Hate Speech Classifiers are Culturally Insensitive

Nayeon Lee, Chani Jung, Alice Oh
School of Computing, KAIST
{nlee0212, 1016chani}@kaist.ac.kr
alice.oh@kaist.edu

Abstract

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

Increasingly, language models and machine translation are becoming valuable tools to help people communicate with others from diverse cultural backgrounds. However, current language models lack cultural awareness because they are trained on data representing only the culture within the dataset. This presents a problem in the context of hate speech classification, where cultural awareness is especially critical. This study aims to quantify the cultural insensitivity of three monolingual (Korean, English, Arabic) hate speech classifiers by evaluating their performance on translated datasets from the other two languages. Our research has revealed that hate speech classifiers evaluated on datasets from other cultures yield significantly lower F1 scores, up to almost 50%. In addition, they produce considerably higher false negative rates, with a magnitude up to five times greater, demonstrating the extent of the cultural gap. The study highlights the severity of cultural insensitivity of language models in hate speech classification.

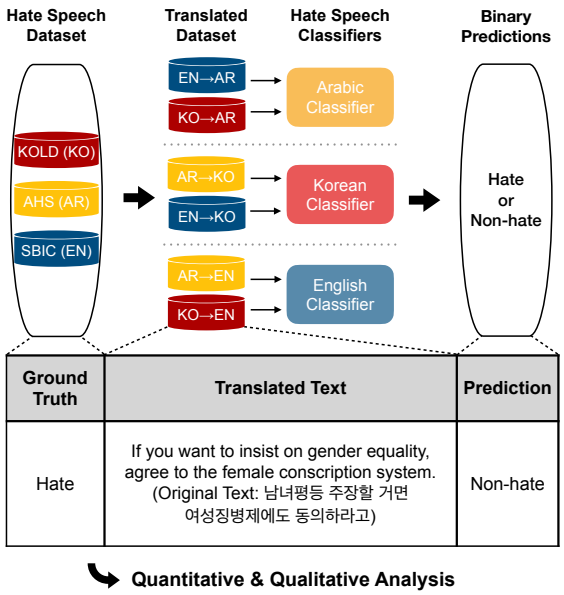

Figure 1: Overview of our cross-cultural evaluation for hate speech classifiers. We translate each of the monolingual datasets (Korean(KO): KOLD, English(EN): SBIC, Arabic(AR): AHS) and evaluate by comparing the ground truth label and the predicted labels of the translated texts and analyzing samples.

1 Introduction

The current NLP models are trained on culturally biased datasets, so they lack sociocultural diversity (Dodge et al., 2021; Callahan and Herring, 2011). There is recent research emphasizing the importance of developing models that are more generalized to other languages and cultures (Hershcovich et al., 2022; Yin and Zubiaga, 2021; Jo and Gebru, 2020).

Hate speech detection poses an extra challenge because it is crucial to consider the impact of inherent social and cultural differences for this task (Ousidhoum, 2021). However, current approaches tend to overlook cultural differences, underscoring the need for more nuanced and culturally sensitive approaches to develop models that can address the challenges posed by diverse languages and cultures. With communication across cultural and linguistic barriers becoming increasingly common in the online landscape, an effective cross-cultural hate speech classifier is necessary. This classifier should identify hate speech that incorporates diverse cultural nuances and variations, regardless of the language. However, to the best of our knowledge, no research has yet addressed this critical necessity.

This study aims to evaluate cross-cultural hate speech classifiers. We investigate cultural disparities in hate speech detection, explicitly focusing on the cultures of Korean, Arabic, and English-speaking countries. To achieve this goal, we develop hate speech classifiers for each language and evaluate their performance on translated datasets from other cultures. The experiment overview can

Proceedings of the First Workshop on Cross-Cultural Considerations in NLP (C3NLP), pages 35–46

be seen in Figure 1. We also perform sample-level analysis within the misclassified texts, providing insights into the reasons for poor classification performance on the datasets from different cultures. Through our analysis, we identify the limitations of current methodologies that fail to address the complexity of cross-cultural communication and perpetuate cultural divides. Our experiment revealed that the F1 scores of hate speech classifiers evaluated on datasets from other cultures decremented by 26% to 48%, and the false negative rate (FNR) increased about two to five times larger. This result shows that models trained in a single language are deficient in detecting hate speeches from other cultures. Deeper examinations of false negative samples showed that the limited performance was likely due to the differences in target groups, sociocultural backgrounds, and even the standards of hate speech.

2 Related Work

Recent research has focused on developing multilingual hate speech detection datasets and models. Several approaches have been proposed to address the scarcity of datasets in different languages, such as building multilingual hate speech corpora (Glavaš et al., 2020; Huang et al., 2020; Ousidhoum et al., 2019) and implementing cross-lingual methods that incorporate translated data or multilingual embeddings (Yin and Zubiaga, 2021; Aluru et al., 2021; Pamungkas et al., 2020; Pamungkas and Patti, 2019; Arango et al., 2019; Sohn and Lee, 2019). Additionally, transfer learning on multilingual models like XLM-R has been utilized to take advantage of large English datasets and cross-lingual contextual word embeddings (Ranasinghe and Zampieri, 2021; Ranasinghe and Zampieri, 2020). However, most of these approaches did not consider the cultural differences among datasets. They did not examine the model's cross-cultural detection ability, where the model could detect hate speech from other cultures.

Challenges in building a hate speech classifier in multilingual or multicultural settings include variations in targets of hate speech among countries and cultures (Ousidhoum, 2021; Billé, 2013), and the need to consider cultural discrepancies and diverse backgrounds. Current studies have not fully addressed these issues, as some have used translated texts and maintained ground truth labeling without considering cultural differences (Glavaš et al., 2020; Pamungkas et al., 2020; Pamungkas and Patti, 2019). Another consideration is that word senses may differ based on dialect, sociolect, language, and culture (Rahman, 2012; Boyle, 2001; Massey, 1992). Therefore, incorporating cultural diversity is crucial in handling linguistically varied and cross-cultural hate speech.

Researchers have proposed various methods for adapting hate speech detection models to different cultural contexts (Sarwar and Murdock, 2022; Chandrasekharan et al., 2017; Nobata et al., 2016), but there is still limited research on cross-cultural hate speech detection. Some methods include using multi-task learning on hate speech datasets from different cultures (Talat et al., 2018) and building new datasets that contain different targets of hate (Arango et al., 2022). While Arango et al. (2022) has evaluated knowledge transfer performance across different datasets from different cultural backgrounds in the same language, it lacked a deeper analysis of the cultural differences behind poor performance. In contrast, this paper includes a thorough analysis of sociocultural backgrounds and differences between hate speech datasets from different cultures and explores the reasons behind the poor performance in various language settings.

3 Datasets from Different Cultures

This study evaluates the cross-cultural performance of hate speech classifiers trained on Korean, Arabic, and English datasets. We translate the datasets to compare the cross-cultural performance of the classifiers in different cultural settings. The datasets represent each culture, allowing for a more nuanced analysis of the performance of hate speech classifiers. We use the training and validation sets of these datasets for training and test sets for evaluations, including the cross-cultural experiment. Since the Korean dataset does not have training, validation, and test sets separated, we divide the entire dataset by the ratio of 8:1:1.

3.1 Korean, English, Arabic Datasets

Korean Dataset: KOLD For the Korean hate speech dataset, we select KOLD (Jeong et al., 2022) as it is large-sized, is collected from sources well reflecting Korean sociological background, and contains carefully curated annotations that provide detailed information on the types of hate speech present in the dataset. The dataset includes a wide range of hate speech types, making it a compre-

hensive resource for studying hate speech in the Korean language.

English Dataset: SBIC For English, we choose SBIC (Sap et al., 2020) since it is extensively collected from diverse online community sites that many English speakers use and includes specific target groups in deep-down hierarchies. It contains diverse target groups that reasonably reflect the sociocultural backgrounds of English-speaking countries.

Arabic Dataset: Arabic Hate Speech (AHS) For the Arabic dataset, we select the Arabic Hate Speech (AHS) dataset from Mubarak et al. (2022), a large-size dataset compared to other Arabic datasets, with offensiveness and hate annotations that lack bias toward specific topics, genres, or dialects. The dataset includes target demographic groups that are specific to Arabic-speaking countries.

3.2 Preprocessing

To ensure the quality of translation and fair evaluation of classifiers on datasets from different cultures, we preprocess the texts of all three datasets to match the form of each other.

Special Token Removal Occasionally, Google Cloud Translation API[1] fails to translate correctly when special tokens such as '@user' are included in the text. An example of a translation error is as below:

- **Original sentence (Arabic):** @user @user @user يقعدن الحريم ان سالفه ع تكذبوها انكم واضح بالارض ولا انا فهمت غلط ودرعمت؟
- **Translated sentence (English):** Replying to @user
- **Translated sentence after removing *@user* (English):** It is clear that you deny it according to its predecessor, that the harems are sitting on the ground, or did I misunderstand and defend?
- **Human-translated sentence:** You lied to your predecessors, that the harems are sitting on the ground, I don't understand, or do I?

[1] https://cloud.google.com/translate

Target Language	Similarity	KOLD	SBIC	AHS
Korean	≥ 0.9	-	57.9	28.8
	≥ 0.8	-	83.9	71.6
	≥ 0.7	-	93.7	87.7
English	≥ 0.9	61.5	-	47.6
	≥ 0.8	85.7	-	83.5
	≥ 0.7	93.0	-	93.6
Arabic	≥ 0.9	55.8	61.4	-
	≥ 0.8	83.5	83.5	-
	≥ 0.7	92.0	92.8	-

Table 1: The percentage of texts from the test dataset according to the cosine similarity score spans of back-translated texts from KOLD, SBIC, and AHS.

	Original		Filtered	
	Size (%)	Hate %	Size (%)	Hate %
KOLD	4045 (100)	31.1	3671 (90.8)	31.8
SBIC	4691 (100)	41.1	4208 (89.7)	42.4
AHS	2451 (100)	10.7	2226 (87.6)	9.6

Table 2: Size and percentage of hate of the original and filtered KOLD, SBIC, and AHS test datasets where each only retained those with cosine similarity scores above 0.7 in both translated languages.

Therefore special tokens are all removed before the translation step and the experiment. The specific preprocessing strategies for each dataset are explained in Appendix A.

3.3 Translation of Test Datasets

The Advanced version of Google Cloud Translation API is utilized for translating the test sets. To ensure the quality of the translation, we use the RTT-SBERT metric proposed in the findings of Moon et al. (2020), demonstrating the cosine similarity of SBERT embeddings (Reimers and Gurevych, 2019) between the input and round-trip translation has a high correlation with human evaluation. In other words, sentences with high cosine similarity scores tend to achieve high scores in the human evaluation. The detailed translation steps are as follows.

3.3.1 Back Translation

After translating each test dataset into two other languages, we translate it back to the original language. For example, for a Korean dataset, we translate it into English and Arabic and translate the English

KOLD		SBIC		AHS	
Target Group Category	**Count (%)**	**Target Group Category**	**Count (%)**	**Target Group Category**	**Count (%)**
Gender	286 (23.9)	Gender	434 (20.3)	Gender	86 (40.2)
Race	290 (24.3)	Race	767 (35.8)	Race/Ethnicity/Nationality	72 (33.6)
Politics	187 (15.6)	Social	95 (4.5)	Ideology	29 (13.6)
Religion	186 (15.6)	Culture	483 (22.5)	Religion/Belief	6 (2.8)
Others	246 (20.6)	Disabled	102 (4.8)	Disability/Disease	2 (0.9)
		Body	50 (2.3)	Social Class	19 (8.9)
		Victim	211 (9.8)		
Total	1179	**Total**	1785	**Total**	214

Table 3: Statistics of each target group category within the entire hate speech in the filtered KOLD, SBIC, and AHS. For KOLD and SBIC, multi-targeted group categories are split into single categories when counting.

and the Arabic version back to Korean.

3.3.2 Cosine Similarity Scores

We utilize SentenceTransformers Python framework[2] for extracting the SBERT embeddings of the texts. Table 1 shows the portion of the test dataset that achieves cosine similarity scores above 0.7 for each of the three datasets and languages.

3.3.3 Filtering

To ensure a fair cross-cultural comparison, we apply a filtering process to the original test sets of each language. Specifically, we only retain texts with RTT-SBERT scores exceeding 0.7 in both translated languages. This approach helps minimize discrepancies in the quality of the translations and ensures that the selected texts are accurately represented in all languages. The data size and the portion of hate of both original and filtered datasets are shown in Table 2, and the target group category distribution for each can be seen in Table 3. The filtered datasets retained over 87% of the original dataset, indicating that the size reduction is unlikely to affect the experiment's results significantly.

3.3.4 Evaluation of Filtered Datasets

We evaluate the actual translation quality of the filtered test datasets with RTT-SBERT scores above 0.7 by manually inspecting the sample texts. We check if the translated text conveys the meaning of the original sentence without leaving out or mistranslating some phrases. As a result, about 70% of the samples properly convey the meaning of the original sentence after translation. Since this portion is acceptable, we maintain the threshold at 0.7.

[2] https://www.sbert.net/

4 Culture Representative Model Training

To ensure that the hate speech classifiers accurately represent the cultures of their respective languages, they must achieve high performance on datasets from their language. To address this, we use monolingual models pretrained in each of the three languages and finetune them. The following sections contain descriptions of each model and the results of finetuning. Specific training details are in Appendix B.1. We use the best model for each language for cross-cultural evaluation in Section 5, and Table 7 shows the performance of all models.

4.1 Model Description and Performance

Korean Pretrained Models For Korean models, we utilize KcELECTRA-base and KcELECTRA-base-v2022 (Lee, 2021) trained on NAVER[3] news comments and nested comments. We also finetune models pretrained on KLUE (Park et al., 2021), the most extensive Korean benchmark dataset, including KLUE-RoBERTa-base, KLUE-RoBERTa-large, and KLUE-BERT-base. KcELECTRA-base-v2022 outperforms all the other Korean pretrained models with an F1 score of 0.81 and is used as the model for cross-cultural hate speech evaluation in Korean.

English Pretrained Models For the English model, we use BERTweet (Nguyen et al., 2020), trained on an 80GB dataset containing 850M Tweets, and Twitter-RoBERTa (Barbieri et al., 2020), trained on the TweetEval benchmark dataset. We also finetune BERT-base, RoBERTa-base, and DistilBERT-base, pretrained on general English data. BERTweet-base exceeds all other English

[3] One of the top three mobile apps used in Korea in 2021. (http://www.koreaherald.com/view.php?ud=20210901001000)

Dataset	Language	F1	FPR	FNR
KOLD	*KO*	0.81	0.08	0.32
	KO → *EN*	0.59	0.04	0.76
	KO → *AR*	0.49	0.02	0.91
SBIC	*EN*	0.87	0.09	0.18
	EN → *KO*	0.56	0.05	0.77
	EN → *AR*	0.45	0.02	0.91
AHS	*AR*	0.81	0.03	0.39
	AR → *KO*	0.56	0.01	0.90
	AR → *EN*	0.60	0.02	0.83

Table 4: Results of cross-cultural evaluation on KOLD, SBIC, and AHS. *KO* (Korean), *EN* (English), *AR* (Arabic) shows prediction results of models on the test dataset from the original dataset for comparison. The KcELECTRA-based classifier was used for classifying test datasets in Korean, the BERTweet-based classifier for datasets in English, and the AraBERT-based classifier for datasets in Arabic.

pretrained models on the English hate speech corpus by achieving an F1 score of 0.86 and is served for cross-cultural hate speech evaluation in English.

Arabic Pretrained Models We use variants of pretrained AraBERT (Antoun et al., 2020). AraBERTv2-base/large are trained on general Arabic datasets, and AraBERTv0.2-Twitter-base/large are trained by continuing the pretraining on 60M Arabic tweets. Among these models, AraBERTv0.2-Twitter-base performs the best with an F1 score of 0.82 when finetuned for Arabic hate speech classification and is used for cross-cultural evaluation of hate speech in Arabic.

5 Cross-Cultural Evaluation

The current study aimed to evaluate the cross-cultural performance of different hate speech classifiers and explore the factors responsible for their poor performances. Table 4 presents the performance of the models on datasets across cultures. It is noteworthy that the cross-cultural performance of the models showed a substantial decrease in overall F1 scores ranging from 0.4 to 0.6 when compared to the models' performance on the original test datasets with F1 scores over 0.8. We experimented to investigate the potential relationship between translation quality and F1 scores, but our findings revealed no discernible correlation between them.

Another common tendency was decreased false positive rate (FPR). This could be due to the lack of understanding of other cultures leading the models to follow the majority label of the training dataset and to predict some instances as non-hate incorrectly. Another possible reason is that hate speech classifiers tend to have identity term bias (Dixon et al., 2018), but they may not have the bias for unknown targets of hate from different cultures.

Our area of interest was the increase in false negative rate (FNR) of the cross-cultural evaluation results, up to five times higher than that of the original dataset. The findings revealed that the poor performance of the models is not only due to differences in the target of hate but also due to variations in the standard across cultures. Table 5 displays false negative examples of cross-cultural evaluation, demonstrating the original text, labeled as hate speech in the original dataset, and translated text predicted as non-hate. Moreover, we evaluated the FNR for each target group category and specific target group for the models trained on different cultures, shown in Table 6. We use the terms **target group category** and **target group** throughout this section, where **target group category** represents the broader category of hate, such as *race* and *gender*, and **target group** refers to a specific type of target group, such as *Asians* and *females*. Note that the target group category named *social class* in AHS includes diverse *social groups* existing in Arabic cultures.

KOLD In the study conducted with translated KOLD, the AraBERT-based classifier had the highest FNR of 0.98 for the *gender* category, while the BERTweet-based classifier had the highest FNR of 0.85 for the *politics* category. The Korean classifier also faced challenges in detecting hate speech for these categories compared to the others, with an FNR of 0.42 for *gender* and the third-highest FNR (0.28) for *politics* category. Sociological complexity of hate speech related to *politics* in Korea, such as mentions of specific names of politicians or political organizations, can explain the high FNRs in both BERTweet-based and AraBERT-based classifiers. In contrast, *gender*-related hate speech tends to target more general demographic groups, which highlights the need to investigate the factors behind the poor performance of the Arabic model in *gender* category.

Sample analysis revealed that the AraBERT-based classifier could not detect any hate speech targeting the *LGBTQ+* community, likely because such tweets were not included in the Arabic training dataset. To our knowledge, there is currently very few available Arabic hate speech dataset that

Language	Original Text	Translated Text
KO → EN	남녀평등 주장할 거면 여성징병제에도 동의하라고ㅋㅋㅋ 그리고 내 말에 그냥 시비만 걸지 말고 혜택은 다 쳐받으면서 왜 차별받는다고 말하는지 말해보라고ㅋㅋㅋ	If you are going to insist on equality between men and women, agree to the female conscription system hahaha And don't just argue with me, tell us why you say you are being discriminated against while receiving all the benefits hahahaha
KO → AR		إذا كنت ستصر على المساواة بين الرجل والمرأة، فوافق على نظام التجنيد للإناث هاهاها ولا تجادلني فقط ، أخبرنا لماذا تقول إنك تتعرض للتمييز أثناء تلقي جميع المز ايا هههههه
EN → KO	So, the Holocaust confuses me... Like, I don't understand why anyone would want to survive it, because they would still be a jew afterwards...	그래서, 홀로코스트가 나를 혼란스럽게 한다... 예를 들어, 왜 누군가가 살아남고 싶어하는지 이해할 수 없다. 왜냐하면 그들은 여전히 유대인일 것이기 때문이다...
EN → AR		لذا ، فإن الهولوكوست يربكني ... مثل ، لا أفهم لماذا يريد أي شخص البقاء على قيد الحياة ، لأنهم سيظلون يهودًا بعد ذلك ...
AR → KO	على زق انتم وإيران المجوسية والمملكه دونها رجال تحمي ارضيها ومقدساتها وجربو حظكم مع سلمان كما جربتوه في اليمن .	당신과 이란, 마기안, 그리고 그것이 없는 왕국은 그 땅과 신성함을 보호하고 예멘에서 시도한 것처럼 살만과 함께 당신의 행운을 시험하는 사람들입니다.
AR → EN		You and Iran, the Magians, and the kingdom without it are men who protect its lands and sanctities, and try your luck with Salman as you tried it in Yemen.

Table 5: Original and translated texts of false negative samples, in which the ground truth is **hate** but the predictions on translated texts are **non-hate**. All of the samples achieved an RTT-SBERT score above 0.9.

includes hate speech explicitly targeting this demographic group. Hence, we express our readiness to replicate the same experiment in the future, provided that a dataset containing plenty of hate speech directed towards *LGBTQ+* in Arabic is available.

In addition, the FNR of the AraBERT-based model for other *gender*-related groups, mainly *females* and *males*, was 0.95 or higher, whereas that from the BERTweet-based model was about 0.77, and that from the KcELECTRA-based model was about 0.41 and 0.23 respectively. *gender* category comprises a significant proportion of hate speech in the Arabic and English training datasets, accounting for 48% and 29% of AHS and SBIC, respectively. Thus, the marked disparity in performance between the two models implies that the standards of hate speech towards *male* and *female* vary between Arabic and English-speaking cultures, in addition to cultural differences in *gender*-targeted hate speech.

The *race* category was a significant challenge for the English hate speech model, with the second-highest FNR among all categories. This was particularly evident for target groups such as *Chinese*, *Korean Chinese*, and *others*, including smaller groups such as *Afghans*, with FNRs exceeding 0.85. Interestingly, although these groups were the main targets of hate speech in KOLD, they were minor targets in the English hate speech corpus. The Korean classifier also had the highest FNR (0.37) for the *others* group within the *race* category, indicating that the classifier may not have been adequately trained to detect all hate speech targeting them. Nevertheless, the Korean and English hate speech classifiers showed varying performances for those target groups, with the KcELECTRA-based classifier achieving FNRs of 0.18 and 0.34 for *Chinese* and *Korean Chinese*, respectively. Notably, the FNR of the English classifier for the *black* group was 0.32, similar to that of the KcELECTRA-based classifier (0.27). This may be attributed to the BERTweet-based classifier having sufficient opportunities to learn to detect hate speech towards *black* people from SBIC, where the primary target group within the *race* category was *black*. These findings highlight the impact of target demographic differences in cross-cultural hate speech detection, indicating that classifiers must be trained on diverse and inclusive datasets to ensure their effectiveness across different cultures and languages.

SBIC Both the AraBERT-based and KcELECTRA-based classifiers exhibited the highest FNRs for *disabled* and *victim* target group categories on the translated SBIC dataset. The Arabic classifier achieved FNRs of 0.98 and 0.96, and the Korean classifier gained 0.90 and 0.88, respectively. Conversely, the BERTweet-based classifier had the highest FNRs for the *social* and *body* target groups. The difference in the FNR rankings can be attributed to the fact that

KOLD				SBIC				AHS			
Target Group Category	***KO***	***KO → EN***	***KO → AR***	**Target Group Category**	***EN***	***EN → KO***	***EN → AR***	**Target Group Category**	***AR***	***AR → KO***	***AR → EN***
Gender	**0.42**	0.78	**0.98**	Gender	0.26	0.70	0.89	Gender	0.41	0.87	0.81
Race	*0.32*	*0.82*	0.88	Race	0.09	0.72	0.92	Race/Ethnicity/Nationality	0.38	0.93	0.82
Politics	0.28	**0.85**	*0.92*	Social	**0.42**	0.69	0.83	Ideology	0.38	0.93	*0.86*
Religion	0.25	0.64	0.91	Culture	0.10	0.82	0.86	Religion/Belief	0.17	0.50	0.50
Others	0.27	0.69	0.86	Disabled	0.23	**0.90**	**0.98**	Disability/Disease	**0.50**	**1.00**	**1.00**
				Body	*0.40*	0.66	0.88	Social Class	*0.47*	*0.95*	**1.00**
				Victim	0.21	*0.88*	*0.96*				

Table 6: False Negative Rate (FNR) of original and translated versions of KOLD, SBIC, and AHS on KcELECTRA-based (Korean (*KO*)), BERTweet-based (English (*EN*)), and AraBERT-based classifiers (Arabic (*AR*)). **Bold** indicates the target group category with the highest FNR, *italic* indicates second-highest, underlined refers to the third highest.

hate speech directed towards *disabled* and *victim* categories, which includes target groups such as *mass shooting victims*, is not prevalent in Arabic and Korean datasets. However, there was a variation in the FNR rankings for specific target groups between the Korean and Arabic models.

For the target group category of *disabled* people, both the AraBERT-based and the KcELECTRA-based classifier had high FNRs (above 0.94) for hate speech targeting *physically disabled* people. For the *mentally disabled* target group, the Arabic classifier displayed a higher FNR (0.98) compared to that of the Korean classifier (0.84). The reason behind their poor performances might have been partially due to the English data's tendency to include posts that mention specific disabilities such as *quadriplegic* or *autistic* patients, or sarcastic metaphors regarding *disabled* people. A rare appearance of these terms in the Arabic and Korean datasets may have led the models to fail to detect them. As the English hate speech classifier was trained on this kind of data, it demonstrated an FNR of 0.25 for *physically disabled* people and 0.12 for *mentally disabled* people. In contrast, this kind of hate speech was rare in the Arabic and Korean datasets, making it difficult for the models to identify.

The detection of hate speech targeting *victim* category also remains a challenge for both AraBERT and KcELECTRA-based classifiers, as indicated by their high FNRs. However, the BERTweet-based classifier had a low FNR (0.21) for the same category. Specifically, hate speech targeting *mass shooting victims* posed difficulty for Arabic and Korean classifiers, with FNRs above 0.95, whereas the English classifier's FNR was only 0.23. Our analysis revealed that *mass shooting events* are more frequent in the United States than in Korean cultures. Also, even though there are *mass shooting events* in Arabic countries, the AHS dataset did not include hate speech targeting *mass shooting victims*. On the other hand, hate speech targeting *terrorism victims* was more challenging for the Korean classifier, with an FNR of 0.97, than the AraBERT-based classifier, with an FNR of 0.90. This was also very different from the English classifier's performance, which showed an FNR of 0.14 for the same group. The prevalence of *terrorism*-related hate speech targeting specific events, such as *9/11 attack*, in America may have accounted for this discrepancy. Additionally, the Arabic classifier had a high FNR (0.98) for the hate speech targeting *assault victims*, whereas the Korean classifier had a relatively low FNR (0.83) for the same group. Through further analysis, we found out that about 80% of the hate speech towards *assault victim* group were about *sexual assaults*. Considering that the FNR of the Arabic classifier on the *gender* category was high (0.89) compared to those of the Korean (0.70) and English classifiers (0.30), the model's tendency towards *gender*-related texts may have affected its performance on the hate speech against *assault victim* group.

Especially for the *gender* category, the AraBERT-based classifier's FNRs for the *trans women*, *gay men*, and *women* groups were greater than or equal to 0.89. In contrast, those of the KcELECTRA-based classifier were below 0.74. The BERTweet-based classifier also had low FNRs of under 0.27 for those groups. The lack of *LGBTQ+*-related hate speech in the AHS dataset, previously mentioned in the analysis regarding the KOLD dataset, could explain the high FNR of the classifier for *trans women* and *gay men*. However, for *women*, as they constitute a more general target group, one of the possible interpretations of the FNR disparity could be the difference in the standard of hate speech between Arabic and Korean-speaking cultures.

The other target groups that the KcELECTRA-based classifier had a high FNR for were *Native American*, *Latino*, and *Jewish* people, which are

not common target groups in Korean society. However, *Christians* were one of the main target groups related to *religion* but still had a high FNR in the Korean classifier. After analyzing hate speech in KOLD and SBIC targeting *Christians*, it was found that those in KOLD tended to include criticism and denouncements of *Christian people*. In contrast, those in SBIC were mainly sarcastic humiliations of Christianity. In contrast, the Arabic hate speech classifier had difficulty detecting hate speech targeting *Christians*, *trans-women*, *Asians*, *Black people*, and *Latinos* due to the lack of hate speech targeting these groups in the Arabic hate speech dataset.

What was common within this experiment was that the classifiers trained in other cultures had difficulty identifying hate speech in English comments due to the language's high use of sarcasm and metaphors that some even embedded societal or cultural background, such as common *mass shootings* in American schools. These nuances were not adequately captured through translations alone, resulting in challenges for the models to understand the context.

AHS The size of the test dataset of AHS was comparatively small, with less than ten examples for the *Religion/Belief* and *Disability/Disease* categories. Therefore, we did not analyze the two categories. The FNR rankings of the BERTweet-based and KcELECTRA-based classifiers were identical for the other categories. However, the AHS dataset only included annotations for target group categories but not their detailed target groups, so the analysis was limited to that scope.

The study revealed that hate speech targeting specific *social class*, such as *Bedouins* (a group of Arabic-speaking nomadic people living primarily in the Middle East and North Africa), posed significant challenges for both the BERTweet-based and KcELECTRA-based classifiers, which were trained on Korean and English datasets, respectively. The classifiers had an FNR of 1.0 and 0.95 for these target groups, respectively. Further analysis of the false negative samples revealed that understanding the context of the target groups required sociological background knowledge of Arabic cultures. In addition, the specific terms were rare or even unknown to the Korean and English models. The content required background knowledge to understand whether the text was hate speech, resulting in incorrect predictions. This characteristic of the category also led to the highest FNR of 0.47 within the Arabic classifier.

Hate speech aimed at particular *ideologies*, such as *partisan*, *intellectual*, or *sports affiliations*, had a high false negative rate (FNR) for both the English and Korean hate speech classifiers. The *ideology* category had an FNR of 0.86 and 0.93 for the English and Korean classifiers, respectively. The difficulty arose due to the culture-dependent nature of these tweets, which included specific names of *football clubs*, *politicians*, and other *ideological terms* that were challenging for classifiers trained on data from different cultures to be aware of. However, the Arabic classifier had a relatively low FNR, achieving a value of 0.38, as it was trained on this type of data.

6 Conclusion

In this paper, we investigated the cross-cultural performance of monolingual hate speech classifiers for Korean, English, and Arabic languages by evaluating the classifiers' performance on translations of hate speech datasets from other languages. Our deep analysis of model performance and false negative samples revealed the limitations of classifiers trained in a single language, including their inability to understand the sociocultural background of other cultures. This lack of understanding resulted in many samples being predicted as non-hate speech, highlighting the need for cross-cultural evaluation of hate speech classifiers. Our research also demonstrated standard differences in hate of general target groups across cultures.

Our findings underscore the importance of cross-cultural evaluation of hate speech classifiers and sample-level analysis to identify their weaknesses in a cross-cultural context. Adopting this approach will enable models to accurately detect hate speech from diverse cultures in global online communities. As such, our research highlights the need for more culturally sensitive approaches to developing hate speech classifiers to address the challenges posed by linguistic and cultural diversity in online spaces.

7 Ethical Considerations

To accurately represent their respective cultures, this paper utilized three publicly available hate speech datasets in Korean, English, and Arabic, with detailed descriptions provided in Section 3.

Regarding user privacy, the Korean dataset KOLD and the Arabic Hate Speech dataset (AHS)

implemented measures to protect user privacy by masking usernames and URLs with their masking tokens. However, the English dataset SBIC did not anonymize texts containing usernames and URLs. To protect user privacy, we anonymized the texts by removing these two attributes.

We relied on multiple resources to comprehend comments from various cultures to avoid any bias resulting from a limited understanding of different cultures. This approach helped ensure that our lack of cultural knowledge did not affect the analysis of cultural differences. Our analysis primarily relied on numerical values from model predictions, and we inspected samples to provide better explanations for the models' performance based on the quantitative results. This approach allowed us to minimize potential biases resulting from cultural misunderstandings and contribute to more culturally sensitive research practices.

8 Limitations

Machine Translation Using machine translation may impact hate speech classifiers' performance on translated data due to challenges in translation quality. To address this, we employed the RTT-SBERT metric from Moon et al. (2020), which correlates well with human evaluation scores, to only leverage the well-translated sentences. However, the classifiers' performance may have been affected because translated texts with high RTT-SBERT scores did not always convey the correct context. Future work should consider carefully performed manual translations by translators with a deep understanding of both languages for more accurate evaluation.

Transfer Learning for Cross-Cultural Hate Speech Classification Our study evaluated a model's cross-cultural ability by testing it on unseen data from different cultures. However, recent research suggests that transfer learning can adapt classifiers to different domains, potentially addressing some limitations of our approach. Future work will explore the effectiveness of transfer learning methods in improving hate speech classifiers' ability to recognize culture-specific terms in monolingual and multilingual settings.

Dependence on Language Models Examining false negative samples to analyze cultural differences can produce incorrect results since they could have been falsely predicted due to model performance instead of cultural differences. To address this issue, we attempted to better understand the reasons for misclassification by examining samples. However, since we are not native speakers of English and Arabic, this approach may not have been sufficient to comprehend cultural differences fully. To address this, future work will use human annotation to analyze hate speech from diverse cultures, with annotators from varying cultural backgrounds to develop a model that understands cultural perception differences in a given context.

Cultural Diversity within a Language The study's Korean, English, and Arabic datasets represent diverse cultural backgrounds. While the Korean dataset (KOLD) contains texts from a relatively homogeneous cultural background, the English (SBIC) and Arabic (AHS) datasets may have texts from various specific cultural backgrounds. English is spoken and written by people from different countries who may not share the same cultural background. Moreover, the AHS dataset contains various dialects, resulting in a mixture of cultures from several Arabic-speaking countries. To ensure accurate cross-cultural studies, it is crucial to constrain the dataset's represented culture or annotate which specific countries or cultures the label represents. This will prevent ignorance of cultural differences, even among countries with the same language.

Human Annotation within Hate Speech Datasets Hate speech classification research relies heavily on annotated datasets that may suffer from subjective and inconsistent labels. Annotation inconsistencies within each dataset may affect hate speech classifier predictions. As a result, the predictions of our hate speech classifiers may have been affected by the annotation inconsistency within datasets. Additionally, our analysis of the results that depend on the ground truth labels of the datasets may also be prone to errors. To alleviate annotation errors' impact, we focused on the performance differences of models on a common dataset rather than the models' performances.

9 Acknowledgements

This work was supported by Institute of Information communications Technology Planning Evaluation (IITP) grant funded by the Korea government(MSIT) (No. 2022-0-00184, Development and Study of AI Technologies to Inexpensively Conform to Evolving Policy on Ethics).

References

Sai Saketh Aluru, Binny Mathew, Punyajoy Saha, and Animesh Mukherjee. 2021. A deep dive into multilingual hate speech classification. In *Machine Learning and Knowledge Discovery in Databases. Applied Data Science and Demo Track: European Conference, ECML PKDD 2020, Ghent, Belgium, September 14–18, 2020, Proceedings, Part V*, pages 423–439. Springer International Publishing.

Wissam Antoun, Fady Baly, and Hazem Hajj. 2020. AraBERT: Transformer-based model for Arabic language understanding. In *Proceedings of the 4th Workshop on Open-Source Arabic Corpora and Processing Tools, with a Shared Task on Offensive Language Detection*, pages 9–15, Marseille, France. European Language Resource Association.

Aymé Arango, Jorge Pérez, and Barbara Poblete. 2019. Hate speech detection is not as easy as you may think: A closer look at model validation. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR'19, page 45–54, New York, NY, USA. Association for Computing Machinery.

Aymé Arango, Jorge Pérez, 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.

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

Franck Billé. 2013. Indirect interpellations: hate speech and "bad subjects" in mongolia. *Asian Anthropology*, 12(1):3–19.

Kevin Boyle. 2001. Hate speech - the united states versus the rest of the world? *Maine Law Review*, 53(2):487–502.

Ewa S. Callahan and Susan C. Herring. 2011. Cultural bias in wikipedia content on famous persons. *Journal of the American Society for Information Science and Technology*, 62(10):1899–1915.

Eshwar Chandrasekharan, Mattia Samory, Anirudh Srinivasan, and Eric Gilbert. 2017. The bag of communities: Identifying abusive behavior online with preexisting internet data. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, CHI '17, page 3175–3187, New York, NY, USA. Association for Computing Machinery.

Lucas Dixon, John Li, Jeffrey Sorensen, Nithum Thain, and Lucy Vasserman. 2018. Measuring and mitigating unintended bias in text classification. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, pages 67–73.

Jesse Dodge, Maarten Sap, Ana Marasović, William Agnew, Gabriel Ilharco, Dirk Groeneveld, Margaret Mitchell, and Matt Gardner. 2021. Documenting large webtext corpora: A case study on the colossal clean crawled corpus. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1286–1305, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Goran Glavaš, Mladen Karan, and Ivan Vulić. 2020. XHate-999: Analyzing and detecting abusive language across domains and languages. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6350–6365, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Daniel Hershcovich, Stella Frank, Heather Lent, Miryam de Lhoneux, Mostafa Abdou, Stephanie Brandl, Emanuele Bugliarello, Laura Cabello Piqueras, Ilias Chalkidis, Ruixiang Cui, Constanza Fierro, Katerina Margatina, Phillip Rust, and Anders Søgaard. 2022. Challenges and strategies in cross-cultural NLP. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6997–7013, Dublin, Ireland. Association for Computational Linguistics.

Xiaolei Huang, Linzi Xing, Franck Dernoncourt, and Michael J. Paul. 2020. Multilingual Twitter corpus and baselines for evaluating demographic bias in hate speech recognition. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 1440–1448, Marseille, France. European Language Resources Association.

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.

Eun Seo Jo and Timnit Gebru. 2020. Lessons from archives: Strategies for collecting sociocultural data in machine learning. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, FAT* '20, page 306–316, New York, NY, USA. Association for Computing Machinery.

Junbum Lee. 2021. Kcelectra: Korean comments electra. https://github.com/Beomi/KcELECTRA.

Calvin R. Massey. 1992. Hate speech, cultural diversity, and the foundational paradigms of free expression. *UCLA Law Review*, 40:103–197.

Jihyung Moon, Hyunchang Cho, and Eunjeong L. Park. 2020. Revisiting round-trip translation for quality estimation. In *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*, pages 91–104, Lisboa, Portugal. European Association for Machine Translation.

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

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.

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*, WWW '16, page 145–153, Republic and Canton of Geneva, CHE. International World Wide Web Conferences Steering Committee.

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.

Nedjma Djouhra Ousidhoum. 2021. *On the Importance and Challenges of the Experimental Design of Multilingual Toxic Content Detection*. Hong Kong University of Science and Technology, 2021, Clear Water Bay, Hong Kong.

Endang Wahyu Pamungkas, Valerio Basile, and Viviana Patti. 2020. Misogyny detection in twitter: a multilingual and cross-domain study. *Information Processing & Management*, 57(6):102360.

Endang Wahyu Pamungkas and Viviana Patti. 2019. Cross-domain and cross-lingual abusive language detection: A hybrid approach with deep learning and a multilingual lexicon. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*, pages 363–370, Florence, Italy. Association for Computational Linguistics.

Sungjoon Park, Jihyung Moon, Sungdong Kim, Won Ik Cho, Ji Yoon Han, Jangwon Park, Chisung Song, Junseong Kim, Youngsook Song, Taehwan Oh, Joohong Lee, Juhyun Oh, Sungwon Lyu, Younghoon Jeong, Inkwon Lee, Sangwoo Seo, Dongjun Lee, Hyunwoo Kim, Myeonghwa Lee, Seongbo Jang, Seungwon Do, Sunkyoung Kim, Kyungtae Lim, Jongwon Lee, Kyumin Park, Jamin Shin, Seonghyun Kim, Lucy Park, Lucy Park, Alice Oh, Jung-Woo Ha (NAVER AI Lab), Kyunghyun Cho, and Kyunghyun Cho. 2021. Klue: Korean language understanding evaluation. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, volume 1.

Jacquelyn Rahman. 2012. The n word: Its history and use in the african american community. *Journal of English Linguistics*, 40(2):137–171.

Tharindu Ranasinghe and Marcos Zampieri. 2020. Multilingual offensive language identification with cross-lingual embeddings. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5838–5844, Online. Association for Computational Linguistics.

Tharindu Ranasinghe and Marcos Zampieri. 2021. Multilingual offensive language identification for low-resource languages. *ACM Trans. Asian Low-Resour. Lang. Inf. Process.*, 22(1).

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. 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 3982–3992, Hong Kong, China. 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.

Sheikh Muhammad Sarwar and Vanessa Murdock. 2022. Unsupervised domain adaptation for hate speech detection using a data augmentation approach. *Proceedings of the International AAAI Conference on Web and Social Media*, 16(1):852–862.

Hajung Sohn and Hyunju Lee. 2019. Mc-bert4hate: Hate speech detection using multi-channel bert for different languages and translations. In *2019 International Conference on Data Mining Workshops (ICDMW)*, pages 551–559.

Zeerak Talat, James Thorne, and Joachim Bingel. 2018. *Bridging the Gaps: Multi Task Learning for Domain Transfer of Hate Speech Detection*, pages 29–55. Springer International Publishing, Cham.

Wenjie Yin and Arkaitz Zubiaga. 2021. Towards generalisable hate speech detection: a review on obstacles and solutions. *PeerJ Computer Science*, 7(e598).

	Model	Metric		
		P	R	F1
KO	KcELECTRA$_{base}$	0.80	0.80	0.80
	KcELECTRA$_{base-v2022}$	0.83	0.80	**0.81**
	KLUE-BERT$_{base}$	0.79	0.78	0.79
	KLUE-RoBERTa$_{base}$	0.79	0.78	0.78
	KLUE-RoBERTa$_{large}$	0.79	0.78	0.79
EN	BERTweet$_{base}$	0.86	0.86	**0.86**
	Twitter-RoBERTa$_{base}$	0.86	0.86	0.86
	BERT$_{base}$	0.85	0.86	0.85
	RoBERTa$_{base}$	0.86	0.86	0.86
	DistilBERT$_{base}$	0.84	0.85	0.85
AR	AraBERTv0.2-Twitter$_{base}$	0.84	0.80	**0.82**
	AraBERTv0.2-Twitter$_{large}$	0.84	0.79	0.81
	AraBERTv2$_{base}$	0.81	0.79	0.80
	AraBERTv2$_{large}$	0.82	0.80	0.81

Table 7: Evaluation results of finetuning on datasets within each of the model's languages (Korean (KO), English (EN), Arabic (AR)). Precision, Recall, and Macro-F1 scores are shown. **Bold** indicates the best performance across the models in each language, and the value in parentheses is the more accurate value to help distinguish the best-performing model.

Appendix

A Preprocessing Strategies for Datasets

KOLD KOLD contained special tokens such as <user>, <url>, and <email>, and very few of the texts included emojis.

SBIC SBIC contained usernames and URLs that were not masked, and some HTML characters such as `&#[numbers];` (emojis) (ex. 😅 as `&#128517;`), `&`(&), and `>`(>). Also, there were substantial line changes, which did not fit other datasets' shapes. Therefore, sequential `\n`s were substituted to '.' as users tended to use a line change to start a new sentence or phrase afterward.

AHS AHS contained special tokens such as @USER, <LF>, URL, and RT. <LF> refers to a line change, so it was substituted to `\n`. As in the SBIC dataset, sequential `\n`s were replaced with '.' Additionally, for all Arabic data, including datasets translated into Arabic, we utilized the ArabertPreprocessor from the arabert python package for cleaning up the Arabic texts.[4]

[4]This was recommended by the authors of AraBERT (Antoun et al., 2020). (https://huggingface.co/aubmindlab/bert-base-arabertv02-twitter)

B Training Hate Speech Classifiers

B.1 Model Training Details

All model training processes were done using the Transformers library from Huggingface[5]. We set the maximum sequence length of texts to 128 except for AraBERT-based models pre-trained on Twitter data, where we set it to 64[6]. We used AdamW as the optimizer with a learning rate of 2e-5 and an epsilon value of 1e-8, used linear scheduling for training, and set batch size as 32 for both training and evaluation steps. For conducting all experiments, 4 GeForce RTX 2080 Ti 10GB were used with CUDA version 11.0, and the experiment for each dataset took up to 3 hours.

B.2 Model Performance

Table 7 shows model performances for each language when finetuned on hate speech datasets. Each monolingual model of each language, Korean, English, and Arabic, was finetuned as a hate speech classifier using the Korean, English, and Arabic datasets, respectively. As a result, the KcELECTRA-base-v2022 model showed the highest performance on KOLD, the BERTweet-base model showed the highest performance on SBIC, and the AraBERTv0.2-Twitter-base model showed the highest performance on AHS. We use these three models for our cross-cultural evaluation in Section 5.

[5]https://github.com/huggingface/transformers

[6]The authors of AraBERT mentioned that these models were trained on texts with a sequence length of 64, and setting the maximum sequence length over this value may lead to performance degrades (https://huggingface.co/aubmindlab/bert-base-arabertv02-twitter, https://huggingface.co/aubmindlab/bert-large-arabertv02-twitter)