Exploring Cross-Cultural Differences in English Hate Speech Annotations: From Dataset Construction to Analysis

Anonymous ACL submission

Abstract

Warning: *this paper contains content that may be offensive or upsetting.*

Most hate speech datasets neglect the cultural diversity within a single language, resulting in a critical shortcoming in hate speech detection. To address this, we introduce **CREHate**, a **CR**oss-cultural **E**nglish **Hate** speech dataset. To construct CREHate, we follow a two-step procedure: 1) cultural post collection and 2) cross-cultural annotation. We sample posts from the SBIC dataset, which predominantly represents North America, and collect posts from four geographically diverse English-speaking countries (Australia, United Kingdom, Singapore, and South Africa) using culturally hateful keywords we retrieve from our survey. Annotations are collected from the four countries plus the United States to establish representative labels for each country. Our analysis highlights statistically significant disparities across countries in hate speech annotations. Only 56.2% of the posts in CREHate achieve consensus among all countries, with the highest pairwise label difference rate of 26%. Qualitative analysis shows that label disagreement occurs mostly due to different interpretations of sarcasm and the personal bias of annotators on divisive topics. Lastly, we evaluate large language models (LLMs) under a zero-shot setting and show that current LLMs tend to show higher accuracies on Anglosphere country labels in CREHate.

1 Introduction

Identifying hate speech is highly subjective and relies heavily on an annotator's understanding and knowledge of the cultural context (Aroyo et al., 2019; Waseem, 2016). Unfortunately, existing English hate speech datasets often overlook the cultural diversity within the posts and the annotators. They are predominantly collected from Twitter (Table 1), reflecting a disproportionate representation

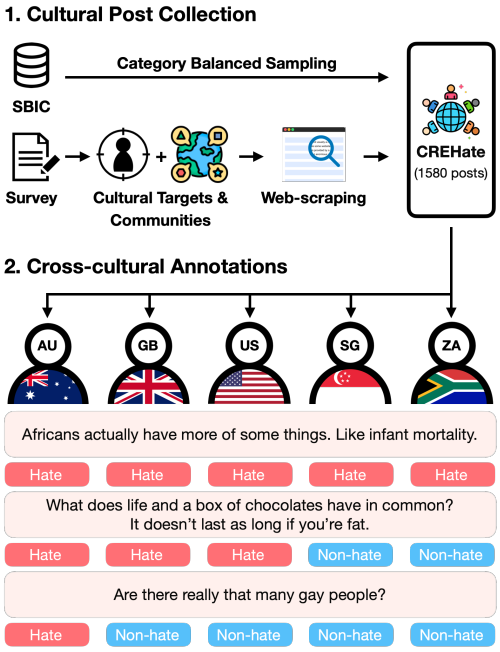


Figure 1: Illustration of the two-step procedure of CREHate construction: 1) cultural post collection and 2) cross-cultural annotation. The examples show how annotations on identical posts differ across countries.

of certain countries, notably the United States [1]. Furthermore, annotators' geographic location is either neglected or limited to only one or two countries, despite English being spoken in over 50 countries [2]. This limitation hinders the datasets' ability to capture diverse viewpoints. Figure 1 illustrates how people from different countries show varying hate speech annotations on identical posts.

In this research, we align culture with nationality when exploring how cultural background influences annotators' interpretations of hate speech. We acknowledge that focusing only on cross-country differences may not fully encompass the multifaceted cultural dynamics within each coun-

[1]The US has the most Twitter users by country (https://datareportal.com/essential-twitter-stats).

[2]The World Factbook, Languages (https://www.cia.gov/the-world-factbook/field/languages/)

try. However, it offers a starting point to understand how annotators' cultural background based on nationality affects language interpretation, particularly in sensitive areas like hate speech. This approach underlines the importance of further, more detailed studies into the complex interplay of cultural identities and their impact on language perception (Kramsch, 2014), especially for enhancing hate speech moderation on global platforms.

To this end, we construct **CREHate** [3]—a **CR**oss-cultural **E**nglish **Hate** speech dataset—comprising 1,580 online posts annotated by individuals from five English-speaking countries: Australia (AU), the United Kingdom (GB), Singapore (SG), the United States (US), and South Africa (ZA) [4]. Construction of CREHate is done in a 2-step procedure: 1) cultural post collection and 2) cross-cultural annotation (Figure 1). For cultural post collection, we collect 600 posts from YouTube and Reddit using keywords gathered from surveys from four countries: AU, GB, SG, and ZA. We also sample 980 posts from SBIC (Sap et al., 2020), a toxic language dataset of social media posts including diverse target groups, primarily reflecting a North American perspective (Table 1) [5]. For cross-cultural annotation, five annotators from each country annotate each post to establish representative labels for each country. Based on cross-cultural considerations, this dataset creation procedure makes CREHate more culturally comprehensive than datasets that ignore cultural differences within English-speaking countries.

We show that cross-cultural annotations of CREHate demonstrate significant differences across countries. Only 56.2% of the entire posts receive unanimous label agreement across all five countries, and the average pairwise agreement between countries is 78.8%, with a maximum label disagreement of 26.0%. The pairwise label agreement distribution among countries exhibits a notable deviation from that of randomly selected annotator groups, with its average being 2.58σ lower than the average pairwise label agreement of the random groups. Furthermore, by conducting a qualitative analysis of potential reasons for label disagreements, we show that the primary contributing factors are likely due to different understandings of sarcasm and the personal bias of annotators on divisive topics.

Datasets	Post Source	Source Country	Annotation Platform (Country)
MLMA (Ousidhoum et al., 2019)	Twitter	US*	MTurk (N/A)
ImplicitHateCorpus (ElSherief et al., 2021)	Twitter	US	MTurk (N/A)
SBIC (Sap et al., 2020)	Twitter, Reddit, Gab, Stormfront	US*	MTurk (US, CA[6])
HateXplain (Mathew et al., 2021)	Twitter, Gab	US*	CrowdFlower (N/A)
OLID (Zampieri et al., 2019)	Twitter	US*	CrowdFlower (N/A)
Davidson et al. (2017)	Twitter	US*	CrowdFlower (N/A)
Founta et al. (2018)	Twitter	US*	CrowdFlower (N/A)
CREHate (Ours)	Twitter, Reddit, Gab, Stormfront, YouTube	AU, GB, SG, US*, ZA	MTurk, Prolific, Tictag (AU, GB, SG, US, ZA)

Table 1: Datasets for toxic language detection annotated using crowdsourcing platforms. Existing datasets neglect or limit the cultural backgrounds of the annotators and posts. 'US*' means there is a high possibility that the post sources are biased towards US due to the platform's skewed user demographics, even if not explicitly targeted during the data collection stage.

Additionally, we show that current LLMs tend to show higher accuracies on core Anglosphere country labels in CREHate. We further identify the limitations of these models in culture-specific hate speech classification, in which they are asked to predict hate speech based on the target country.

Our main contributions are as follows:

- We build CREHate, a cross-cultural English hate speech dataset including posts and annotations from diverse cultural backgrounds.
- Through quantitative and qualitative analysis, we identify significant variations in hate speech annotations attributed to the cultural backgrounds of the posts and the annotators.
- We show LLMs' higher accuracies on core Anglosphere country labels in hate speech classification and limitations in making culture-specific predictions.

2 Related Work

Impact of Annotator Demographics. Annotator demographics, such as gender, affect their annotations in NLP datasets (Biester et al., 2022). Hate

[3]Our dataset and codes used are available at: https://anonymous.4open.science/r/CREHate-6BFE

[4]Two-letter ISO country codes (https://www.iso.org/iso-3166-country-codes.html).

[5]Reddit and Gab's users are mainly from the US (https://www.semrush.com/website/reddit.com/overview/, https://www.semrush.com/website/gab.com/overview/), as well as Twitter.

[6]CA refers to Canada.

speech detection is particularly a subjective task where the demographics can affect the annotations, inter-annotator agreement (IAA), and classifier performance (Waseem, 2016; Sap et al., 2022; Goyal et al., 2022; Larimore et al., 2021; Binns et al., 2017).

Cultural Considerations in Hate Speech Detection. Recent research in offensive language examined cultural differences and built datasets in diverse languages (Lee et al., 2023; Jeong et al., 2022; Jin et al., 2023; Arango Monnar et al., 2022; Deng et al., 2022; Demus et al., 2022; Mubarak et al., 2022), but these papers assume that a single language reflects a single culture. However, languages such as English are spoken by a culturally diverse population, necessitating the consideration of cultural differences among language speakers. Arango Monnar et al. (2022) built the first hate speech dataset for Chilean Spanish to enrich the cultural diversity of Spanish datasets. They evaluated knowledge transfer performance on another Spanish dataset with a different cultural background, but the impact of cultural background on annotations was unexplored. We aim to conduct a thorough study of how hate speech and its annotations vary across English-speaking countries.

Multiple Cultures in English NLP. Frenda et al. (2023) developed a corpus for irony detection, focusing on which annotator demographic group's perspectives are more represented by majority voting. They collected posts and gathered annotators from five English-speaking countries: Ireland, India, AU, GB, and US. Our study, focusing on hate speech detection, extends the scope by collecting posts as well as annotations from different cultures and investigating the annotation disparities stemming from cultural variations.

3 Dataset Construction

This section describes the construction process of CREHate, an English hate speech dataset with both posts and annotations collected from five different countries to analyze the country-level divergences when it comes to hate speech. We follow a 2-step procedure: 1) cultural post collection and 2) cross-cultural annotation. The dataset consists of 1,580 posts, each with five labels representing five countries, resulting in a total of 7,900 labels. Dataset statistics are shown in Table 2.

English-speaking Countries. We choose one country from each continent to ensure geographical diversity while also considering cultural differences within and outside the Anglo-American sphere of influence (Cox and O'Connor, 2020; Gamble, 2021). Specifically, we select three core Anglosphere countries—AU, GB, and US (Davies et al., 2013)—and two countries with English as official language but not necessarily the primary language—SG and ZA (Khokhlova, 2015; Tan, 1997).

Data		Source	# Posts
CREHate	CC-SBIC	Reddit	568
		Twitter	273
		Gab	80
		Stormfront	59
		subtotal	*980*
	CP	Reddit	311
		YouTube	289
		subtotal	*600*
		total	**1,580**

Table 2: Data statistics and sources of CREHate. CC-SBIC refers to cross-culturally re-annotated SBIC posts. CP refers to additionally collected cultural posts from four countries (AU, GB, SG, and ZA), which are also cross-culturally annotated.

3.1 CREHate Post Collection

3.1.1 Sampling from SBIC

To incorporate hate speech targeting diverse groups, we sample posts from the SBIC dataset (Sap et al., 2020), which contains annotations of offensive posts targeted towards different demographic groups and minorities. From SBIC, we sample 980 posts while balancing the target group categories. The details of SBIC and the sampling process are specified in Appendix A.1.1. This set of sampled posts is referred to as CC-SBIC (**C**ross-**C**ultural SBIC) throughout the paper, as it is cross-culturally re-annotated as mentioned in §3.2.

3.1.2 Collecting Cultural Samples

The sources of SBIC's posts are culturally skewed towards the US, resulting in a bias towards prevalent target groups and the cultural context of the US. To address this issue, we collect and annotate 150 cultural online posts each (a total of 600 posts) from four English-speaking countries: AU, GB, SG, and ZA. The posts are collectively referred to as CP, and the country-specific posts are called CP_{AU}, CP_{GB}, CP_{SG}, and CP_{ZA}, respectively.

Keyword Collection. To efficiently gather hate speech posts, we use words that refer to specific demographic groups that are often subjected to hate as queries. We recruit workers whose nationality and current residency match our target country and

Data	Post	AU	GB	SG	US	ZA
CC-SBIC	Why did they plant trees in harlem? For free public transit.	○	○	○	○	×
CP_{AU}	The special treatment is for real abbos only..... equality.	○	○	×	×	○
CP_{GB}	@USER A pikey calling someone uneducated. The irony is painful.	×	○	○	○	○
CP_{SG}	The attitude of Malays being "holier than thou" is kinda sickening. Have experienced it many times, sadly it's not gonna stop.	○	○	×	×	○
CP_{ZA}	I wonder when people will recognize the EFF for what it really is. It's not a political party, it's a terrorist organization.	×	×	×	×	○

Table 3: Examples of annotated posts in CREHate. ○ refers to 'hate', and × refers to 'non-hate' label. The example posts require culture-specific knowledge on target minorities (abbos, Malays), locations (Harlem), public organizations (EFF), and slang (pikey). People from different countries provide different hate speech annotations for posts including culture-specific context.

who have spent most of their lives in their respective countries to obtain the most appropriate and culturally relevant keywords. We ask them to provide commonly targeted groups and possible hateful keywords that may refer to them within their culture. We collect target groups in *race/ethnicity*, *gender/sexuality*, and *religion/culture* categories, the three main categories within the original SBIC dataset. We continue collecting until we gather at least 20 keywords per country.

Post Collection. We gather popular social media and news sites from the workers in their countries and select Reddit as our primary social media platform for collecting comments, as it is widely used across all countries. We also crawl comments from the YouTube channels of news sites in each country. To ensure that we have enough potentially hateful posts in our dataset, we go through a pre-annotation stage, gathering only two annotations from the country the post originated from. Based on the pre-annotation results, we finalize 150 posts to be annotated from each country. As a result, we maintain the ratio of posts labeled as hate between 39.8% and 48.5% for each country [7]. As a result, the posts from each culture contain some unique topics and keywords, such as '*abo*' or '*lebs*' in CP_{AU}, '*gypsy*' or '*paki*' in CP_{GB}, '*malay*' or '*pinoy*' in CP_{SG}, and '*boer*' or '*EFF*' in CP_{ZA}.

3.2 Cross-Cultural Annotation

Annotator Recruitment. We recruit annotators from five countries, applying the same annotator qualifications as we used for keyword collection, from Prolific [8] (AU, GB, ZA), Amazon Mechanical Turk [9] (US), and Tictag [10] (SG) depending on annotator recruitment availability of the desired country. As a result, we have 1,061 annotators, balancing their gender but not restricting others for a broader representation of demographics (Frenda et al., 2023). Table 10 shows a detailed demographic distribution of annotators.

Annotation Process. Before annotating, annotators are required to review the definitions [11] and examples of hate and non-hate speech. Examples are selected among posts with identical labels across all countries from the pilot study. The task is to annotate posts as either *Hate* or *Non-hate*, with an additional option of *I don't know* [12]. We obtain five *Hate* or *Non-hate* labels for each post from each country. The specific annotation process and quality control methods are in Appendix A.3.

Label Finalization. After gathering all five annotations, we use majority voting to finalize the representative labels for each country. Examples of posts with labels from each country are presented in Table 3.

4 Analysis on the Annotations

In this section, we show that varying cultural backgrounds of annotators and posts lead to a significant disparity in hate speech annotation.

4.1 Significance of Cultural Backgrounds

To analyze the role of an annotator's cultural background in hate speech detection, we obtain labels representative of different demographic categories[13] using majority voting. We only collect labels from groups with at least three annotators per

[7] Specific post crawling and sampling process is provided in Appendix A.1.2.

[8] https://www.prolific.co/

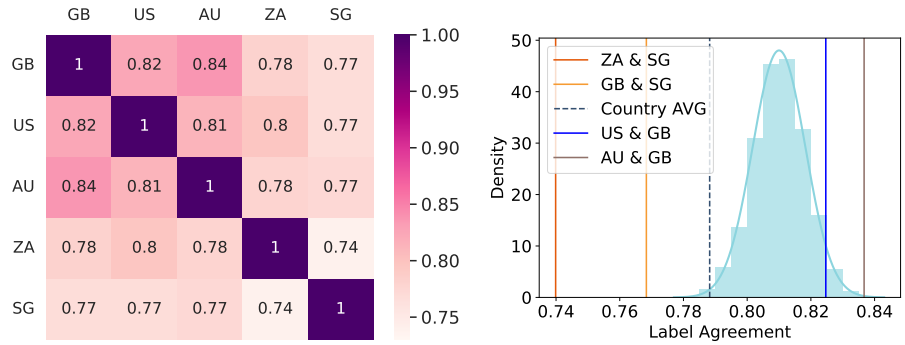
[9] https://www.mturk.com/

[10] https://www.tictagkr.com/

[11] https://www.un.org/en/hate-speech/understanding-hate-speech/what-is-hate-speech

[12] The *I don't know* labels took up about 3-7% of the raw annotations, and more analysis on these labels are mentioned in Appendix B.

[13] For more details on the demographic categories analyzed and their statistics, please refer to Table 10 in the Appendix.



(a) (b)

Figure 2: (a) Pairwise label agreements across countries ordered by the average agreement with others. Labels from Singapore tend to be the most different. (b) Comparison of the label agreements among country pairs and random ones. The histogram and its density function show the distribution of pairwise label agreements among randomly selected annotator groups. The solid lines indicate country pairs with top-2 and bottom-2 label agreement scores, and the dashed line indicates the average of label agreements of all country pairs. Countries that are closely related exhibit high label agreements compared to the random annotator groups, whereas culturally distant countries show significantly low label agreements compared to label agreements from random annotator groups.

post on average. Labels from each group are subjected to chi-squared tests, and the results indicate significant disparities in annotations across country ($p = 0.000$), race ($p = 0.002$), gender ($p = 0.006$), and education level ($p = 0.000$), while there were no significant differences for other groups. Several studies have shown the importance of race or gender of annotators (Pei and Jurgens, 2023; Sachdeva et al., 2022), whereas the impact of annotators' cultural background has been underexplored.

4.2 Label Agreement among Countries

Pairwise Country Label Agreement. Overall, only 56.2% of the posts achieve unanimous agreement across all countries, with 25.5% of the posts showing agreement across four countries. To further explore the label differences across cultures, we examine the label agreements between all pairs of countries, as shown in Figure 2a. It suggests pairwise label agreements among core Anglosphere countries are greater than those observed in other country pairs. Among all countries, AU and GB exhibit the highest label agreement at 83.7%, while SG and ZA show the lowest agreement at 74.0%.

We compare these results to the cultural distance index (Kogut and Singh, 1988) [14] between countries, which measures the degree to which cultural norms in two countries differ (Table 12). The cultural distance and the hate speech label agreements among the countries show a high negative Pearson correlation with $r = -0.658$ ($p = 0.039$). This implies that country pairs with more considerable cultural distances have lower label agreement. SG and ZA, the country pair with the lowest label agreement, show a higher cultural distance (2.178) than AU and GB (0.144), the country pair with the highest agreement.

Furthermore, to investigate the pairwise label differences on identical posts across different countries, we employ the McNemar Test (McNemar, 1947). The results indicate significant pairwise label disparity between 8 out of 10 country pairs.

Comparison with Random Annotator Groups. To show that label disparities stem from the annotators' cultural backgrounds rather than random variations among individuals, we compare the pairwise country label agreements with the distribution of label agreements between randomly organized annotator groups. For each post, we create two groups of five randomly selected annotations out of 25 (5 from each country) and construct representative labels from each group via majority voting. We calculate the label agreement of the two groups for the whole dataset and repeat this process 10^5 times. The outcomes are illustrated in Figure 2b, which includes a histogram and the corresponding estimated normal distribution.

Based on the D'Agostino-Pearson normality test (D'Agostino and Pearson, 1973), the label agreements among random annotators follow a normal distribution with $\mu = 0.81$ and $\sigma = 0.008$. The two highest label agreements, shown between the core Anglosphere countries, are larger than the average label agreement among random annotator groups by 1.82σ and 2.36σ. The two lowest agreements fall 4.97σ and 8.41σ below this average. The

[14] A value of 0 indicates identical cultural norms, while a value close to 1 indicates average distance among all countries.

	Agreement	H-F1	N-F1
CREHate	0.7882	0.7636	0.8077
CC-SBIC	0.8045	0.8034	0.8050
CP	0.7617	0.6762	0.8108
CP$_{AU}$	0.7293	0.6937	0.7565
CP$_{GB}$	0.7493	0.6851	0.7913
CP$_{SG}$	0.7827	0.6583	0.8390
CP$_{ZA}$	0.7853	0.6565	0.8433

Table 4: The average pairwise label agreements F1 scores for hate (H-F1) and non-hate (N-F1) labels among countries on subsets of CREHate. Our cultural posts (CP) show lower average pairwise country label agreement and lower F1 scores for hate labels.

average label agreement between pairs of different countries falls 2.58σ below the average agreement of the random annotator groups. Through this, we demonstrate that label agreements between culturally distant countries significantly differ from those between random annotator groups.

Label Agreements on Subsets of CREHate. We also analyze label agreements among countries on different subsets of CREHate (Table 4). Firstly, we compare the label agreements on two disjoint subsets of CREHate, CC-SBIC and CP. Our findings reveal that CP has a lower average pairwise label agreement than CC-SBIC. Although the two divisions show similar average pairwise F1 scores for non-hate labels, the F1 score for hate labels on CP significantly lags behind CC-SBIC's. This implies that CP derives more considerable label disparities for identifying one post as hate compared to CC-SBIC. This trend is consistent across all sets of posts collected from different countries.

Annotator Agreement. Krippendorf's α is used to calculate IAA in US ($\alpha = 0.462$), GB ($\alpha = 0.425$), AU ($\alpha = 0.408$), ZA ($\alpha = 0.351$), and SG ($\alpha = 0.344$). These are higher than or comparable to those achieved in previous work in toxic language detection (Ross et al., 2016; Sap et al., 2020).

4.3 Annotators' Disagreement Analysis

We analyze the main factors behind label disagreements across countries using the taxonomy of reasons for annotators' disagreement for subjective tasks proposed by Sandri et al. (2023). Appendix D shows detailed definitions and examples for each reason type. Some of the annotated examples are shown in Table 5.

Disagreement Reason Annotation. Among the 1,580 posts in CREHate, 692 posts exhibit label discrepancies across countries. To conduct a thorough analysis, we randomly sample 400 posts, including 200 posts from CC-SBIC and 50 posts from each of the four country's CP posts. After a norming session, in which we clarify category definitions and apply them to our task, two authors annotate all sampled posts. The initial Cohen's Kappa score from the two authors is 0.556, which is comparable to that of the annotations in Sandri et al. (2023) (0.591), done by two linguists. After that, the authors go through a discussion stage to establish a consensus on all labels. As a result, the labels on the reasons for disagreement are finalized based on a unanimous agreement between the authors.

Possible Factors behind Disagreement. Overall, *ambiguity* and *subjectivity* of the posts contributed the most to the disagreements, taking up 44.3% and 37.5%, respectively. Among the lower-level subtype reasons, *sarcasm* was the most frequently observed, followed by *personal bias*, *swearing*, and *not complete* as shown in Figure 3. A detailed analysis comparing CC-SBIC and CP's main disagreement reasons are shown in Appendix D.2.

Data	Post	Reason Type
CP$_{AU}$	You're Irish and you support your pals the Provisional IRA murderers of the innocent Australian civilians Stephen Melrose and Nick Spanos. You support terrorists.	personal bias
CP$_{GB}$	Gays have cards? are they sparkly?	sarcasm
CP$_{SG}$	So with the repeal of 377A, married men are going to church to have gay sex?	not complete

Table 5: Examples of disagreement reason annotation. For a sampled set of posts that countries disagree on, we annotate the possible reasons behind the disagreements following the disagreement reason taxonomy for subjective tasks by Sandri et al. (2023).

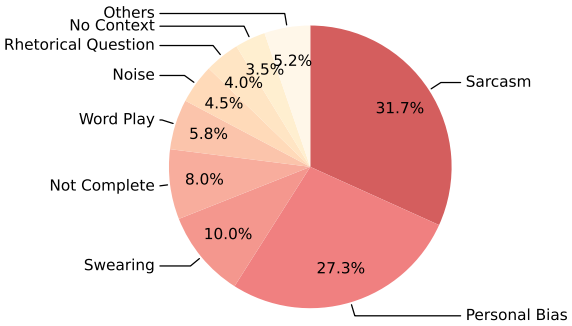


Figure 3: Ratio of disagreement reasons within posts. Differing interpretations of sarcasm and personal bias on divisive topics contribute to the main factors of disagreement.

Sarcasm heightens challenges in intercultural agreement in hate speech annotation, as annotators' sensitivity to sarcasm may vary depending on the topic and the annotators' cultural backgrounds. Furthermore, sarcasm referring to a specific culture-specific context may be difficult for annotators from different backgrounds to accurately identify.

Personal bias also plays a significant role in label disagreements, as they may arise when annotators hold differing opinions about specific topics, especially divisive issues. For example, if the post is about divisive topics within the annotator's culture, their personal bias would have a larger impact on the annotation.

Swearing is important in label disagreement since annotators' perceived offensiveness of a swear word can vary depending on their backgrounds. Different cultures may have varying perceptions of swear words based on their usage and social context, resulting in label disagreements on the text containing them.

Not complete indicates insufficient information for annotators to fully comprehend the post. Annotators from diverse cultures may struggle to label posts involving cultural references or nuances from other cultures when crucial information is missing, requiring extra cultural background knowledge.

5 Experiments

This section evaluates the performance of current LLMs in hate speech classification on CREHate, with a specific focus on analyzing their performance with respect to country-specific annotations.

Experimental Settings. We conduct zero-shot experiments using a multiple-choice question format. We use five variations of the question prompt, each presented with two answer options: 'Hate' and 'Non-hate.' We test both answer option orders for all prompts, resulting in 10 formats [15].

The experiments are divided into two parts. In §5.1, we ask the model to predict whether a given post is hate speech. In §5.2, we add a target country to the question prompt to examine the model's ability to make culture-specific predictions in hate speech detection. For all results, we present the average values for all variations and perform a Wilcoxon rank-sum test to determine any statistically significant differences between the highest and the lowest values.

The models evaluated include GPT-4 (gpt-4-1106-preview), GPT-3.5 (gpt-3.5-turbo-1106), Orca 2 (7B) (Mitra et al., 2023), Flan-T5-XXL (Chung et al., 2022), and OPT (Iyer et al., 2022) [16].

5.1 Zero-shot Predictions and Country Labels

We obtain the predictions of LLMs using prompts such as:

> Answer if this post is hate or not with a single alphabet letter among given answer choices a and b.

Table 6 shows that the accuracy scores of all models, except for GPT models, are below 73% when evaluated against all country labels on CREHate. We also assess whether LLM-based hate speech predictions are biased towards particular cultures by comparing the accuracy of model predictions against the labels from each country. Our analysis reveals that GPT models exhibit cultural bias, as their predictions display higher accuracy to the core Anglosphere cultures' labels. While GPT-4 shows the highest overall accuracy across all country labels with an average value of 78.2%, it also exhibits a significant performance gap with a maximum value of 6.79%, most prominently between US labels (highest accuracy) and SG labels (lowest accuracy). These findings suggest that high model accuracy does not necessarily equate to fairness, highlighting the need for more diverse training datasets and methods to mitigate cultural biases.

To determine if the IAA differences among countries (as shown in §4.2) are the primary cause of varying accuracies, we examine the model accuracy on posts with unanimous annotator agreement within each country [17]. Our analysis reveals that GPT-4 shows higher accuracy for US labels (95.25%) and lower accuracy for SG labels (87.11%), even on unanimously agreed posts. This suggests that the bias is inherent to the model's processing rather than a reflection of annotation quality.

Furthermore, we observe that the overall accuracy for CP posts is lower than that of CC-SBIC across all countries. Even for posts with unanimous annotator agreement within each country for the two dataset divisions, accuracies for CC-SBIC are higher than those on CP for most models. This indicates difficulties in models classifying hate speech

[15] The prompts we use are listed in Appendix E.1.

[16] For GPT models, we use the OpenAI API and set the temperature as 0 to use greedy decoding (https://platform.openai.com/docs/models). For other models, we use the Huggingface Transformers (Wolf et al., 2020) library (https://github.com/huggingface/transformers).

[17] Please refer to Table 14 for the results for all models in the Appendix.

Model	Data	GB	US	AU	ZA	SG
GPT-4	CREHate	79.66	**80.64***	78.02	78.03	74.65
	CC-SBIC	80.74	**82.13***	79.28	80.63	75.34
	CP	77.91	**78.21***	75.96	73.79	73.54
GPT-3.5	CREHate	**72.47***	70.62	72.39	69.28	71.94
	CC-SBIC	75.73	75.00	**75.75***	73.20	75.10
	CP	**67.13***	63.47	66.90	62.87	66.77
Orca 2	CREHate	**69.99**	69.09	69.80	68.80	68.61
	CC-SBIC	72.19	**72.58**	72.13	72.15	70.87
	CP	**66.38**	63.38	65.98	63.32	64.92
Flan T5	CREHate	**68.58**	67.49	68.28	68.35	68.15
	CC-SBIC	72.49	**72.86***	71.84	71.63	70.35
	CP	62.18	58.72	62.48	62.98	**64.55***
OPT	CREHate	66.25	**69.29**	64.68	66.94	64.11
	CC-SBIC	65.22	**68.75**	64.27	67.68	63.06
	CP	67.93	**70.18***	65.36	65.72	65.83

Table 6: Accuracy of the models in terms of each country's labels in each dataset division. The highest score is highlighted in **bold**, while the lowest score is underlined. The asterisk (*) means the two values differ significantly ($p < 0.05$).

in CP posts explicitly sourced from countries other than the US.

5.2 Culture-Specific Hate Speech Classification

In this section, we explore the culture-specific hate speech classification capability of GPT-4, which is the model that performed better overall in our previous experiment. We aim to evaluate the model's ability to make culturally tailored predictions regarding hate speech classification using prompts that target specific countries:

> Answer if this post is hate or not **in {*country*}** with a single alphabet letter among given answer choices a and b.

Our analysis indicates that GPT-4's performance remains consistent, regardless of the inclusion of specific country information in the prompts, as shown in Table 7. The inclusion of country names in the prompts led to only a marginal variation in predictions, with a 6.5-8% difference from the predictions obtained using the original, non-country-specific prompts. This finding implies that providing country context alone does not significantly enhance GPT-4's ability to identify hate speech accurately across different cultural contexts. Consequently, this underscores a limitation in the model's capability to adapt its hate speech detection to specific cultural nuances merely through the explicit mention of a country in the prompt. However, introducing cultural background information or other extra knowledge about the target country, or even using different prompts, may show different results.We leave the exploration of prompt engineering that could enhance culture-specific hate speech detection in LLMs for future work.

Prompt	GB	US	AU	ZA	SG
Original	79.66	**80.64***	78.02	78.03	74.65
+ in GB	79.66	**80.28***	77.97	77.36	73.52
+ in US	79.27	**80.26***	77.34	77.09	73.32
+ in AU	**79.62***	79.59	77.95	77.40	73.48
+ in ZA	79.07	**79.61***	77.38	77.44	72.91
+ in SG	**79.70***	79.56	78.02	77.53	73.27

Table 7: Accuracy of GPT-4 in terms of each country labels when asked to predict whether a given post is hateful within specific countries (e.g., "*Answer if this post is hate or not in Australia.*"). The highest score is highlighted in **bold**, while the lowest score is underlined. The asterisk (*) means the two values differ significantly ($p < 0.05$).

6 Conclusion

In this paper, we analyze how cultural differences across English-speaking countries affect hate speech annotations. To this end, we develop CREHate, a cross-cultural English hate speech dataset comprising 1,580 posts from five English-speaking countries—AU, GB, SG, US, and ZA. Our work shows that there are notable variations in hate speech interpretations between these countries through various statistical methods. The overall agreement on hate speech identification across all countries is only 56.2%, with an average pairwise country disagreement of 21.2%. Qualitative analysis suggests these differences stem from varied understandings of sarcasm and annotators' biases on divisive topics. We also discover that GPT models display higher accuracies with labels from Anglosphere cultures and fail to make culturally tailored predictions when the target country is given.

This research establishes a foundational framework for continuously evaluating and adapting hate speech models and datasets. We suggest expanding CREHate to include more countries and posts to create a comprehensive tool for assessing cultural biases in model predictions and enhancing culturally tailored hate speech detection. We urge collaborative efforts in constructing datasets with broad cultural references and contextual nuances. Annotators with relevant cultural knowledge should be employed to construct a more representative cultural dataset. Such a comprehensive approach is crucial for developing more effective, culturally sensitive hate speech classifiers and promoting safer and more inclusive online communication.

Limitations

CREHate consists of 1,580 posts, making it relatively small compared to other existing English hate speech datasets. Moreover, the collection of culture-specific posts was limited to Reddit and YouTube based on fixed hate-related keywords, which may introduce bias into the collected posts. Also, employing a single crowdsourcing platform for collecting each country's annotation may lead to annotator bias, as different platforms possess varying user demographics. To enhance the representativeness and generalizability of our findings, we anticipate future efforts to expand our dataset by using diverse platforms and post collection methods.

Considering that many countries are multicultural, it is also essential to examine within-country annotation differences. For instance, Singapore has a diverse population, including Chinese, Malaysians, and Indians. Exploring hate speech annotation differences across different ethnicities within a country presents another avenue for investigation. Moreover, although we recruit annotators from countries where English is one of their official language(s), this may not be enough to cover all English-speaking cultures. Further study is needed to include English as a Foreign Language (EFL) learners in cross-cultural hate speech detection. Moreover, the same approach could be extended to languages other than English (e.g., Spanish) spoken in various countries.

There are other subjective tasks that are affected by cultural context, such as common sense reasoning. Future research could extend the scope of our study to other tasks by constructing datasets tailored towards specific cultures, both within and across countries with diverse languages.

Ethics Statement

This research was conducted with full approval from the Institutional Review Board (IRB). The instructions that were given to the annotators, including the disclaimer, can be seen in Figure 4 in the Appendix. We made sure to inform the annotators from the crowdsourcing platforms that the contents they encounter during the annotation task may potentially be offensive or distressing. We also provided access to online therapy platforms and encouraged the annotators to seek help in case they experience any strong negative reactions or mental distress.

We conducted our crowd worker recruitment without any discrimination based on age, ethnicity, disability, or gender. Our workers are compensated at a rate higher than Prolific's ethical standards. Our payment principles are based on the ethical standards of Prolific, and we ensure that our workers are compensated at a rate that is higher than the minimum wage of £9.00 per hour. It is worth noting that this amount exceeds the federal minimum wage in the United States and Singapore, where the annotation process was held on other crowdsourcing platforms.

We are aware of the potential risk involved in releasing a dataset containing hate speech or offensive language. We will explicitly state the terms of usage, emphasizing our unequivocal disapproval of any form of malicious exploitation. We urge researchers and practitioners to harness this dataset only for constructive purposes. We expect our dataset to contribute to developing more equitable and culturally sensitive automated content moderation systems. We emphasize our unequivocal disapproval of any form of malicious exploitation of our dataset, including any misuse of our dataset for generating hateful language. We demand that researchers and practitioners use this dataset solely for constructive purposes.

We used AI assistants — ChatGPT [18], Google Translate [19], and Grammarly [20] — to assist with editing and translating sentences in our paper writing.

References

Ayme Arango Monnar, Jorge Perez, Barbara Poblete, Magdalena Saldaña, and Valentina Proust. 2022. Resources for multilingual hate speech detection. In *Proceedings of the Sixth Workshop on Online Abuse and Harms (WOAH)*, pages 122–130, Seattle, Washington (Hybrid). Association for Computational Linguistics.

Lora Aroyo, Lucas Dixon, Nithum Thain, Olivia Redfield, and Rachel Rosen. 2019. Crowdsourcing subjective tasks: The case study of understanding toxicity in online discussions. In *Companion Proceedings of The 2019 World Wide Web Conference*, WWW '19, page 1100–1105, New York, NY, USA. Association for Computing Machinery.

Francesco Barbieri, Jose Camacho-Collados, Luis Espinosa Anke, and Leonardo Neves. 2020. TweetEval: Unified benchmark and comparative evaluation for tweet classification. In *Findings of the Association*

[18] https://chat.openai.com
[19] https://translate.google.com
[20] https://app.grammarly.com

for Computational Linguistics: EMNLP 2020, pages 1644–1650, Online. Association for Computational Linguistics.

Laura Biester, Vanita Sharma, Ashkan Kazemi, Naihao Deng, Steven Wilson, and Rada Mihalcea. 2022. Analyzing the effects of annotator gender across NLP tasks. In *Proceedings of the 1st Workshop on Perspectivist Approaches to NLP @LREC2022*, pages 10–19, Marseille, France. European Language Resources Association.

Reuben Binns, Michael Veale, Max Van Kleek, and Nigel Shadbolt. 2017. Like trainer, like bot? Inheritance of bias in algorithmic content moderation. In *Social Informatics*, pages 405–415, Cham. Springer International Publishing.

Luke Breitfeller, Emily Ahn, David Jurgens, and Yulia Tsvetkov. 2019. Finding microaggressions in the wild: A case for locating elusive phenomena in social media posts. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1664–1674, Hong Kong, China. Association for Computational Linguistics.

Tommaso Caselli, Valerio Basile, Jelena Mitrović, and Michael Granitzer. 2021. HateBERT: Retraining BERT for abusive language detection in English. In *Proceedings of the 5th Workshop on Online Abuse and Harms (WOAH 2021)*, pages 17–25, Online. Association for Computational Linguistics.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Eric Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Sharan Narang, Gaurav Mishra, Adams Yu, Vincent Y. Zhao, Yanping Huang, Andrew M. Dai, Hongkun Yu, Slav Petrov, Ed H. Chi, Jeff Dean, Jacob Devlin, Adam Roberts, Denny Zhou, Quoc V. Le, and Jason Wei. 2022. Scaling instruction-finetuned language models. *CoRR*, abs/2210.11416v5.

Lloyd Cox and Brendon O'Connor. 2020. That "Special Something": The U.S.-Australia Alliance, Special Relationships, and Emotions. *Political Science Quarterly*, 135(3):409–438.

Ralph D'Agostino and E. S. Pearson. 1973. Tests for departure from normality. Empirical results for the distributions of b2 and $\sqrt{\text{b1}}$. *Biometrika*, 60(3):613–622.

Thomas Davidson, Dana Warmsley, Michael Macy, and Ingmar Weber. 2017. Automated hate speech detection and the problem of offensive language. *Proceedings of the International AAAI Conference on Web and Social Media*, 11(1):512–515.

Andrew Davies, Graeme Dobell, Peter Jennings, Sarah Norgrove, Andrew Smith, Nic Stuart, and Hugh White. 2013. Keep calm and carry on: Reflections on the anglosphere. Technical report, Australian Strategic Policy Institute.

Ona de Gibert, Naiara Perez, Aitor García-Pablos, and Montse Cuadros. 2018. Hate speech dataset from a white supremacy forum. In *Proceedings of the 2nd Workshop on Abusive Language Online (ALW2)*, pages 11–20, Brussels, Belgium. Association for Computational Linguistics.

Christoph Demus, Jonas Pitz, Mina Schütz, Nadine Probol, Melanie Siegel, and Dirk Labudde. 2022. Detox: A comprehensive dataset for German offensive language and conversation analysis. In *Proceedings of the Sixth Workshop on Online Abuse and Harms (WOAH)*, pages 143–153, Seattle, Washington (Hybrid). Association for Computational Linguistics.

Jiawen Deng, Jingyan Zhou, Hao Sun, Chujie Zheng, Fei Mi, Helen Meng, and Minlie Huang. 2022. COLD: A benchmark for Chinese offensive language detection. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 11580–11599, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

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, Minneapolis, Minnesota. Association for Computational Linguistics.

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 2021 Conference on Empirical Methods in Natural Language Processing*, pages 345–363, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Antigoni Founta, Constantinos Djouvas, Despoina Chatzakou, Ilias Leontiadis, Jeremy Blackburn, Gianluca Stringhini, Athena Vakali, Michael Sirivianos, and Nicolas Kourtellis. 2018. Large scale crowdsourcing and characterization of Twitter abusive behavior. *Proceedings of the International AAAI Conference on Web and Social Media*, 12(1).

Simona Frenda, Alessandro Pedrani, Valerio Basile, Soda Marem Lo, Alessandra Teresa Cignarella, Raffaella Panizzon, Cristina Marco, Bianca Scarlini, Viviana Patti, Cristina Bosco, and Davide Bernardi. 2023. EPIC: Multi-perspective annotation of a corpus of irony. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13844–13857, Toronto, Canada. Association for Computational Linguistics.

Andrew Gamble. 2021. 4: The Anglo–American World View1. In *After Brexit and other essays*, pages 75 – 90. Bristol University Press, Bristol, UK.

Nitesh Goyal, Ian D. Kivlichan, Rachel Rosen, and Lucy Vasserman. 2022. Is your toxicity my toxicity? Exploring the impact of rater identity on toxicity annotation. *Proceedings of the ACM on Human-Computer Interaction*, 6(CSCW2):1–28.

G. Hofstede. 1984. *Culture's Consequences: International Differences in Work-Related Values*. Cross Cultural Research and Methodology. SAGE Publications.

Srinivasan Iyer, Xi Victoria Lin, Ramakanth Pasunuru, Todor Mihaylov, Daniel Simig, Ping Yu, Kurt Shuster, Tianlu Wang, Qing Liu, Punit Singh Koura, Xian Li, Brian O'Horo, Gabriel Pereyra, Jeff Wang, Christopher Dewan, Asli Celikyilmaz, Luke Zettlemoyer, and Ves Stoyanov. 2022. OPT-IML: Scaling language model instruction meta learning through the lens of generalization. *CoRR*, abs/2212.12017v3.

Younghoon Jeong, Juhyun Oh, Jongwon Lee, Jaimeen Ahn, Jihyung Moon, Sungjoon Park, and Alice Oh. 2022. KOLD: Korean offensive language dataset. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 10818–10833, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Jiho Jin, Jiseon Kim, Nayeon Lee, Haneul Yoo, Alice Oh, and Hwaran Lee. 2023. KoBBQ: Korean bias benchmark for question answering. *CoRR*, abs/2307.16778v1.

Irina Khokhlova. 2015. Lingua franca english of south africa. *Procedia - Social and Behavioral Sciences*, 214:983–991.

Bruce Kogut and Harbir Singh. 1988. The effect of national culture on the choice of entry mode. *Journal of International Business Studies*, 19(3):411–432.

Claire Kramsch. 2014. Language and culture. *Research methods and approaches in Applied Linguistics*, 27:30–55.

Savannah Larimore, Ian Kennedy, Breon Haskett, and Alina Arseniev-Koehler. 2021. Reconsidering annotator disagreement about racist language: Noise or signal? In *Proceedings of the Ninth International Workshop on Natural Language Processing for Social Media*, pages 81–90, Online. Association for Computational Linguistics.

Nayeon Lee, Chani Jung, and Alice Oh. 2023. Hate speech classifiers are culturally insensitive. In *Proceedings of the First Workshop on Cross-Cultural Considerations in NLP (C3NLP)*, pages 35–46, Dubrovnik, Croatia. Association for Computational Linguistics.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692v1.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *International Conference on Learning Representations*.

Binny Mathew, Punyajoy Saha, Seid Muhie Yimam, Chris Biemann, Pawan Goyal, and Animesh Mukherjee. 2021. HateXplain: A benchmark dataset for explainable hate speech detection. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(17):14867–14875.

Quinn McNemar. 1947. Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika*, 12(2):153–157.

Arindam Mitra, Luciano Del Corro, Shweti Mahajan, Andrés Codas, Clarisse Simões, Sahaj Agrawal, Xuxi Chen, Anastasia Razdaibiedina, Erik Jones, Kriti Aggarwal, Hamid Palangi, Guoqing Zheng, Corby Rosset, Hamed Khanpour, and Ahmed Awadallah. 2023. Orca 2: Teaching small language models how to reason. *CoRR*, abs/2311.11045v2.

Aida Mostafazadeh Davani, Mark Díaz, and Vinodkumar Prabhakaran. 2022. Dealing with disagreements: Looking beyond the majority vote in subjective annotations. *Transactions of the Association for Computational Linguistics*, 10:92–110.

Hamdy Mubarak, Sabit Hassan, and Shammur Absar Chowdhury. 2022. Emojis as anchors to detect arabic offensive language and hate speech. *CoRR*, abs/2201.06723v2.

Dat Quoc Nguyen, Thanh Vu, and Anh Tuan Nguyen. 2020. BERTweet: A pre-trained language model for English tweets. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 9–14, Online. Association for Computational Linguistics.

Nedjma Ousidhoum, Zizheng Lin, Hongming Zhang, Yangqiu Song, and Dit-Yan Yeung. 2019. Multilingual and multi-aspect hate speech analysis. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4675–4684, Hong Kong, China. Association for Computational Linguistics.

Jiaxin Pei and David Jurgens. 2023. When do annotator demographics matter? measuring the influence of annotator demographics with the POPQUORN dataset. In *Proceedings of the 17th Linguistic Annotation Workshop (LAW-XVII)*, pages 252–265, Toronto, Canada. Association for Computational Linguistics.

Björn Ross, Michael Rist, Guillermo Carbonell, Benjamin Cabrera, Nils Kurowsky, and Michael Wojatzki.

2016. Measuring the Reliability of Hate Speech Annotations: The Case of the European Refugee Crisis. In *Proceedings of NLP4CMC III: 3rd Workshop on Natural Language Processing for Computer-Mediated Communication*, volume 17 of *Bochumer Linguistische Arbeitsberichte*, pages 6–9, Bochum.

Pratik S. Sachdeva, Renata Barreto, Claudia von Vacano, and Chris J. Kennedy. 2022. Assessing annotator identity sensitivity via item response theory: A case study in a hate speech corpus. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '22, page 1585–1603, New York, NY, USA. Association for Computing Machinery.

Marta Sandri, Elisa Leonardelli, Sara Tonelli, and Elisabetta Jezek. 2023. Why don't you do it right? Analysing annotators' disagreement in subjective tasks. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 2428–2441, Dubrovnik, Croatia. Association for Computational Linguistics.

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 58th Annual Meeting of the Association for Computational Linguistics*, pages 5477–5490, Online. Association for Computational Linguistics.

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, Seattle, United States. Association for Computational Linguistics.

Jason Tan. 1997. Education and colonial transition in singapore and hong kong: Comparisons and contrasts. *Comparative Education*, 33(2):303–312.

Zeerak Waseem. 2016. Are you a racist or am I seeing things? Annotator influence on hate speech detection on Twitter. In *Proceedings of the First Workshop on NLP and Computational Social Science*, pages 138–142, Austin, Texas. Association for Computational Linguistics.

Zeerak Waseem and Dirk Hovy. 2016. Hateful symbols or hateful people? Predictive features for hate speech detection on Twitter. In *Proceedings of the NAACL Student Research Workshop*, pages 88–93, San Diego, California. Association for Computational Linguistics.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Marcos Zampieri, Shervin Malmasi, Preslav Nakov, Sara Rosenthal, Noura Farra, and Ritesh Kumar. 2019. Predicting the type and target of offensive posts in social media. 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 1415–1420, Minneapolis, Minnesota. Association for Computational Linguistics.

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*, KDD '23, page 5597–5607, New York, NY, USA. Association for Computing Machinery.

Xuhui Zhou, Maarten Sap, Swabha Swayamdipta, Yejin Choi, and Noah Smith. 2021. Challenges in automated debiasing for toxic language detection. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 3143–3155, Online. Association for Computational Linguistics.

Appendix

A Dataset Construction Details

A.1 Post Collection

A.1.1 SBIC

Category	Original Test Set (%)	Sampled Dataset (%)
Race/Ethnicity	819 (17.5)	150 (15.3)
Gender/Sexuality	503 (10.7)	150 (15.3)
Religion/Culture	495 (10.6)	150 (15.3)
Victims	215 (4.6)	150 (15.3)
Disability	112 (2.4)	112 (11.4)
Social/Political	104 (2.2)	104 (10.6)
Body/Age	58 (1.2)	58 (5.9)
Non-hate	2765 (58.9)	327 (33.4)
Total	4691	980

Table 8: Category distribution within the original and the sampled SBIC test set. CC-SBIC posts are comprised of randomly sampled 980 posts from the original SBIC test set, maintaining balance among target group categories. Multi-labeled group categories are split into multiple individual categories when counting.

Source	Reddit	YouTube
AU	r/australia, r/Australian, r/melbourne, r/sydney, r/perth, r/brisbane, r/Adelaide	Sky News Australia
GB	r/unitedkingdom, r/CasualUK, r/england, r/Scotland, r/Wales, r/northernireland	SkyNews, GBNews
SG	r/singapore, r/SingaporeRaw, r/singaporehappenings, r/singapuraa	CNA, The Straits Times
ZA	r/southafrica, r/RSA, r/capetown, r/johannesburg, r/Durban, r/Pretoria	SABC News, eNCA

Table 9: Data sources for each country. We crawled comments from country-specific subreddits and news platforms' YouTube channels.

Posts in SBIC originate from subReddits, microaggressions corpus (Breitfeller et al., 2019), Twitter (Founta et al., 2018; Davidson et al., 2017; Waseem and Hovy, 2016), and hate sites (Gab [21] and Stormfront (de Gibert et al., 2018)). The dataset contains offensive posts targeted towards diverse demographic group categories, including *race/ethnicity*, *gender/sexuality*, *religion/culture*, *victims*, *disability*, *social/political*, and *body/age*. We maintain a 2:1 ratio between hateful and non-hateful posts in our sampled SBIC data to prioritize our analysis on hate speech rather than non-hate speech. The sampled SBIC data statistics are shown in Table 8.

A.1.2 Cultural Posts

Specific subReddits and news sites used for post crawling are shown in Table 9. There is only one news site for Australia, as no other YouTube channels of news sites provided by the workers allow comments. On Reddit, we extract all comments on the posts that include the target group names or the keywords provided by workers. On YouTube, we search using the query, '<media name> + <target group name>', to locate comments related to the target groups (e.g., 'BBC news pakistani'). We only include comments and posts written in 2020 or later for an up-to-date dataset.

After crawling cultural posts from the four countries, we go through a pre-annotation stage in order to balance hate and non-hate speech in our dataset to the extent possible. The process begins by randomly selecting 300 comments from each country, balancing those from Reddit and YouTube. We then obtain two annotations per comment from the source country of the comments. Subsequently, we curate a collection of 150 comments by selecting 50 from each of the three hate annotation counts, ranging from 0 to 2. With this procedure, we get 600 cultural posts from four countries.

A.1.3 Post-processing of Posts

SBIC posts and crawled Reddit and YouTube comments contained usernames and URLs that were not masked. To anonymize all posts, we mask the usernames as `@USER`, and URLs as `URL`.

A.1.4 Terms of Use

Our research is performed in the public interest under GDPR, as we meet the substantial public interest conditions as an academic research. The SBIC dataset is licensed under CC BY 4.0. We use Reddit's official data API, following the terms of use mentioned in 'Data API Terms' [22]. We use YouTube API from Google for Developers site, following the terms of use mentioned in Complying with YouTube's Developer Policies page [23]. The CREHate dataset is licensed under CC BY-SA 4.0.

A.2 Annotator Demographics

Table 10 shows the total number of annotators and the proportion of all types of demographic groups among annotators from each country. The first three demographic categories—gender, ethnicity, and level of education—are shown to be factors that significantly affect hateful post annotations.

A.3 Annotation Process

Disclaimer and instruction are first shown to the annotators, as shown in Figure 4. Each annotator is then asked to answer a demographic survey. If the annotator matches our target group mentioned in Section §3.1.2, annotators proceed to the guideline page shown in Figure 5. After reading the guideline for a minimum of 30 seconds, annotators are asked to annotate 15 posts (Figure 6) that are randomly assigned among the remaining ones.

We include two explicit and two implicit attention check questions among the annotation questions to ensure the dataset's quality. The implicit attention check questions are selected from the samples on which all annotators from all countries agree in previously completed annotations. For the first round of the actual survey, we choose samples with total agreement from the pilot study. As the study progresses, we update them with the new samples on which annotators agreed. The two explicit attention checks instruct the annotators to

[21] https://files.pushshift.io/gab/GABPOSTS_CORPUS.xz

[22] https://www.redditinc.com/policies/data-api-terms

[23] https://developers.google.com/youtube/terms/developer-policies-guide?hl=en

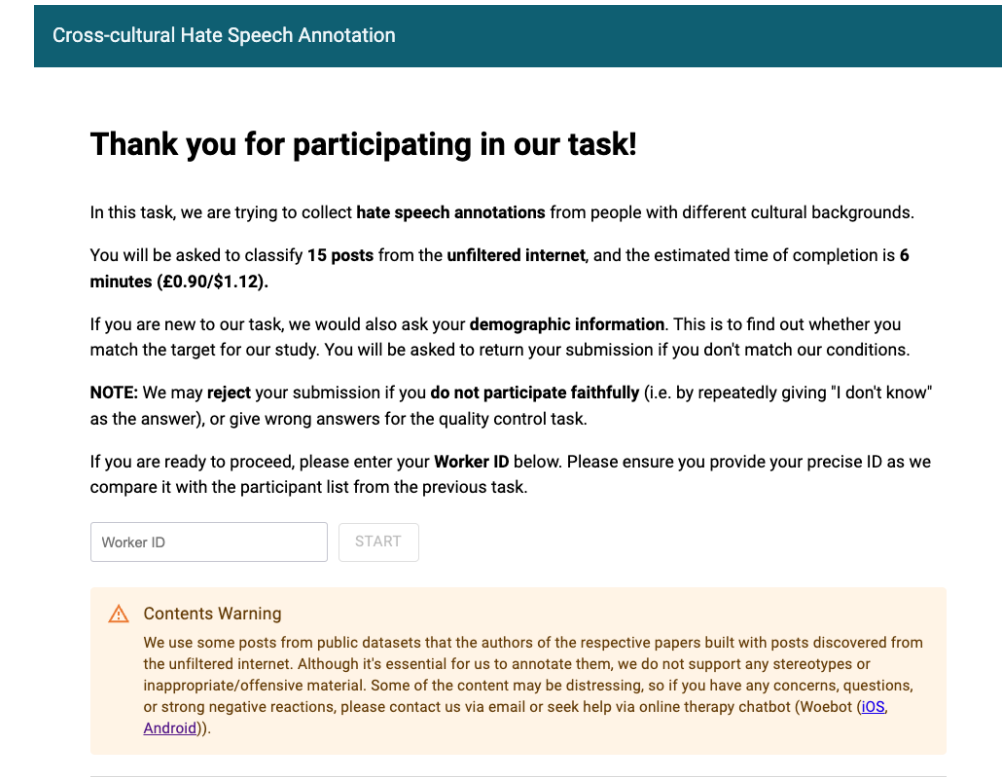


Figure 4: Disclaimer and instruction shown to the annotators.

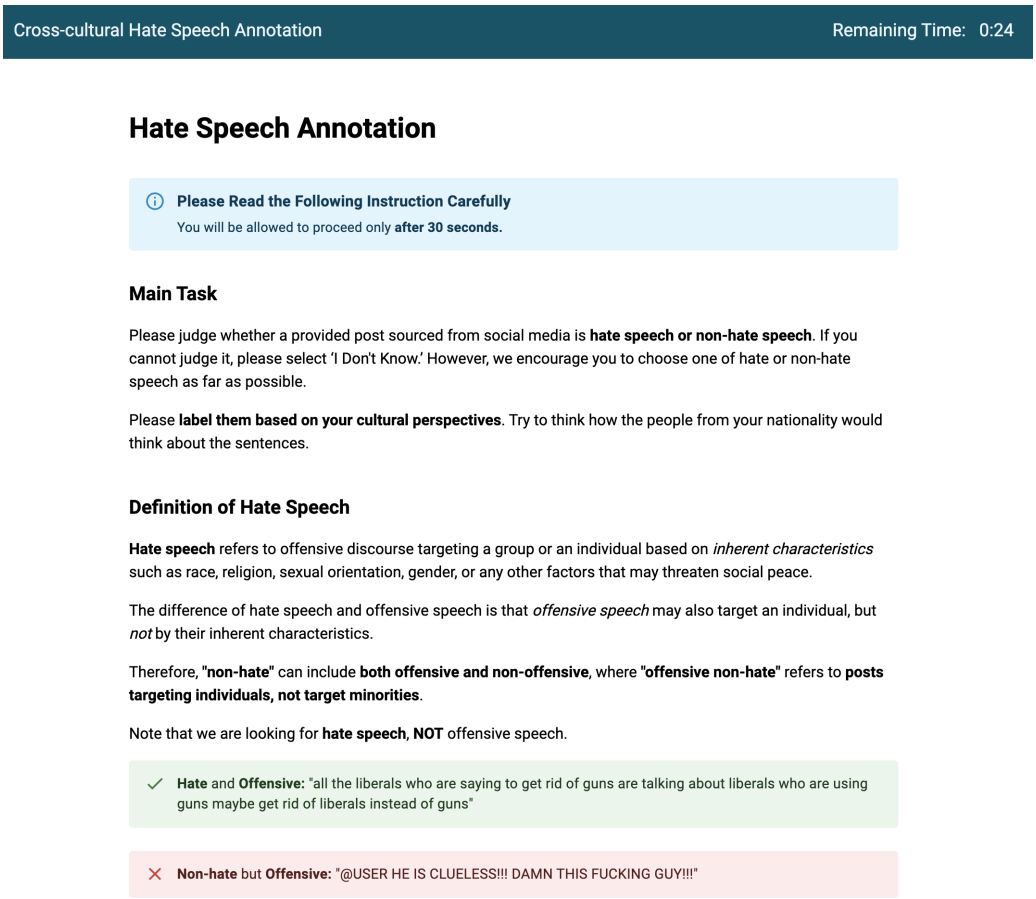


Figure 5: Guideline page of the hate speech annotation task shown to the annotators.

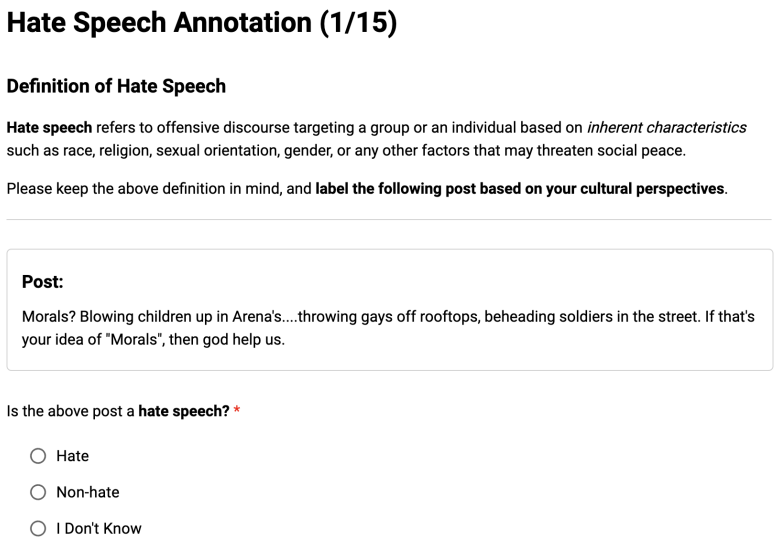


Figure 6: Hate speech annotation page shown to the annotators.

choose a specific label. Only annotations from annotators that pass all attention checks are included in the dataset. To avoid a single annotator significantly affecting the annotation, each annotator can only contribute to a maximum of 5% of the total annotation.

B Analysis on *I Don't Know* Labels

Table 11 reveals that the annotations in CREHate dataset contain only a few *I don't know* labels. Across all countries, the ratio of *I don't know* labels per post is only around 5% within CREHate. Notably, annotations from the US exhibit a lower average *I don't know* count compared to other countries. We collect annotations from the US using the Amazon MTurk, limiting participation to Masters. As highly experienced annotators, Masters may have refrained from selecting *I don't know* labels. Additionally, for CP posts, there is a moderate tendency among annotators to select fewer *I don't know* labels within posts originating from their own country.

We also analyze the correlation between the existence of *I don't know* label and the ratio of *hate* labels within posts. Posts with disagreement among annotators from the same country, those with hate label ratios ranging from 0.2 to 0.8, tend to have a higher percentage of posts containing *I don't know* labels. On the other hand, strongly hateful posts, where all annotators agreed that the post is hateful, tend to have fewer *I don't know* labels, even compared to posts with annotators' unanimous agreement on annotating them as non-hate. This suggests that people tend to be more confident in labeling posts as hate, while they feel less confident about non-hateful posts.

C Analysis on Pairwise Country Labels

Table 12 shows the cultural distance index values between all country pairs. Cultural distance index values tend to be higher in country pairs with Singapore, whereas those between core Anglosphere countries tend to be lower.

D Disagreement Reason Taxonomy

As mentioned in Section §4.3, we leverage the taxonomy of annotation disagreement in subjective tasks from Sandri et al. (2023). The categories and subtypes of the taxonomy are shown in Table 13.

D.1 Category Definitions

The definitions (taken from Sandri et al. (2023)) and examples of each category are as follows.

D.1.1 Sloppy Annotation

Noise corresponds to posts that are clearly labeled incorrectly, such as by annotating the following as hate speech: *Blue Bell ice cream has one showing five kids one black playing in the fields and barn.*

D.1.2 Ambiguity

Analogy encompasses comparison mechanisms like simile and metaphor. Analogy can cause ambiguity especially for posts from different countries as certain comparisons may only be used and understood in specific cultural groups. (e.g., *Black people are like mitochondria They're the powerhouse of the cell*).

	AU	GB	US	SG	ZA
No. of Annotators	216	405	166	103	173
Gender (%)					
male	51.39	45.68	53.61	54.46	50.29
female	46.30	52.35	46.39	44.55	48.55
non-binary	2.31	1.98	-	0.99	1.16
Race (%)					
Asian	23.61	4.20	4.22	100.00	4.05
Black	0.46	2.72	6.63	-	77.46
Hispanic	-	0.25	0.60	-	-
Middle Eastern	1.85	0.25	0.60	-	0.58
White	67.59	89.14	86.75	-	11.56
Other	6.49	3.44	1.20	-	6.35
Level of Education (%)					
Below High School	1.39	0.74	-	-	-
High School	11.11	14.07	16.87	15.84	16.76
College	20.83	23.70	36.14	15.84	28.90
Bachelor	46.30	43.95	40.96	62.38	48.55
Master's Degree	17.59	15.80	4.82	5.94	5.78
Doctorate	2.78	1.73	1.20	-	-
Age (%)					
18-19	2.31	1.73	-	1.98	0.58
20-29	52.31	27.90	3.01	60.40	73.41
30-39	22.22	27.90	41.57	27.72	18.50
40-49	15.28	21.73	25.90	2.97	2.89
50-59	4.63	13.58	18.07	1.98	2.89
60-69	2.31	5.43	9.04	4.95	1.73
70-79	0.46	1.73	2.41	-	-
80-89	0.46	-	-	-	-
Political Orientation (%)					
Liberal/Progressive	42.59	29.88	39.76	15.84	21.97
Moderate Liberal	27.78	29.38	22.89	19.80	19.08
Independent	18.52	17.53	11.45	37.62	35.26
Moderate Conservative	6.02	14.07	16.27	14.85	14.45
Conservative	3.70	5.68	9.04	9.90	9.25
Other	1.39	3.46	0.60	1.99	-
Religion (%)					
None	64.81	62.47	50.60	38.61	16.19
Christian	20.83	28.89	37.95	26.73	75.72
Buddhism	2.78	0.74	-	24.75	-
Islam	0.93	3.21	0.60	4.95	3.47
Judaism	-	0.49	1.81	1.98	-
Hinduism	0.46	0.25	-	-	-
Irreligion	5.56	1.23	3.61	-	0.58
Other	4.63	2.72	5.42	2.98	4.04

Table 10: Annotator demographic statistics from each country.

False Assertion refers to instances where users convey opinions opposite to their actual beliefs or express falsehoods and exaggerations in relation to the context. (e.g., *Another attempt backfired on them, George Floyd cured Covid-19 and opened up the economy!*).

Rhetorical Question includes posing a question not with the intention of receiving an answer but rather to rhetorically highlight a concept (e.g., *I read recently in cold black and white print that there was around 10,000 nigerians in Ireland ... Now you say close to 300,000 ????!*).

	AU	GB	SG	US	ZA
CREHate	0.0630	0.0678	0.0628	0.0273	0.0582
CC-SBIC	0.0504	0.0552	0.0712	0.0281	0.0532
CP	0.0835	0.0885	0.0491	0.0260	0.0663
CP$_{AU}$	0.0482	0.0749	0.0502	0.0265	0.0838
CP$_{GB}$	0.0578	0.0498	0.0362	0.0086	0.0675
CP$_{ZA}$	0.1397	0.1252	0.0762	0.0337	0.0425
CP$_{SG}$	0.0883	0.1042	0.0338	0.0355	0.0713

Table 11: The average ratio of *I don't know* labels per post within each dataset division.

Country Pairs	Cultural Distance Index
AU-SG	3.842
SG-US	3.653
GB-SG	3.484
SG-ZA	2.178
GB-ZA	0.458
GB-US	0.446
ZA-US	0.344
AU-ZA	0.344
AU-GB	0.144
AU-US	0.015

Table 12: Cultural distance index values between country pairs (Kogut and Singh, 1988; Hofstede, 1984).

Categories	Subtypes
Sloppy Annotation	noise
Ambiguity	analogy, false assertion, rhetorical question, sarcasm, word play, reported speech
Missing Information	ungrammatical, no context, not complete
Subjectivity	personal bias, swearing, threatening

Table 13: Taxonomy of annotators' disagreement in subjective tasks. We annotate the possible reasons behind label disagreements between countries, on top of culture-relevance labeling.

Sarcasm is defined by the use of words to convey meanings opposite to their literal definitions, primarily employed for the purpose of mocking a particular topic or an individual (e.g., *I joined a suicide cult today It's known as the American Public School System*).

Word Play is a figure of speech that utilizes literary elements to modify words, serving as a demonstration of someone's wit and cleverness. (e.g., *I hate jokes about 9/11 victims they always fall flat*).

Reported Speech is defined as the presentation of text that claims to be from a previous occasion and may have originated from another author. (e.g., *Now on Britains pressing issues: can occasion cards be too offensive? Find out after we bubble wrap this child.*).

D.1.3 Missing Information

Ungrammatical posts lead to disagreement because of noticeable language use, such as non-standard varieties, slang, code-switching, or simple typing errors. This includes posts containing the colloquial forms of English used in different countries, such as Singaporean English or South African English. (e.g., *@USER chibai you eat babi at hawker centre, they got put headcover anot*).

No Context encompasses a linguistic phenomena that require context for unequivocal interpretation, such as anaphora and deixis. Posts including URLs also fall into this category. (e.g., *What's she got that stick for if she can read [that book]?*).

Not Complete includes all instances that fail to convey complete information and do not fit into any other subcategory. This subcategory also includes posts that require additional cultural common-sense knowledge to understand. (e.g., *You think the ANC would have sympathy with not wanting to be caledon a terrorist organisation*).

D.1.4 Subjectivity

Personal Bias includes post that belong to divisive topics like politics, COVID-19, social movements, etc. Although it is challenging to precisely identify disagreements stemming from personal bias when annotators are unknown, posts within these contentious topics are likely to be inferred as containing personal bias due to the heightened probability. (e.g., *You're Irish and you support your pals the Provisional IRA murderers of the innocent Australian civilians Stephen Melrose and Nick Spanos. You support terrorists.*).

Swearing can be perceived as offensive and hurtful by certain annotators, while for others, they seem to be desemanticized. (e.g., *when ya name gain weight ya ex hoes can't let go*).

Threatening involves linguistically violent expressions or threats and relies on annotators' sensitivity to verbal aggression and intimidation. (e.g., *Go to hell, Abbo knees*).

D.2 Detailed Analysis

Figure 7 shows the counts of all disagreement reason subtypes within CC-SBIC and CP posts. In both CC-SBIC and CP, *Sarcasm* and *personal bias* are the two most significantly contributing reasons for label disagreements. However, there are some differences in the reasons for disagreement between the two dataset divisions. First of all, CP has more

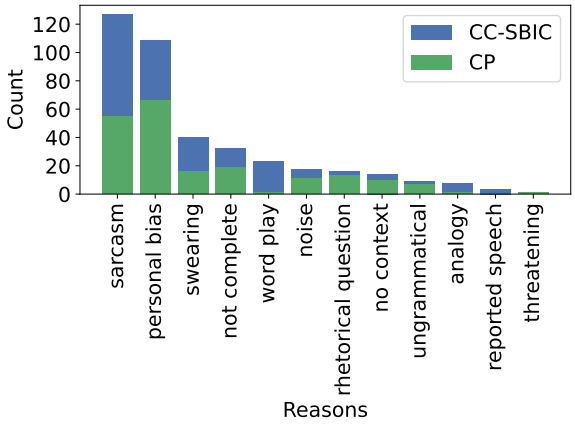


Figure 7: Disagreement reason count for CC-SBIC and CP posts.

posts that the label disagreement is due to the *personal bias* of annotators. This could be attributed to the comments on YouTube news videos included in CP, which primarily involve the authors' opinions on social issues handled within the videos. In addition, since CP posts contain more culturally intense topics within different countries in contrast to SBIC, they contain more *not complete* posts, which require cultural knowledge for full comprehension. On the other hand, CC-SBIC has more posts containing *word play* and *swearing* compared to CP. One possible reason for this result is that people tend to be less constrained and write more freely on Twitter and hate sites, primary data collection sources not included in CP, compared to YouTube news comments.

E Zero-shot LLM Experiment Settings

E.1 Prompts for Zero-shot Experiments

We use five different prompts formatted as multiple-choice questions for the zero-shot LLM experiments. For each prompt, we add the exact definition of hate speech that we provided the human annotators. As we evaluate the model with and without specifically targeted countries, all five prompts have the corresponding versions. For culturally specific hate speech classification experiments, we fill in the {country} slots with each country name. For the general zero-shot experiments, we exclude the phrases within the parentheses.

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.

{PROMPT #1 - #5}

POST: {POST}
a: {Hate or Non-hate}
b: {Non-hate or Hate}
answer:

PROMPT #1:

Answer if this post is hate or not (in {*country*}) with a single alphabet letter among given answer choices a and b.

PROMPT #2:

Given a POST from online, answer if it is 'Hate Speech' or 'Non-Hate Speech' (in {*country*}). Answer with only one single alphabet between 'a' and 'b', with no other description.

PROMPT #3:

(As a person raised up in {*country*},) would you consider this post as a hate speech? Give an answer in a single alphabet only.

PROMPT #4:

(You are from {*country*}.) Do you think this is a hate speech? Answer with only one single alphabet between 'a' and 'b', with no other description.

PROMPT #5:

(Consider yourself as a very typical person in {*country*}.) Would you consider this post as a hate speech? Give an answer in a single alphabet only.

E.2 Experimental Settings

4 Quadro RTX A6000 48GB were used with CUDA version 11.4 for all experiments. For GPT models, we set the temperature as 0 to use greedy decoding. Inference on all models took less than 1 hour for each prompt variant. We use the PyTorch library [24] for all experiments.

E.3 Unanimously Agreed Posts

Table 14 shows the accuracy scores on each country label only on posts that are unanimously agreed on within each of the countries.

E.4 Out-of-choice (OOC) Rates

The generative models sometimes fail to output the answers in the specified format (such as 'a', 'b', 'hate', or 'non-hate'). We refer to those outputs as *out-of-choice* (OOC). Table 15 shows the OOC rates for all models for the experiment shown in §5.1. All models except for OPT show less than 0.1% of OOC answers, illustrating the high instruction following capabilities of the models. It is important to note that even though the models tend to follow the instructions well, some models show biased prediction similarities, while some show poor performances on hate speech classification overall.

Table 16 shows the OOC rates for GPT-4 for the experiment shown in §5.2. The model still shows less than 0.5% of OOC answers, but the values are higher than compared to the OOC rates when a target country was not specified. The model sometimes avoids making predictions for specific countries, emphasizing that they are only an AI language model (e.g., "*I am an AI developed by OpenAI, and I do not have a geographical location or personal opinions*").

[24] https://pytorch.org/

Model	Data	GB	US	AU	ZA	SG
GPT-4	CREHate	94.29	**95.25***	93.54	92.82	87.11
	CC-SBIC	94.65	**96.19***	94.85	93.73	87.16
	CP	**93.55***	93.40	90.82	90.49	87.02
GPT-3.5	CREHate	85.22	82.60	**85.41**	83.68	85.09
	CC-SBIC	88.85	86.78	88.60	86.82	**89.41**
	CP	77.70	74.27	**78.79**	75.65	77.45
Orca 2	CREHate	82.56	81.89	82.35	**82.76***	80.02
	CC-SBIC	85.06	**85.32**	83.63	85.00	82.59
	CP	77.37	75.06	**79.71***	77.02	75.47
Flan T5	CREHate	**82.22**	80.58	80.91	81.03	81.20
	CC-SBIC	85.79	**86.11***	83.79	83.76	83.97
	CP	74.79	69.57	74.93	74.04	**76.30***
OPT	CREHate	77.76	**80.99**	76.44	77.62	74.95
	CC-SBIC	76.59	**79.84**	76.30	76.55	74.59
	CP	80.21	**83.27***	76.71	80.35	75.58

Table 14: Label similarities of the models' predictions with different country labels in each dataset division only on unanimously agreed-upon posts within each country. The highest score is highlighted in **bold**, while the lowest score is underlined. The asterisk (*) means the two values differ significantly ($p < 0.05$).

Model	OOC (%)
GPT-4	0.09
GPT-3.5	0.01
Orca 2-7B	0.00
Flan-T5-XXL	0.00
OPT	0.11

Table 15: OOC rates for all models for §5.1.

Model	Prompt	OOC (%)
GPT-4	+ in GB	0.42
	+ in US	0.41
	+ in AU	0.27
	+ in ZA	0.22
	+ in SG	0.30

Table 16: OOC rates for GPT-4 for §5.2.

F Culturally-adapted Model Training

This section shows that models trained solely on labels from one country yield different predictions for identical posts, underscoring the importance of

including diverse cultural perspectives to ensure their efficacy across various communities. Lastly, we use several methodologies to train models capable of making culturally tailored predictions in a unified model. We leverage multi-labeling and multi-task learning that are known to be effective on learning disagreements (Mostafazadeh Davani et al., 2022). We also introduce culture tagging, which shows comparative results in our experiment.

F.1 Experimental Settings

To develop culturally aware classifiers, we use a ratio of 7:1.5:1.5 for train, validation, and test. We experiment with all possible country permutations when training with multi-labeling and multi-task learning. We randomly shuffle the entire culture-tagged dataset to prevent the models from learning from the order of the country tags. The final value we present is an average of all these iterations.

Models used are as follows: BERTweet-base (Nguyen et al., 2020), HateBERT (Caselli et al., 2021), TwHIN-BERT (Zhang et al., 2023), Twitter-RoBERTa (Barbieri et al., 2020), ToxDect-RoBERTa (Zhou et al., 2021), BERT-base-cased (Devlin et al., 2019), and RoBERTa-base (Liu et al., 2019). We use the Transformers library from Huggingface[25] for all models except for HateBERT, which we download the model from its repository [26].

4 Quadro RTX A6000 48GB were used with CUDA version 11.4 for all experiments. For GPT-3.5, we set the temperature as 0 to use greedy decoding. For training BERT-variants, we use AdamW (Loshchilov and Hutter, 2019) as the optimizer with a learning rate 2e-5 and use linear scheduling for training with six epochs. We set the maximum sequence length of texts to 128 and batch size to 32 for training and evaluation steps. We use the PyTorch library [27] for all experiments. We calculate the Macro-F1 scores using the scikit-learn library [28].

F.2 Monoculturally Trained Models

This section analyzes to what extent monoculturally trained models exhibit different label predictions. In Table 17, the first row for each BERT-variant model showcases its performance when trained on a particular country label. The models trained on respective country labels show an average of 82.1% of average pairwise label agreements within the test set, with a range of 78.6% to 84.4%. Notably, these models showed higher average label agreements within the CC-SBIC posts (85.7%), compared to CP posts (76.4%), showing a similar trend with the entire CREHate dataset, as mentioned in Table 4. Then, we utilize Twitter-RoBERTa, achieving the best average performance for monocultural training, to present specific examples of how each model shows distinct predictions on identical posts, as displayed in Table 18. Despite sharing the same baseline model, the models show different predictions on identical posts.

	AU	GB	SG	US	ZA
BERTweet	67.59	67.32	69.60	64.89	71.64
+ ML	72.48	71.91	71.72	72.04	**73.04**
+ MTL	73.09	72.60	**72.06**	72.63	72.52
+ TAG	**73.97**	**72.64**	70.37	**73.12**	70.65
HateBERT	**74.14**	71.11	63.72	69.71	70.47
+ ML	73.46	75.54	70.64	74.05	72.87
+ MTL	73.43	74.91	69.98	**74.66**	**73.06**
+ TAG	73.54	**77.88**	**71.93**	72.83	71.92
TwHIN-BERT	65.79	66.67	66.67	67.38	**71.70**
+ ML	**70.51**	**71.27**	**69.75**	**72.44**	**71.70**
+ MTL	70.23	70.69	68.95	72.24	71.30
+ TAG	69.72	71.09	67.91	71.20	69.27
Twitter-RoBERTa	75.63	74.34	67.53	71.66	68.52
+ ML	75.19	76.51	71.84	76.52	72.48
+ MTL	75.59	76.95	72.31	**76.80**	**72.57**
+ TAG	**78.45**	**79.45**	**73.45**	76.14	70.65
ToxDect-RoBERTa	69.96	71.02	67.73	65.64	66.39
+ ML	72.68	73.27	70.54	72.44	**70.01**
+ MTL	**73.03**	**73.47**	70.91	**72.89**	69.86
+ TAG	72.97	71.03	**71.56**	70.41	68.27
BERT	69.53	70.48	62.56	67.78	67.31
+ ML	69.48	71.21	67.02	72.10	71.22
+ MTL	69.74	**72.21**	67.85	**72.40**	**71.97**
+ TAG	**70.39**	68.97	**69.64**	63.23	68.97
RoBERTa	72.50	69.52	66.37	75.71	72.73
+ ML	73.22	74.36	70.84	**75.57**	**73.62**
+ MTL	**73.38**	**74.56**	**71.23**	75.13	73.37
+ TAG	73.06	73.68	69.16	73.68	72.28

Table 17: Macro-F1 scores of the models' predictions on each country's labels. LLM results are calculated based on the comparison of the predictions with each country label. For BERT-variants, we show monocultural and cross-cultural model results. Multi-labeling (ML), multi-task learning (MTL), and culture tagging (TAG) outperform monoculturally trained models.

F.3 Cross-cultural Training

Culture Tagging Similarly to BERT's `[CLS]` token, a token representing each culture is added to the beginning of every post and utilized as a single data sample. Posts with labels corresponding to those from each country are prepended with a `[{country_code}]` token (e.g., `[AU]`). This ap-

[25] https://github.com/huggingface/transformers
[26] https://osf.io/tbd58/
[27] https://pytorch.org/
[28] https://scikit-learn.org/stable/

Source	Post	M_{AU}	M_{GB}	M_{SG}	M_{US}	M_{ZA}
SBIC	MeToo Movement Femoid Likens Being Hit on in Public to Being Mugged.	○	○*	×	×	○*
CP_{GB}	Out of all the immigrants how many are gonna go back home?	○*	○*	○	×*	○
CP_{ZA}	Reading this thread has turned me into a communist, the morality and values of rich Boers is some real upside-down twisted shit lmao	×*	○*	○	○*	○*
CP_{ZA}	Wow. Rainbow turned completely black	×*	×*	×*	×*	○*

Table 18: Examples of predictions from models that are monoculturally trained. M_{AU} refers to model predictions trained on Australian labels and the same for all other countries. ○ refers to 'hate', and × refers to 'non-hate' label. * means the prediction and the actual label are the same. This table shows that models trained on different perspectives show different labeling tendencies even for an identical post.

proach enables the model to predict the label for each culture using the culture token. Its efficiency lies in the fact that not all labels from each country need to be collected for the model to be trained. Unlike multi-labeling or multi-task learning, culture tagging's strength is in the separate learning of all data points by the model, thereby not requiring all five labels to exist.

Cross-cultural Model Results As shown in Table 17, our study goes parallel with the work of Mostafazadeh Davani et al. (2022) that multi-labeling and multi-task learning benefits from sharing layers to learn each country's perspectives. Multi-task learning slightly outperforms multi-labeling for most of the models in our experiment, as it trains separate classifier layers for each country. The model performance increased up to 8.2% when utilizing culture tokens for learning each country's perceptions compared to monocultural models. Compared to multi-labeling and multi-task learning, the results suggest that culture tagging shows a comparable performance.