Seeing Hate Differently: Modeling Culture-Based Hate Perception for Hate Speech Detection

Anonymous ACL submission

Abstract

Hate speech detection has been widely studied, yet existing methods often overlook a key real-world challenge: annotations are subjective, and perceptions of hate vary across individuals with different cultural backgrounds. We first analyze three major challenges in culture-based hate speech detection, namely data sparsity, complex interactions between cultural factors, and ambiguous labeling. To address these challenges, we propose a culture-based framework that models individuals' hate perception through combinations of cultural attributes. By modeling cultural combinations rather than isolated factors, the proposed approach alleviates data sparsity and enables structured analysis of cultural influences. We further introduce a label propagation mechanism to aggregate annotation signals across related combinations, mitigating the effect of ambiguous labels. Experimental results demonstrate that our approach not only improves classification performance, but also provides an exploratory modeling perspective for analyzing how cultural factors shape hate perception.

1 Introduction

Hate speech detection aims to determine whether a text contains hateful content. Traditional approaches primarily rely on textual features, such as lexicon cues and syntactic patterns (Nobata et al., 2016; Burnap and Williams, 2014, 2016), while recent methods focus on fine-tuning pre-trained language models (PLMs) (Caselli et al., 2020; Koufakou et al., 2020), achieving strong performance with F1 scores of 0.8-0.9 on benchmark datasets.

However, such results can be misleading, as ground-truth labels are typically obtained through majority voting among annotators, which introduces bias and oversimplifies the inherently subjective nature of hate speech (Sap et al., 2019). In practice, hate perception varies across individuals with different beliefs (Sap et al., 2022), and

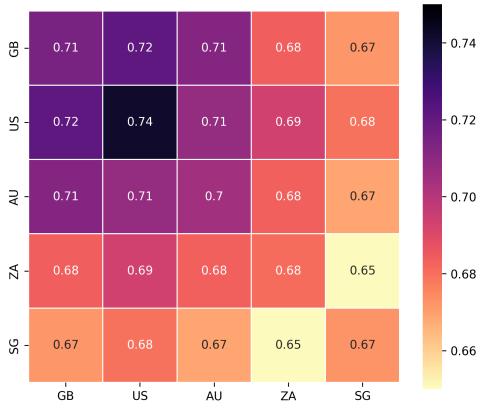

Figure 1: Pairwise hate-speech label agreement ratios between users from five countries: United Kingdom (GB), United States (US), Australia (AU), South Africa (ZA), and Singapore (SG).

across cultural backgrounds (Lee et al., 2023a; Davani et al., 2024). For example, annotators from the United States and the United Kingdom tend to exhibit higher agreement than those from the United States and Singapore (Figure 1). Moreover, even within the same cultural group, perceptions can diverge substantially. As shown in Figure 1, pairwise label agreement ratios indicate that annotators from the same country may exhibit lower agreement than those from different countries (e.g., SG-SG < SG-US). These observations suggest that hate perception is too complex to be explained by a single cultural factor. To better understand hate perception and to enable personalized hate speech detection, it is therefore crucial to uncover the diverse factors that shape individual's judgments.

Prior work has explored related directions such as modeling annotator representations (Deng et al., 2023; Mokhberian et al., 2023; Fleisig et al., 2023). One line of research (Deng et al., 2023; Mokhberian et al., 2023) learns annotator representations using trainable vectors or one-hot encodings, but largely overlooks cultural background information. Another line of work incorporates demographic information as prompts for language models; how-

ever, this design limits interpretability and makes it difficult to analyze which cultural factors drive specific annotations. Other studies leverage side information about hate speech creators (Vijayaraghavan et al., 2021) or analyze cross-culture variations in hate definitions (Korre et al., 2025), while the modeling of receivers remains underexplored.

Modeling culture-based hate speech detection presents three major challenges: ①**Data sparsity.** Hate perception is shaped by multiple factors, such as religion and gender, which leads to an exponential number of possible background combinations. For instance, the CREHate dataset (Lee et al., 2023a) includes eight background attributes. Excluding the continuous feature 'age', the remaining seven categorical backgrounds yield 91,045,500 possible combinations in theory, whereas the dataset contains annotations from only 1,064 annotators, covering merely 1.17×10^{-5} of the theoretical space. ②**Lack of structured modeling of cultural interactions.** Cultural backgrounds do not influence hate perception in isolation; instead, their effects often depend on how multiple backgrounds interact. However, existing models lack a structured way to analyze how judgments change when cultural backgrounds are added, removed or modified. For instance, if an annotator with background `<Country=United States, Religion=Christian>` considers a post hateful, it remains unclear how this judgment would change after adding `<Sex=Male>` or replacing `<Religion=Christian>` with `<Religion=Buddhism>`. As also reflected in our experiments, current LLM-based approaches struggle to effectively leverage such background information (Table 1). ③**Ambiguous labeling and attribution.** Cultural background information in datasets is often incomplete, which introduces label noise. Even when full cultural attributes are available, it remains unclear *which* cultural factors contribute to a particular judgment. For example, when an annotator with background `<Country=United States, Religion=Christian, Sex=Male>` labels a post as hateful, the perception may arise from nationality, religion, or their joint effect. This ambiguity further complicates the interpretation and analysis of hate perception.

In this work, we propose a culture hate speech detection framework that models individual's hate perception through combinations of cultural backgrounds to address these challenges. To alleviate data sparsity, we model cultural background combinations rather than treating attributes in isolation. Although interactions between backgrounds cannot be perfectly modeled and also label ambiguity cannot be fully resolved, we introduce a label propagation mechanism from higher-level cultural combinations to their subsets and construct a weight matrix to differentiate their contributions. Finally, each individual's hate perception is represented by aggregating information from the combinations of their cultural backgrounds. Our contributions are summarized as follows:

- We identify key challenges in culture-based hate speech detection and propose a simple yet effective framework that models individuals' hate perception based on interactions between cultural backgrounds and posts, rather than relying solely on textual features.
- Extensive experiments demonstrate that our approach consistently outperform state-of-the-art baselines, achieving an average improvement of 1.05% across all metrics. Moreover, our analysis shows that considering only about 10% of cultural combinations is sufficient to achieve strong performance, highlighting the exploratory nature of the proposed model.

2 Problem Statement

Definition. Culture-Based Hate Speech Detection. Let $\mathcal{U} = \{(u_1, \mathbf{c}_1), (u_2, \mathbf{c}_2), \ldots, (u_n, \mathbf{c}_n)\}$ denote a set of n users, where each user u_i is associated with a set of k cultural background attributes $\mathbf{c}_i = \{c_{i1}, c_{i2}, \ldots, c_{ik}\}$. Given a post p, the goal of culture-based hate speech detection is to estimate the probability that user u_i would perceive p as hateful, conditioned on the user's cultural background, i.e., $P(\text{hate} \mid u_i, \mathbf{c}_i, p)$.

3 Method

Our method models hate perception in three stages. We first model how cultural background combinations interact with posts (Section 3.1), then aggregate these signals to represent individual hate perception (Section 4), and finally perform hate speech classification (Section 3.3).

3.1 Culture-Post Interaction Matrix

To model individual hate perception and alleviate data sparsity, we shift the modeling unit from individual users to combinations of cultural background attributes, which allows supervision to be

shared across users with overlapping attributes. For a user u_i with cultural background $\mathbf{c}_i = \{c_{i1}, \ldots, c_{ik}\}$, we consider all subsets of attributes, denoted by the power set $\mathcal{P}(\mathbf{c}_i)$. Each element in $\mathcal{P}(\mathbf{c}_i)$ represents a cultural background combination. For example, if $c_i = (\text{Male}, \text{US})$ then $\mathcal{P}(c_i) = \{\text{Male}, \text{US}, \{\text{Male}, \text{US}\}\}$. Our goal is to model how these cultural combinations interact with posts, and how their aggregated effects characterize an individual's hate perception. By operating at the combination level rather than the individual level, we reduce sparsity by increasing effective supervision and enable structured analysis of cultural factors within a shared representation space. To aggregate annotation signals, we collect labels at the combination level. For each cultural combination $comb_l$ and post p_j, we record their co-occurrence:

$$U_{l,j} = \{(u, \mathbf{c}) \in \mathcal{U} \mid comb_l \in \mathcal{P}(\mathbf{c}), \; u \in Label(p_j)\} \quad (1)$$

which represents users who possess combination $comb_l$ and annotated post p_j. This aggregation implements a label propagation mechanism: labels provided by users with richer combinations are propagated to their subsets. Intuitively, while a single annotation does not reveal which attribute caused a judgment, aggregating labels from users sharing a combination yields more reliable combination-level tendencies. We then construct a culture–post interaction matrix $Y \in \mathbb{R}^{z \times m}$, where rows correspond to cultural combinations and columns correspond to posts. Each entry is weighted using TF–IDF based on U, reducing the influence of frequent but uninformative signals.

3.2 Hate Perception Representation

We learn latent representations for cultural combinations and posts by factorizing the culture post interaction matrix Y. Specifically, we learn embeddings $P \in \mathbb{R}^{z \times d}$ for cultural combinations and $Q \in \mathbb{R}^{m \times d}$ for posts, along with bias terms $B_c \in \mathbb{R}^z$ and $B_w \in \mathbb{R}^m$, where d denotes the embedding dimension. The interaction score between a combination $comb_l$ and a post p_j is estimated as

$$\hat{Y}_{l,j} = \mu + b_{c_l} + b_{w_j} + q_j^\top p_l, \quad (2)$$

where μ is a global bias, b_{c_l} and b_{w_j} are combination and post biases, and p_l and q_j denote their corresponding embeddings. Parameters are learned by minimizing squared reconstruction error with ℓ_2 regularization, which penalizes large parameter values and helps prevent overfitting:

$$\sum_{l,j} (Y_{l,j} - \hat{Y}_{l,j})^2 + \lambda (b_{c_l}^2 + b_{w_j}^2 + \|p_l\|^2 + \|q_j\|^2) \quad (3)$$

An individual's hate perception reflects multiple cultural aspects rather than a single attribute. Because all cultural combinations are embedded in a shared latent space, we model an individual's hate perception by aggregating the embeddings of all combinations associated with that user:

$$HP(u_i) = \sum_{comb_l \in \mathcal{P}(\mathbf{c}_i)} \alpha_l \begin{bmatrix} p_l \\ b_{c_l} \end{bmatrix}, \quad (4)$$

where α_l is a learnable coefficient that captures the relative influence of combination $comb_l$ on user u_i.

3.3 Classification

We integrate the individual hate perception embedding with post features for classification. Given an individual u_i and a post p_j, the prediction is:

$$P(\text{hate} \mid u_i, \mathbf{c_i}, p_j) = f_\theta(HP(u_i), q_j, s_j) \quad (5)$$

where q_j is the post's interaction feature from Eq. 2, s_j is the text embedding extracted by the CLIP text encoder (Radford et al., 2021), and $f_\theta(\cdot)$ is a classifier with parameters θ. The post is predicted as hateful if $P(\text{hate} \mid u_i, \mathbf{c_i}, p_j) \geq 0.5$, and as non-hateful otherwise.

4 Experiments

Dataset. we conduct experiments on CREHate dataset (Lee et al., 2023b), where each annotator has 8 different backgrounds. We randomly split the data at the post level into training/validation/test with a ratio of 70%/15%/15%. Additional dataset details are provided in Appendix A.4.

Baselines. We compare our method against two groups of baselines: (1) Pretrained language models (PLMs) (Devlin et al., 2019; Nguyen et al., 2020; Caselli et al., 2020; Zhang et al., 2023; Barbieri et al., 2020; Zhou, 2020): To incorporate cultural background information, we prepend learnable background tokens (e.g., "[male]" to indicate gender) to the post text, following prior work (Lee et al., 2023a) and a similar usage in (Fleisig et al., 2023). (2) Zero-Shot Prompting: We further test LLMs in zero-shot setting, including *LLama-2-7b-chat-hf, Mistral-7B-v0.3, Qwen2-7B* and GPT-5. Besides, *LLama-2-SPT* applies soft prompt turning (Lester et al., 2021) by inserting ten trainable

Table 1: Classification Performance

Model	Accuracy	Precision	Recall	F1
HateBERT	76.23±0.15	75.97±0.15	76.10±0.24	76.01±0.18
Twin-BERT	76.26±0.27	75.98±0.27	76.04±0.32	76.00±0.29
Twitter-Roberta	76.33±0.14	76.05±0.15	76.10±0.15	76.06±0.14
ToDect-Roberta	75.90±0.26	75.61±0.26	75.65±0.23	75.63±0.24
BERT	*76.38*±0.28	*76.11*±0.28	*76.20*±0.35	*76.14*±0.31
BERTweet	76.15±0.14	75.89±0.14	76.05±0.14	75.95±0.14
LLama-2-7b-chat-hf	56.79	55.94	52.89	47.80
LLama-2-SPT	73.79	73.49	73.57	73.52
Mistral-7B-v0.3	58.19	58.08	54.68	51.07
Qwen2-7B	61.59	66.15	63.87	60.82
GPT-5	71.08	70.72	70.56	70.63
Ours	**77.37**±0.14	**77.14**±0.15	**77.33**±0.23	**77.19**±0.17

Table 2: Ablation Study

Model	Accuracy	Precision	Recall	F1
Ours	**77.37**±0.14	**77.14**±0.15	**77.33**±0.23	**77.19**±0.17
Ours (sum)	76.06±0.31	75.95±0.19	76.18±0.26	75.92±0.28
Ours (mean)	76.25±0.34	76.04±0.28	76.23±0.21	76.07±0.29
Ours (anno)	76.37±0.17	76.17±0.12	76.40±0.09	76.21±0.13
$-HP(u_i)$	76.20±0.22	76.00±0.12	76.17±0.11	76.01±0.15
$-q_j$	*76.84*±0.23	*76.64*±0.23	*76.88*±0.24	*76.69*±0.22
$-s_j$	76.33±0.44	76.09±0.40	76.05±0.25	76.03±0.36

vectors in the input sequence. Prompt templates and details are provided in Appendix A.1 A.2.

4.1 Classification Evaluation

We evaluate whether models can effectively capture the relationship between text and cultural backgrounds. As shown in Table 1, our proposed method outperform the best baseline by an average margin of 1.05% across all metrics. Since all model are trained on the same set of posts, differences in text encoding ability are minimal, which explains comparable performance of PLMs. This also suggests that PLMs share similar limitations in culture-based modeling under standard fine-tuning. Although GPT-5 achieves relatively strong zero-shot performance, it still falls behind fine-tuned models by a substantial margin.

4.2 Ablation Study

We conduct two groups of ablation experiments to assess our design: (1) Hate perception representation. we compare different strategies for building hate perception. Specifically, we replace the weighted aggregation in Eq 4 with sum and mean pooling, denoted as *Ours (sum)* and *Ours (mean)*. We also replace cultural combinations with direct annotator representations, denoted as *Ours (anno)*. The result show that modeling cultural combination, and aggregating them to represent individuals' hate perception are both beneficial. (2) Each component in Eq 3.3. We evaluate the importance of each input feature by removing one component at a time. Results indicate that individual hate perception is the most critical factor, while all components contribute positively to overall performance.

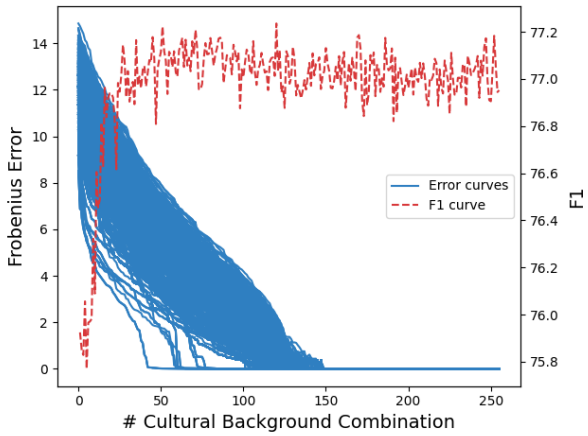

Figure 2: Effect of Number of Cultural Combinations on Hate Subspace and Classification

4.3 The Analysis of Hate Perception Subspace

Our framework represents hate perception using cultural combinations, which can result in a large number of combinations. The embeddings of these combinations form a vector space of individual hate perception. To analyze the effectiveness of combinations, we compute leverage scores for each combination and rank their relative importance and progressively add them in descending order. We measure the Frobenius reconstruction error during this process and find that fewer than half of the combinations are sufficient to reconstruct the space (Figure 2). We further evaluate classification performance and observe that performance converges after using only about 10% of combinations, while adding more combinations may introduce noise.

5 Conclusion

In this paper, we analyze key challenges in culture-based hate speech detection, including data sparsity, interactions among cultural backgrounds, and ambiguous labeling. To address these challenges, we model interactions between cultural background combinations and posts by constructing a culture-post interaction matrix. We then apply matrix factorization to learn hate perception representations for each combination and aggregate them to represent individual hate perception. We further analyze the role of cultural combinations and show that only a small subset is sufficient to achieve strong performance. Extensive experiments validate the effectiveness of our framework, with consistent improvements across evaluation metrics.

6 Limitation

Our model adopts a culture-based modeling perspective; however, verifying whether the learned representations truly capture cultural information remains challenging. To probe this issue, we design a diagnostic experiment using the template "These disgusting [object]", where "[object]" is replaced with a target group, such as female or male. If the model reflects cultural sensitivity, different target groups are expected to induce differences in predicted hate scores across groups. Using the $-q_j$ variant of our model, we compute scores for female and male groups. When "[object]" is set to female, the mean score is 0.6142 for the female group and 0.5794 for the male group. When "[object]" is set to male, the corresponding scores are 0.6256 and 0.6033. In both cases, the female group exhibits higher scores than the male group, regardless of the target group. This result may indicate that the model does not aware cultural information. Alternatively, it may also reflect a general tendency for female annotators to exhibit higher sensitivity to offensive content. Moreover, as shown in Figure 1, distinguishing perceptions of one single background is challenging. These observations suggest that reliably evaluating cultural sensitivity remains an open challenge. Therefore, we deliberately refer to the proposed framework as *culture-based* rather than *culture-aware*, to accurately reflect its modeling scope and avoid overclaiming generalization.

Our framework also inherits inherent limitations from matrix factorization. First, the method treats each combination as a distinct entity, which limits its ability to explicitly model intersectionality (Hancock, 2007). Although we model all observed combinations and apply label propagation to introduce dependencies among them, this limitation cannot be fully eliminated. Second, the model operates only on observed cultural combinations, restricting its ability to generalize to unseen combinations or users. While this design supports interpretable and exploratory analysis over observed data, extending this framework to enable stronger generalization remains an important direction for future work. Overall, although we identify three key challenges in modeling hate perception, the proposed framework alleviates rather than fully resolve these issues.

7 Ethical Considerations

The dataset used in this work is publicly available and anonymous. We do not annotate any data on our own. All the models employed are publicly accessible and their use is consistent with their intended purposes. The proposed framework is intended to support socially beneficial applications, such as improving the understanding and detection of harmful content.

Despite improved performance in culture-based hate speech detection, the model may still misclassify content, which could lead to unintended harm. Moreover, because the framework explicitly models cultural factors, there is a potential risk of misuse, for example, generating or amplifying hate targeting specific cultural or demographic groups. These risks highlight the importance of responsible deployment. Future work should further examine fairness, robustness, and broader societal impacts. All model training and inference were conducted using an NVIDIA Quadro RTX 6000 GPU.

8 Use of AI Assistants

We acknowledge the use of AI language models, such as ChatGPT, for assistance in improving writing clarity and grammar. All research content is solely authored by the human authors.

References

Francesco Barbieri, Jose Camacho-Collados, Leonardo Neves, and Luis Espinosa-Anke. 2020. Tweeteval: Unified benchmark and comparative evaluation for tweet classification. *arXiv preprint arXiv:2010.12421*.

Pete Burnap and Matthew L Williams. 2016. Us and them: identifying cyber hate on twitter across multiple protected characteristics. *EPJ Data science*, 5(1):11.

Peter Burnap and Matthew Leighton Williams. 2014. Hate speech, machine classification and statistical modelling of information flows on twitter: Interpretation and communication for policy decision making.

Tommaso Caselli, Valerio Basile, Jelena Mitrović, and Michael Granitzer. 2020. Hatebert: Retraining bert for abusive language detection in english. *arXiv preprint arXiv:2010.12472*.

Aida Davani, Mark Diaz, Dylan Baker, and Vinodkumar Prabhakaran. 2024. D3code: Disentangling disagreements in data across cultures on offensiveness detection and evaluation. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 18511–18526.

Naihao Deng, Xinliang Zhang, Siyang Liu, Winston Wu, Lu Wang, and Rada Mihalcea. 2023. You are what you annotate: Towards better models through

annotator representations. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 12475–12498.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. 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)*, pages 4171–4186.

Eve Fleisig, Rediet Abebe, and Dan Klein. 2023. When the majority is wrong: Modeling annotator disagreement for subjective tasks. *arXiv preprint arXiv:2305.06626*.

Ange-Marie Hancock. 2007. When multiplication doesn't equal quick addition: Examining intersectionality as a research paradigm. *Perspectives on politics*, 5(1):63–79.

Katerina Korre, Arianna Muti, Federico Ruggeri, and Alberto Barrón-Cedeño. 2025. Untangling hate speech definitions: A semantic componential analysis across cultures and domains. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 3184–3198.

Anna Koufakou, Endang Wahyu Pamungkas, Valerio Basile, Viviana Patti, et al. 2020. Hurtbert: Incorporating lexical features with bert for the detection of abusive language. In *Proceedings of the fourth workshop on online abuse and harms*, pages 34–43. Association for Computational Linguistics.

Nayeon Lee, Chani Jung, Junho Myung, Jiho Jin, Jose Camacho-Collados, Juho Kim, and Alice Oh. 2023a. Exploring cross-cultural differences in english hate speech annotations: From dataset construction to analysis. *arXiv preprint arXiv:2308.16705*.

Nayeon Lee, Chani Jung, and Alice Oh. 2023b. Hate speech classifiers are culturally insensitive. In *Proceedings of the first workshop on cross-cultural considerations in NLP (C3NLP)*, pages 35–46.

Brian Lester, Rami Al-Rfou, and Noah Constant. 2021. The power of scale for parameter-efficient prompt tuning. *arXiv preprint arXiv:2104.08691*.

N Mokhberian, MG Marmarelis, FR Hopp, V Basile, F Morstatter, and K Lerman. 2023. Capturing perspectives of crowdsourced annotators in subjective learning tasks. arxiv preprint. *arXiv preprint arXiv:2311.09743*.

Dat Quoc Nguyen, Thanh Vu, and Anh Tuan Nguyen. 2020. Bertweet: A pre-trained language model for english tweets. *arXiv preprint arXiv:2005.10200*.

Chikashi Nobata, Joel Tetreault, Achint Thomas, Yashar Mehdad, and Yi Chang. 2016. Abusive language detection in online user content. In *Proceedings of the 25th international conference on world wide web*, pages 145–153.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR.

Maarten Sap, Saadia Gabriel, Lianhui Qin, Dan Jurafsky, Noah A Smith, and Yejin Choi. 2019. Social bias frames: Reasoning about social and power implications of language. *arXiv preprint arXiv:1911.03891*.

Maarten Sap, Swabha Swayamdipta, Laura Vianna, Xuhui Zhou, Yejin Choi, and Noah A Smith. 2022. Annotators with attitudes: How annotator beliefs and identities bias toxic language detection. In *Proceedings of the 2022 conference of the north american chapter of the association for computational linguistics: Human language technologies*, pages 5884–5906.

Prashanth Vijayaraghavan, Hugo Larochelle, and Deb Roy. 2021. Interpretable multi-modal hate speech detection. *arXiv preprint arXiv:2103.01616*.

Xinyang Zhang, Yury Malkov, Omar Florez, Serim Park, Brian McWilliams, Jiawei Han, and Ahmed El-Kishky. 2023. Twhin-bert: A socially-enriched pre-trained language model for multilingual tweet representations at twitter. In *Proceedings of the 29th ACM SIGKDD conference on knowledge discovery and data mining*, pages 5597–5607.

Xuhui Zhou. 2020. *Challenges in automated debiasing for toxic language detection*. University of Washington.

A Appendix

A.1 Prompts for Zero-shot Experiments

For open-source LLMs we use following prompts:

> Hate speech is defined as language that attacks or demeans a person or group based on attributes like race, religion, gender, etc. Personal background: [all backgrounds]. Based on the given definition, determine whether people, based on the provided personal background information, are likely to consider the following text as hate speech: POST: [post]. Is this text hateful or not? You have to answer with only one word, "Yes" or "No", do not provide any explanation or irrelevant content. Answer:

For GPT-5, We set instructions as "Perform personalized hate speech classification." to let LLMs aware its role during inferences. Then we first give a definition of hate speech follow with a description for the task and background information:

Definition of Hate Speech: Hate speech refers to offensive discourse targeting a group or an individual based on inherent characteristics such as race, religion, sexual orientation, gender, or any other factors that may threaten social peace.

Answer if this post is hate or not (for people with following backgrounds:[nationality], [age], [education], [ethnicity], [gender], [politic], [religion], [gender_sexual_orientation] with a single alphabet letter among given answer choices a and b.

POST: [post]
a: Hate
b: Non-hate
answer:

A.2 Culturally-adapted PLMs

Analogous to the [CLS] token in BERT, we prepend each post with trainable culture-specific tokens that serves as the representation of the corresponding cultural context. Specifically, posts associated with a given nationality are prefixed with a [nationality] token (e.g., [Singapore]). In our scenario, since we consider multiple backgrounds, we concatenate every tokens in front of the post text, such as, "[864] [Singapore] [100] [2] [Asian] [male] [Moderate_liberal] [Buddhism] [heterosexual]" follow with post text, where "[864]" denotes annotator id.

A.3 Hyperparameters

We experimented with several hyperparameter settings for fine-tuning PLMs and selected the optimal configuration: learning rate $lr = 5\text{e}{-6}$ and $\epsilon = 1\text{e}{-8}$. All experiments were run five times, and we report the mean and standard deviation. The batch size was set to 32 for all experiments. we set $lr = 0.01, \lambda = 0.01$ in Eq 3, and $d = 128$.In label propagation, we assign a constant weight of 1. Specifically, if the combination <US, male> labels post p_j as hateful, the co-occurrence count between <US> and p_j is increased by one.

A.4 Dataset Details

The CREHate splits are reported in Table 3, with background distributions shown in Figures 3–5. The attribute *Education* is a discrete variable, and we map its string values to integer labels ranging from 1 to 7, each corresponding to a distinct education level. In addition, we discretize age into intervals of ten years; for example, individuals aged 27–36 are assigned to the same group, with values 100–107 denoting the corresponding mapped group indices. As shown in Figure 6, background combinations are largely balanced across splits.

Table 3: Class distribution of the dataset.

Split	#Records	Hateful (%)	Non-hateful (%)
Train	26,850	44.90	55.10
Valid	4,738	45.10	54.90
Test	7,898	44.54	55.46

A.5 Complexity Situation

The proposed method mainly consists of three components:

(1) Aggregation: This step is well-controlled. We do not need to propagate labels to all subsets of backgrounds. For example, restricting propagation to specific subsets can effectively trade off performance and preprocessing time (Our Figure 2 shows that only a small number of combinations are needed.

(2) Matrix factorization: This step is highly efficient—it takes only approximately 9 seconds and 1.3 MB on a Quadro RTX 6000.

(3) Learning annotator weights: This takes about 60 seconds and 33.61 MB on the same GPU.

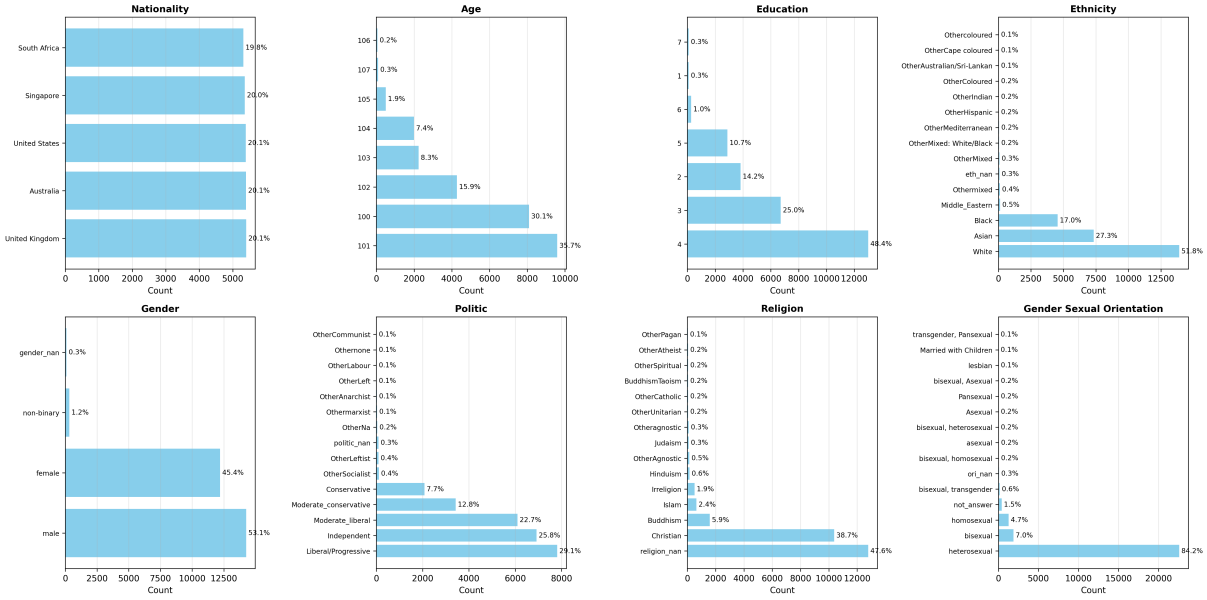

Figure 3: Background distribution in Train set.

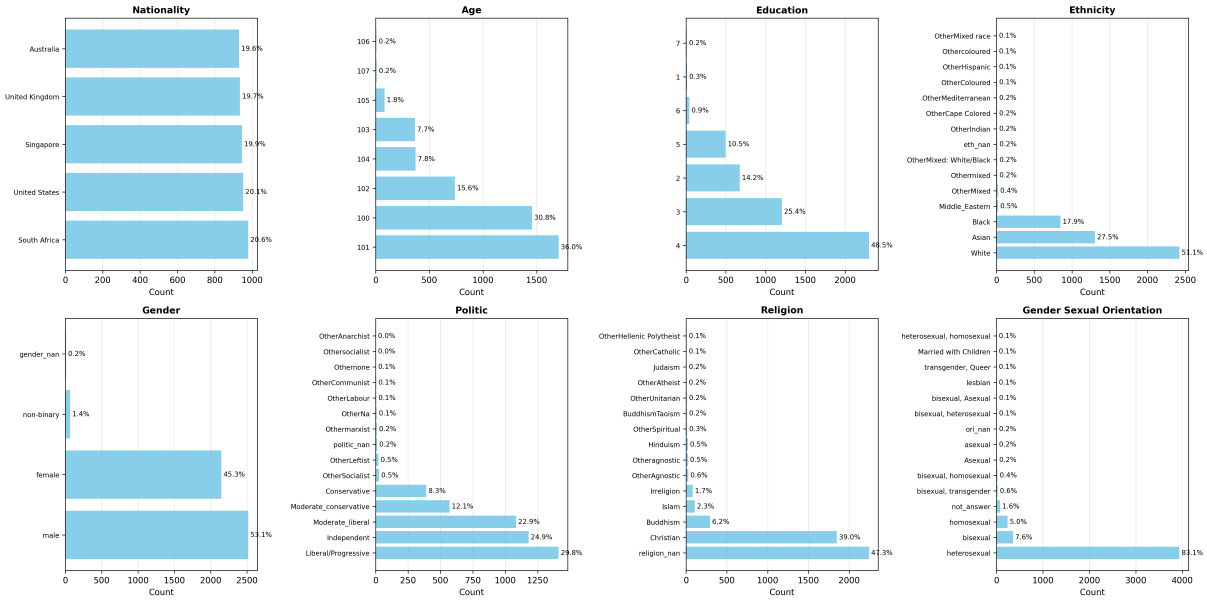

Figure 4: Background distribution in Valid set.

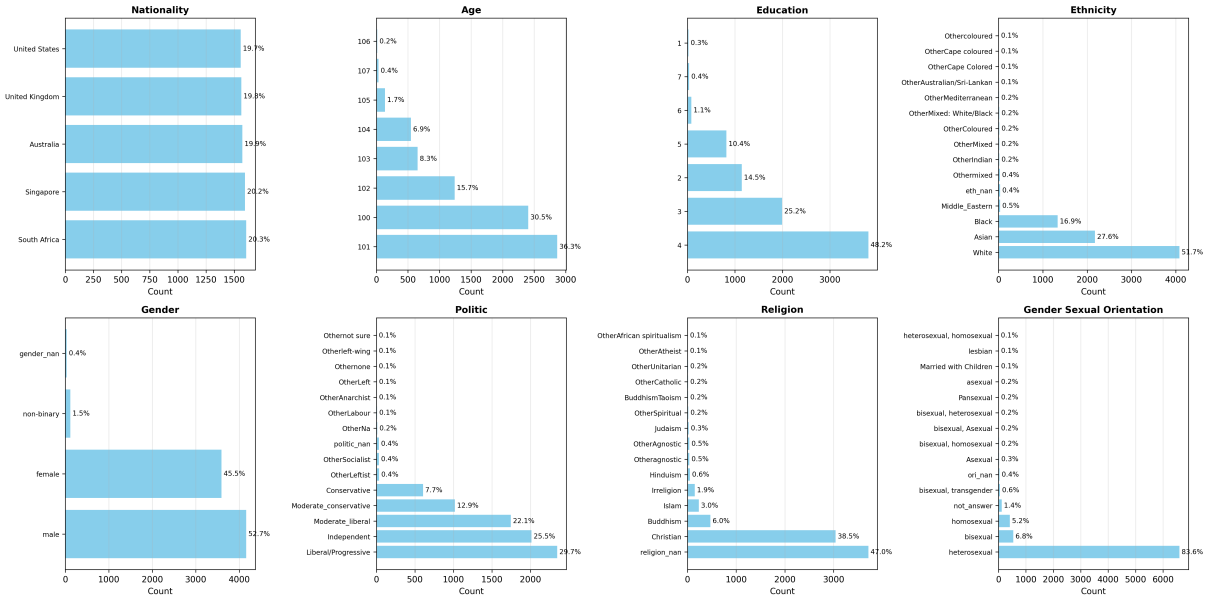

Figure 5: Background distribution in Test set.

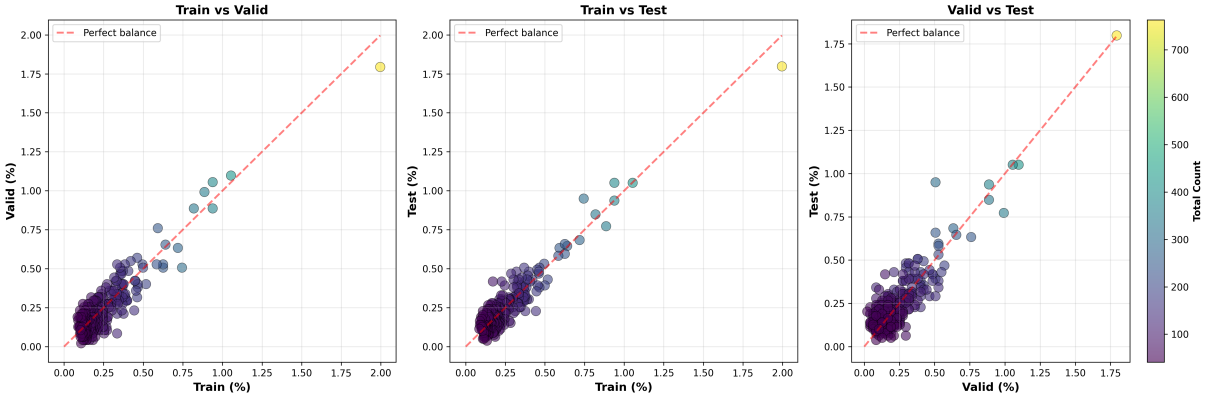

Figure 6: Dataset Balance Analysis. Each circle represents a unique combination of backgrounds. Circle color indicates the frequency of the corresponding combination in the dataset, with brighter colors denoting higher frequencies. The value along each axis represents the percentage of that combination in the dataset. Circles closer to the red reference line indicate a more balanced dataset.