Learn from Failure: Causality-guided Contrastive Learning for Generalizable Implicit Hate Speech Detection

Tianming Jiang
School of Information Management,
Central China Normal University, Wuhan, China
tmjiang@ccnu.edu.cn

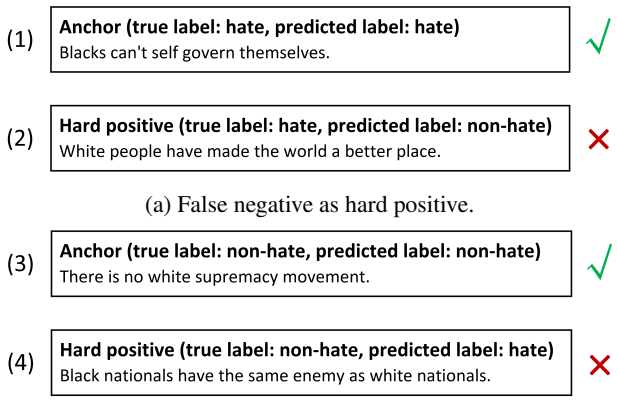
Abstract

Implicit hate speech presents a significant challenge for automatic detection systems due to its subtlety and ambiguity. Traditional models trained using empirical risk minimization (ERM) often rely on correlations between class labels and spurious attributes, which leads to poor performance on data lacking these correlations. In this paper, we propose a novel approach using causality-guided contrastive learning (CCL) to enhance the generalizability of implicit hate speech detection. Since ERM tends to identify spurious attributes, CCL works by aligning the representations of samples with the same class but opposite spurious attributes, identified through ERM's inference failure. This method reduces the model's reliance on spurious correlations, allowing it to learn more robust features and handle diverse, nuanced contexts better. Our extensive experiments on multiple implicit hate speech datasets show that our approach outperforms current state-of-the-art methods in cross-domain generalization.

1 Introduction

With the proliferation of social media platforms, tremendous hate speeches are created and spread (Fortuna and Nunes, 2018). Hate speech targeting individuals based on religion, gender, or other characteristics not only causes mental distress to its victims but also leads to real-world violence (Arviv et al., 2021). Due to the insufficient efficiency of manual content review, deep learning-based methods have been used to construct automatic hate speech detection models (Gandhi et al., 2024). However, these approaches often struggle to detect implicit hate speech which lacks of explicit lexical signals (Kim et al., 2022, 2024).

Recently, contrastive learning has emerged as an effective approach for detecting implicit hate speech. For instance, Kim et al. (2022) used the implications of anchor sentences as positive samples, applying contrastive loss to improve detection. Similarly, ConPrompt (Kim et al., 2023) leveraged machine-generated statements to enhance performance, using example sentences from the original prompt as positive samples. To reduce the need for additional data construction, Ahn et al. (2024) clustered training data and used shared semantics as anchor for contrast learning. Additionally, label-aware hard negative sampling was introduced to optimize the learning of hard negatives by fully utilizing label information (Kim et al., 2024). However, most of these methods use only one positive sample per anchor and fail to address spurious correlations, limiting their effectiveness.



(a) False negative as hard positive.

(b) False positive as hard positive.

Figure 1: Our research motivations. Hard positives have same ground truth label but opposite predicted label with the anchor, including (a) false negatives and (b) false positives.

Spurious correlations refer to misleading associations between features and labels that frequently appear in training data but are incorrectly generalized as patterns, leading to diminished model performance. In the context of implicit hate speech detection, the subtle nature of implicit language exacerbates these spurious correlations. As illustrated in Figure 1, case (2) demonstrates how semantic

Proceedings of the 31st International Conference on Computational Linguistics, pages 8858–8867
January 19–24, 2025.

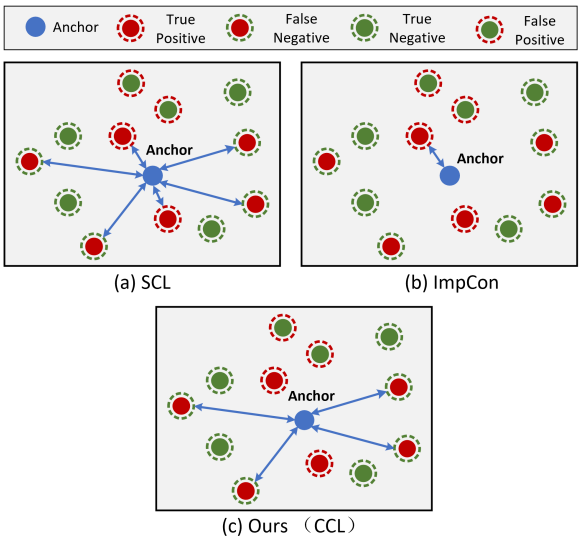


Figure 2: Illustration of the positives sampling strategy within three contrastive learning methods in a situation where the class of the anchor is hate speech while predicted correctly. (a) Supervised contrastive learning (SCL) selects all the same ground truth labels with the anchor as positive samples. (b) ImpCon only uses an implication of the anchor as a positive sample. (c) CCL selects posts with same ground truth labels while opposite predicted labels with the anchor as positive samples (i.e., false negatives in this case).

subtlety can result in false negatives, while case (4) shows how term bias can lead to false positives. These insights inspire our approach to identifying hard positive samples with the same ground truth label as the anchor but opposite predicted labels, including both false negatives and false positives.

We introduce a novel approach called Causal Contrastive Learning (CCL) for implicit hate speech detection, grounded in causal reasoning. Drawing from the observation that Empirical Risk Minimization (ERM) is effective at predicting spurious correlations, we use prediction errors to guide the selection of hard positives in contrastive learning. The core idea is to align representations of samples from the same class but with opposite spurious attributes (Figure 2), using contrastive learning to bring these representations closer together. Specifically, we leverage both false positives and false negatives as hard positive pairs. The strength of CCL lies in its ability to incorporate multiple positives while utilizing label information, effectively addressing two challenges at once. First, causality-guided hard positives fully exploit both ground truth labels and the spurious correlations predicted by ERM. Second, this approach allows for multiple positives per anchor, alleviating the issues posed by coarse-grained labels in binary hate speech detection (Suresh and Ong, 2021a; Kim et al., 2022).

To sum up, we make the following three contributions:

- We introduce a novel causality-guided contrastive learning method, namely CCL, to enhance representation learning by focusing on hard positives samples.
- CCL is implemented by leveraging the inference failures of ERM, which serve as a indicators of spurious correlation, without relying on external knowledge or incurring addition costs.
- We demonstrate the effectiveness of CCL through cross-dataset evaluation, achieving state-of-the-art performance on three widely-used benchmark datasets for implicit hate speech detection.

The remainder of this paper is organized as follows. Section 2 summarizes the related work. Section 3 presents the proposed approach. The experimental results are presented in Section 4, followed by conclusion in Section 5.

2 Related Work

2.1 Implicit Hate Speech Detection

Current models perform well in detecting explicit hate speech but struggle with implicit content with semantic subtlety (Ocampo et al., 2023). A significant challenge in implicit hate speech detection is the lack of high-quality datasets. To avoid the over-reliance on spurious correlations of term bias and topic bias, ElSherief et al. (2021) developed a theoretically-justified six-class taxonomy and built a large-scale implicit hate speech dataset based on random samples 22K Twitter messages. In contrast to manual annotation, Hartvigsen et al. (2022) used large language models (LLMs) to create a large-scale, machine-generated, balanced dataset for implicit toxic language, which led to improved model performance.

In addition to building large-scale implicit hate speech datasets, researchers have developed dedicated methods for implicit hate speech detection. Considering the coded or indirect characteristic of implicit hate speech, Lin (2022) incorporated knowledge graphs to integrate real world knowledge into the detection process. Building on the

observation that many implicit hate posts share a common underlying implication, Kim et al. (2022) augmented hate posts with their corresponding implication. Then they applied contrastive learning to align these posts with their implications in representation space. To better capture user and conversational context in online interactions, CoSyn (Ghosh et al., 2023) introduced a context synergized neural network for implicit hate speech detection.

With the growing prominence of LLMs, recent work has investigated their ability to detect implicit hate speech. For example, Huang et al. (2023) utilized ChatGPT to detect implicit hateful tweets in zero-shot setting and showed it could correctly identify implicit hateful tweets. To further exploit the capability of LLMs, Yun (2023) incorporated chain-of-thought (CoT) explanations into LLM training. Despite these advancements, LLMs still encounter challenges in both accuracy and efficiency of nuanced tasks like hate speech detection (Sheth et al., 2024). In many cases, they remain less efficient compared to fine-tuned language models that have been specifically trained on specific datasets (Kim et al., 2024).

2.2 Methods to Enhance Hate Speech Generalization

Traditional models often fail to generalize across different contexts or evolving forms of hate speech. By focusing on the subtle differences between hateful and non-hateful content, contrastive learning has emerged as an effective method for improving generalization in implicit hate speech detection. For instance, ImpCon (Kim et al., 2022) leveraged external knowledge of implications as positive samples to train models using contrastive loss, enhancing model sensitivity to context-dependent expressions of hate. Similarly, ConPrompt (Kim et al., 2023) applied contrastive learning to machine-generated statements, further improving their ability to detect implicit hate speech. However, these previous methods still have limitations due to their reliance on predefined external knowledge or the high costs of text generation. To mitigate additional data construction, Ahn et al. (2024) clustered the training data and leveraging shared semantics as anchor for contrast learning. To fully utilize label information and semantic information, Kim et al. (2024) proposed Label-aware Hard negative sampling strategies (LAHN) to mine hard negatives that are semantically similar to the anchor but have opposite labels. In contrast to naive negative samples in random batch, LAHN focuses more on distinguishing between the anchor and hard negatives, mitigating over-fitting to the context of the text or specific words.

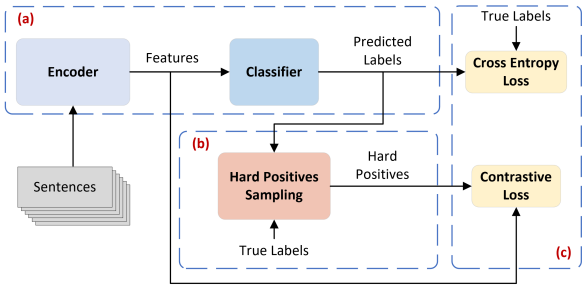


Figure 3: The overview of our CCL. (a) shows the encoder transforming sentences to features, followed by a classifier inferring predicted labels. (b) is the causality-guided hard positives sampling, which mines hard positives based on both ground truth labels and predicted labels achieved by (a). (c) is the overall training objective, including a cross-entropy loss and a contrastive loss.

Another major challenge in hate speech detection is spurious correlation, where models incorrectly classify content based on superficial cues, such as identity terms (e.g., "Black" or "Asian"). Inspired by the invariance of causal relationships (Lv et al., 2022), some causality-based methods have been proposed to alleviate spurious correlations of hate speech detection. For example, (Sheth et al., 2023) summarized two causal clues, i.e., the overall sentiment and the aggression, that are commonly present in hate speech to learn generalizable representations. Furthermore, to explicitly cutoff the spurious correlations between target object and hate speech, (Sheth et al., 2024) leveraged multi-task learning to disentangle the input representations into invariant and target-dependent features. From the perspective of inherent forms of spurious correlations, (Zhang et al., 2023) recognized spurious correlated features automatically via mutual information measurement.

In summary, contrastive learning methods and causality-based methods have shown promise in enhancing the generalizability of hate speech detection models. In this work, we introduce a novel approach that integrates causality in contrastive learning to enhance the generalization performance of hate speech detection.

3 Approach

We now present CCL, a causality-guided contrastive learning method designed to enhance the

generalization and robustness for implicit hate speech detection. As illustrated in Figure 3, CCL consists of three key components: (a) It first predicts labels based on features generated by pre-trained language models such as BERT. (b) Next, it employs causality-guided hard positives sampling, which identifies challenging positive samples using both ground truth and predicted labels. (c) Finally, the overall training objective combines cross-entropy loss with contrastive loss to optimize performance. In the following sections, we will provide a detailed explanation of each component.

3.1 Inferring predicted labels

We utilize a pre-trained language model as the encoder to transform a sentence x into a representation $r = Encoder(x)$, where r is a vector in R^{D_E}. Then, a classification head is applied to the normalized representation r, mapping it to a binary label $y = Classifier(r)$, where $y \in \{0, 1\}$, indicates whether the sample is classified as hate or non-hate. During training, both the encoder and classifier are updated to dynamically infer the predicted labels.

3.2 Causality-guided hard positives sampling

Next, we train a robust model using supervised contrastive learning, leveraging the prediction errors from the ERM model. While our approach follows standard supervised contrastive learning, we introduce new strategies for sampling hard positives and capping to enhance generalizability.

As previously discussed, two major challenges in detecting implicit hate speech detection are spurious correlations and semantic subtlety. Specifically, spurious correlations often result in false positives, while the nuanced nature of implicit language increases the likelihood of false negatives. To address both types of prediction errors, our hard positives sampling selects samples that share the same ground truth label as the anchor but have an opposite prediction from the ERM model (Figure 2-(c)).

We also limit the number of positive samples, as binary classification with coarse-grained labels can produce too many positives, potentially hindering contrastive learning (Suresh and Ong, 2021b). We calculate the cosine similarity between each positive and the anchor, rank them, and select the top k positives with the lowest similarity. The rationale behind this is that hard positives are semantically distinct from the anchor but share the same label, making them challenging for the model to represent effectively.

3.3 Overall Training objective

Generally, hate speech detection models are fine-tuned using supervised learning with the cross-entropy loss, which is calculated as following:

$$\mathcal{L}_{\text{ce}} = -\frac{1}{N}\sum_{i=1}^{N}[y_i \log \hat{y}_i + (1 - y_i)\log(1 - \hat{y}_i)] \tag{1}$$

where N is the number of input posts in a batch, y_i and $\hat{y}_i$ are the ground truth label and predicted label of input x_i, respectively.

To improve robustness against spurious correlations, our approach focuses on learning aligned representations. Specifically, we jointly train the encoder with a supervised contrastive learning loss. Following the approach of Khosla et al. (2020), within in a batch I, all samples except for the anchor itself, i.e., $A(i) \equiv I \setminus \{i\}$, are treated as negative samples. The final supervised contrastive learning loss is computed as following:

$$\mathcal{L}_{\text{cl}} = \sum_{i \in I} \frac{-1}{|P(i)|} \sum_{p \in P(i)} \log \frac{\exp(z_i \cdot z_p/\tau)}{\sum_{a \in A(i)} \exp(z_i \cdot z_a/\tau)} \tag{2}$$

where z_i, z_p and z_a are the representations from the encoder for the inputs x_i, x_p and x_a, respectively. P_i is the set of indices of positive samples obtained in Section 3.2 and τ is a temperature parameter specified in Section 4.3.

Using the available class labels, we update the model's encoder with contrastive learning loss and train the full model, including both encoder and classifier, with cross-entropy loss. The overall training objective is a weighted sum of cross-entropy loss and contrastive learning loss, calculated as following:

$$L_{total} = \lambda \mathcal{L}_{\text{ce}} + (1 - \lambda)\mathcal{L}_{\text{cl}} \tag{3}$$

where λ is pre-defined weighting hyperparameter that controls the balance between the cross-entropy loss and the contrastive learning loss.

4 Experimental Results

4.1 Datasets

We conduct a binary classification task aimed at detecting hateful language on implicit hate datasets. For evaluation, we use three implicit hate speech datasets, as Kim et al. (2022) and Kim et al. (2024).

Pre-trained Language Model	Objective	DA	IHC (Cross-dataset)	DynaHate (Cross-dataset)	SBIC (In-dataset)
BERT	Cross-Entropy Loss (CE)	✗	61.5	58.2	<u>78.6</u>
	SCL (Gunel et al., 2021)	✗	63.1	58.3	77.3
	ImpCon (Kim et al., 2022)	✓	63.6	59.9	77.6
	LAHN (Kim et al., 2024)	✓	<u>64.0</u>	60.0	**78.9**
	CCL w/o DA (ours)	✗	63.3	<u>60.5</u>	77.4
	CCL w/ DA (ours)	✓	**65.3**	**62.7**	78.4
HateBERT	Cross-Entropy Loss (CE)	✗	60.8	57.6	76.4
	SCL (Gunel et al., 2021)	✗	63.6	60.5	77.3
	ImpCon (Kim et al., 2022)	✓	64.9	61.8	77.5
	LAHN (Kim et al., 2024)	✓	64.4	<u>62.2</u>	**77.9**
	CCL w/o DA (ours)	✗	<u>65.1</u>	61.5	76.7
	CCL w/ DA (ours)	✓	**66.4**	**63.1**	<u>77.6</u>

Table 2: Cross-dataset and in-dataset evaluation results for different training objectives trained on IHC dataset. Boldfaced values denote the best performance and the underline denotes the second-best performance among different training objectives. (DA: Data Augmentation.)

- Implicit Hate Speech Corpus (IHC) (ElSherief et al., 2021): This dataset consists of 18,666 tweets collected from Twitter (i.e., X). Of these, 5,450 tweets are labeled as implicit hate speech and their targets and implications are also given.
- Social Bias Inference Corpus (SBIC) (Sap et al., 2020): The dataset features hierarchical annotations of stereotypes and social bias, including target and implied statement.
- Dynamically Generated Hate Speech Dataset (DynaHate) (Vidgen et al., 2021): The dataset was created through a human-and-model-in-the-loop process.

A summary of these datasets is presented in Table 1. We use the macro F1-score measure for validation to ensure a balanced evaluation across classes.

Datasets	No. of Posts	Hateful Posts	Hate %
IHC	18,666	5,460	29.3
SBIC	44,391	22,964	51.7
DynaHate	41,245	22,257	54.0

Table 1: Datasets statistics

4.2 Baselines

- Cross-Entropy (CE) loss: This method is a widely used approach for general classification tasks, including hate speech detection.
- Supervised Contrastive Learning (SCL) with CE loss (Gunel et al., 2021): This method refines CE loss by incorporating supervised contrastive learning, which enhances the model's ability to distinguish between subtle class differences–crucial for tasks like hate speech detection.
- Contrastive Learning using Implication (ImpCon) with CE loss (Kim et al., 2022): ImpCon further improves upon CE loss by integrating implication-based contrastive learning. It enhances the model's understanding of contextual relationships by introducing common implications associated with implicit hate speech.
- Label-aware Hard Negative Sampling Strategies with Momentum Contrastive Learning (LAHN) (Kim et al., 2024): LAHN leverages label information and semantic similarity for hard negative sampling in momentum contrast learning, focusing on challenging negative samples. Unlike ImpCon, it avoids relying on external information, making it more broadly applicable.

4.3 Implementation Details

For the convenience of comparison, as in Kim et al. (2022), we employed the pre-trained language model BERT-base-uncased and HateBERT as the sentence encoder. For all the models, we utilized a NVIDIA Tesla V100 GPU (32GB).

Pre-trained Language Model	Objective	DA	IHC (Cross-dataset)	DynaHate (Cross-dataset)	SBIC (In-dataset)
BERT	Cross-Entropy Loss (CE)	✗	59.9	60.6	84.1
	SCL (Gunel et al., 2021)	✗	59.4	61.0	83.9
	ImpCon (Kim et al., 2022)	✓	<u>60.7</u>	60.7	<u>84.3</u>
	LAHN (Kim et al., 2024)	✓	60.5	<u>61.1</u>	**84.8**
	CCL w/o DA (ours)	✗	60.3	61.0	83.3
	CCL w/ DA (ours)	✓	**61.3**	**62.1**	<u>84.3</u>
HateBERT	Cross-Entropy Loss (CE)	✗	61.0	60.0	<u>85.0</u>
	SCL (Gunel et al., 2021)	✗	58.6	60.1	84.6
	ImpCon (Kim et al., 2022)	✓	61.0	61.0	**85.1**
	LAHN (Kim et al., 2024)	✓	<u>61.3</u>	<u>61.2</u>	84.9
	CCL w/o DA (ours)	✗	61.1	61.0	84.1
	CCL w/ DA (ours)	✓	**61.5**	**61.9**	84.8

Table 3: Cross-dataset and in-dataset evaluation results for different training objectives trained on SBIC dataset. Boldfaced values denote the best performance and the underline denotes the second-best performance among different training objectives. (DA: Data Augmentation.)

Using a 80-10-10 split for each task, we trained, validated, and test all the models. For hyperparameter, we used the Adam optimizer with a learning rate 2e-5, and search the hyper-parameters temperature τ from 0.1, 0.3, 0.5, weights λ from 0.25, 0.5, 0.75, batch size from 8, 16, 32, 64, capping size from 1, 3, 5, 7, 9. We chose the best model score with macro F1-score in the validation set and report the macro F1-score on the test set.

4.4 Overall Performance

Table 2 and Table 3 present the evaluation results for both cross-dataset setting and in-dataset setting across three datasets for the training models on the IHC and SBIC datasets respectively.

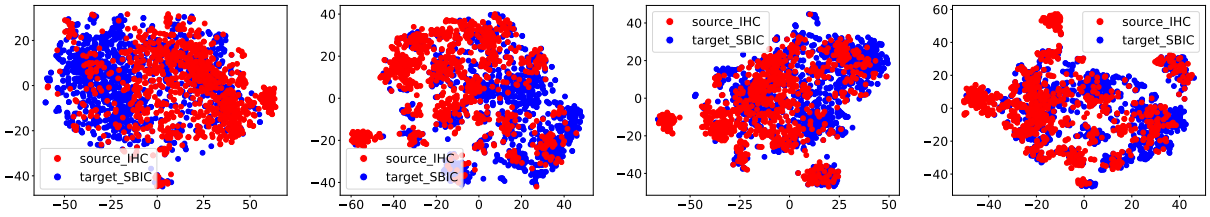
Compared to prior studies, our CCL consistently outperforms in cross-dataset evaluations. Simply adding augmented posts to the training set in Cross-Entropy based method proves ineffective. Additionally, using label information in Supervised Contrastive Learning (SCL) is less effective than our approaches, likely due to the task's coarse-grained labels (only two classes), consistent with earlier research findings. CCL, by leveraging both ground truth and predicted label information, produces finer-gradient distinctions. While ImpCon, which uses implication relationships, performs well and LAHN effectively combines label information with semantic similarity, CCL's superior performance may be attributed to its use of multiple positive samples, whereas ImpCon and LAHN rely on a single positive samples. CCL also mitigates spurious correlations by incorporating predicted labels, leading to more invariant representations with stronger generalization capabilities.

A key advantage of CCL is its effectiveness without data augmentation, a trait it inherits from SCL. To access the impact of data augmentation, we tested CCL without data augmentation and explored three augmentation methods: implication and synonym substitution (Kim et al., 2022), as well as dropout noise applied to anchors (Kim et al., 2024). While CCL performs competitively without data augmentation, incorporating external knowledge through augmentations further improves performance. This highlights CCL's versatility and effectiveness across diverse learning conditions.

In terms of in-dataset evaluation, the differences between the various methods are minimal, consistent with the findings of Kim et al. (2022). However, it is important to recognize that in-dataset performance may be inflated due to the presence of spurious correlations. On the other hand, cross-dataset evaluation offers a more accurate assessment of a model's true generalization capabilities. In conclusion, CCL enhances cross-dataset generalizability without compromising in-dataset performance, demonstrating its robustness and adaptability across different evaluation setting.

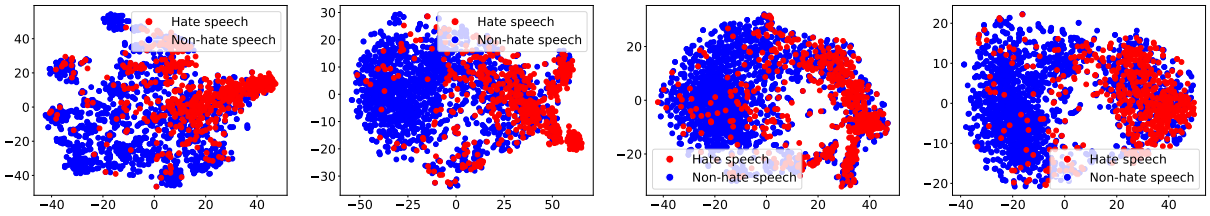
4.5 Qualitative Analysis

The ability to generalize largely depends on a model's capacity to learn invariant features. Our hypothesis is that, since causal features remain con-



(a) SCL (b) ImpCon (c) LAHN (d) CCL (ours)

Figure 4: Visualization of the representations from different models to verify invariance across datasets. (Red: IHC as source dataset, Blue: SBIC as target dataset.)



(a) SCL (b) ImpCon (c) LAHN (d) CCL (ours)

Figure 5: Visualization of implicit hate speech and non-hate sentences from the SBIC dataset in the zero-shot setting. (Red: Implicit hate speech, Blue: Non-hate speech.)

sistent across datasets, a model capable of learning these invariant features should exhibit significant overlap in its representations of those features. To test this, we trained CCL on the IHC dataset and evaluated its generalization on SBIC. We used t-SNE to visualize representations from 1,000 samples in each dataset to assess the model's representation invariance. For a fair comparison, we extend this experiment to three baseline models: SCL, ImpCon, and LAHN. The resulting visualizations are shown in Figure 4.

As we can see, CCL's representations display substantial overlap, demonstrating its strong ability to learn invariant features across datasets. In comparison, SCL and ImpCon show less overlap, with ImpCon performing better than SCL. This is may be due to ImpCon aligning posts with their implications. However, despite this advantage, ImpCon still fall short of both CCL or LAHN, possibly because it relies on only one positive sample per anchor and overlooks label information. LAHN, while demonstrating some overlap, is outperformed by CCL, likely due to its vulnerability to biases arising from label information and spurious correlations. In contrast, CCL excels by leveraging both ground truth and predicted label information and using multiple positive samples per anchor, enhancing its generalization ability

Another indicator of strong generalization is a clear margin between different classes (Bengio et al., 2013; Gunel et al., 2021). To assess the boundary between hate speech and non-hate speech, we randomly sampled 2,000 instances of each from the SBIC dataset. Figure 5 illustrates a visualization of the embeddings for these samples, with the models fine-tuned on the IHC dataset. The embeddings produced by CCL create a notably sharper distinction between hate speech and non-hate speech compared to the other three models. This suggests that CCL enables the model to generate more generalizable representations than the alternative methods.

4.6 Ablation Study

To assess the effectiveness of the causality-guided positive sampling, we contrast it with standard supervised positive sampling in SCL, where all samples sharing the same ground truth label as the anchor are selected as positive samples. Additionally, we conduct ablation studies to examine the impact of varying the number of positive samples in mini-batch across different batch sizes. It is important to highlight that, due to the uneven distribution of data, the number of positive and negative samples in each mini-batch may not be balanced. This can result in cases where the number of positive pairs in a batch is lower than the predicted cap size. For example, if the cap size is set to 9 but a mini-batch contains only 7 positive pairs, the final number of positive pairs will be 7 rather than 9.

Positives Sampling Methods	Capping Methods	Batch Size	Number of Positives					
			1	3	5	7	9	No cap
SCL	Random	8	60.2	60.8	60.5	59.6	60.0	60.5
		16	62.1	62.5	60.7	61.2	61.4	60.4
		32	62.3	62.7	62.5	62.1	**63.1**	62.6
		64	62.0	61.3	61.7	61.0	61.9	61.3
	Similarity-based	8	60.3	60.9	60.7	59.8	59.7	60.5
		16	61.4	63.3	62.2	61.2	61.7	60.4
		32	62.8	63.1	62.6	**63.6**	62.0	62.6
		64	61.1	61.3	62.3	61.5	61.1	61.3
Causality-guided	Random	8	61.1	61.9	62.7	61.3	60.0	60.8
		16	61.3	64.0	63.2	62.2	61.7	61.9
		32	63.7	**64.3**	63.1	63.4	62.4	63.6
		64	62.4	62.1	61.7	62.2	63.5	62.6
	Similarity-based	8	61.3	62.2	63.0	62.9	61.7	60.8
		16	62.5	64.7	64.4	65.0	64.6	61.9
		32	64.9	64.4	65.7	65.4	65.9	63.6
		64	63.0	62.5	63.0	**66.4**	65.1	62.6

Table 4: Ablation study results for positives sampling methods and capping methods.

As shown in Table 4, causality-guided models consistently outperform SCL methods. A possible reason is that we only treat posts with opposite predicted labels as well as the same ground truth, rather than all posts with the same ground truth, as positive samples. This introduces fine-grained label information for supervised contrast learning, allowing the model to focus on more challenging cases where predictions diverge from ground truth labels. This demonstrates the robustness and effectiveness of incorporating spurious correlations mitigation into the learning process for hate speech detection.

When comparing random sampling to similarity-based sampling, we find that similarity-based sampling yields significantly better performance. By selecting samples with lower similarity to the anchor, the model is encouraged to learn more diverse and robust representations, rather than relying on overly similar or redundant information. These findings highlight the critical role of carefully selecting relevant and challenging examples during training, as it helps improve the model's ability to generalize across varied and nuanced contexts, ultimately leading to more accurate and reliable detection outcomes.

Regarding the batch size, we observed that increasing batch size does not always enhance model performance. That is, there is an optimal batch size beyond which performance plateaus or declines. Similarly, increasing the number of positive samples does not result in linear improvement. Since hard positive samples are selected based on similarity to the anchor, choosing too many risks increasing similarity. As a result, increasing the capping threshold beyond a certain point introduces noise, limiting the model's effectiveness in detecting hate speech. This findings emphasize the need for careful selection of hard positive samples, batch size, and hyperparameters to optimal performance in implicit hate speech detection.

5 Conclusions

In this paper, we propose a causality-guided contrastive learning framework aimed at improving the generalization capabilities of implicit hate speech detection models. By incorporating causality into the contrastive learning process, our approach effectively mitigates spurious correlations, leading to more robust representations. Particularly, we leverage prediction errors from ERM model for hard positives sampling. The empirical results confirm that our method significantly enhances the generalization capabilities in cross-dataset setting. Future work could explore the extension of this framework to other types of toxic language and its integration with multi-modal data.

6 Limitations

While our causality-guided contrastive learning approach has shown improvements in detecting implicit hate speech, several limitations must be addressed. First, like many machine learning models, our method requires careful tuning of multiple hyperparameters, which can be computationally expensive. Second, we sampled only the hard positives while ignoring the benefit of hard negatives sampling. Exploring causality-guided hard negatives sampling would further enhance the model's generalizability.

7 Ethics Statement

Our work on implicit hate speech detection aims to address the challenges of identifying content with subtle semantics and spurious correlations. We use publicly available datasets to ensure transparency and reproducibility while prioritizing user privacy. All data employed in this research is anonymized, and no personal information is used or disclosed. Our goal is to enhance the detection of implicit hate speech without exacerbating harm.

Acknowledgments

The research is supported by the Fundamental Research Funds for the Central Universities under grant No. CCNU24ZZ149, the China Postdoctoral Science Foundation under grant No. 2021M701367 and the Basic Scientific Research of China University under grant No. CCNU21XJ020 and No. CCNU22QN016.

We would also like to thank all the reviewers for their knowledgeable reviews.

References

Hyeseon Ahn, Youngwook Kim, Jungin Kim, and Yosub Han. 2024. SHAREDCON : Implicit Hate Speech Detection using Shared Semantics. In *Findings of the Association for Computational Linguistics*, pages 10444–10455.

Eyal Arviv, Simo Hanouna, and Oren Tsur. 2021. It'sa thin line between love and hate: Using the echo in modeling dynamics of racist online communities. In *Proceedings of the International AAAI Conference on Web and Social Media*, pages 61–70.

Yoshua Bengio, Aaron Courville, and Pascal Vincent. 2013. Representation learning: A review and new perspectives. *IEEE transactions on pattern analysis and machine intelligence*, 35(8):1798–1828.

Mai ElSherief, Caleb Ziems, David Muchlinski, Vaishnavi Anupindi, Jordyn Seybolt, Munmun De Choudhury, and Diyi Yang. 2021. Latent Hatred: A Benchmark for Understanding Implicit Hate Speech. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 345–363.

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

Ankita Gandhi, Param Ahir, Kinjal Adhvaryu, Pooja Shah, Ritika Lohiya, Erik Cambria, Soujanya Poria, and Amir Hussain. 2024. Hate speech detection: A comprehensive review of recent works. *Expert Systems*, page e13562.

Sreyan Ghosh, Manan Suri, Purva Chiniya, Utkarsh Tyagi, Sonal Kumar, and Dinesh Manocha. 2023. CoSyn: Detecting Implicit Hate Speech in Online Conversations Using a Context Synergized Hyperbolic Network. In *Conference on Empirical Methods in Natural Language Processing, Proceedings*, pages 6159–6173.

Beliz Gunel, Jingfei Du, Alexis Conneau, and Veselin Stoyanov. 2021. Supervised contrastive learning for pre-trained language model fine-tuning. In *Proceedings of the 9th International Conference on Learning Representations*, pages 1–17.

Thomas Hartvigsen, Saadia Gabriel, Hamid Palangi, Maarten Sap, Dipankar Ray, and Ece Kamar. 2022. TOXIGEN: A Large-Scale Machine-Generated Dataset for Adversarial and Implicit Hate Speech Detection. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, pages 3309–3326.

Fan Huang, Haewoon Kwak, and Jisun An. 2023. Is chatgpt better than human annotators? potential and limitations of chatgpt in explaining implicit hate speech. In *Companion proceedings of the ACM web conference 2023*, pages 294–297.

Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. 2020. Supervised contrastive learning. *Processing of the Advances in Neural Information Processing Systems*, 33:18661–18673.

Jaehoon Kim, Seungwan Jin, Sohyun Park, Someen Park, and Kyungsik Han. 2024. Label-aware hard negative sampling strategies with momentum contrastive learning for implicit hate speech detection. In *Findings of the Association for Computational Linguistics*, pages 16177–16188.

Youngwook Kim, Shinwoo Park, and Yo Sub Han. 2022. Generalizable Implicit Hate Speech Detection using Contrastive Learning. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 6667–6679.

Youngwook Kim, Shinwoo Park, Youngsoo Namgoong, and Yo-Sub Han. 2023. ConPrompt: Pre-training a Language Model with Machine-Generated Data for Implicit Hate Speech Detection. In *Findings of the Association for Computational Linguistics: EMNLP*, pages 10964–10980.

Jessica Lin. 2022. Leveraging world knowledge in implicit hate speech detection. In *Proceedings of the Second Workshop on NLP for Positive Impact (NLP4PI)*, pages 31–39.

Fangrui Lv, Shuang Li, Bin Zang, Chi Harold Liu, Ziteng Wang, and Di Liu. 2022. Causality Inspired Representation Learning for Domain Generalization. In *in Proceedings of the Conference on Computer Vision and Pattern Recognition*, pages 8036–8046.

Nicolas Ocampo, Ekaterina Sviridova, Elena Cabrio, and Serena Villata. 2023. An In-depth Analysis of Implicit and Subtle Hate Speech Messages. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 1997–2013.

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

Paras Sheth, Tharindu Kumarage, Raha Moraffah, Aman Chadha, and Huan Liu. 2023. PEACE: Cross-Platform Hate Speech Detection - A Causality-guided Framework. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 559–575.

Paras Sheth, Tharindu Kumarage, Raha Moraffah, Aman Chadha, and Huan Liu. 2024. Causality Guided Disentanglement for Cross-Platform Hate Speech Detection. In *Proceedings of the 17th ACM international conference on web search and data mining*, pages 626–635.

Varsha Suresh and Desmond C. Ong. 2021a. Not All Negatives are Equal: Label-Aware Contrastive Loss for Fine-grained Text Classification. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 4381–4394.

Varsha Suresh and Desmond C. Ong. 2021b. Not All Negatives are Equal: Label-Aware Contrastive Loss for Fine-grained Text Classification. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 4381–4394.

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

James Thorne Se-young Yun. 2023. HARE: Explainable Hate Speech Detection with Step-by-Step Reasoning. In *Findings of the Association for Computational Linguistics*, pages 5490–5505.

Zhehao Zhang, Jiaao Chen, and Diyi Yang. 2023. Mitigating Biases in Hate Speech Detection from A Causal Perspective. In *Findings of the Association for Computational Linguistics: EMNLP*, pages 6610–6625.