. Evaluating the evaluation of diversity in natural language generation. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 326–346, Online. Association for Computational Linguistics.

United Nations. What is hate speech? | United Nations — un.org. https://www.un.org/en/hate-speech/understanding-hate-speech/what-is-hate-speech. [Accessed 08-10-2025].

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.

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.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. *ArXiv*, abs/1904.09675.

Zhixue Zhao, Ziqi Zhang, and Frank Hopfgartner. 2022. Utilizing subjectivity level to mitigate identity term bias in toxic comments classification. *Online Soc. Networks Media*, 29:100205.

Yi Zheng, Björn Ross, and Walid Magdy. 2023. What makes good counterspeech? a comparison of generation approaches and evaluation metrics. In *Proceedings of the 1st Workshop on CounterSpeech for Online Abuse (CS4OA)*, pages 62–71, Prague, Czechia. Association for Computational Linguistics.

Wanzheng Zhu and Suma Bhat. 2021. Generate, prune, select: A pipeline for counterspeech generation against online hate speech. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 134–149, Online. Association for Computational Linguistics.