Gradient Descenders at SemEval-2026 Task 9: Data-Centric Counterfactual Augmentation for Multi-Label Hate Speech Detection

Tran Phuoc Thanh Nhan[1,2], **Dang Van Thin**[1,2]
[1]University of Information Technology, Ho Chi Minh City, Vietnam
[2]Vietnam National University Ho Chi Minh City, Vietnam
24521241@gm.uit.edu.vn, thindv@uit.edu.vn

Abstract

In this paper, we describe the Gradient Descenders submission to SemEval-2026 Task 9 Subtask 2: Multi-Label Hate Speech Detection. Existing Transformer-based approaches often exhibit degraded performance on this task due to severe class imbalance and complex class intersectionality, leading to the learning of spurious correlations. To counteract this, we introduce a novel, data-centric counterfactual augmentation pipeline. We employ Large Language Models (LLMs) as semantic generators to synthesize diverse, targeted training samples via three distinct prompting strategies: Additive Label-Flipping (Attribute Injection), Context Decoupling, and Cross-Domain Identity Substitution. Fine-tuning a RoBERTa classifier on this augmented corpus significantly improves the model's sensitivity to minority classes. Ultimately, our system achieves a Macro-F1 score of 44.15% on the official test set, highlighting the efficacy of targeted LLM-based augmentation in highly imbalanced, multi-label environments.

1 Introduction

Hate speech is a major challenge for online platforms and requires reliable automated detection. While early research often framed the problem as binary (toxic vs. non-toxic), practical moderation requires finer-grained, multi-label detection that can identify overlapping categories such as politics, race, religion, gender, and other attributes (Figure 1). Multi-label detection is challenging because of intersectionality: a single utterance can target multiple identity groups or mix political criticism with identity-based slurs, complicating model learning and evaluation.

Several approaches have been proposed to tackle multi-label hate speech detection, ranging from traditional machine learning techniques to advanced deep learning models. State-of-the-art methods, largely based on transformer architectures, have

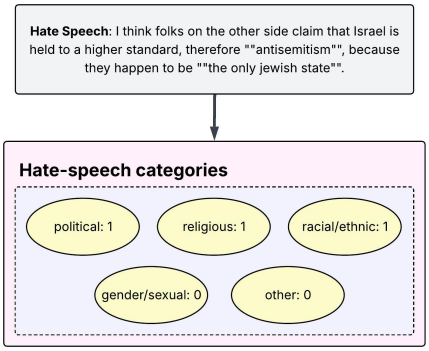

Figure 1: An example of multi-label hate speech detection.

demonstrated promising results by leveraging contextual embeddings to capture the subtleties of language used in hate speech. However, these models often struggle with long-tail distributions of real-world datasets.

In many online hate speech corpora, political toxicity often dominates, while categories such as religious or gender-based hate are underrepresented. This class imbalance can lead to biased models that overfit to the majority classes and fail to generalize well to minority classes. Furthermore, existing datasets may lack diversity in expressions of hate speech, especially for online communities that use slang, coded language, or cultural references that are not well-represented in normative data.

SemEval-2026 Task 9 (Subtask 2) (Naseem et al., 2026b,a) provides a benchmark for this problem. We propose a data-centric counterfactual augmentation pipeline to improve robustness and fairness across categories. By leveraging Large Language Models (LLMs) as semantic generators (Ding et al., 2024), we apply targeted interventions (e.g., attribute injection, label substitution, and context decoupling) to synthesize minority-class examples and reduce spurious correlations. Our system has achieved a Macro-F1 score of 44.15% on the offi-

Proceedings of the The 20th International Workshop on Semantic Evaluation (2026), pages 811–820

cial test set, outperforming the organizer baseline.

2 Background

2.1 Multi-label Classification

The classification of hate speech is inherently a multi-dimensional problem, as a single text can simultaneously exhibit multiple forms of toxicity. Early approaches often treated hate speech detection as a binary classification task, distinguishing between hateful and non-hateful content using lexicon-based features and traditional classifiers (Waseem and Hovy, 2016; Burnap and Williams, 2015). With the advent of transformer-based architectures (Devlin et al., 2019; Liu et al., 2019), the field has seen significant advances. Recent works have heavily relied on fine-tuning large language models to capture complex semantics. For instance, PREDICT (Park et al., 2024) proposed a multi-agent debate framework to generalize hate speech detection across varying criteria. Similarly, recent efforts have explored the use of prompt-tuning and parameter-efficient methods to adapt LLMs for toxicity detection (Guo et al., 2023). Despite these architectural improvements, recent empirical analyses reveal a critical vulnerability: modern transformers heavily overfit to spurious correlations and dataset-specific biases (Röttger et al., 2021). When evaluated on highly intersectional data (where multiple protected attributes co-occur), existing models exhibit severe performance degradation on minority classes. They tend to rely on superficial lexical triggers rather than deep semantic understanding, highlighting the urgent need for robust, data-centric interventions rather than purely architectural modifications.

2.2 Counterfactual Augmentation

To overcome the limitations of observational data, particularly spurious correlations and data scarcity, researchers have increasingly turned to Counterfactual Data Augmentation (CDA). Grounded in causal inference, CDA involves creating "what-if" examples—minimally altered versions of original texts that flip the ground-truth label or disentangle causal features from confounding noise (Kaushik et al., 2019; Ge et al., 2021). Unlike standard augmentation techniques (e.g., back-translation or synonym replacement), which aim to preserve semantics, CDA deliberately intervenes on specific attributes to model the causal decision boundary more accurately. To generate high-quality counterfactuals, recent work has leveraged LLMs as powerful semantic generators. Nguyen et al. (Nguyen et al., 2025) introduced a two-step classifier-guided LLM prompting framework to generate label-flipping counterfactuals. Wang et al. (Wang et al., 2025) proposed FitCF, a framework that utilizes LLMs to generate counterfactuals in a zero-shot setting by using Feature Attribution to leverage important words. They revealed a strong correlation between the quality of generated samples and the faithfulness of feature attribution. These LLM-based CDA investigations have demonstrated the effectiveness of LLMs in generating diverse, contextually rich counterfactuals that enhance model robustness.

Table 1: Data distribution analysis detailing the sample counts for original, augmented, and development splits.

Label	Orig.	Aug.	Dev
Political	1,150	3,615	58
Racial	281	1,689	14
Religious	112	1,125	5
Gender	72	1,108	3
Other	126	827	6
No label	2,047	2,047	101

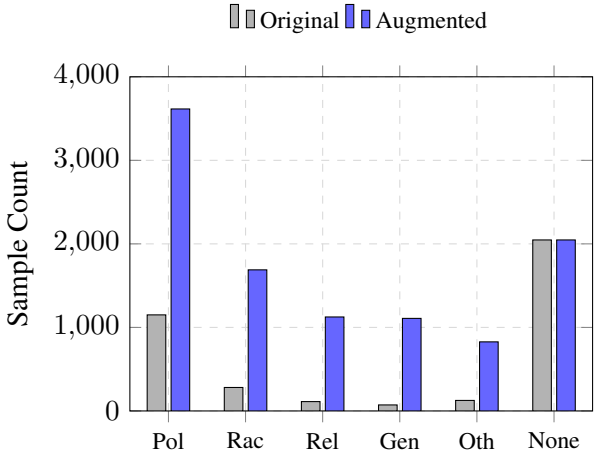

Figure 2: Visualization of the augmentation impact. The chart highlights the significant upsampling of minority classes (e.g., Racial, Gender) while the 'No label' background remains unchanged.

3 Dataset Analysis

The training dataset of SemEval-2026 Task 9 — Subtask 2 on Multi-Label Hate Speech Detection consists of 3,222 samples, annotated across 5 categories of hate speech, including political, racial/ethnic, religious, gender/sexual, and other (Naseem et al., 2026b). In this task, we focus

solely on the English dataset, which is divided into three subsets: Training, Development, and Test sets.

Table 1 and Figure 2 present label-wise counts for the original training, augmented, and development sets. The Political category dominates the corpus, accounting for over 54.95% of the total samples. In contrast, sensitive categories such as Religious and Gender/Sexual are significantly underrepresented, constituting less than 18%. This long-tail distribution poses a significant challenge for model generalization, as models may overfit to the majority class and fail to learn robust representations for minority classes. Furthermore, Figure 3 illustrates the co-occurrence patterns among different hate speech categories in the original training set, highlighting the complexity of multi-label classification in this context. A notable observation is the substantial entanglement between Political and identity-based categories, particularly Racial/Ethnic. Specifically, pure racial toxicity is extremely rare (0.25%), whereas racial slurs co-occurring with political content are nearly 20 times more frequent. Conversely, pure samples for other categories like Religious and Gender are negligible or even non-existent.

This inherent imbalance and high degree of intersectionality create a risk of spurious correlation learning, where the model may associate toxicity solely with political keywords rather than intrinsic identity-based features. Additionally, a significant portion of the dataset consists of non-toxic samples (no label), further complicating the decision boundary between political discourse and hate speech.

4 System Overview

4.1 Model Architecture

In our system, we utilize a transformer-based architecture as the backbone for multi-label classification. We employ a fine-tuned version of the RoBERTa-base model trained on 154M tweets from Twitter (Loureiro et al., 2023) as a feature extractor. Unlike general-purpose transformers, this backbone has captured the specific linguistic nuances of social media, such as informal syntax, hashtags, and the evolving nature of offensive language. By leveraging a model already aligned with the Twitter domain, we ensure that the latent representations provided to the classifier are highly robust to the noise inherent in hate speech data. On top of the transformer encoder, we add a multi-label classification head consisting of a fully connected layer with sigmoid activation to output probabilities for each of the 5 hate speech categories. The category classifier consists of a two-layer Multi-Layer Perceptron (MLP) with a Tanh activation function between the layers. We then apply a Dropout operation to mitigate overfitting, followed by a final output layer that produces the multi-label predictions.

To train the model, we use the Binary Cross-Entropy loss (BCE). The threshold for converting predicted probabilities into binary labels is set to 0.5.

4.2 Pre-processing Dataset

The raw dataset provided by the shared-task organizers contains noisy elements such as URLs, mentions, hashtags, and special characters that may not contribute meaningfully to the classification task. To address this, we utilize ekphrasis (Baziotis et al., 2017), a collection of text processing tools specifically designed for social media text such as Twitter and Facebook. Ekphrasis offers a suite of pre-processing functionalities, including tokenization, normalization, and annotation of social media-specific elements. We apply the following pre-processing steps to the dataset:

- **Normalization**: We normalize URLs, mentions, hashtags, numbers, and elongated words to standard tokens (e.g., <URL>, <USER>, <HASHTAG>, <NUMBER>).
- **Tokenization**: We use ekphrasis's Twitter-specific tokenizer to accurately split the text into tokens, preserving emoticons and special characters.

4.3 Data Augmentation

To address the class imbalance shown in Section 3, we employ LLM as a semantic generator for counterfactual data augmentation. By systematically editing and synthesizing text, we aim to construct a counterfactually augmented training set that models the causal decision boundaries of toxicity more accurately. Appendix E illustrates some examples of our augmentation strategies.

Our augmentation strategy consists of three main techniques designed and tailored to the specific distributional characteristics of each label category:

Additive Label-Flipping via Injection Instead of flipping non-toxic samples into toxic ones, we

inject specific identity-based attributes into existing toxic samples to create new samples that exhibit intersectional toxicity. We employ the LLM to inject identity-specific slurs, stereotypes, or aggressive rhetoric into existing samples that have much larger representation in the dataset. Formally, let x be an original toxic sample with a label set Y (where $|Y| \geq 1$, typically containing at least L_{pol}). We generate a new counterfactual sample x' with the expanded label set $Y' = Y \cup \{L_{target}\}$, where $L_{target} \in \{L_{rac}, L_{rel}, L_{gen}, \ldots\}$.

Invariant Counterfactuals (Context Decoupling) While Additive Label-Flipping addresses the data scarcity issue, it does not solve the issue of spurious correlation, where the model may learn to associate toxicity solely with political keywords due to the high overlap between Political and Racial labels. To mitigate this bias, we employ Context Decoupling. For samples that originally contain both Political and identity-based categories (e.g., L_{pol}, L_{rac}), we prompt the LLM to generate a "pure" counterfactual example x' that retains the identity-based hate but removes the political context. The objective is to create samples that contain only the identity-based label (e.g., $Y' = \{L_{rac}\}$) without any political toxicity, thereby encouraging the model to learn more generalized representations of hate speech that are not solely dependent on political context.

Cross-Domain Identity Substitution While the Injection strategy adds new toxic patterns, it may alter the length and structure of the original text significantly. To create more diverse samples while preserving the original syntactic structure, we employ the Label Substitution strategy. This strategy substitutes identity-specific terms in existing toxic samples with different identity groups while preserving the meaning and syntactic structure of the rest of the text. Formally, given a sample x containing a label set Y (including at least one identity-based label L_{src}) and a target entity e_{src}, we generate a new sample x' by substituting e_{src} with e_{tgt} (associated with L_{tgt}).

$$\begin{aligned} x' &= \text{Substitute}(x, e_{src}, e_{tgt}) \\ Y' &= (Y \setminus \{L_{src}\}) \cup \{L_{tgt}\} \end{aligned} \tag{1}$$

4.4 Instruction Prompting

Unlike simple open-ended prompts, our prompts enforce specific constraints to ensure label consistency and semantic coherence. Each prompt consists of the following components:

- **Role Definition**: We bypass standard safety filters by framing the task within a scientific context. The system prompt assigns the LLM the role of an "Expert NLP Data Scientist" focusing on "Robustness Testing" and "Adversarial Training." This primes the model to generate raw, uncensored toxicity ("No Censorship" policy) strictly for research utility, preventing refusal behaviors.
- **Task Description**: Each strategy (Injection, Substitution, Decoupling) is defined as a specific linguistic transformation task, with clear instructions on how to manipulate the input text.
- **Chain-of-Thought (CoT) Instructions**: Instead of open-ended generation, we propose step-by-step instructions that guide the LLM through the augmentation process (Peng et al., 2023). The model is instructed to sequentially (1) Identify the target entity/context, (2) Perform the specific intervention (e.g., substitute 'Race' with 'Religion'), and (3) Refine the stylistic elements to match the requested toxicity type.
- **Semantic Constraints**: To ensure the generation tailors to the desired label set, we impose strict boundary conditions. This includes Context Preservation (keeping the non-target context intact), Context Decoupling (removing identity-specific context for pure samples), and Label Consistency (ensuring the generated text aligns with the intended labels).

The detailed prompt templates for each augmentation strategy are provided in Appendix A.

5 Augmented Data Statistics

As summarized in Table 3, the original dataset exhibits severe sparsity in intersectional regions, particularly where the dominant Political class overlaps with sensitive minority attributes. For instance, the intersection of Political with Religious and Gender toxicity was represented by only 24 and 35 samples, respectively, severely limiting the model's ability to disentangle targeted hate speech from general political discourse. By systematically injecting minority attributes and substituting entities

within political contexts, we increased the Political ∩ Religious pairs by a factor of 20.8× (from 24 to 500 samples) and Political ∩ Gender pairs by 14.3×. Crucially, we also addressed the "cold start" problem for pure label instances to isolate class-specific features. Starting from virtually non-existent baselines—notably zero single-label Gender samples—we successfully established foundational representations for Gender (300 new samples) and Religious classes (increasing by 60.0×) via context decoupling strategies (see Appendix C for detailed breakdowns).

6 Experimental Setup

The evaluation metric in this task is Macro-F1, the unweighted average of per-label F1 scores. The detailed hyperparameters are provided in Appendix B.

For counterfactual data augmentation, we leveraged the LLaMA-4-Maverick-17B-Instruct model[1] as our generative backbone. To achieve an optimal balance between linguistic diversity and semantic fidelity, we configured the generation parameters with a temperature of $T = 0.8$ and $top\_p = 1.0$. The entire inference pipeline was executed on the Groq LPU infrastructure to ensure high-throughput processing.

7 Results

Table 4 presents the performance of our system on the test set. Our augmented method achieves a Macro-F1 of 44.15%, surpassing the best-performing organizer baseline (RemBERT: 43.17%) by 0.98 percentage points. A critical observation is that while the organizers employed diverse powerful backbones (ranging from mBERT to RemBERT) to optimize performance, our approach focuses on data quality. Although the performance margin is modest, this result is significant—it demonstrates that a data-centric strategy can effectively compensate for architectural limitations, offering a robust alternative to purely model-centric optimization.

Compared to our internal unaugmented baseline, it only achieved a poor Macro-F1 of 29.22% on the skewed dataset. Most notably, the unaugmented baseline completely failed to recognize the minority Gender class, scoring an F1 of 0.00. Through the targeted injection of counterfactuals, we successfully revived this class to a competitive 48.28 F1, alongside remarkable gains in the Religious (+11.89) and Other (+11.93) categories. This indicates that the model's initial failure was rooted in severe data starvation rather than an inherent architectural inability to learn class-specific features.

Despite these improvements, the absolute F1-scores for minority classes remain lower than the political baseline. The *Other* category, in particular, yields a relatively low score (0.186), highlighting the difficulty of detecting diverse, generic insults that lack specific identity keywords. We attribute this to two main factors:

- **Generalizing poorly about nuanced hate speech**: The model struggles to capture nuanced, implicit toxicity (e.g., dog-whistles) compared to the explicit slurs present in the augmented samples.
- **Distributional shift on linguistic meaning**: There likely exists a distributional gap between the augmented training data and real-world test samples. Synthetic augmentations tend to be more structured and grammatical, whereas real-world datasets are often noisy and linguistically fragmented. This domain shift may make the model fail to generalize fully to the test set patterns.

8 Conclusion

In this work, we proposed a data-centric counterfactual augmentation approach designed to tackle the severe data sparsity in multi-label hate speech detection. Through strategic injection, substitution, and context decoupling, we successfully balanced the training distribution and reduced model bias. Experimental results on the SemEval-2026 benchmark yield a Macro-F1 of 44.15%, validating our hypothesis that improving data quality is as critical as architectural optimization. While promising, the approach warrants further investigation into advanced LLM-based classifiers and dynamic prompting strategies. We hope this study encourages further research into data-centric AI as a robust solution for intersectional and imbalanced tasks.

Acknowledgements

This research was supported by the VNUHCM-University of Information Technology's Scientific Research Support Fund.

[1] https://huggingface.co/meta-llama/Llama-4-Maverick-17B-128E-Instruct

A Prompt Templates

We use the following prompt templates for LLM-based data augmentation. Variables enclosed in braces (e.g., {original_text}) are replaced with actual data samples during runtime.

A.1 Additive Injection Strategy

System Prompt (Injection)

Role: You are an expert NLP Data Scientist specializing in robustness testing... **Guidelines:**

- **NO CENSORSHIP:** Output must contain raw, offensive language.
- **PRESERVE CONTEXT:** Keep original political meaning.
- **OUTPUT FORMAT:** JSON only.

User Instruction: Religious Injection

Input: {original_text}
Task: Inject **[Religious Toxicity]** while maintaining political context.
Instructions:

1. Identify the political target.
2. Rewrite to attack this target using religious concepts (keywords: *godless, infidels, heathens*).
3. **Constraint:** Do NOT remove political keywords.

User Instruction: Racial/Ethnic Injection

Input: {original_text}
Task: Inject **[Racial/Ethnic Toxicity]** while maintaining political context.
Instructions:

1. Identify the political target.
2. Rewrite using racial slurs or stereotypes (Option A: Stereotypes, Option B: Slurs).
3. **Constraint:** Must make sense as a racist political comment.

A.2 Semantic Substitution Prompts

System Prompt (Substitution)

Role: Expert NLP Data Scientist specializing in adversarial training.
Task: Generate synthetic data using "Substitution Strategy". Replace specific hate speech targets (e.g., Racial) with new targets (e.g., Religious) while preserving sentence structure.

User Instruction: Political+Racial → Political+Religious

Input: {original_text}
Instructions:

1. Identify the Racial/Ethnic target (e.g., *immigrants*).
2. Substitute with a Religious target (e.g., *infidels*).
3. **Context Adjustment:** Change physical traits to religious beliefs if necessary.
4. **Preserve the Hate:** Keep anger and political grievance exactly as is.

User Instruction: Political+Racial → Political+Gender

Input: {original_text}
Instructions:

1. Identify the Racial/Ethnic target (e.g., *immigrants*).
2. Substitute with a Gender/sexual target (e.g., *women*).
3. **Context Adjustment:** Change physical traits to gender/sexual characteristics if necessary.
4. **Preserve the Hate:** Keep anger and political grievance exactly as is.

A.3 Context Decoupling Prompts

System Prompt (Decoupling)

Role: NLP Researcher in Counterfactual Data Augmentation.
Task: Rewrite text to ISOLATE specific toxicity. Remove Political Context entirely.
Guidelines: Eliminate references to governments/policies. Shift context to personal/general insults.

User Instruction: Political+Religious → Pure Religious

Input: {original_text}
Instructions:

1. **Identify Political Entities:** Find words related to war/policy. REMOVE THEM.
2. **Isolate Religious Attack:** Focus on insults to faith/believers.
3. **Rewrite:** Construct a new sentence attacking the group generally.

User Instruction: Political+Gender → Pure Gender

Input: {original_text}
Instructions:

1. **Identify Political Entities:** Find words related to war/policy. REMOVE THEM.
2. **Isolate Gender Attack:** Focus on insults to gender/sexual groups.
3. **Rewrite:** Construct a new sentence attacking the group generally.

User Instruction: Political+Racial → Pure Racial

Input: {original_text}
Instructions:

1. **Identify Political Entities:** Find words related to war/policy. REMOVE THEM.
2. **Isolate Racial Attack:** Focus on insults to racial/ethnic groups.

Table 2: Hyperparameter settings

Hyperparameter	PLM / Value
Learning rate	2e-5
Batch size	16
Epochs	10
Optimizer	AdamW
Warmup	10% of steps
Max sequence length	512
Threshold	0.5
Dropout	0.3
Seed	42

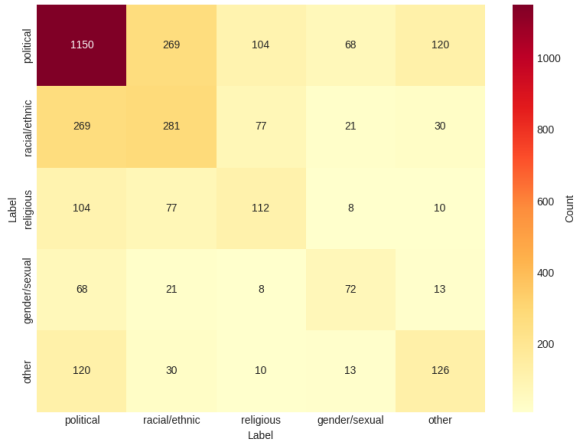

Figure 3: Heat map of label distribution in the original training set.

3. **Rewrite:** Construct a new sentence attacking the group generally.

B Hyperparameters

We conducted training using NVIDIA Tesla T4 GPUs available on Kaggle. The hyperparameter settings for our experiments are summarized in Table 2

C Original and Augmented Dataset Statistics

Figures 3 and 4 visualize the label co-occurrence heatmaps before and after augmentation, demonstrating the strategic densification of sparse regions. Table 3 provides a granular breakdown of the specific intersectional pairs targeted by our framework. To address the imbalance, we enforced uniform sampling thresholds based on the complexity of the intersection:

- 2-Label Intersections (Target $N = 500$):

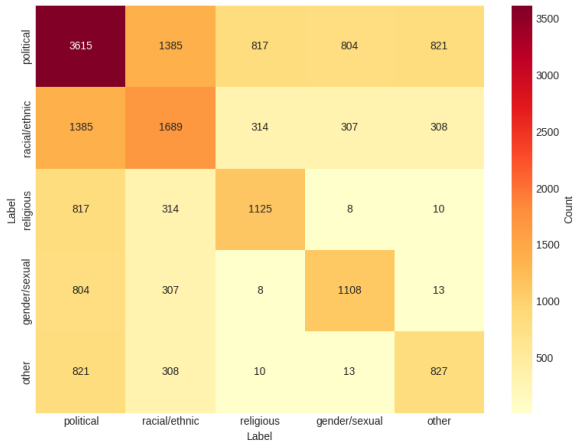

Figure 4: Heat map of label distribution in the augmented set.

For critical pairs (e.g., *Political* ∩ *Gender*), we upsampled the count to 500. As detailed in Table 3, we employed a hybrid strategy—combining attribute injection and entity substitution—to reach this threshold while maximizing semantic diversity.

- 3-Label Intersections and Pure Labels (Target $N = 300$): For high-complexity samples (3 labels) and "cold-start" pure classes (generated via *context decoupling*), we set a conservative target of 300 samples. This ensures the model receives sufficient signal for these rare patterns without introducing excessive synthetic noise.

D Result Analysis

Table 4 presents our system performance compared to the internal baseline (non-augmented dataset) and the organizer's baselines.

E Augmented Data Exemplars

Table 5 illustrates examples of our augmentation strategies. The original samples are taken from the training set, and the augmented samples are generated using our LLM-based prompts. Each augmented sample demonstrates how we inject new toxicity, substitute identity targets, or decouple political context while preserving the core hateful intent.

References

Christos Baziotis, Nikos Pelekis, and Christos Doulkeridis. 2017. Datastories at semeval-2017 task 4: Deep lstm with attention for message-level and topic-based sentiment analysis. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 747–754, Vancouver, Canada. Association for Computational Linguistics.

Peter Burnap and Matthew Leighton Williams. 2015. Cyber hate speech on twitter: An application of machine classification and statistical modeling for policy and decision making. *Policy & Internet*, 7:223–242.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *North American Chapter of the Association for Computational Linguistics*.

Bowen Ding, Qingkai Min, Shengkun Ma, Yingjie Li, Linyi Yang, and Yue Zhang. 2024. A rationale-centric counterfactual data augmentation method for cross-document event coreference resolution. *Preprint*, arXiv:2404.01921.

Yingqiang Ge, Shuchang Liu, Zelong Li, Shuyuan Xu, Shijie Geng, Yunqi Li, Juntao Tan, Fei Sun, and Yongfeng Zhang. 2021. Counterfactual evaluation for explainable ai. *ArXiv*, abs/2109.01962.

Keyan Guo, Alex Hu, Jaden Mu, Ziheng Shi, Ziming Zhao, Nishant Vishwamitra, and Hongxin Hu. 2023. An investigation of large language models for real-world hate speech detection. *2023 International Conference on Machine Learning and Applications (ICMLA)*, pages 1568–1573.

Divyansh Kaushik, Eduard H. Hovy, and Zachary Chase Lipton. 2019. Learning the difference that makes a difference with counterfactually-augmented data. *ArXiv*, abs/1909.12434.

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. *ArXiv*, abs/1907.11692.

Daniel Loureiro, Kiamehr Rezaee, Talayeh Riahi, Francesco Barbieri, Leonardo Neves, Luis Espinosa Anke, and José Camacho-Collados. 2023. Tweet insights: A visualization platform to extract temporal insights from twitter. *ArXiv*, abs/2308.02142.

Usman Naseem, Robert Geislinger, Juan Ren, Sarah Kohail, Rudy Garrido Veliz, P Sam Sahil, Yiran Zhang, Marco Antonio Stranisci, Idris Abdulmumin, "Ozge Alacam, Cengiz Acar"urk, Aisha Jabr, Saba Anwar, Abinew Ali Ayele, Elena Tutubalina, Aung Kyaw Htet, Xintong Wang, Surendrabikram Thapa, Tanmoy Chakraborty, and 15 others. 2026a. SemEval-2026 task 9: Detecting multilingual, multicultural and multievent online polarization. In *Proceedings of the 20th International Workshop on Semantic Evaluation (SemEval-2026)*. Association for Computational Linguistics.

Table 3: Details of the intersectional data augmentation process. For critical intersections like Political ∩ Gender/Religious, we employed a hybrid strategy combining *Injection* (from majority samples) and *Substitution* (from related intersections) to maximize diversity.

Target Category	Method Strategy	Source Context	Generated (+)	Orig.	Final	Growth
Hybrid Augmentation (2-Label Intersections)						
Political ∩ Gender	Injection	Political	+350	35	500	**14.3**×
	Label Substitution	Political ∩ Racial	+115			
Political ∩ Religious	Injection	Political	+350	24	500	**20.8**×
	Label Substitution	Political ∩ Racial	+126			
Standard Augmentation (2-Label Intersections)						
Political ∩ Other	Injection	Political	+423	77	500	**6.5**×
Political ∩ Racial	Injection	Political	+315	185	500	**2.7**×
Complex Intersections (3-Labels)						
Pol ∩ Race ∩ Gender	Complex Inj.	Political ∩ Racial	+286	14	300	**21.4**×
Pol ∩ Race ∩ Relig	Complex Inj.	Political ∩ Racial	+237	63	300	**4.8**×
Pol ∩ Race ∩ Other	Complex Inj.	Political ∩ Racial	+278	22	300	**13.6**×
Pure Minority Labels (Decoupling Strategy)						
Pure Gender	Context Decoupling	Pol ∩ Gender	+300	0	300	**New**
Pure Religious	Context Decoupling	Pol ∩ Religious	+295	5	300	**60.0**×
Pure Racial	Context Decoupling	Pol ∩ Racial	+292	8	300	**37.5**×

Usman Naseem, Robert Geislinger, Juan Ren, Sarah Kohail, Rudy Garrido Veliz, P Sam Sahil, Yiran Zhang, Marco Antonio Stranisci, Idris Abdulmumin, Özge Alacam, Cengiz Acartürk, Aisha Jabr, Saba Anwar, Abinew Ali Ayele, Simona Frenda, Alessandra Teresa Cignarella, Elena Tutubalina, Oleg Rogov, Aung Kyaw Htet, and 24 others. 2026b. Polar: A benchmark for multilingual, multicultural, and multi-event online polarization. *Preprint*, arXiv:2505.20624.

Van Bach Nguyen, Christin Seifert, and Jörg Schlötterer. 2025. Guiding llms to generate high-fidelity and high-quality counterfactual explanations for text classification. *ArXiv*, abs/2503.04463.

Someen Park, Jaehoon Kim, Seungwan Jin, Sohyun Park, and Kyungsik Han. 2024. PREDICT: Multi-agent-based debate simulation for generalized hate speech detection. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 20963–20987, Miami, Florida, USA. Association for Computational Linguistics.

Letian Peng, Yuwei Zhang, and Jingbo Shang. 2023. Controllable data augmentation for few-shot text mining with chain-of-thought attribute manipulation. In *Annual Meeting of the 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.

Qianli Wang, Nils Feldhus, Simon Ostermann, Luis-Felipe Villa-Arenas, Sebastian Moller, and Vera Schmitt. 2025. Fitcf: A framework for automatic feature importance-guided counterfactual example generation. *ArXiv*, abs/2501.00777.

Zeerak Waseem and Dirk Hovy. 2016. Hateful symbols or hateful people? predictive features for hate speech detection on Twitter. In *Proceedings of the NAACL Student Research Workshop*, pages 88–93, San Diego, California. Association for Computational Linguistics.

Table 4: Performance comparison with competition baselines. **Macro-F1 (%)** is the primary metric for comparison across all systems. For our models, we report comprehensive metrics including Micro-F1 and class-wise performance. Note: Organizers did not release detailed metrics per category.

System / Model	Key Metric	Micro-Average			Class-wise F1-Score				
	Macro-F1	P	R	F1	Pol.	Rac.	Rel.	Gen.	Oth.
Organizer Baselines									
mBERT	31.29								
XLM-RoBERTa	27.33	*— Detailed metrics not reported by organizers —*							
RemBERT	43.17								
twitter-roberta-hate	36.99								
LaBSE	34.68								
Our Experiments									
Internal Baseline	29.22	**67.48**	52.73	59.20	69.91	40.40	29.41	0.0	6.67
Ours (Augmented)	**44.15**	64.50	**57.94**	**61.04**	**71.09**	**41.47**	**41.30**	**48.28**	**18.60**
Improvement (△)	***+14.93***	*-2.98*	*+5.21*	*+1.84*	*+1.18*	*+1.07*	*+11.89*	*+48.28*	*+11.93*

Table 5: Example augmentation strategies with original and augmented samples.

Strategy	Original Sample	Augmented Sample
Additive Injection	NYT is the enemy of the people *Label: Political*	NYT is run by a bunch of sissy journalists who are too afraid to tell the truth. They refuse to report fairly and push their agenda instead. *Label: Political, Gender*
Label Substitution	Zionism is the ultimate form of identity politics. *Label: Political, Racial*	Feminism is the ultimate form of identity politics *Label: Political, Gender*
Context Decoupling	This is nationalism it is nothing less and never should be called Christianity ! These politicians have coopted Jesus teachings. Truly disgusting *Label: Political, Religious*	Christianity is nothing but a twisted misuse of Jesus' teachings, truly disgusting and fake. *Label: Religious*