Seeing Hate Differently: Hate Subspace Modeling for Culture-Aware Hate Speech Detection

Weibin Cai
Data Lab, EECS Department
Syracuse University
weibin44@data.syr.edu

Reza Zafarani
Data Lab, EECS Department
Syracuse University
reza@data.syr.edu

Abstract

Hate speech detection has been extensively studied, yet existing methods often overlook a real-world complexity: training labels are biased, and interpretations of what is considered hate vary across individuals with different cultural backgrounds. We first analyze these challenges, including data sparsity, cultural entanglement, and ambiguous labeling. To address them, we propose a culture-aware framework that constructs individuals' hate subspaces. To alleviate data sparsity, we model combinations of cultural attributes. For cultural entanglement and ambiguous labels, we use label propagation to capture distinctive features of each combination. Finally, individual hate subspaces, which in turn can further enhance classification performance. Experiments show our method outperforms state-of-the-art by 1.05% on average across all metrics.

1 Introduction

Hate speech detection aims to determine whether a text contains hateful content. Traditional approaches primarily focus on textual features, such as critical lexicon cues and syntactic patterns (Nobata et al., 2016; Burnap and Williams, 2014, 2016), while recent works rely 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, these results can be misleading as ground-truth labels are often obtained through a majority voting among some annotators, which introduces bias and oversimplifies the problem (Sap et al., 2019). In reality, **individuals from different cultural background may perceive the same text differently**. Prior work has shown cross-cultural disagreement in hate speech annotation (Lee et al., 2023a); for example, annotators from the United States and United Kingdom exhibit higher agreement than those from the United States and Singapore (see Figure 1). Even within the same cultural group, perceptions can diverge considerably. As illustrated in Figure 1, pairwise label agreement ratios reveal that even annotators from the same country can have lower agreement scores compared to those of with other countries (e.g., SG-SG < SG-US). These findings suggest that hate perception is too complex to be explained by a single cultural factor. To better understand hate perception and build personalized hate speech detection systems, it is essential to uncover the diverse factors shaping individual's hate perception.

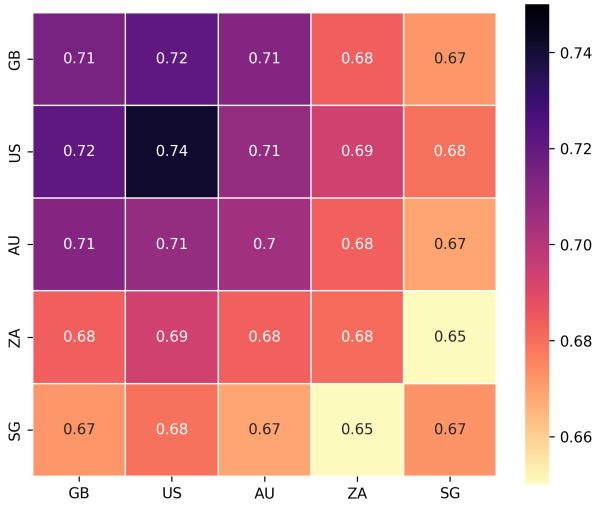

Figure 1: Pairwise hate-speech label agreement ratios between users from five countries: United Kingdom (GB), United States (US), Australia (AU), South Africa (ZA), and Singapore (SG).

Modeling culture-aware hate speech detection presents three major challenges: ①**Data sparsity.** Hate perception is shaped by numerous factors——such as religion and gender——resulting in an exponential number of possible cultural background combinations. For instance, the CREHate dataset (Lee et al., 2023a) provides eight background attributes. Excluding the continuous feature 'age', the remaining seven categorical backgrounds yield 91,045,500 possible combinations (in the ideal scenario), whereas the

dataset contains annotations from only 1,064 annotators, representing merely 1.17×10^{-5} of the theoretical population. ②**Complex and abstract cultural entanglement.** It is difficult to measure how judgments shift when attributes are added or altered. For instance, suppose an annotator with background <Country=United States, Religion=Christian> considers a post hateful, how does this perception change if we add <Sex=Male>, or replace <Religion=Christian> with <Religion=Buddhism>? Current models lack the ability to capture such nuanced entanglements, a limitation also reflected in our experiments showing that LLMs fail to effectively leverage background information (Table 1). ③**Ambiguous labeling.** Cultural attributes in datasets are often incomplete, introducing label noise. Even when labels 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 stem from nationality, religion, or their joint effect. Importantly, even the annotator themselves may not be able to disentangle these factors explicitly.

In this work, we propose a culture-aware hate speech detection framework that models individual's *hate subspaces* to address these challenges. To alleviate data sparsity, we model each cultural background combination in the dataset rather than individual factors. To capture the influence between cultural factors, we introduce a one-way label propagation mechanism from a cultural background combination to its subsets. Although label ambiguity remains difficult to fully resolve, we mitigate its effect by aggregating labels from higher-level combinations and constructing a weight matrix to differentiate between them. Finally, we represent each individual's hate perception using the combinations of their cultural backgrounds:

- We identify key challenges in culture-aware hate speech detection and propose a simple yet effective framework that models individuals' hate subspaces 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.

Table 1: LLM's ability in culture-aware hate speech detection. Details are provided in Appendix A.1

Model	Accuracy	Precision	Recall	F1
One for all	57.17	56.05	55.51	55.15
+background	56.79	55.94	52.89	47.80
+historical labeling	**61.47**	**61.49**	**58.97**	**57.83**
+both	*60.31*	*59.75*	*58.09*	*57.26*

2 Problem Statement

Definition. Culture-Aware Hate Speech Detection. Let $\mathcal{U} = \{(u_1, \mathbf{c}_1), (u_2, \mathbf{c}_2), \ldots, (u_n, \mathbf{c}_n)\}$ be a set of n users, where each user u_i is associated with k cultural background attributes (e.g., nationality, religion), denoted as $\mathbf{c}_i = \{c_{i1}, c_{i2}, \ldots, c_{ik}\}$. Given a post p, the goal of culture-aware hate speech detection is to predict $P(\text{hate}|u_i, p)$ or equivalently $P(\text{hate}|\mathbf{c}_i, p)$, i.e., the likelihood that user u_i would consider post p as hateful, conditioned on the user's cultural background $\mathbf{c}_i$.

3 Method

We begin by modeling cultural hate perception (Section 3.1), which is then used to construct individual hate perception (Section 3.2), and is finally employed for classification (Section 3.3).

3.1 Culture-Post Interaction Matrix

To better model individual hate perception and alleviate data sparsity, we shift our modeling target from individual to specific cultural background combinations. Let $\mathcal{P}(c_i)$ denote the power set of user u_i's cultural background $\mathbf{c}_i$, i.e., $\mathcal{P}(\mathbf{c}_i) = \{\, s \mid s \subseteq \{c_{i1}, c_{i2}, \ldots, c_{ik}\} \,\}$ with $|\mathcal{P}(c_i)| = 2^k$. For example, if $c_i =$ (Male, US, Christian) then

$$\mathcal{P}(c_i) = \left\{ \begin{matrix} \text{Male, US, Christian, \{Male, US\}} \\ \text{\{US, Christian\}, \{Male, Christian\},} \\ \text{\{Male, US, Christian\}} \end{matrix} \right\};$$

Our goal is to predict $P(\text{hate}|g(\mathcal{P}(\mathbf{c}_i)), p_j)$, where $g(\cdot)$ aggregates the effects of all combinations to model an individual's hate perception. Modeling combinations offers a practical benefit: even when an unseen user u_i has incomplete or unseen background information, we can approximate their prediction by utilizing overlapping combinations observed in the dataset, i.e.

$$P(\text{hate}|g(\mathcal{P}(\mathbf{c}_l) \cap (\bigcup_{i=1}^{n} \mathcal{P}(\mathbf{c}_i)), p_j). \quad (1)$$

To further alleviate data sparsity and label ambiguity, we aggregate annotation signals at combination-level. Concretely, for each background combination $comb_l$ and post p_j we collect

$$U_{l,j} = \{\, (u, \mathbf{c}) \in \mathcal{U} \mid comb_l \in \mathcal{P}(\mathbf{c}), \; u \in Label(p_j) \,\} \quad (2)$$

i.e., the set of users who possess combination $comb_l$ and labeled p_j. This aggregation implements a single-direction label propagation from observed (higher-order) combinations toward their constituent combinations: labels provided at a richer (upper) combination inform estimates for its subsets. The intuition is that while a single annotator's label does not reveal which attribute cased the judgment, pooling labels across users who share a combination yields more robust estimates of combination-level tendencies.

To further distinguish combinations, we treat each combination as a "document", and its contributing $(u, \mathbf{c})$ pairs as "words". Under this view we build a culture-post interaction matrix $Y \in \mathbb{R}^{z \times m}$ (with z total combinations and m posts) using TF–IDF weighting derived from aggregated labels and co-occurrence information.

3.2 Individual Hate Perception Embedding

We factorize Y to derive latent features of combinations and posts. Specifically, we initialize an embedding matrix for combinations $P \in \mathbb{R}^{z \times d}$, for posts $Q \in \mathbb{R}^{m \times d}$, and bias terms $B_c \in \mathbb{R}^z$ and $B_w \in \mathbb{R}^m$, where d is the embedding dimension. The predicted score $\hat{Y}_{l,j}$ is estimated as:

$$\hat{Y}_{l,j} = \mu + b_{c_l} + b_{w_j} + q_j^T p_l \quad (3)$$

where μ is the global mean, b_{c_l} and b_{w_j} are biases, and p_l, q_j are latent vectors of combinations and posts. To learn these embeddings, we minimize the following objective:

$$\sum_{l,j} (Y_{l,j} - \hat{Y}_{l,j})^2 + \lambda (b_{c_l}^2 + b_{w_j}^2 + \|q_j\|^2 + \|p_l\|^2) \quad (4)$$

where λ controls regularization and prevents overfitting by penalizing large parameter values.

Individuals interpret hate speech from multiple perspectives, depending on which subset of their cultural attributes is active. Since embeddings of all combinations are aligned in the same latent space, we represent an individual's hate perception embedding as a linear combination of all its combinations:

$$HP(u_i) = \sum_{comb_l \in \mathcal{P}(\mathbf{c_i})} \alpha_l \begin{bmatrix} p_l \\ b_{c_l} \end{bmatrix} \quad (5)$$

where α_l is a learnable coefficient reflecting the relative influence of combination $comb_l$ on 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:

$$P(\text{hate} | u_i, p_j) = f_\theta(HP(u_i), q_j, s_j) \quad (6)$$

where q_j is the post's interaction feature from Eq. 3, 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 model predicts that u_i perceives p_j as hateful if $P(hate \mid u_i, p_j) \geq 0.5$, and 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%.

Baselines. To enable a comprehensive comparison, we evaluate 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 adopt these models to our setting, we introduce additional learnable background tokens (e.g., "[male]" to indicate the user's gender) and prepend them to the post text before fine-tuning, following prior work (Lee et al., 2023a). (2) Zero-Shot Prompting: We further test LLMs in zero-shot setting, including *LLama-2-7b-chat-hf* and GPT-5. The specific prompt templates and more details are described 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 2, our proposed method outperform the best baseline by an average margin of 1.05% across all metrics. Since every model has access to the same set of posts during training, their differences in text encoding capability are minimal, which explains why PLMs achieve comparable results. This also highlights that PLMs share similar limitations in culturally-aware modeling, at least under the standard fine-tuning paradigm. Besides, although GPT-5 in the

Table 2: 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
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 3: 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

zero-shot setting yields relatively strong performance, it still lags behind fine-tuned models by a substantial margin.

4.2 Ablation Study

To validate the effectiveness of our approach, we design two sets of ablation experiments: (1) Construction of hate subspace. we compare different strategies for building individual hate perception. Specifically, we replace the weighted sum in Eq 3.2 with sum and mean pooling, denoted as *Ours (sum)* and *Ours (mean)*, respectively. In addition, we replace cultural combinations with direct annotators, denoted as *Ours (anno)*. The result demonstrate that modeling cultural combination, and constructing individuals' hate subspaces are helpful. (2) Contribution of each component in Eq 3.3. We further assess the importance of each input feature by removing one component at a time (denoted by "-"). The results show that an individual's hate subspace is the most critical factor. All components contribute positively to the overall performance.

4.3 The Analysis of Hate Subspace

In this paper, we assume that an individual's hate subspace is constructed by all possible cultural background combinations. However, the size of these combinations can be extremely large. For instance, in CREHate dataset, although there are only 1,064 annotators, the induced cultural combinations exceed 60,000. This raises both computational and modeling concerns. To analyze the complexity and effectiveness of combinations, we first compute

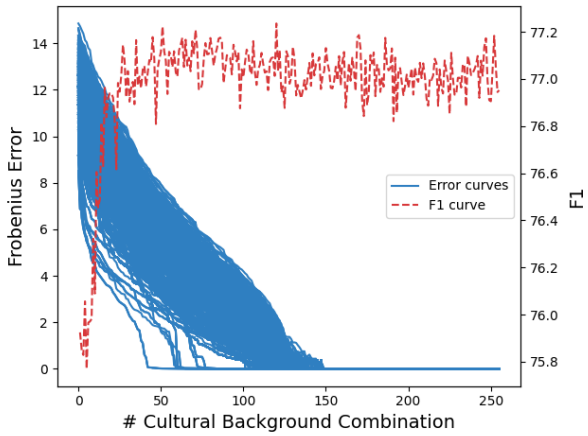

Figure 2: Effect of Number of Cultural Combinations on Hate Subspace and Classification

the leverage scores of each combination within an individual's hate subspace. We then calculate the Frobenius reconstruction error when progressively adding combinations in descending order of leverage scores. As shown in Figure 2, fewer than half of the combinations are sufficient to reconstruct most of the hate subspaces with minimal error. We further evaluate classification performance by gradually accumulating combinations. The red curve reveals a consistent trend: performance saturates after incorporating roughly 50 combinations, and adding more combinations may introduce noise. This observation suggests that a large number of combinations are redundant and highlights an important direction for future work——understanding which combinations truly matter and how to effectively select them when constructing the hate subspace.

5 Conclusion

In this paper, we analyzed the challenges in culture-aware hate speech detection, focusing on three main challenges: data sparsity, the complex interplay between cultural backgrounds, and ambiguous labeling. To address these challenges, we model the interaction between cultural background combinations and posts. Specifically, we construct a culture-post interaction matrix using label propagation and apply matrix factorization to derive the hate perception of each combination. These perceptions are further used to form individuals' hate subspace, which can be leveraged to enhance classification performance. Finally, we examine the effect of the number of cultural combinations on classification. Extensive experiments demonstrate the effectiveness of our approach, yielding an average improvement of 1.05% across all metrics.

6 Limitation

In this paper, we propose a culture-aware model; however, it remains uncertain whether the model truly captures cultural information, and evaluating this is challenging. To investigate this, we designed an experiment using the template "These disgusting [object]", where "[object]" is replaced with a specific group, such as female or male. If the model is genuinely culture-aware, inserting a target group into "[object]" should lead to higher sensitivity from that group toward others, as reflected in the classifier's output.

Using the $-q_j$ version of the model, we obtain predicted scores for the female and male groups. When "[object]" was set to female, the mean hateful score is 0.6142 for the female group and 0.5794 for the male group. When "[object]" was set to male, the female group scored 0.6256 and the male group 0.6033. In both cases, the female group's scores are higher than the male group's, regardless of the target group. This may suggest that the approach fails to aware culture. However, an alternative explanation is that females may generally exhibit higher sensitivity than males. Another factor contributing to these similar scores may be the dataset itself; as shown in Figure 1, distinguishing groups from a single cultural background is challenging. Therefore, while evaluating cultural awareness is important, it remains difficult to conduct reliably.

7 Ethical Considerations

The dataset we use is publicly available and anonymous. We do not annotate any data on our own. All the models used in this paper are publicly accessible. Their usage are consistent with their intended use. The usage of our proposed work should be used for social good.

Although our model improves cross-cultural hate speech detection, it may still misclassify text, potentially causing unintended harm. Besides, since this work focuses on culture-aware modeling, there is a potential risk that the model could be misused to generate hate speech targeting specific cultural or demographic groups. Future work should address fairness and societal implications. The inference and finetuning of models are performed on Quadro RTX 6000.

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*.

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.

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.

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*.

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

In Table 1, we use LLaMA-2-7b-chat-hf to perform zero-shot experiments to test LLM's reasoning ability in personalized hate speech detection. *One for all*: post content only, prediction shared by all annotators. *+background*: adds annotator's cultural background to the prompt. *+historical labeling*: adds 5 previously labeled posts of the annotator. *+both*: includes both cultural background and historical labeling. Details are provided in Appendix A.1 Here are prompts for different prompts:

Prompt for "One for all":

Hate speech is defined as language that attacks or demeans a person or group based on attributes like race, religion, gender, etc. Based on the given definition, determine whether people 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.

Prompt for "+background":

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:

Prompt for "+historical labeling":

Hate speech is defined as language that attacks or demeans a person or group based on attributes like race, religion, gender, etc. Historical labeling: [5 historical posts and corresponding labels]. Based on the given definition, determine whether people, based on their historical labeling of hate speech, 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:

Prompt for "+both":

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]. Historical labeling: [5 historical posts and corresponding labels]. Based on the given definition, determine whether people, based on their background and historical labeling of hate speech, 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 latest LLMs, 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 4, and $d = 128$.