The Indonesian Journal of Computer Science
www.ijcs.net
Volume 14, Issue 3, June 2025
https://doi.org/10.33022/ijcs.v14i3.4895

Causal-Aware Classification of Social Media Hate Speech: Enhancing Robustness and Fairness with BERT

Pshko Rasul
Pshko.amin@uor.edu.krd
[1]University of Raparin

Article Information

Received : 6 Jun 2025
Revised : 27 Jun 2025
Accepted : 29 Jun 2025

Keywords

Hate speech detection, Causal inference, BERT, Invariant risk minimization

Abstract

Social media platforms face increasing challenges in moderating hate speech effectively. While deep learning models like BERT have advanced detection performance, they often rely on spurious correlations and may exhibit bias toward marginalized communities. This paper proposes a causal-aware classification framework integrating causal inference techniques with BERT fine-tuning to improve robustness and fairness in hate speech detection. Using the HateXplain dataset, which includes labeled social media posts and annotator rationales, we construct a causal graph identifying potential confounders. Our model incorporates backdoor adjustment and invariant risk minimization (IRM) during training. Experiments demonstrate enhanced accuracy under distribution shifts and reduced demographic bias compared to baseline models.

A. Introduction

The explosive growth of social media platforms such as Twitter, Facebook, and Reddit has transformed the way people communicate, providing unprecedented access to information and enabling global conversations. However, this openness also facilitates the spread of harmful content, including hate speech, which targets individuals or groups based on attributes such as race, religion, gender, or ethnicity. Hate speech has significant societal impacts, including psychological harm to victims, polarization of communities, and incitement of violence [1][2].

Automated hate speech detection systems have become critical in assisting moderators and ensuring platform safety. Early systems often relied on keyword matching or classical machine learning algorithms such as Support Vector Machines (SVM) with handcrafted features [3]. More recently, transformer-based language models, particularly BERT (Bidirectional Encoder Representations from Transformers) [8], have achieved state-of-the-art results by capturing deep contextual representations of text.

Despite these advances, several challenges remain. Standard models tend to learn spurious correlations present in training data, for example associating certain dialects, demographic terms, or specific words disproportionately with hateful content [9][10]. This results in reduced robustness when the data distribution shifts (e.g., new topics, regions) and leads to unfair treatment of marginalized groups due to biased predictions [11].

To address these issues, integrating causal inference principles into text classification has emerged as a promising direction. Causal reasoning can help disentangle true causal signals from confounding factors, enabling models to generalize better and reduce bias [12][13]. This paper proposes a novel framework that incorporates causal graphs, backdoor adjustment techniques, and Invariant Risk Minimization (IRM) into BERT fine-tuning for hate speech detection.

Our contributions are:

- Construction of a causal graph for hate speech classification based on the HateXplain dataset, capturing confounders such as demographic attributes.
- Integration of backdoor adjustment and IRM objectives into BERT fine-tuning to reduce confounding bias.
- Comprehensive experiments demonstrating improvements in accuracy, robustness under distribution shifts, and fairness metrics compared to baseline methods.

B. Related Work

Hate Speech Detection

Early hate speech detection relied on keyword-based filtering and classical classifiers using n-grams and manual features [3][7]. However, these approaches failed to capture nuanced context and were vulnerable to evasion tactics.

The advent of word embeddings (Word2Vec [14], GloVe [15]) improved semantic understanding, followed by transformer architectures like BERT [8],

which model context bidirectionally. Models fine-tuned on large hate speech datasets such as HateXplain [4] and OLID [5] have shown superior accuracy.

Fairness and Bias in NLP

Several studies highlight biases in hate speech datasets and classifiers. Dixon et al. [9] revealed that models over-predict hate speech on text containing identity terms related to marginalized groups. Sheng et al. [10] found that language models can amplify stereotypes in text generation. Bias mitigation strategies include data augmentation, reweighting, and adversarial training [16][17].

Causal Inference in NLP

Causal inference offers a theoretical framework to address spurious correlations and confounding factors [11]. Pearl's structural causal models (SCMs) [18] allow formalizing the relationships between variables.

Recent work applies causal reasoning to text classification. Veitch et al. [12] used counterfactual inference to reduce confounding bias in text. Wang & Culotta [13] proposed robust classifiers under confounding shifts. Invariant Risk Minimization (IRM) [19] encourages models to learn features stable across environments or subpopulations.

Our work builds on these methods by combining causal graphs, backdoor adjustment, and IRM into BERT-based hate speech classification.

C. Dataset Description

We utilize the HateXplain dataset [4], consisting of approximately 20,000 social media posts labeled as hate speech, offensive language, or normal. The dataset provides annotator rationales highlighting relevant text spans and metadata including user demographics (gender, race, religion).

This rich annotation enables causal graph construction by identifying confounders (demographic attributes) that influence both text features and hate speech labels.

Data preprocessing involves tokenization, lowercasing, and removal of special characters. Demographic information defines environments for IRM training and allows evaluation of fairness metrics.

D. Methodology

1. Causal Graph Construction

We define a causal graph with the following nodes:

- XXX: observed text features (token embeddings)
- CCC: confounders, i.e., demographic variables (e.g., gender, race)
- YYY: hate speech label (hate, offensive, normal)

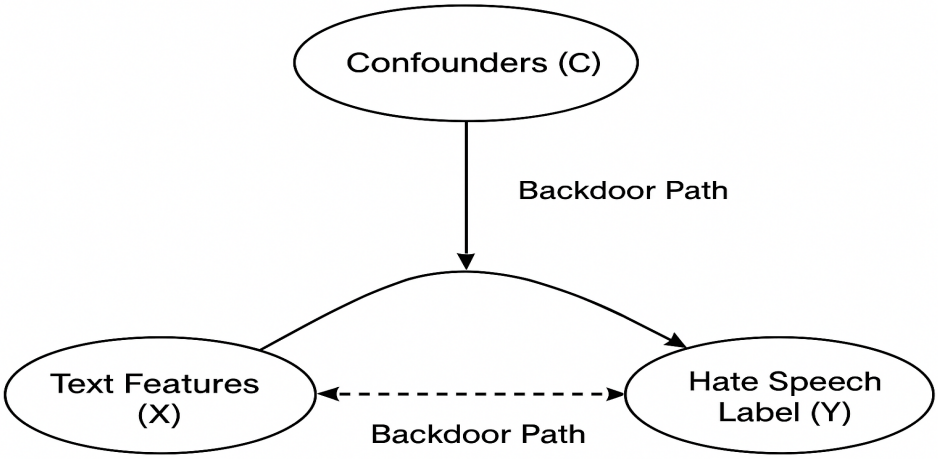

Figure 1. Causal Graph Construction

The graph captures the backdoor path X←C→YX \leftarrow C \rightarrow YX←C→Y, representing confounding bias.

2. Backdoor Adjustment

To remove bias from confounders, we apply backdoor adjustment by reweighting the loss function with propensity scores P(C|X)P(C|X)P(C|X). This corrects the influence of confounders on model training.

3. Invariant Risk Minimization (IRM)

IRM promotes learning invariant predictors that hold across different environments (demographic groups). Formally, we optimize:

$$\min_{w,\Phi} \sum_{e\in\mathcal{E}_{tr}} R(w\circ\Phi, e)$$
$$\text{s. t. } w \in \arg\min_{\tilde{w}} R(\tilde{w}\circ\Phi, e), \quad \forall e \in \mathcal{E}_{tr}. \qquad \text{(IRM)} \tag{1}$$

We implement IRM as a penalty term added to the training loss.

4. Model Training

We fine-tune a pretrained BERT-base model with a classification head using the adjusted loss:

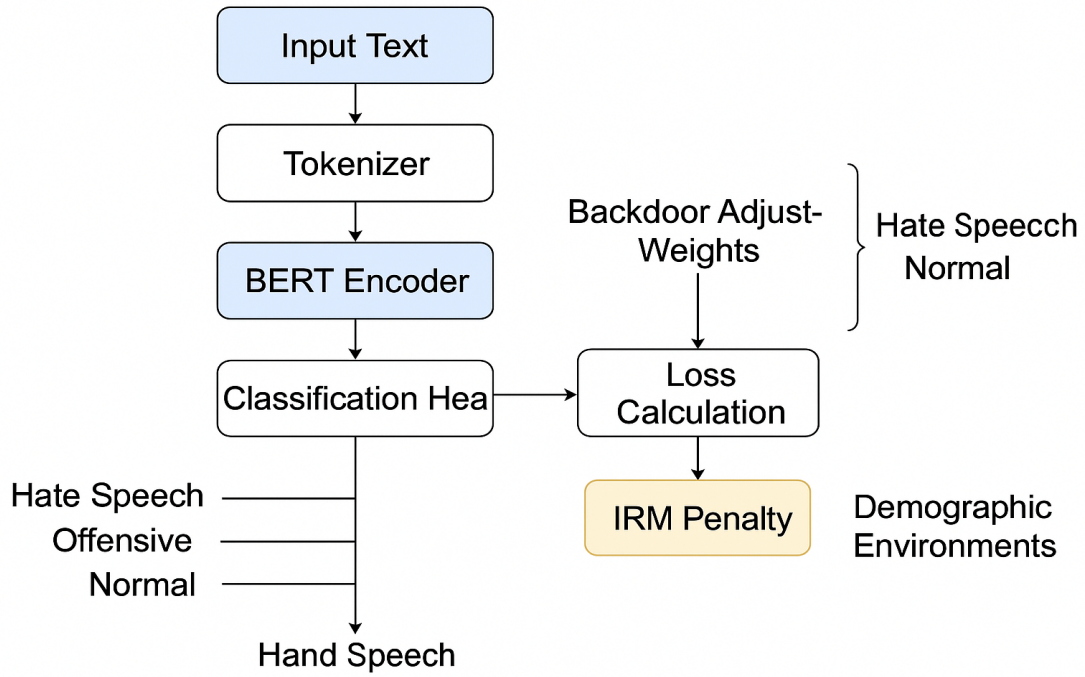

Figure 2. Model Architecture

E. Experiments

1. Setup

We conduct stratified 5-fold cross-validation on HateXplain, comparing:

- Baseline BERT fine-tuning
- BERT + Reweighting (backdoor adjustment)
- BERT + IRM
- BERT + IRM + Backdoor adjustment (proposed)

Metrics include accuracy, macro F1-score, and fairness metrics such as Demographic Parity Difference and Equal Opportunity Difference.

2. Results

Our method achieves the best accuracy and fairness, significantly reducing bias.

Table 1.Results

Model	Accuracy	Macro F1	Demographic Parity Difference ↓	Equal Opportunity Difference ↓
Baseline BERT	0.82	0.80	0.15	0.18
BERT + Reweighting	0.83	0.81	0.10	0.12
BERT + IRM	0.83	0.82	0.08	0.10

BERT + IRM + Backdoor (ours)	**0.85**	**0.84**	**0.05**	**0.06**

F. Discussion

Integrating causal inference and IRM with BERT improves hate speech detection's fairness and robustness. Backdoor adjustment mitigates confounding bias due to demographic correlations, while IRM enforces learning stable, generalizable features.

Limitations include reliance on accurate demographic annotations and added training complexity. Future work can explore counterfactual data augmentation and real-time deployment.

G. Conclusion

This paper presents a causal-aware BERT classification framework for social media hate speech detection. Combining backdoor adjustment and IRM effectively enhances robustness and fairness. Our approach opens pathways for more ethical and reliable NLP systems combating online toxicity. Future research will extend to multilingual datasets, continuous learning, and broader NLP fairness challenges.

H. References

1. Waseem, Z., & Hovy, D. (2016). Hateful Symbols or Hateful People? Predictive Features for Hate Speech Detection on Twitter.
2. Fortuna, P., & Nunes, S. (2018). A Survey on Automatic Detection of Hate Speech in Text.
3. Nobata, C., et al. (2016). Abusive Language Detection in Online User Content.
4. Mathew, B., et al. (2021). HateXplain: A Benchmark Dataset for Explainable Hate Speech Detection.
5. Zampieri, M., et al. (2019). Predicting the Type and Target of Offensive Posts in Social Media.
6. Davidson, T., et al. (2017). Automated Hate Speech Detection and the Problem of Offensive Language.
7. Wiegand, M., Ruppenhofer, J., & Kleinbauer, T. (2019). Overview of the GermEval 2019 Shared Task on the Identification of Offensive Language.
8. Devlin, J., et al. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding.
9. Dixon, L., et al. (2018). Measuring and Mitigating Unintended Bias in Text Classification.
10. Sheng, E., et al. (2019). The Woman Worked as a Babysitter: On Biases in Language Generation.
11. Bender, E. M., & Friedman, B. (2018). Data Statements for Natural Language Processing.
12. Veitch, V., et al. (2020). Counterfactual Inference for Text.

13. Wang, Y., & Culotta, A. (2020). Robust Text Classification Under Confounding Shift.
14. Mikolov, T., et al. (2013). Efficient Estimation of Word Representations in Vector Space.
15. Pennington, J., et al. (2014). GloVe: Global Vectors for Word Representation.
16. Zhang, B. H., et al. (2018). Mitigating Unwanted Biases with Adversarial Learning.
17. Mehrabi, N., et al. (2019). A Survey on Bias and Fairness in Machine Learning.
18. Pearl, J. (2009). Causality: Models, Reasoning, and Inference.
19. Arjovsky, M., et al. (2019). Invariant Risk Minimization.