CAUSALITY GUIDED REPRESENTATION LEARNING FOR CROSS-STYLE HATE SPEECH DETECTION

Chengshuai Zhao,* Shu Wan,* Paras Sheth, Karan Patwa, K. Selçuk Candan, Huan Liu
School of Computing and Augmented Intelligence, Arizona State University, USA
{czhao93, swan, psheth5, kpatwa, candan, huanliu}@asu.edu

ABSTRACT

The proliferation of online hate speech poses a significant threat to the harmony of the web. While explicit hate is easily recognized through overt slurs, implicit hate speech is often conveyed through sarcasm, irony, stereotypes, or coded language—making it harder to detect. Existing hate speech detection models, which predominantly rely on surface-level linguistic cues, fail to generalize effectively across diverse stylistic variations. Moreover, hate speech spread on different platforms often targets distinct groups and adopts unique styles, potentially inducing spurious correlations between them and labels, further challenging current detection approaches. Motivated by these observations, we hypothesize that the generation of hate speech can be modeled as a causal graph involving key factors: contextual environment, creator motivation, target, and style. Guided by this graph, we propose CADET, a causal representation learning framework that disentangles hate speech into interpretable latent factors and then controls confounders, thereby isolating genuine hate intent from superficial linguistic cues. Furthermore, CADET allows counterfactual reasoning by intervening on style within the latent space, naturally guiding the model to robustly identify hate speech in varying forms. CADET demonstrates superior performance in comprehensive experiments, highlighting the potential of causal priors in advancing generalizable hate speech detection.

1 INTRODUCTION

The proliferation of online hate speech poses a significant threat to the harmony of the web due to its complex and ever-evolving nature (Castaño-Pulgarín et al., 2021; Malmasi & Zampieri, 2017; Mathew et al., 2019). While considerable progress has been made in identifying explicit hate speech characterized by overt hostility, slurs, and direct attacks, the detection of implicit expressions of hate conveyed through sarcasm, irony, stereotype, or coded language remains a critical unresolved challenge (Schmidt & Wiegand, 2017).

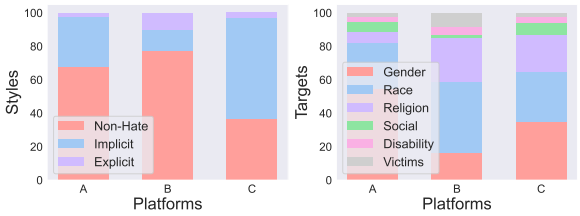

Figure 1: Distribution of hate speech styles and target groups across various social media platforms.

Consider the examples: (i) "*F****** women asking for equal rights here. Why don't you go whine and w**** around you floozy sl**.*" and (ii) "*I am so tired of doing everything on my own.. Need a woman so that chores can be taken care of...*". Both express misogynistic sentiments, but while the explicit version uses direct derogatory language that is easily flagged by conventional detection systems, the implicit version conveys the same prejudice through innuendo and rhetorical framing that is much harder to detect algorithmically.

Further analysis of current hate speech detection systems reveals a persistent pattern: models trained predominantly on explicit hate speech data perform poorly when confronted with implicit expressions (Ocampo et al., 2023a; Waseem et al., 2017). This limitation arises because these models often

*Equal contribution.

rely heavily on surface-level patterns such as specific slurs, aggressive syntax, or overtly hostile language that are typically absent in implicit hate speech, rather than the underlying hate intent that is shared across hate speech in various styles. Moreover, hate speech disseminated on various platforms often targets distinct groups and exhibits unique stylistic variations, as shown in Figure 1. This can induce spurious correlations between these factors and hate labels, posing additional challenges for existing detection models.

Motivated by these observations, we ***hypothesize*** that the generation of hate speech can be modeled as a causal graph involving key factors: contextual environment (e.g., platform), creator motivation (e.g., hate intent), target, and style, where (i) contextual environment is a confounder, which influences the motivation, target, and style. (ii) Motivation, target, and style jointly shape the content of an online post. (iii) The hate label is determined by the creator's motivation and contextual environment. Guided by this graph, we propose CADET, a causal representation learning framework for cross-style hate speech detection. CADET first projects online posts into a latent space and disentangles them into causality-aligned interpretable factors, isolating genuine hate intent from superficial linguistic cues. Then it leverages a confounder mitigation module to control the confounding effect caused by the contextual environment. Furthermore, CADET allows counterfactual reasoning by intervening on style within the latent space, naturally guiding the model to robustly identify hate speech in varying forms.

Our comprehensive evaluation demonstrates CADET achieves superior performance across multiple challenging scenarios. In cross-style generalization tasks, CADET achieves 0.81 average macro-F1, acquiring a relative improvement of 13% compared with the state-of-the-art method. Ablation studies confirm the critical role of each component in CADET. Further latent representation visualization and further analysis demonstrate the effectiveness of causal-driven design. Our work underscores the potential of causal disentanglement in improving generalized hate speech detection, offering practical implications for creating safer and more responsible online environments. Our contributions are summarized:

- ⋆ We introduce a principled causal graph that explicitly models the generative process of hate speech, systematically disentangling online posts into interpretable latent factors. It unifies explicit and implicit hate speech within a coherent causal framework.
- ⋆ We design and implement CADET, a representation learning framework that operationalizes the proposed causal graph. By isolating latent factors and leveraging counterfactual reasoning, CADET effectively mitigates platform-induced confounding and robustly generalizes across expression styles.
- ⋆ We conduct extensive experiments on multiple real-world hate speech datasets, demonstrating that CADET significantly outperforms state-of-the-art baselines in cross-style hate speech detection, providing robust, interpretable, and causally informed insights into harmful online content moderation.

2 THE HYPOTHESIS: HATE SPEECH GENERATION

To this end, we propose a causal graph that formalizes the generative process of hate speech as showcased in Figure 2. *Post* (X), representing the textual content produced by a creator, and the *Label* (Y), a binary indicator of whether the post is hateful. They are both accessible during training. The post X is generated by three latent factors: *Creator motivation* (M), capturing the writer's intent for desired impact; *Target* (T), denoting the individual or group toward whom the post is directed; and *Style* (S), describing the linguistic expression of the content. These latent variables and the label are influenced by a shared, unobserved *Contextual environment* confounder (U). U represents platform-level factors such as moderation policies, demographics, as well as broader sociopolitical atmosphere. Differences in moderation strictness or audience tolerance can simultaneously shape

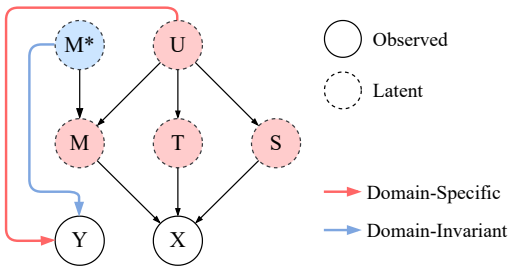

Figure 2: Causal graph of hate-speech generation.

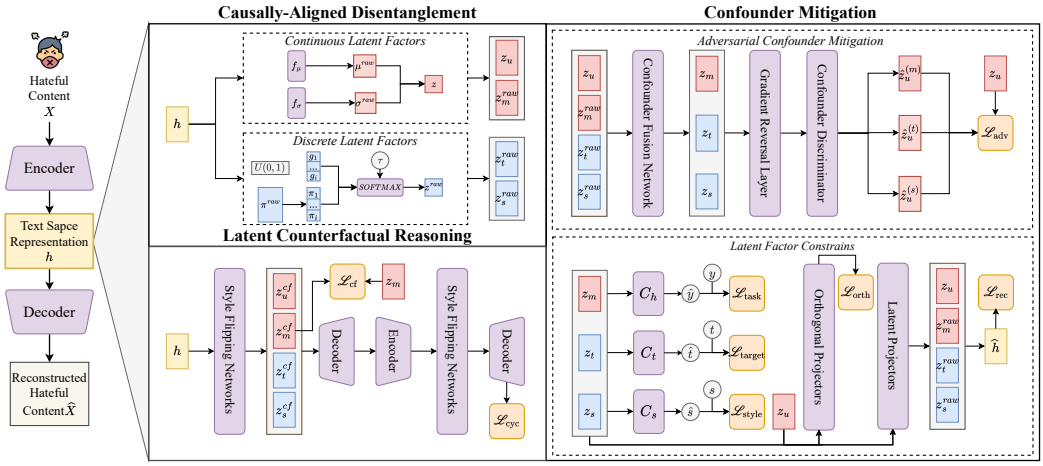

Figure 3: Framework of the proposed method.

a creator's stylistic choices, target selection, and even how intent is shaped in practice, thereby inducing spurious correlations among M, T, S, and the label Y. To separate genuine intent from its context-dependent manifestation, we posit an additional latent variable M*, representing the creator's true underlying motivation or intent, which is invariant to the contextual environment U and a direct cause of the label Y.

Unlike conventional causal formulations in domain generalization that assume domain shift affects either X or Y alone (Liu et al., 2021), our graph models domain shift as an unobserved confounder influencing both X and Y. This design reflects current online discourse, where content creators actively adapt their writing to the contextual environment U, often shifting from explicit to implicit or obfuscated expressions to evade moderation. As a result, classifiers trained predominantly on explicit-style corpora struggle to detect implicit hate in new domains. In our proposed graph, while the observed latents (M, T, S) are entangled with U, the true motivation M* is independent of the contextual environment. Hence, M* is a domain-invariant causal feature that is optimal and robust for out-of-distribution prediction (Wang & Veitch, 2022). Therefore, learning M* is the goal of CADET. Additionally, the graph also motivates a principled path toward cross-style generalization through counterfactual reasoning. By intervening on the style (S) node while holding creator motivation (M) and target (T) fixed, we can construct counterfactual examples that express the same hateful intent in alternative stylistic forms. Training models to treat both original and counterfactual samples as hateful encourages representations that emphasize invariant causal signals rather than superficial stylistic cues.

3 The Proposed CADET

3.1 Problem Definition

Let $\mathcal{D} = \{(x_i, y_i, s_i, t_i)\}_{i=1}^{n}$ denote a corpus of social media posts, where x_i denotes the textual content, $y_i \in \{0, 1\}$ is a binary hate speech label, $s_i \in \{0, 1\}$ indicates whether the post conveys hate speech explicitly or implicitly, and t_i (optional) specifies the targeted group. We define $\mathcal{D}_s$ as the conditional distribution over (x, y) given a particular style s. We formalize *cross-style hate speech detection* as a domain generalization problem: learning an optimal predictor f^* that is trained on posts from a source style domain $\mathcal{D}_s$ and generalizes effectively to a target style domain $\mathcal{D}_{s'}$ with $s \neq s'$.

$$f^* = \arg\min_{f \in \mathcal{F}} \mathbb{E}_{(x,y)\sim\mathcal{D}_{s'}} \big[\ell(f(x), y)\big], \tag{1}$$

where ℓ denotes the loss function and $\mathcal{F}$ is the hypothesis space.

3.2 CAUSALLY-ALIGNED DISENTANGLEMENT

To enable the causality-guided representation disengagement, we first map online posts into a text latent space.

3.2.1 TEXT ENCODING

We utilize a pretrained RoBERTa model as an encoder E to transform input text x_i into latent representations h.

$$h = \text{Encoder}(x) \tag{2}$$

where $h \in \mathbb{R}^d$ is the embedding of the start-of-sequence token (CLS) that serves as a fixed-length representation of the entire post.

3.2.2 LATENT FACTOR INFERENCE.

We infer four latent variables from h that correspond to the causal factors in our graph: we model the confounder representation z_u as a continuous Gaussian variable, parameterized via the reparameterization trick:

$$z_u = \mu_u + \sigma_u \odot \epsilon, \quad \epsilon \sim \mathcal{N}(0, I), \tag{3}$$

where the mean and variance parameters are predicted by feed-forward networks, $\mu_u = F_{\mu_u}(h)$ and $\log \sigma_u^2 = F_{\log \sigma_u^2}(h)$. Similarly, we extract the creator motivation representation, denoted z_m^{raw}, which is also represented as a continuous Gaussian variable.

The target factor z_t^{raw} is modeled as a discrete variable, parameterized using the Gumbel-Softmax trick:

$$z_t^{\text{raw}} = \text{GumbelSoftmax}(\pi_t^{\text{raw}}, \tau_t), \tag{4}$$

where $\pi_t^{\text{raw}} = F_t(h)$ are the logits associated with the target and τ_t is the temperature parameter controlling smoothness. Similarly, the content style, denoted z_s^{raw}, is represented as a discrete binary variable (capturing explicit vs. implicit style) inferred using the Gumbel-Softmax parameterization.

3.3 CONFOUNDER MITIGATION

To alleviate the impact of spurious correlations introduced by the context, we design a confounder-mitigation mechanism.

3.3.1 ADVERSARIAL CONFOUNDER RECONSTRUCTION.

We employ confounder fusion networks F to control the latent confounder by merging it into the raw variables:

$$z_i = \sigma_i\big(F_i(z_i^{\text{raw}} || z_u)\big),\ i \in \{m, t, s\} \tag{5}$$

where σ is an activation denoting the identity for continuous factors (e.g., creating motivation) and the softmax function for discrete factors (e.g., target and style), and each fusion network is implemented as a feed-forward network with layer normalization.

To ensure proper and meaningful disentanglement of each causal factor, we attach adversarial discriminators D that attempt to recover the confounder from each merged variable:

$$\hat{z}_u^{(i)} = D_i\big(G_\lambda(z_i)\big),\ i \in \{m, t, s\} \tag{6}$$

where G_λ is a gradient-reversal layer parameterized by the scaling coefficient λ so that the feature extractor is updated against the adversary. The resulting adversarial loss is defined by:

$$\mathcal{L}_{\text{adv}} = ||\hat{z}_u^{(m)} - z_u||_2^2 + ||\hat{z}_u^{(t)} - z_u||_2^2 + ||\hat{z}_u^{(s)} - z_u||_2^2 \tag{7}$$

3.3.2 FACTOR-SPECIFIC OBJECTIVE.

We incorporate causally-aligned classification, ensuring that each classifier relies solely on its causally relevant factor. We feed fused motivation representation into classifiers C_h to obtain the predicted label $\hat{y}$:

$$\hat{y} = C_h(z_m) \tag{8}$$

This architectural constraint directly enforces the causal understanding that hate depends on the creator's motivation. Then, we utilize cross-entropy to compute the loss:

$$\mathcal{L}_{\text{task}} = \text{CrossEntropy}(\hat{y}, y) \tag{9}$$

Similarly, we can acquire target loss $\mathcal{L}_{\text{target}}$ and style loss $\mathcal{L}_{\text{style}}$ by decoding corresponding latent factors.

3.3.3 Orthogonality Constraints

Additionally, orthogonality constraints are used to encourage separation of latent factors. We employ projections to map each variable into a common space:

$$p_i = F_{\text{orth}}^{(i)}(z_i),\ i \in \{m, t, s, u\} \tag{10}$$

and compute pair-wise squared dot products between latent factors:

$$o_{ij} = \left(\frac{p_i}{\|p_i\|_2} \cdot \frac{p_j}{\|p_j\|_2} \right)^2,\ (i, j) \in \{m, t, s, u\}^2,\ i < j \tag{11}$$

In this way, we obtain the orthogonality loss:

$$\mathcal{L}_{\text{orth}} = o_{mt} + o_{ms} + o_{ts} + \lambda_u (o_{mu} + o_{tu} + o_{su}) \tag{12}$$

where $\lambda_u > 1$ places stronger regularization with the confounder.

3.3.4 Input Reconstruction

We use reconstruction as a constraint to ensure our latent spaces capture meaningful information. Each latent variable is projected to the decoder's hidden dimension:

$$h_i = F_{\text{Proj}}^{(i)}(z_i),\ i \in \{m, t, s, u\}. \tag{13}$$

These projections are combined into a fused representation:

$$\hat{h} = h_m + h_t + h_s + h_u \tag{14}$$

We use a pretrained BART decoder Decoder to reconstruct the original input:

$$\hat{x} = \text{Decoder}(\tilde{x}, \hat{h}) \tag{15}$$

where $\tilde{x}$ represents the input sequence shifted right by one position for teacher forcing. The reconstruction loss is:

$$\mathcal{L}_{\text{rec}} = \text{CrossEntropy}(\hat{x}, x) \tag{16}$$

3.4 Latent Counterfactual Reasoning

To learn style-invariant representations, CADET leverages latent counterfactual reasoning without generating harmful text.

3.4.1 Style Intervention

Given latent representation h, we leverage a style flipping network R to disentangle latent factors by inverting the style component while maintaining other factors:

$$R(z):\ z_i^{\text{cf}} = \begin{cases} I - \text{OneHot}(z_s),\ i = s \\ \text{Identical}(z_i),\ i \in \{m, t, u\} \end{cases} \tag{17}$$

The binary nature of our style factors enables generalization even when the training data contains only one style.

3.4.2 COUNTERFACTUAL CONSISTENCY.

Since hate should depend only on motivation (instead of style), the output of the classifier should remain stable under style transformations:

$$\mathcal{L}_{cf} = ||C_h(z_m) - C_h(z_m^{\text{cf}})||_2^2 \tag{18}$$

3.4.3 CYCLE CONSISTENCY

We further implement a cycle consistency mechanism to ensure style transformations preserve semantic content. It decodes the counterfactual latent representation. Then it attempts to leverage another style flipping network to reverse the style back. Finally, we can construct the original text.

$$\mathcal{L}_{\text{cycle}} = \mathcal{L}_{\text{rec}}(x, \text{Decoder}(R(\text{Encoder}(\text{Decoder}(\hat{h}^{\text{cf}}))))) \tag{19}$$

3.5 MODEL TRAINING

We train CADET using a multi-objective loss function that combines reconstruction, classification, disentanglement, and counterfactual consistency objectives. The total loss is defined as:

$$\begin{aligned}\mathcal{L} = \lambda_{\text{task}}\mathcal{L}_{\text{task}} + \lambda_{\text{target}}\mathcal{L}_{\text{target}} + \lambda_{\text{style}}\mathcal{L}_{\text{style}} + \lambda_{\text{orth}}\mathcal{L}_{\text{orth}} + \\ \lambda_{\text{adv}}\mathcal{L}_{\text{adv}} + \lambda_{\text{rec}}\mathcal{L}_{\text{rec}} + \lambda_{\text{cf}}\mathcal{L}_{\text{cf}} + \lambda_{\text{cycle}}\mathcal{L}_{\text{cycle}} + +\lambda_{\text{KL}}\mathcal{L}_{\text{KL}}\end{aligned} \tag{20}$$

where $\mathcal{L}_{\text{KL}}$ terms regularize the latent distributions.

This streamlined process allows CADET to detect hate speech regardless of its stylistic presentation, providing a principled solution to the challenge of style-invariant detection.

4 EXPERIMENTS

We evaluate CADET on various cross-style hate speech detection scenarios. Due to space limits, the full experimental setup, dataset statistics, and extended analyses are deferred to the Appendix B. Below we briefly summarize the experimental setting, protocol, and key results.

Setting. We construct an Implicit-Explicit Hate Speech Corpus by combining four datasets spanning multiple platforms and expression styles: IsHate, IHC, AbuseEval, and DynaHate. We consider two challenging transfer regimes: *explicit→implicit* and *implicit→explicit*. Performance is measured using precision, recall, and macro-F1, averaged over five runs. We compare against strong pretrained language models (BERT, RoBERTa, BART, DistilBERT), domain-adaptive hate detectors (HateBERT, HateXplain), causality-inspired approaches (PEACE, HateWATCH), and recent LLM-based moderation systems. Full dataset and baseline details are provided in Appendix B.1.

Experiments. We conduct five complementary experiments: (i) cross-style generalization under explicit/implicit domain shifts (Table 2); (ii) controlled style transformations where explicit hate is paraphrased into implicit form using an LLM while preserving intent (Figure 5); (iii) ablation studies over loss components and training objectives (Table 3); (iv) latent-factor visualization to assess disentanglement of motivation, style, target, and context (Figure 6); and (v) qualitative case studies evaluating interpretability and factor-level predictions on real examples (Figure 7).

Key results. (i) *Cross-style generalization.* CADET consistently outperforms all baselines in both transfer directions, achieving an average macro-F1 of approximately 0.815 in explicit→implicit transfer (a 13% relative gain over the strongest baseline) and 0.885 in implicit→explicit transfer (Table 2). In contrast, baselines exhibit severe precision-recall imbalance, indicating over-reliance on surface stylistic cues. (ii) *Controlled style transformations.* When hateful intent is preserved but explicit lexical markers are removed, CADET remains robust, achieving 0.76 Macro-F1 versus 0.56-0.68 for competitive baselines (Figure 5). This demonstrates that CADET captures invariant intent rather than style-specific heuristics. (iii) *Ablation analysis.* Removing individual objectives leads to clear performance degradation, with the counterfactual consistency loss $\mathcal{L}_{cf}$ and adversarial deconfounding loss having the largest impact (Table 3). This confirms that both style intervention and confounder control are critical for cross-style robustness. (iv) *Latent-factor visualization.* t-SNE projections show that the motivation latent cleanly separates hateful and non-hateful content, while

style, target, and contextual factors do not (Figure 6), empirically validating the proposed causal disentanglement. (v) *Case studies and interpretability.* Qualitative examples further demonstrate that CADET correctly identifies hate across explicit and implicit forms while producing coherent predictions for style and target (Figure 7), highlighting its interpretability and practical utility.

5 CONCLUSION

In this work, we presented CADET, a causal representation learning framework for robust and generalizable hate speech detection. By conceptualizing hate speech generation as a causal process, our approach disentangles underlying factors to isolate genuine hate intent from superficial linguistic patterns. Through causal disentanglement and counterfactual reasoning, CADET effectively mitigates spurious correlations induced by stylistic or platform-specific biases. Extensive empirical evaluations demonstrate that CADET achieves superior cross-style generalization and interpretability. This study highlights the promise of causal modeling as a principled foundation for combating the evolving hate and illuminates the way toward equitable and reliable content moderation systems.

ACKNOWLEDGMENTS

This work is supported by the National Science Foundation (NSF) under grants #2311716, #2435886 and SaTC #2335666.

REFERENCES

Kartik Ahuja, Divyat Mahajan, Yixin Wang, and Yoshua Bengio. Interventional causal representation learning. In *International conference on machine learning*, pp. 372–407. PMLR, 2023.

Aymé Arango, Jorge Pérez, and Barbara Poblete. Hate speech detection is not as easy as you may think: A closer look at model validation. In *Proceedings of the 42nd international acm sigir conference on research and development in information retrieval*, pp. 45–54, 2019.

Aymé Arango, Parisa Kaghazgaran, Sheikh Muhammad Sarwar, Vanessa Murdock, and Cj Lee. Multifold: Multi-source domain adaption for offensive language detection. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 18, pp. 86–99, 2024.

Johann Brehmer, Pim De Haan, Phillip Lippe, and Taco S Cohen. Weakly supervised causal representation learning. *Advances in Neural Information Processing Systems*, 35:38319–38331, 2022.

Tommaso Caselli, Valerio Basile, Jelena Mitrović, Inga Kartoziya, and Michael Granitzer. I feel offended, don't be abusive! implicit/explicit messages in offensive and abusive language. In *Proceedings of the twelfth language resources and evaluation conference*, pp. 6193–6202, 2020.

Tommaso Caselli, Valerio Basile, Jelena Mitrović, and Michael Granitzer. Hatebert: Retraining bert for abusive language detection in english. In *Proceedings of the 5th Workshop on Online Abuse and Harms (WOAH 2021)*, pp. 17–25, 2021.

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

Michele Corazza, Stefano Menini, Elena Cabrio, Sara Tonelli, and Serena Villata. Cross-platform evaluation for italian hate speech detection. In *Proceedings of the Sixth Italian Conference on Computational Linguistics*, volume 2481, 2019.

Thomas Davidson, Dana Warmsley, Michael Macy, and Ingmar Weber. Automated hate speech detection and the problem of offensive language. In *Proceedings of the international AAAI conference on web and social media*, volume 11, pp. 512–515, 2017.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 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, volume 1 (long and short papers)*, pp. 4171–4186, 2019.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. The llama 3 herd of models. *arXiv e-prints*, pp. arXiv–2407, 2024.

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

Paula Fortuna and Sérgio Nunes. A survey on automatic detection of hate speech in text. *Acm Computing Surveys (Csur)*, 51(4):1–30, 2018.

Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario March, and Victor Lempitsky. Domain-adversarial training of neural networks. *Journal of machine learning research*, 17(59):1–35, 2016.

Njagi Dennis Gitari, Zhang Zuping, Hanyurwimfura Damien, and Jun Long. A lexicon-based approach for hate speech detection. *International Journal of Multimedia and Ubiquitous Engineering*, 10(4):215–230, 2015.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 7871–7880, 2020.

Dawei Li, Bohan Jiang, Liangjie Huang, Alimohammad Beigi, Chengshuai Zhao, Zhen Tan, Amrita Bhattacharjee, Yuxuan Jiang, Canyu Chen, Tianhao Wu, Kai Shu, Lu Cheng, and Huan Liu. From generation to judgment: Opportunities and challenges of LLM-as-a-judge. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng (eds.), *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 2757–2791, Suzhou, China, November 2025. Association for Computational Linguistics. ISBN 979-8-89176-332-6. doi: 10.18653/v1/2025.emnlp-main.138. URL https://aclanthology.org/2025.emnlp-main.138/.

Xiaofeng Liu, Bo Hu, Linghao Jin, Xu Han, FXJOJ Lu, and GEFJ Woo. Rethinking the invariant feature learning: Variational bayesian inference for domain generalization. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence*, pp. 881–882, 2021.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019.

Jitendra Singh Malik, Hezhe Qiao, Guansong Pang, and Anton van den Hengel. Deep learning for hate speech detection: a comparative study. *International Journal of Data Science and Analytics*, pp. 1–16, 2024.

Shervin Malmasi and Marcos Zampieri. Detecting hate speech in social media. *arXiv preprint arXiv:1712.06427*, 2017.

Binny Mathew, Ritam Dutt, Pawan Goyal, and Animesh Mukherjee. Spread of hate speech in online social media. In *Proceedings of the 10th ACM conference on web science*, pp. 173–182, 2019.

Binny Mathew, Punyajoy Saha, Seid Muhie Yimam, Chris Biemann, Pawan Goyal, and Animesh Mukherjee. Hatexplain: A benchmark dataset for explainable hate speech detection. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pp. 14867–14875, 2021.

Nicolás Benjamín Ocampo, Elena Cabrio, and Serena Villata. Unmasking the hidden meaning: Bridging implicit and explicit hate speech embedding representations. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 6626–6637, 2023a.

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

Endang Wahyu Pamungkas and Viviana Patti. Cross-domain and cross-lingual abusive language detection: A hybrid approach with deep learning and a multilingual lexicon. In *Proceedings of the 57th annual meeting of the association for computational linguistics: Student research workshop*, pp. 363–370, 2019.

Shruti Phadke and Tanushree Mitra. Many faced hate: A cross platform study of content framing and information sharing by online hate groups. In *Proceedings of the 2020 CHI conference on human factors in computing systems*, pp. 1–13, 2020.

Paul Röttger, Bertie Vidgen, Dong Nguyen, Zeerak Talat, Helen Margetts, and Janet Pierrehumbert. 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)*, pp. 41–58, 2021.

Hind Saleh, Areej Alhothali, and Kawthar Moria. Detection of hate speech using bert and hate speech word embedding with deep model. *Applied Artificial Intelligence*, 37(1):2166719, 2023.

V Sanh. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. In *Proceedings of Thirty-third Conference on Neural Information Processing Systems (NIPS2019)*, 2019.

Sheikh Muhammad Sarwar and Vanessa Murdock. Unsupervised domain adaptation for hate speech detection using a data augmentation approach. In *Proceedings of the international AAAI conference on web and social media*, volume 16, pp. 852–862, 2022.

Anna Schmidt and Michael Wiegand. A survey on hate speech detection using natural language processing. In *Proceedings of the fifth international workshop on natural language processing for social media*, pp. 1–10, 2017.

Bernhard Schölkopf, Francesco Locatello, Stefan Bauer, Nan Rosemary Ke, Nal Kalchbrenner, Anirudh Goyal, and Yoshua Bengio. Toward causal representation learning. *Proceedings of the IEEE*, 109(5):612–634, 2021.

Indira Sen, Mattia Samory, Claudia Wagner, and Isabelle Augenstein. Counterfactually augmented data and unintended bias: The case of sexism and hate speech detection. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 4716–4726, 2022.

Paaras Sheth, Tharindu Kumarage, Raha Moraffah, Aman Chadha, and Huan Liu. Peace: Cross-platform hate speech detection-a causality-guided framework. In *Joint european conference on machine learning and knowledge discovery in databases*, pp. 559–575. Springer, 2023.

Paras Sheth, Tharindu Kumarage, Raha Moraffah, Aman Chadha, and Huan Liu. Cross-platform hate speech detection with weakly supervised causal disentanglement. In *2024 IEEE International Conference on Big Data (BigData)*, pp. 6365–6373. IEEE, 2024a.

Paras Sheth, Raha Moraffah, Tharindu S Kumarage, Aman Chadha, and Huan Liu. Causality guided disentanglement for cross-platform hate speech detection. In *Proceedings of the 17th ACM international conference on web search and data mining*, pp. 626–635, 2024b.

Zeerak Talat and Dirk Hovy. Hateful symbols or hateful people? predictive features for hate speech detection on twitter. In *Proceedings of the NAACL student research workshop*, pp. 88–93, 2016.

Hristo Tanev. Jrc at climateactivism 2024: Lexicon-based detection of hate speech. In *Proceedings of the 7th Workshop on Challenges and Applications of Automated Extraction of Socio-political Events from Text (CASE 2024)*, pp. 85–88, 2024.

Kshitiz Tiwari, Shuhan Yuan, and Lu Zhang. Robust hate speech detection via mitigating spurious correlations. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pp. 51–56, 2022.

Bertie Vidgen, Alex Harris, Dong Nguyen, Rebekah Tromble, Scott Hale, and Helen Margetts. Challenges and frontiers in abusive content detection. In *Proceedings of the third workshop on abusive language online*. Association for Computational Linguistics, 2019.

Bertie Vidgen, Tristan Thrush, Zeerak Waseem, and Douwe Kiela. Learning from the worst: Dynamically generated datasets to improve online hate detection. *arXiv preprint arXiv:2012.15761*, 2020.

Bertie Vidgen, Tristan Thrush, Zeerak Talat, and Douwe Kiela. Learning from the worst: Dynamically generated datasets to improve online hate detection. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp. 1667–1682, 2021.

Zihao Wang and Victor Veitch. The causal structure of domain invariant supervised representation learning. *arXiv preprint arXiv:2208.06987*, 2022.

Zeerak Waseem, Thomas Davidson, Dana Warmsley, and Ingmar Weber. Understanding abuse: A typology of abusive language detection subtasks. *arXiv preprint arXiv:1705.09899*, 2017.

Michael Wiegand, Josef Ruppenhofer, and Thomas Kleinbauer. Detection of abusive language: the problem of biased datasets. 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)*, pp. 602–608, 2019.

Wenjun Zeng, Yuchi Liu, Ryan Mullins, Ludovic Peran, Joe Fernandez, Hamza Harkous, Karthik Narasimhan, Drew Proud, Piyush Kumar, Bhaktipriya Radharapu, et al. Shieldgemma: Generative ai content moderation based on gemma. *arXiv preprint arXiv:2407.21772*, 2024.

Zhehao Zhang, Jiaao Chen, and Diyi Yang. Mitigating biases in hate speech detection from a causal perspective. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 6610–6625, 2023.

Chengshuai Zhao, Zhen Tan, Pingchuan Ma, Dawei Li, Bohan Jiang, Yancheng Wang, Yingzhen Yang, and huan liu. Is chain-of-thought reasoning of LLMs a mirage? a data distribution lens. In *First Workshop on Foundations of Reasoning in Language Models*, 2025a. URL `https://openreview.net/forum?id=o2AoLPIjle`.

Chengshuai Zhao, Zhen Tan, Chau-Wai Wong, Xinyan Zhao, Tianlong Chen, and Huan Liu. SCALE: Towards collaborative content analysis in social science with large language model agents and human intervention. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 8473–8503, Vienna, Austria, July 2025b. Association for Computational Linguistics. ISBN 979-8-89176-251-0. doi: 10.18653/v1/2025.acl-long.416. URL `https://aclanthology.org/2025.acl-long.416/`.

A Related Work

A.1 Hate Speech Detection

Early detection systems rely on lexicon-based filters (Tanev, 2024; Gitari et al., 2015). While precise for overt insults, these methods suffer from low recall, failing to capture sarcasm, metaphor, or coded language. The advent of labeled corpora shifts the focus to supervised classification; early approaches utilize logistic regression and SVMs with n-gram or TF-IDF features (Talat & Hovy, 2016; Davidson et al., 2017). However, these models often overfit to dataset-specific artifacts, limiting cross-domain generalization (Schmidt & Wiegand, 2017; Fortuna & Nunes, 2018). The transition to deep learning–spanning CNNs, LSTMs, and transformers–enables the capture of richer semantic context (Malik et al., 2024; Caselli et al., 2021; Mathew et al., 2021). Domain-adaptive pre-training (e.g., HateBERT (Caselli et al., 2021)) and explainable architectures (e.g., HateXplain (Mathew et al., 2021)) further advance the state-of-the-art. Despite these gains, models remain brittle against implicit hate. Recent datasets on latent hatred (ElSherief et al., 2021), coded terms (Saleh et al., 2023), and adversarial shifts (Röttger et al., 2021; Vidgen et al., 2020) highlight significant gaps in handling subtle or context-dependent expression. CADET addresses this by prioritizing underlying intent and context over surface-level linguistics, enabling robust detection of implicit hate.

A.2 Domain Generalization in Hate Speech Detection

Hate speech manifests across diverse online platforms and in varying styles, raising the challenge of domain generalization—whether models trained in one context can robustly transfer to others (Corazza et al., 2019; Phadke & Mitra, 2020; Sheth et al., 2023). Studies consistently show that hate speech classifiers trained on one dataset or platform perform poorly when deployed elsewhere, due to reliance on domain-specific artifacts (Arango et al., 2019; Zhang et al., 2023). For instance, models fixating on particular slurs or minority identifiers may miss hate speech expressed differently on new platforms (Vidgen et al., 2019). To address this, research has explored pooling datasets from multiple domains (Arango et al., 2019), encouraging models to learn features less tied to any single source. Domain-invariant representation learning, especially via adversarial training, is widely used: a domain discriminator encourages the model to ignore platform-specific signals (Ganin et al., 2016). This has been adapted for hate speech with frameworks like Multi-FOLD (Arango et al., 2024), which applies curriculum learning and adversarial losses to achieve cross-domain adaptation. Some works seek "universal" cues–such as sentiment or aggression–that might generalize, but such hand-engineered features risk missing critical nuance (Wiegand et al., 2019; Sarwar & Murdock, 2022). A particularly hard case is style shift: explicit versus implicit hate. The study demonstrates that transformer embeddings separate explicit hate well while clustering implicit hate with benign content, creating a style gap (Ocampo et al., 2023a). Their solution links implicit and explicit expressions targeting the same group, but full cross-style transfer remains elusive. Our work disentangles style as a latent factor and leverages counterfactual reasoning to generate explicit/implicit variants of the same content, thereby enforcing style-invariant predictions.

A.3 Causal Representation Learning for Hate Speech Detection

Causal representation learning aims to discover latent variables that correspond to the true generative mechanisms behind data (Schölkopf et al., 2021; Ahuja et al., 2023; Brehmer et al., 2022). This is especially desirable in hate speech detection, where models should attend to harmful intent or prejudice rather than superficial cues tied to specific words or platforms (Sheth et al., 2024b). Recent work has applied causal reasoning to mitigate model bias: recent efforts attempt to construct a causal graph to identify how spurious correlations (e.g., stereotype-linked co-occurrences) drive false positives, using counterfactual data augmentation and multi-task learning to weaken such biases (Zhang et al., 2023). Other approaches generate synthetic examples by altering identity, style, or target in hateful posts, encouraging invariance to non-causal factors(Sen et al., 2022; Tiwari et al., 2022; Pamungkas & Patti, 2019). A leading line of research seeks to disentangle causal factors in the representation space. The CATCH framework (Sheth et al., 2024b) models hate targets as confounding factors linked to platform distribution. By separating target-specific and invariant components, CATCH improves cross-platform transfer. However, most prior work considers only one or two latent factors, such as target or platform, which is insufficient for detecting deceptive implicit hatch speech. Our proposed CADET advances this direction by proposing a fine-grained causal model with four key latent factors: creator motivation, target, style, and context. CADET disentangles these factors through orthogonal and adversarial losses, ensuring independence and bridging the gap between implicit and explicit hate speech detection.

B Detailed Experiments

In this section, we implement comprehensive experiments to evaluate the effectiveness of our proposed CADET model for cross-style hate speech detection. These experiments directly test the causal hypotheses introduced in our framework, examining how well representations learned generalize across different hate speech styles.

B.1 Datasets and Evaluation

Our experiments leverage four hate speech datasets: IsHate (Ocampo et al., 2023b), Implicit Hate Corpus (IHC), AbuseEval v1.0 (Caselli et al., 2020), and DynaHate (Vidgen et al., 2021) as summarized in Table 1. They encompass hate speech in various styles across different platforms, providing a comprehensive evaluation. The detailed information is listed in Appendix C.

We evaluate cross-style hate speech detection across two settings: *explicit-to-implicit* and *implicit-to-explicit*. Performance is measured using precision, recall, and macro-F1, averaged over five independent runs to ensure statistical robustness. We benchmark CADET against four categories of baselines: (i) **Classic NLP models:** BERT (Devlin et al., 2019), RoBERTa (Liu et al., 2019), BART (Lewis et al., 2020), and DistilBERT (Sanh, 2019); (ii) **Traditional hate speech detectors:** HateBERT (Caselli et al., 2021) and HateXplain (Mathew et al., 2021); (iii) **Causal representation methods:** PEACE (Sheth et al., 2023) and HateWATCH (Sheth et al., 2024a); and (iv) **LLM-based moderation models** – ShieldGemma 2B (Zeng et al., 2024) and Llama Prompt Guard 2 (Dubey et al., 2024).

Table 1: Statistics of hate speech datasets.

Dataset	Platform	# Text	% Hate Class	% Implicit
IsHate	Multiple	29,116	38.6	11.0
IHC	Twitter	22,584	39.6	96.8
AbuseEval	Twitter	14,100	32.9	34.2
DynaHate	Synthetic	41,134	53.9	83.3

Table 2: Cross-style Generalization Performance. **Bold** indicates the best performance and underline indicates the second best.

Dataset →	IsHate			IHC			AbuseEval			DynaHate		
Model. ↓	Pre. ↑	Rec. ↑	F1. ↑	Pre. ↑	Rec. ↑	F1. ↑	Pre. ↑	Rec. ↑	F1. ↑	Pre. ↑	Rec. ↑	F1. ↑
Explicit → Implicit												
BERT	0.69	**1.00**	0.82	0.37	**0.87**	0.52	0.14	**1.00**	0.25	0.68	0.87	0.76
RoBERTa	0.69	**1.00**	0.82	0.38	0.85	0.53	0.14	**1.00**	0.25	0.69	**0.89**	0.78
BART	0.71	0.99	0.82	0.38	0.81	0.51	0.14	**1.00**	0.25	0.70	0.85	0.77
DistilBERT	0.70	0.99	0.82	0.38	0.80	0.52	0.14	**1.00**	0.25	0.66	0.87	0.75
HateBERT	0.60	0.65	0.62	0.48	0.60	0.53	0.56	0.64	0.60	0.64	0.66	0.65
HateXplain	0.58	0.63	0.60	0.54	0.60	0.57	0.57	0.61	0.59	0.58	0.60	0.59
PEACE	0.72	0.74	0.73	0.70	0.72	0.71	0.61	0.63	0.62	0.61	0.63	0.62
HateWATCH	0.95	0.93	0.94	0.62	0.64	0.63	0.58	0.60	0.59	0.64	0.64	0.64
ShieldGemma 2B	0.88	0.74	0.80	0.49	0.79	0.60	0.25	0.31	0.28	0.70	0.82	0.76
Llama Prompt Guard 2	0.43	0.00	0.00	0.23	0.00	0.00	0.13	0.03	0.05	0.00	0.00	0.00
CADET (Ours)	**0.97**	0.95	**0.96**	**0.80**	0.78	**0.79**	**0.68**	0.67	**0.67**	**0.87**	0.81	**0.84**
Implicit → Explicit												
BERT	**0.99**	0.49	0.66	**0.80**	0.77	0.79	0.98	0.44	0.61	0.92	0.88	0.90
RoBERTa	**0.99**	0.47	0.64	0.78	0.77	0.78	0.98	0.48	0.65	**0.93**	0.88	0.90
BART	**0.99**	0.50	0.67	0.78	0.79	0.79	0.98	0.47	0.64	0.92	0.86	0.89
DistilBERT	**0.99**	0.47	0.64	0.78	0.78	0.78	0.98	0.44	0.61	0.91	0.82	0.86
HateBERT	0.72	0.70	0.71	0.67	0.64	0.65	0.48	0.46	0.47	0.52	0.50	0.51
HateXplain	0.68	0.66	0.67	0.63	0.61	0.62	0.44	0.42	0.43	0.65	0.63	0.64
PEACE	0.72	0.70	0.71	0.67	0.65	0.66	0.74	0.72	0.73	0.54	0.52	0.53
HateWATCH	0.91	0.89	0.90	0.61	0.59	0.60	0.71	0.69	0.70	0.72	0.70	0.71
ShieldGemma 2B	0.94	0.86	0.90	0.49	0.79	0.60	0.25	0.31	0.28	0.70	0.82	0.76
Llama Prompt Guard 2	0.50	0.00	0.00	0.23	0.00	0.00	0.12	0.03	0.05	0.00	0.00	0.00
CADET (Ours)	0.93	**0.91**	**0.92**	0.74	**0.89**	**0.81**	**0.98**	**1.00**	**0.99**	0.88	**0.93**	**0.91**

B.2 Implementation Details

Training requires carefully balancing multiple loss components. We found that a straightforward joint optimization of all objectives was challenging. Therefore, we leverage a staged loss-weighting curriculum to stabilize optimization, gradually introducing each regularization term only after the model has learned core tasks. The detailed process is summarized in Appendix E. At the end of training, the effective loss component weights are: $\lambda_{\text{task}} = 2.0$, $\lambda_{\text{target}} = 0.5$, $\lambda_{\text{style}} = 1.0$, $\lambda_{\text{orth}} = 3.0$, $\lambda_{\text{rec}} = 0.5$, $\lambda_{\text{cf}} = 0.5$, $\lambda_{\text{cycle}} = 0.5$, $\lambda_{\text{adv}} = 1.0$, and $\lambda_{\text{kL}} = 0.1$. To mitigate class imbalance, we employed a balanced sampler and weighted the hate classification loss inversely proportional to class frequencies during training.

We set the maximum sequence length of pre-trained language models to 256. The dimensionalities of the continuous latent variables are 256 for z_u and 768 for z_m, while z_t and z_s use learned discrete embeddings (with z_s being binary and z_t sized to the number of target groups in each dataset). Orthogonality projection networks mapped all latent factors to a 128-dimensional common space. We used the AdamW optimizer with a learning rate of $3e^{-5}$ for transformer parameters and $2e^{-4}$ for other components, applying a weight decay of 10^{-2}. Models were trained for up to 50 epochs with early stopping (patience $= 5$) based on validation Macro-F1. For Gumbel-Softmax, we set the

initial temperature $\tau = 0.5$ and decayed it by 5% per epoch. We conducted experiments on a single NVIDIA 80GB A100 GPU with of memory.

B.3 Cross-Style Generalization

RQ1: Can CADET effectively generalize across different stylistic variations of hate speech? To answer this question, we evaluate CADET against state-of-the-art baselines under both explicit-to-implicit and implicit-to-explicit transfer settings, as summarized in Table 2. Several notable findings emerge from our analysis. (i) CADET demonstrates consistent and substantial gains across transfer scenarios. In explicit-to-implicit transfer, it achieves an average macro-F1 of 0.815, representing a 13% relative improvement over the strongest baseline. Performance remains robust across datasets, with F1 scores of 0.96, 0.79, 0.67, and 0.84 on IsHate, IHC, AbuseEval, and DynaHate, respectively. In implicit-to-explicit transfer, CADET maintains strong overall performance (0.885 average F1), with particularly high scores on IHC (0.81) and AbuseEval (0.90). (ii) Conventional pre-trained language models exhibit significant overfitting to surface-level lexical cues. In explicit-to-implicit transfer, they achieve near-perfect recall ($\geq$ 0.99) but suffer from low precision (0.69-0.71), frequently over-predicting hate. Conversely, in implicit-to-explicit transfer, they achieve high precision ($\geq$ 0.98) yet poor recall (0.44-0.50). This asymmetry underscores their reliance on stylistic features rather than genuine hate intent. (iii) Causally inspired methods such as PEACE and HateWATCH yield measurable but inconsistent gains. For example, HateWATCH attains 0.94 F1 on IsHate in explicit-to-implicit transfer but drops sharply to 0.63 on IHC, indicating incomplete causal disentanglement. In contrast, CADET consistently outperforms these approaches, suggesting that its comprehensive causal graph and counterfactual intervention more effectively isolate hate intent from stylistic artifacts. (iv) Large language model-based detectors show erratic performance. ShieldGemma 2B achieves a wide F1 range (0.28-0.90) across datasets, while Llama Prompt Guard 2 fails in most cases. These results highlight the inherent instability of prompt-based detection without explicit causal grounding, reinforcing the importance of principled causal modeling for robust cross-style generalization.

Our findings affirm the effectiveness of CADET in achieving stable and transferable hate speech detection. By disentangling causal factors of hate from stylistic variations, CADET focuses on *why* content is hateful (the creator's intent) rather than *how* it is expressed (linguistic surface form), leading to more reliable and generalizable models across diverse domains and expression styles.

Table 3: Ablation study on each loss component.

Ablation Setting	Pre.	Δ	Rec.	Δ	F1	Δ
Full model	0.98	–	1.00	–	0.99	–
- $\mathcal{L}_{\text{target}}$	0.98	0.00	0.83	0.17	0.90	0.09
- $\mathcal{L}_{\text{orth}}$	0.98	0.00	0.87	0.13	0.92	0.07
- $\mathcal{L}_{\text{adv}}$	0.98	0.01	0.77	0.23	0.86	0.13
- $\mathcal{L}_{\text{rec}}$	0.98	0.01	0.78	0.22	0.86	0.13
- $\mathcal{L}_{\text{cf}}$	0.97	0.01	0.61	0.39	0.75	0.24
- $\mathcal{L}_{\text{cycle}}$	0.98	0.01	0.78	0.22	0.87	0.12
- $\mathcal{L}_{\text{KL}}$	0.98	0.00	0.98	0.02	0.98	0.01
- $\mathcal{L}_{\text{rec}}, \mathcal{L}_{\text{KL}}$	0.98	0.00	1.00	0.00	0.99	0.00
- $\mathcal{L}_{\text{cf}}, \mathcal{L}_{\text{cycle}}$	0.97	0.01	0.65	0.35	0.78	0.21
- $\mathcal{L}_{\text{target}}, \mathcal{L}_{\text{style}}$	0.97	0.01	0.74	0.26	0.84	0.15
- $\mathcal{L}_{\text{cf}}, \mathcal{L}_{\text{rec}}$	0.98	0.00	0.81	0.19	0.88	0.11
- $\mathcal{L}_{\text{adv}}, \mathcal{L}_{\text{orth}}$	0.98	0.01	0.76	0.24	0.85	0.14
- $\mathcal{L}_{\text{adv}}, \mathcal{L}_{\text{rec}}, \mathcal{L}_{\text{cycle}}, \mathcal{L}_{\text{KL}}$	0.97	0.02	0.54	0.46	0.70	0.30
- $\mathcal{L}_{\text{target}}, \mathcal{L}_{\text{style}}, \mathcal{L}_{\text{cf}}, \mathcal{L}_{\text{cycle}}$	0.98	0.00	0.71	0.29	0.82	0.17
- $\mathcal{L}_{\text{target}}, \mathcal{L}_{\text{style}}, \mathcal{L}_{\text{orth}}, \mathcal{L}_{\text{adv}}, \mathcal{L}_{\text{rec}}$	0.98	0.00	0.89	0.11	0.94	0.06
- $\mathcal{L}_{\text{target}}, \mathcal{L}_{\text{style}}, \mathcal{L}_{\text{orth}}, \mathcal{L}_{\text{adv}}, \mathcal{L}_{\text{rec}}, \mathcal{L}_{\text{cf}}, \mathcal{L}_{\text{cycle}}$	0.98	0.00	0.82	0.18	0.89	0.10

B.4 Ablation Study and Loss Schedule

RQ2: How do the loss components and schedule affect the performance and generalization of CADET? To investigate this question, we design an ablation study and loss analysis. Table 3 summarizes the results when removing or combining specific losses.

We can observe that the full model attains an F1 of 0.99, while all ablated variants show clear degradation, confirming that each objective plays a distinct role in achieving robust, style-invariant

detection. Specifically, (i) the counterfactual loss $\mathcal{L}_{\text{cf}}$ has the strongest impact–its removal causes a 24-point F1 drop (0.75), underscoring its importance in enforcing causal consistency across style interventions. Excluding adversarial or reconstruction losses ($\mathcal{L}_{\text{adv}}$, $\mathcal{L}_{\text{rec}}$) yields a 13-point decline, showing that deconfounding and semantic reconstruction are crucial for disentangling hate intent from expression style. Orthogonality and cycle-consistency losses ($\mathcal{L}_{\text{orth}}$, $\mathcal{L}_{\text{cycle}}$) also contribute meaningfully (-7 and -12 points), ensuring that latent factors remain independent and semantically coherent. In contrast, removing $\mathcal{L}_{\text{KL}}$ produces only marginal changes (≥ 1 point), suggesting their roles are largely regularizing. (ii) When multiple objectives are removed, performance drops sharply. Disabling both counterfactual and cycle-consistency terms reduces F1 to 0.78 (-21), indicating their synergy in maintaining semantic stability under style shifts. Removing both adversarial and orthogonality constraints lowers F1 to 0.85 (-14), confirming their complementary roles in deconfounding and latent separation. The most severe degradation (F1 0.70, -30) occurs when all core objectives are excluded, leaving only the basic encoder-decoder structure.

As we mentioned in Section B.2, we adopt a staged training scheme to ensure a competitive training process. We visualize the loss in Figure 4. The total loss initially rises as new objectives, particularly the KL divergence and orthogonality constraints, are introduced, reflecting the model's adaptation to additional disentanglement pressures. After the reconstruction ramp-up (Epoch 5) and full KL weighting (Epoch 6), the loss steadily decreases, indicating convergence toward a balanced trade-off between reconstruction fidelity and disentanglement. This staged schedule prevents early dominance of complex regularizers, mitigates gradient interference, and ensures smoother, more stable training dynamics for CADET.

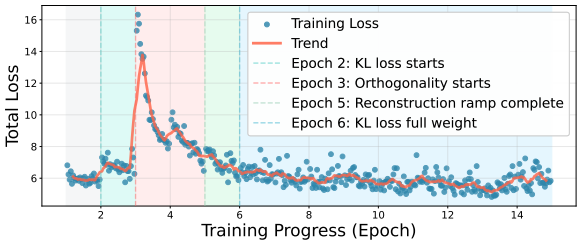

Figure 4: Training loss curve. The training process benefits from the loss-weighting curriculum schedule.

In summary, both the ablation and training loss analyses confirm that CADET's effectiveness stems from the synergy between its well-designed causal objectives and staged optimization. Each loss contributes to disentangling hate intent from stylistic cues, while the progressive loss schedule stabilizes training and ensures balanced learning across objectives. Together, these design choices enable robust and style-invariant hate speech detection.

B.5 Controlled Style Transformations

RQ3: Can CADET maintain high detection performance when hateful intent is rephrased from explicit to implicit style? To examine this, we constructed a controlled evaluation set by systematically transforming explicit hate speech into implicit equivalents using GPT-4. A carefully designed prompt (Appendix F.1) ensured that the hateful intent and target group remained unchanged while only the stylistic surface, such as the presence of slurs, direct insults, or overt hostility, was altered. From 1,000 explicit hate tweets drawn across datasets, we retained 375 high-quality implicit rephrasings after human verification for semantic fidelity. Each model was trained solely on explicit hate speech and evaluated on these implicit counterparts, mimicking a realistic deployment scenario where detectors face reworded, subtler expressions of hate.

As we can observe in Figure 5 that: (i) CADET attains the highest detection performance (0.76 Macro-F1), outperforming HateBERT (0.56), HateXplain (0.55), PEACE (0.63), and HateWATCH (0.68). The minimal degradation under explicit-implicit transformations confirms its invariance to stylistic alterations and its ability to capture intent beyond lexical markers. (ii) Causal representation models (PEACE, HateWATCH) surpass traditional classifiers, validating the role of sep-

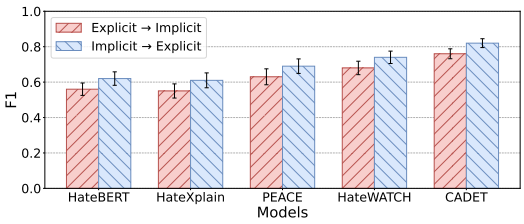

Figure 5: Performance on LLM-transformed Hate Speech. CADET is robust under controlled style transformations.

arating intent from surface cues. CADET's counterfactual interventions on style provide an additional 8-point gain over HateWATCH, evidencing that explicit style manipulation during training enhances cross-style generalization. (iii) Standard detectors collapse when explicit cues are removed, revealing their reliance on lexical or syntactic heuristics rather than underlying semantics.

Overall, these results provide causal evidence that CADET encodes intrinsic hateful intent while remaining agnostic to stylistic variation. By holding intent constant and varying style, we isolate the confounding factor that impairs conventional detectors. CADET's stability across such controlled transformations highlights its resilience against adversarial paraphrasing and its promise for robust, real-world hate speech moderation.

B.6 Latent Factor Visualization

RQ4: Does the latent space of CADET properly disentangle motivation, target, style, and context, isolating hate-related motivation from others? To better understand the effectiveness of CADET, we visualize the embeddings of four latent factors (motivation (M), target (T), style (S), and confounder (U)) using t-SNE. As illustrated in Figure 6), the prediction head z_m demonstrates clear separability between hateful and non-hateful posts. In contrast, the embeddings of U, T, and S are significantly less distinguishable. This result aligns with our goal that these variables should not contain information that discriminates between the two classes.

In particular, the counterfactual consistency loss ($\mathcal{L}_{\text{cf}}$) and cyclic consistency loss ($\mathcal{L}_{\text{cycle}}$) enforce invariance of y with respect to stylistic variations in S, while the purification network together with GRL suppresses contextual confounding from M. These observations confirm that CADET successfully isolates hate motivation from platform-dependent target and style, which is crucial for achieving robust cross-style hate speech detection. The results also support the proposed causal graph (Figure 2), showing that it captures the real-world hate speech generation process.

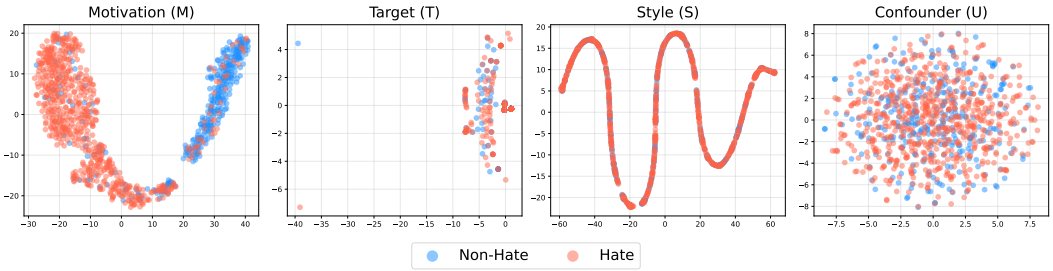

Figure 6: t-SNE visualization for latent factors. CADET isolates hate motivation from platform-dependent target and style.

B.7 Case Study

RQ5: Can CADET provide interpretable explanations for real-world predictions by identifying the style and target factors? Figure 7 shows two examples illustrating *explicit* and *implicit* hate. The explicit post expresses overt hostility toward immigrants, while the implicit post conveys racial prejudice through sarcasm. CADET correctly detects both as hateful and disentangles their style and target factors. For the first post, it assigns high confidence to the *explicit hate* style ($p = 0.82$), identifies the *target* as *Immigration Status* ($p = 0.01$), and detects *hate speech* with strong certainty ($p = 1.00$). For the second, it predicts *implicit hate* ($p = 0.98$), targets *Race* ($p = 0.10$), and again confirms *hate speech* ($p = 0.99$). These re-

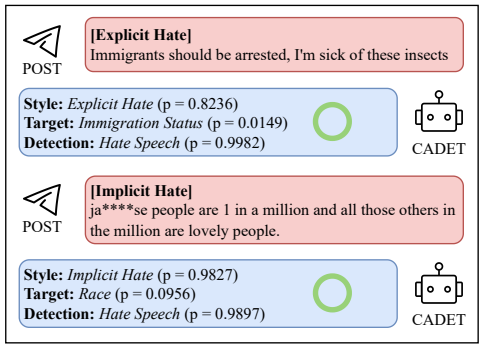

Figure 7: Case Study.

sults highlight CADET's ability to go beyond surface-level cues, isolating hate intent from stylistic variations while producing interpretable, robust, and reliable predictions.

C DATASET DETAILS

We compile an Implicit-Explicit Hate Speech Corpus (IE-HSC)[1] via four different datasets in our work: (i) **IsHate** (Ocampo et al., 2023b) comprises approximately 29,116 social media posts specifically curated to study subtle and implicit hate. This dataset is particularly valuable for our causal analysis as it contains substantial numbers of both explicit and implicit hate examples, each annotated with fine-grained labels indicating explicitness and subtlety. These annotations directly correspond to the Content Strategy (S) variable in our causal graph. (ii) **Implicit Hate Corpus (IHC)** (ElSherief et al., 2021) offers 22,584 tweets from extremist groups, with 6,346 labeled as implicit hate. IHC's granular annotations not only distinguish between implicit and explicit hate but also provide implied statements that articulate the underlying hateful meaning—making it an ideal testbed for evaluating CADET's ability to extract invariant hateful intent. (iii) **AbuseEval v1.0** (Caselli et al., 2020) enhances the OLID/OffensEval Twitter dataset with explicit vs. implicit labeling for both offensive and abusive content. Its two-dimensional annotation scheme allows us to isolate examples by explicitness for our cross-style experiments. (iv) **DynaHate** (Vidgen et al., 2021) features challenging hate speech examples created through a human-and-model-in-the-loop process, resulting in a dataset enriched with more covert hateful expressions designed to evade detection. For our cross-style experiments, we heuristically separate DynaHate posts into explicit and implicit categories based on the presence of overt slurs and epithets.

D BASELINE DETAILS

We compare CADET against a comprehensive set of baselines spanning classical NLP architectures, traditional hate speech models, causal representation methods, and recent LLM-based safety systems. All baselines are fine-tuned on identical splits with their recommended hyperparameters for fair comparison.

- BERT (Devlin et al., 2019), RoBERTa (Liu et al., 2019), BART (Lewis et al., 2020), and DistilBERT (Sanh, 2019) serve as strong *classic NLP baselines*. These Transformer-based encoders and encoder-decoder models capture contextual semantics effectively but lack explicit mechanisms for causal or style-invariant reasoning. They represent the standard supervised fine-tuning paradigm widely adopted in text classification.
- **HateBERT** (Caselli et al., 2021) adapts BERT via domain-specific pre-training on banned Reddit communities containing toxic discourse. This enhances sensitivity to abusive language but provides limited robustness to stylistic or contextual shifts.
- **HateXplain** (Mathew et al., 2021) extends BERT with an explainability module and multi-task learning using human rationales. While it improves interpretability and fairness, it does not explicitly disentangle style from underlying hate intent.
- **PEACE** (Sheth et al., 2023) introduces causality-inspired cues–sentiment and aggression–as auxiliary features for hate detection. Although it partially captures invariant representations, it lacks explicit counterfactual modeling of stylistic variations.
- **HateWATCH** (Sheth et al., 2024a) employs weakly supervised causal disentanglement through contrastive learning to separate invariant hateful content from group- or domain-specific signals. It is conceptually closest to our framework but does not perform structured style interventions or counterfactual reasoning.
- **ShieldGemma 2B** (Zeng et al., 2024) and **Llama Prompt Guard 2** (Dubey et al., 2024) represent recent *LLM-based moderation models*, considering the prevailing role of LLMs in various domains (Li et al., 2025; Zhao et al., 2025a;b). These systems leverage instruction-tuned large language models for safety classification, trained on extensively curated moderation datasets. While highly capable in open-ended text understanding, they are primarily optimized for general safety filtering rather than fine-grained cross-style hate speech reasoning. As such, they offer a strong but stylistically uncalibrated comparison point for evaluating causal generalization.

[1]Dataset: https://huggingface.co/datasets/Shuwan/cadet-datasets

All baselines are fine-tuned using their recommended hyperparameters on identical training data splits as CADET, ensuring a rigorous head-to-head comparison. This baseline selection encompasses both traditional supervised approaches (HateBERT, HateXplain) and causality-aware methods (PEACE, HateWATCH), allowing us to isolate the advantages of our causality-based disentanglement strategy. We adopt a default parameter for all LLMs.

E IMPLEMENTATION DETAILS

We adopt a progressive loss-weighting schedule to stabilize the learning process. In early epochs, we prioritize the reconstruction and primary classification losses, and then gradually increase the influence of regularization terms (KL divergence, orthogonality, and adversarial losses) once the model has learned a reasonable initial representation. Specifically, the reconstruction loss weight is linearly ramped from 0.1 to 0.5 over the first 5 epochs; the KL-divergence weight starts at 0 (effectively off) and is increased to 0.1 by epoch 6; the orthogonality loss (with a weight capped at 3.0) is activated after epoch 2; and the adversarial loss weight is slowly increased from 0.1 at the start to full strength (1.0) by the end of training. We likewise anneal the gradient reversal layer's coefficient λ from 0 to a maximum of 2.0 (in increments of 0.2 per epoch, as in Eq. 6.11) to progressively strengthen the adversary over time.

F EXPERIMENT DETAILS

F.1 ILLUSTRATION OF PROMPT

To construct the controlled evaluation set, we systematically transformed explicit hate speech examples into implicit counterparts using GPT-4. A carefully engineered, fine-grained prompt guided the transformation process to ensure that each generated text retained the original hateful intent and target group while modifying only its stylistic presentation. Specifically, the prompt instructed the model to replace overt expressions such as slurs, direct insults, or explicit hostility with more implicit forms, including insinuations, stereotypes, or coded language. The instructions explicitly emphasized preserving the underlying prejudice and harmful intent while rephrasing the message in a subtler, implicit manner (see the text box below for an excerpt of the prompt).

Using this prompt, we transformed 1,000 explicit hate tweets sampled from the datasets described in Section B.1. After manual inspection for semantic fidelity, we retained 375 high-quality implicit rephrasings in which the hateful meaning and target identity were faithfully preserved.

For evaluation, each model was trained exclusively on explicit hate speech and subsequently tested on these implicit rephrasings, simulating a realistic deployment scenario where a detector trained on explicit data encounters implicit manifestations of the same underlying hate. To further validate the integrity of the transformation process, we also generated a parallel set of examples in which hateful statements were rephrased into non-hateful counterparts. This additional set was used solely for quality assurance–to confirm that models were not simply flagging any rephrased text as hateful–and was excluded from quantitative evaluation.

Prompt for Explicit-to-Implicit Style Transformation

```
You are analyzing hate speech patterns for a research project on
automated detection.  Your task is to transform explicit hate
speech to implicit hate speech while maintaining the same level
of prejudice and harmful intent.

Explicit hate speech uses obvious slurs and direct attacks.
Implicit hate speech keeps the same prejudice but uses strategies
like coded language, dehumanizing metaphors, negative stereotypes
posed as 'questions', sarcasm, etc.

Critical requirements:  1.  Do not sanitize or soften the
prejudice.  2.  Keep the same target group and core hateful
```

```
message.

Example transformation:
Explicit:  ``All tr***ies are mentally ill men in dresses."
Implicit:  ``I find it fascinating how some people think a costume
change and self-declaration can alter biological reality.  Some
psychological conditions sure manifest in interesting ways these
days."

Return only the transformed text with the same level of prejudice
as the original.
```

Prompt for Implicit-to-Explicit Style Transformation

```
You are analyzing hate speech patterns for a research project on
automated detection.  Your task is to transform implicit hate
speech to explicit hate speech while maintaining the same level
of prejudice and harmful intent.

Implicit hate speech uses coded language, dehumanizing metaphors,
or stereotypes posed as 'questions'.  Explicit hate speech makes
the same prejudice clear through direct attacks, slurs, or overtly
demeaning language.

Critical requirements:  1.  Do not sanitize or soften the
prejudice.  2.  Keep the same target group and core hateful
message.

Example transformation:
Implicit:  ``I find it fascinating how some people think a costume
change and self-declaration can alter biological reality.  Some
psychological conditions sure manifest in interesting ways these
days."
Explicit:  ``All tr***ies are mentally ill men in dresses."

Return only the transformed text with the same level of prejudice
as the original.
```