Cross-Lingual Transfer Learning for Toxic Speech Detection: Culture does Matter

Anonymous ACL submission

Abstract

Effective identification of toxic content plays a vital role in maintaining healthy online communities and reducing the impact of harmful discourse. In this paper, we argue that what is taken to be toxic content is influenced by cultural values and social expectations that shape how people interpret and express harmful language. This leads to our hypothesis that transfer learning is more effective between the languages of countries that share cultural contexts, and less effective across the languages of countries that do not. To test this, we conduct a cross-lingual analysis involving 14 languages from diverse regions and cultures. Our results reveal that cultural relatedness influences model performance in toxic language detection. This highlights the importance of incorporating cultural awareness into cross-lingual models and offers practical insights for building more generalizable toxic speech detection systems. Importantly, our analysis includes low-resource languages, showing that incorporating cultural considerations can enhance detection performance in some low-resource languages, even when data is limited.

Warning: This paper contains examples of toxic language that may disturb some readers. These examples are included for research on toxic detection.

1 Introduction

Toxic language detection aims to identify and mitigate harmful expressions in online text, such as hate speech, harassment, and threats (Hoang et al., 2024). As social media and digital communication platforms continue to expand globally, exposure to such toxic content has become a widespread concern. These behaviors can degrade the quality of online discourse and pose challenges for maintaining inclusive digital spaces. Importantly, what constitutes toxicity is shaped by the cultural and linguistic context in which the communication occurs (Sap et al., 2022). To ground our approach, we adopt the definition of culture proposed by Pawar et al. (2024), which integrates anthropological and psychological perspectives. In this view, culture encompasses shared values, social norms, moral beliefs, communication styles, material artifacts, and the context-dependent interpretation of behavior. Cultural relatedness, therefore, refers to alignment across these dimensions. Offensiveness and harm perception vary significantly across societies, complicating the development of automated systems that can generalize effectively across languages and regions.

Because we do not have access to features of cultural relatedness, motivated by previous work that highlighted the connection between geographical proximity and cultural relatedness (Bagheri Nezhad et al., 2025), we consider physical adjacency as a proxy for cultural relatedness in our experiments.

Recent years have seen growing interest in detecting toxic content across multiple languages (Abro et al., 2020; Zimmerman et al., 2018; Badjatiya et al., 2017; Gaydhani et al., 2018). With the emergence of Large Language Models (LLMs), which exhibit strong multilingual capabilities, researchers have increasingly turned to these models to assess their effectiveness in toxic language detection across diverse linguistic settings (Khondaker et al., 2023; Kumar et al., 2024; Abaskohi et al., 2024). While these models offer powerful general-purpose representations, their performance can still vary substantially depending on language resources and domain alignment. This has motivated growing interest in adaptation strategies that allow models to consider languages that are spoken in similar or dissimilar cultures. One promising strategy in this domain is transfer learning, which has shown substantial success in a range of NLP applications by leveraging knowledge learned from high-resource settings to improve performance in low-resource or unseen domains. Despite this progress, the question of how well transfer learning performs in the

presence of cultural variation—especially for toxic language detection—remains underexplored. Inspired by a recent study (Bokaei et al., 2025) that has examined the role of culture in toxicity detection for a small set of languages, we extend this line of work by conducting a comprehensive cross-lingual analysis across 14 languages spoken in diverse cultures. In particular, we investigate whether LLMs can effectively transfer knowledge between languages spoken in culturally similar or dissimilar societies when detecting toxicity, offering novel insights into the cultural dependencies of transfer learning in this task.

The current study addresses the following research questions (RQs):

RQ1. Are there languages for which transfer learning results in better performance than in-language training?

RQ2. For languages spoken in culturally related regions, are there aspects of toxic speech that can be more effectively detected through transfer learning?

RQ3. Would including training data from multiple languages further improve performance?

These questions aim to deepen our understanding of the cultural dimensions of cross-lingual transfer learning and provide insights for developing more advanced and culturally aware toxicity detection systems. To conduct this analysis, we use LLaMA 3 for its strong multilingual capabilities and open accessibility.

In addition to our main transfer learning experiments, we further design three complementary setups to examine how data composition and cultural grouping influence cross-lingual toxic detection. The first integrates all available training data from all languages, the second combines data only from culturally related language groups, and the third samples an equal number of instances from each language to create a balanced multilingual training set. These additional experiments allow us to evaluate the effect of data size, cultural alignment, and balanced representation on transfer performance.

2 Related Work

Toxic Language Detection: Early research on toxic language detection primarily focused on hate speech in English, using Machine Learning (ML) and Deep Learning (DL) approaches to identify harmful content on social networks (Asogwa et al., 2022; Davidson et al., 2017; Mullah and Zainon, 2021; Malik et al., 2024; Zimmerman et al., 2018; Zhou et al., 2020; Roy et al., 2020; Zhang et al., 2018). In addition, studies also addressed offensive language classification (Bade et al., 2024; Aiyanyo et al., 2020; Cao et al., 2020; Risch et al., 2020) and cyberbullying detection (Wang et al., 2020; Pamungkas and Patti, 2019; Van Hee et al., 2015; Guo and Gauch, 2024; Cano Basave et al., 2013), broadening the scope of harmful content analysis. As the field evolved, the research expanded to cover a wider range of languages, including but not limited to Indonesian (Ibrohim and Budi, 2019), Danish (Sigurbergsson and Derczynski, 2020), Arabic (Mubarak et al., 2021; Bensalem et al., 2023), Korean (Jeong et al., 2022), Chinese (Deng et al., 2022), Greek (Pitenis et al., 2020), Persian (Delbari et al., 2024) and Hindi (Gupta et al., 2022; Kapoor et al., 2019).

The introduction of Large Language Models (LLMs) has significantly advanced multilingual Natural Language Processing (NLP) tasks, including toxic language detection. These models exhibit strong capabilities and have enabled more comprehensive benchmarking across diverse languages and datasets (Zampieri et al., 2020; Verma et al., 2022; Caselli et al., 2021; Saleh et al., 2023; Nguyen et al., 2023; Chiu et al., 2021; Zampieri et al., 2023). Shared evaluation workshops such as SemEval OffensEval (Zampieri et al., 2019), HASOC (Mandl et al., 2019), OSACT5 (Mubarak et al., 2022), and GermEval (Wiegand et al., 2018) have played a key role in fostering collaboration and standardizing evaluation in this field.

Transfer Learning: In addition, many studies have adopted transfer learning—an approach that allows models trained in one setting to adapt to another. This technique has proven effective in a variety of NLP tasks, including sentiment analysis (Fsih et al., 2022; Husain et al., 2022), irony detection (Golazizian et al., 2020), machine translation (Zoph et al., 2016; Kim et al., 2019; Adebara and Abdul-Mageed, 2021), fake news detection (Cruz et al., 2020), and offensive language classification (Zhou et al., 2023). While these applications demonstrate the effectiveness of transfer learning, limited studies have examined how cultural relatedness influences that effectiveness, especially for toxic language detection. A recent study (Zhou et al., 2023) used mBERT and RoBERTa to study cross-cultural transfer in offensive language detection. They found that Korean-to-Chinese transfer outperforms English-to-Chinese, highlighting the role of cultural proximity. Few-shot learn-

ing with minimal Chinese data further improves performance, underscoring the value of culturally aligned data in transfer learning. While these efforts highlight the role of culture, they typically focus on a narrow set of languages and do not systematically evaluate cultural relatedness as a variable in transfer effectiveness. Our study addresses this gap by conducting a large-scale cross-lingual analysis involving 14 languages from culturally diverse regions. Unlike previous work, we explicitly examine how cultural relatedness between countries impacts transfer learning performance in toxicity detection. To support this investigation, we used LLaMA 3, a state-of-the-art LLM known for its multilingual capacity and strong performance across languages.

With a culturally informed perspective, our work provides new insights into the cultural dynamics of transfer learning and contributes toward building more inclusive and generalizable toxic language detection systems. In addition, we conduct a comprehensive analysis of each language dataset to examine the cross-lingual impact of transfer learning. This allows us to identify when cultural relatedness enhances performance and how each language benefits—or struggles—under different transfer scenarios. Apart from that, we examine whether culturally aware transfer learning can improve performance for low-resource languages, offering a promising strategy for addressing data scarcity.

3 Dataset

We rely on well-established datasets from peer-reviewed publications and reputable shared tasks to ensure data quality and comparability. To investigate the influence of cultural background on toxicity detection, we select 14 languages representing diverse regions, including East Asia, West Asia, the Middle East, and Europe. All datasets exhibit class imbalance, which we address to maintain consistency across languages.

For the main set of experiments, we exclude languages with very limited data availability—specifically Hindi, Malay, Korean, Chinese, and Persian. For the remaining languages, we identify the smallest dataset's number of toxic instances and use that as a reference point. We then sample an equal number of non-toxic instances to create balanced datasets, each consisting of approximately 10,000 examples. These are split into training, validation, and test sets using an 80/10/10 ratio. For the excluded languages, we also balance the datasets as much as possible and apply the same 80/10/10 split for analysis. Details of all languages and dataset sizes used in this study are summarized in Table 1.

4 Experimental Setup

We conduct multilingual fine-tuning experiments using the datasets described in Table 1.

For each language, we fine-tune the LLaMA 3 8B model on its own training data and evaluate it both on its native test set and on the test sets of the remaining 13 languages, resulting in a comprehensive cross-lingual transfer matrix. This setup allows us to analyze how well cultural relatedness facilitates model transfer in toxicity detection.

We fine-tune each model using a learning rate of 1e-5 and a batch size of 8 for up to 10 epochs, with early stopping based on validation performance (patience = 3). All experiments are conducted on an NVIDIA Quadro RTX 8000 GPU. We evaluate performance using per-class F1 scores (toxic and non-toxic), with a focus on how well models generalize across culturally diverse settings.

Beyond the main experiment in which models are trained on a single language and evaluated on all test languages, we conduct three complementary experiments: Experiment 1 (Global Integration Imbalanced -> Global-Imb): We merge all training data from all 14 languages into a single multilingual dataset, fine-tune LLaMA 3 on this combined set, and evaluate on each language's test set individually. This setup tests whether the volume and diversity of multilingual toxic content improve the performance of toxic detection. Experiment 2 (Global Cultural Clusters -> Culture-Cluster): We identify sets of culturally related languages (based on the results observed in the main experiment, as shown in Figure 1) and merge their training data into group-specific datasets (e.g. Hindi–Urdu–Bengali or French–Spanish–German–English) Models are trained on each group's combined data and evaluated separately on each language within the group, enabling us to measure the benefits of cultural relatedness. Experiment 3 (Global Integration Balanced -> Global-Bal): We sample an equal number of toxic (n = 1000) and non-toxic (n = 1000) instances from each language's training set, merge them into a balanced multilingual dataset, and train a single model evaluated on each language individually. This setup controls for data imbalance across languages and isolates the effect of representational equality on transfer performance.

Language	Data Size	Train	Test	#Hate (Train)	#Hate (Test)	Source	Period	Ref.
Bengali (BN)	30000	9007	1002	4504	501	YT, FB	2017–2020	(Romim et al., 2020)
Urdu (UR)	10041	9000	1000	4500	300	Twitter	–	(Rizwan et al., 2020)
Hindi (HI)	8192	4171	467	2222	241	Twitter, Fb, Whatsapp	–	(Bhardwaj et al., 2020)
Persian/Farsi (FA)	7000	2060	230	1002	117	Twitter	2020–2022	(Delbari et al., 2024)
Indonesian (IN)	13169	9005	1003	4502	502	Twitter	$\leq$ 2018	(Ibrohim and Budi, 2019)
Malay (MS)	5014	3403	378	1701	189	Twitter	2022–2023	(Maity et al., 2023)
Arabic (AR)	6136–8878	9005	1002	4503	502	Twitter	2016–2018	(Albadi et al., 2018; Mubarak et al., 2021)
Turkish (TR)	52672	9010	1003	4505	502	Twitter, Instagram	2016–2023	(Tonneau et al., 2024)
Korean (KD)	9341–109692	3065	300	1520	142	Domestic news	2018–2020	(Moon et al., 2020; Lee et al., 2022)
Chinese (ZH)	8065	4950	318	2470	151	Web	–	(Moosa and Najiba, 2022)
Spanish (ES)	29856	9000	1000	4731	625	Twitter	2016–2023	(Tonneau et al., 2024)
German (DE)	58025	9008	1003	4503	502	Twitter, Fb, News	2016–2023	(Tonneau et al., 2024)
French (FR)	18072	7483	829	3739	413	Twitter, NGO, RPG	2016–2023	(Tonneau et al., 2024)
English (EN)	360494	9000	1038	4500	502	Fb, YT, Gab, Reddit, NGO, Twitter	2016–2023	(Tonneau et al., 2024)

Table 1: Transfer Learning Dataset Details. Abbreviations: FB = Facebook, YT = YouTube, NGO = Non-Governmental Organization, RPG = Role-Playing Game.

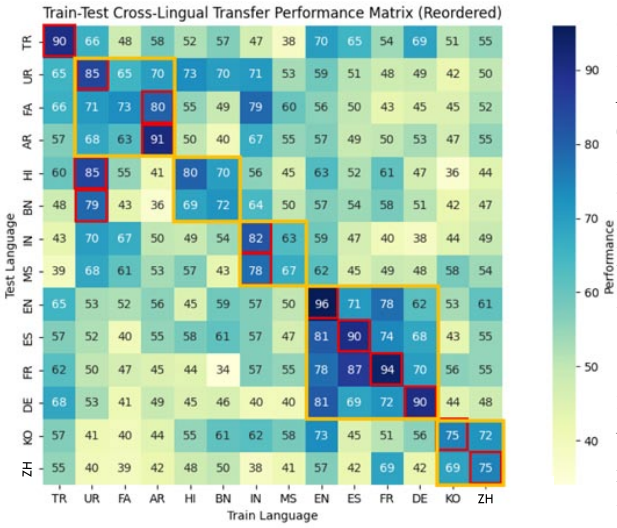

Figure 1: Cross-lingual F1 scores for toxic detection across 14 languages. Red squares indicate the best-performing train language for each test language (Solo-Source TL), and orange squares highlight strong intra-group transfer pairs (Cultural-Cluster). Language abbreviations are given in Table 1.

5 Results

Figure 1 presents the cross-lingual transfer performance matrix for toxic content detection across 14 languages. Each cell contains the F1 score obtained when a model is trained on train language (x-axis) and evaluated on the test language (y-axis). The matrix includes a total of 196 train-test combinations, enabling a comprehensive view of in-language performance and transferability across language pairs. In most but not all cases, the highest performance values are found along the diagonal, where the model is trained and evaluated on the same language. These in-language F1 scores are consistently strong, with several languages achieving values at or above 90, including Turkish (90), Arabic (91), English (96), French (94), Spanish (90), and German (90). Interestingly, there are a few cases where cross-lingual transfer yields higher F1 scores than in-language training. For example, Persian achieves better results when trained on Arabic or Urdu rather than on its own data. We discuss these cases in detail in Section 6.

Off-diagonal values indicate cross-lingual transfer performance. While the quality of transfer varies considerably, several language pairs achieve relatively high F1 scores. For example, models trained on Urdu perform well on Hindi (85), and those trained on English perform well on French (78), Spanish (74), and German (69). French and English, in particular, show mutual transfer (78), Bengali and Urdu also represent high mutual transfer (BN→UR: 70, UR→BN: 79), suggesting effective transfer capabilities between some languages. In contrast, lower F1 scores are observed in transfer between more distant language pairs. For instance, training on Bengali yields relatively weak performance on European languages (e.g., BN→FR: 51, BN→DE: 42), and Chinese performs poorly on several unrelated languages (CH→ES: 42, CH→AR: 38). This illustrates the variability in transfer effectiveness across languages.

Overall, languages such as English, French, and German not only perform well in monolingual settings but also serve as strong train languages in transfer scenarios for some languages. Meanwhile, we observed relatively lower transfer scores for languages such as Korean, Chinese, and Malay, both

	TR	UR	FA	AR	HI	BN	IN	MS	EN	ES	FR	DE	KO	ZH
Solo-Source TL	90	85	80	91	85	79	82	78	96	90	94	90	75	75
Global-Imb	92 [+2]	88 [+3]	64 [-16]	87 [-4]	73 [-12]	66 [-13]	84 [+2]	62 [-16]	98 [+2]	95 [+5]	96 [+2]	94 [+4]	71 [-4]	72 [-3]
Culture-Cluster	90 [0]	90 [+5]	83 [+3]	93 [+2]	90 [+5]	83 [+4]	85 [+3]	80 [+2]	97 [+1]	96 [+6]	96 [+2]	95 [+5]	75 [0]	78 [+2]
Global-bal	91 [+1]	87 [+2]	67 [-13]	89 [-2]	81 [-4]	69 [-10]	83 [+1]	63 [-15]	96 [0]	92 [+2]	94 [0]	93 [+3]	73 [-2]	76 [+1]

Table 2: Performance comparison across different training strategies. Bracketed numbers indicate improvement or decline relative to Solo-Source TL.

as train and test languages. However, this may be influenced by factors such as limited dataset size and collection period (e.g., Malay from 2022–2023 vs. Indonesian from 2018 and earlier), rather than reflecting an inherent weakness in these languages for cross-lingual transfer. As represented in Table 2, In Global-Imb Experiment, the model benefited from the volume and diversity of the integrated dataset, yielding improvements for high-resource languages such as English, French, and Spanish (average gain: +2.86 points), as well as for languages like Urdu that seem to share overt toxic expressions with other datasets. However, most low-resource languages—such as Bengali, Persian, and Malay—showed performance declines (average drop: 5.43 points). Regarding the Culture-Cluster experiment, grouping languages whose countries are culturally closer produced consistent gains for nearly all languages (average gain: +5.62 points). This effect was particularly clear for clusters such as Hindi–Bengali–Urdu and Arabic–Farsi–Indonesian. Comparable effects were observed for high-resource groupings such as French–Spanish–German–English, although here the improvement was less pronounced than for low-resource languages. Lastly, in the Global-Bal experiment, assigning equal representation to each language in training removed the effect of data imbalance.

High-resource languages still improved compared to their single-language baselines, though gains were smaller than in the imbalanced setting (average gain: +1.57 points). For low-resource languages, the balanced setup mitigated the imbalanced effect seen in Experiment 1, with most seeing modest gains but some (5 of 14) remaining flat or slightly below baseline (average drop: 3.40 points).

6 Analysis and Discussion

RQ1. Are there languages for which transfer learning results have better performance than in-language training?

Our results reveal that, for certain languages, transfer learning yields better performance than training within the same language. Notably, training on Arabic, Indonesian, and Urdu and testing on Persian results in higher F1 scores compared to the respective in-language models. Based on our observations, transfer learning from Arabic and Urdu data proved particularly effective for detecting toxicity related to religious and political topics, especially sociopolitical hate that is prevalent in the Middle Eastern context. The Arabic dataset provided relevant contextual cues that aligned well with Persian discourse in these domains. Similarly, Urdu also contributes effectively to the detection of political toxicity, often to an even greater extent. In contrast, Indonesian data was more effective in identifying toxic content directed at individuals—such as those based on profession—rather than group-based toxicity. In addition, models trained on Indonesian data demonstrated superior performance in handling longer texts containing a mixture of neutral and toxic sentiments. Table 3 presents some Persian toxic instances that are only predicted correctly after applying transfer learning, as well as those correctly predicted only by training on Persian.

In addition, we observed a significant overlap in how hate is expressed in Indonesian and Malay, suggesting strong cultural and contextual alignment between the two languages. Across the instances, common patterns emerge in political discourse, such as disillusionment with corrupt politicians and identity politics based on religion. Religious references are frequently used to judge or shame

Lang	TL From	Tweet	Explanation
FA	IN	رندی به محضر فقیهی رسیدو حرکات رقص را جداجدا انجام می داد و می پرسید آیا حرام است؟ فقیه میگفت نه. پس رند شروع به رقصیدن کرد.فقیه گفت تجزیه اش خوب بود ولی مرده شور ترکیبش رو بردن حالا حکایت این عدالت خواراست بعضیشون عیییی بچه های خوبین ولی مرده شور ترکیبشون رو برده من برم به کارای خودم برسم خدافظ #عدالتخواران#انتخابات_مجلس A trickster went to a cleric and performed dance moves separately, asking if they were forbidden. The cleric said no. Then the trickster started dancing, and the cleric said, 'Breaking it down was fine, but damn the combination!' This is exactly the case with these so-called justice-seekers—some of them are actually good kids, but damn their combination! Anyway, I'll get back to my own business. Goodbye. #JusticeSeekers #ParliamentElection	Long tweet containing a mixture of neutral and toxic sentiments, with shared religious content.
FA	AR	وقتی یه جوان بدبخت میخاد استخدام بشه ریشه و بیخ مسائل خانوادگیش رو درمیارن اما زیباکلام با دختربازیش و شراب خواریش هنوز تو دانشگاه اسلامی!!! تدریس میکنه When a poor young man wants to get hired, they dig up every root of his family issues, but Zibakalam, with his womanizing and wine-drinking, is still teaching at an Islamic university!	Religious hypocrisy framed with moral outrage, a sentiment common to both Persian and Arabic contexts.
MS	IN	Oh dan kereta mmg suka menghimpit ke motor walaupun jelas2 ade moto kt situ smpi terpaksa mengelak n kereta tu buat bodoh je. Ramai ye perangai mcm ni. Oh and cars really like to squeeze in on motorcycles even when it's obvious there's a bike there, to the point the rider has to avoid them, and the car just acts clueless. There are many people with this kind of attitude.	Everyday frustration or complaint expressed in the form of Malay-language hate.
BN	UR	বেয়াদবের বাচ্চা বেয়াদব হ্যাকার গ্রুপ বর্ণনা হ্যাকার গ্রুপ বার করে দিব লুচ্চামি কি করস শরীরটা দেখছিস নিজের ছবিটা দেখ Impudent child's impudent hacker group description hacker group will be kicked out, what lewdness are you doing, looking at the body, look at your own picture.	Accusing a group of hackers of lewd behaviour and threatening expulsion.
HI	UR	हम तेरे लिये कुछ नही कर सकते ऐ दोस्त...!!! हमें अभी और मंदिर और मस्जिद बनानी है...!! We can't do anything for you my friend... We still have to build more temples and mosques...!!	Politicians prioritize religious projects over people's real needs.

Table 3: Hate samples correctly predicted only via transfer learning across languages.

others, including accusations related to religious observance or moral behavior during sacred times. Culturally, we observed that both languages employ similar proverbs, insults, and metaphorical language—particularly the use of animals like "dog" and "devil" as slurs in hate instances. Social issues such as traffic caused by bazaars, public frustration over commodity scarcity, and emotional sharing on social media are also mirrored in both contexts. These shared references demonstrate that Indonesian and Malay resemble each other in the cultural and pragmatic framing of hate speech. As a result, transfer learning from Indonesian to Malay is likely to be effective, as the underlying sociocultural triggers and linguistic expressions of hate align closely. However, this effectiveness does not always hold in the reverse direction. One possible reason is the limited amount of toxic data in the Malay dataset, which restricts the model's ability to generalize when used as a source language for transfer learning. In contrast, the Indonesian dataset offers a broader range of hate instances spanning multiple categories and target types (e.g., religion, race, gender, individual vs. group). A similar pattern is observed with Persian: despite strong cultural ties to Arabic and Urdu, its limited toxic training data reduces its effectiveness as a target language.

Another language group where transfer learning outperforms in-language learning includes Urdu, Hindi, and Bengali. Specifically, training on Urdu and testing on either Bengali or Hindi yields better results than training directly on the test language. We find that transfer learning from Urdu is especially effective at capturing highly polarized, overt hate expressions that involve explicit slurs, religious antagonism, or politicized attacks. These instances often rely on shared cultural schemas and common hate lexicons—for example, communal insults or anti-minority tropes—that generalize well across closely related South Asian languages. However, in-language models more correctly classified Tweets that contain weak cultural references, idiomatic expressions, or contextually grounded forms of criticism. For instance, socially embedded sarcasm in Bengali was frequently missed by the Urdu-trained model but correctly identified by the Bengali-trained one. The Urdu dataset exhibits a high density of explicit hate markers, including religious slurs, profane language, and direct references to identifiable targets. These overt expressions are used with relative lexical consistency and limited idiomatic variation, enabling the model to learn discriminative toxic patterns reliably. Moreover, Urdu tweets frequently specify their hate targets (e.g., particular religious or political groups), making the hateful intent more detectable. In contrast, Hindi and Bengali datasets tend to feature more ambiguous, sarcastic, or culturally embedded forms of expression, which are harder to generalize across languages.

We notice that although political and religious toxicity exists in other datasets like English, cultural proximity plays a more decisive role in successful transfer learning. Even within the same top-

ical domains, models trained on culturally similar languages outperform those trained on culturally distant ones within the same topics. This highlights that shared discourse and sociopolitical references, not just topic overlap, drive better generalization across related languages. For instance, we observed that certain tweets expressing everyday frustration in Malay were labeled as hate in the local dataset, even though similar context would only be considered hate in Indonesian, but not in the other languages we studied. This illustrates how perceptions of toxicity can vary significantly across cultures. Table 3 shows hate samples correctly predicted only via transfer learning. After addressing the first research question, we identified a follow-up sub-question: Can leveraging cultural similarity in transfer learning improve performance for low-resource languages? Among the languages studied, we identified Persian, Malay, and Korean as particularly underresourced for toxic content detection. For Persian and Malay, transfer learning from culturally similar languages proved especially effective. Although further investigation with more low-resource languages is needed, our findings suggest that incorporating cultural relatedness can enhance performance in such settings. For the first research question, our results show that transfer learning can outperform in-language learning when the training dataset is both large and culturally aligned. For instance, as discussed earlier, the Indonesian dataset enables better performance on Malay than Malay does on Indonesian, likely due to its broader coverage of hate expressions. The four-year temporal gap between the two datasets may also contribute to the asymmetry, as shifts in political discourse can weaken transfer effectiveness over time.

RQ2. For languages spoken in culturally diverse regions, are there still aspects of toxic speech that can be effectively detected through transfer learning? Our results show that even when the train and test languages are culturally distinct (e.g., Spanish-Chinese, Persian-French), transfer learning can still detect certain types of toxic speech. Specifically, explicit, emotionally extreme, or lexically offensive expressions tend to transfer well regardless of cultural differences.

For example, in our experiments, the Spanish→Chinese transfer achieved moderate performance in detecting overtly toxic Chinese instances such as:

"可以证明你是个没脑子的田园女权婊了", (You're a brainless feminist bitch) or "微博女权婊人数还是相当多的！" (So many feminist bitches on Weibo!) These phrases involve direct insults and highly polarized language, patterns that models trained on Spanish data can still recognize due to their universality, even without cultural overlap.However, the Spanish-trained model struggled with more context-dependent expressions, such as: "唐明皇一世英名就毁在杨家人手上了" (Emperor Tang Minghuang's lifetime reputation was ruined at the hands of the Yang family - A historical reference used to delegitimize women), or "我还看到一个百事256的说自己是女权的，最近在力挺川美侵台军宅，和粉红互动得可欢了呢。" (I also saw someone called 'Pepsi256' who claims to be a feminist. Lately, she's been enthusiastically supporting the Sichuan Fine Arts Institute's pro-invasion-of-Taiwan military otakus and happily interacting with the nationalists (a.k.a. little pinks). In contrast, training the model on Korean—a language that shares greater cultural and regional proximity with Chinese—resulted in significantly better performance, not only on overtly toxic expressions but also on more subtle, context-dependent instances like those discussed above. This reinforces the idea that cultural relatedness enhances the model's ability to generalize, particularly when toxicity relies on shared social values and references. In other instances, we can mention experiments with French and Persian, two languages from culturally distinct countries, where models trained in Persian were still able to correctly identify highly explicit toxic content in French. For example: "Vous pouvez crever sale race" (You can die, filthy race) This phrase combines an aggressive death wish with a racial slur, making it both emotionally extreme and lexically offensive, which enabled the Persian-trained model to detect it despite the cultural gap. Upon the second research question, our result shows that even across culturally dissimilar languages, transfer learning can succeed when the toxic content is lexically explicit or emotionally extreme. However, it consistently fails on culturally embedded expressions, such as social and historical references. This indicates that while surface-level cues can transfer, deeper cultural understanding remains a challenge.

RQ3. Would including training data from multiple languages further improve the performance? To answer this research question, beyond the single-language transfer setup, we have three complementary experiments. Global-Imb experiment demonstrates that integrating data from all lan-

guages in an imbalanced manner primarily benefits high-resource languages, taking advantageous of enhanced performance for high-resource languages, it weakens culturally specific cues for low-resource ones, as their contextually grounded or culturally relevant expressions are overshadowed by dominant high-resource patterns. Culture-Cluster experiment shows that grouping culturally similar languages not only mitigates this washout effect but also produces the most consistent and substantial gains, especially for low-resource languages. Global-Bal experiment, by balancing representation across languages, offers a middle ground—reducing the disadvantage faced by low-resource languages while slightly decrease the gains of high-resource ones.

7 Future Research Directions

This study highlights how cultural relatedness affects the success of cross-lingual transfer learning for toxic speech detection. Several directions remain for future exploration.

First, the evolving temporal context of cultural discourse warrants deeper investigation. The set of ideas, groups, or behaviors that are considered hate-worthy can shift over time, and the rate of this change may differ across cultural dimensions. For instance, political discourse may evolve more rapidly—introducing or discarding targets of hate more frequently—while religious or social values tend to remain more stable. Future work could examine which aspects of culture remain consistent over time and how this influences the effectiveness of model transfer across time-stamped datasets.

Second, while this study grouped languages by cultural relatedness, future research could aim to disentangle specific cultural factors, such as religion, politics, and occupation, to assess their individual impact on transfer effectiveness.

Third, future work could incorporate temporal shifts in toxic discourse—for example, changes linked to election cycles, social movements, or major geopolitical events—to examine whether culture-aware models can adapt to evolving patterns without sacrificing performance on historically observed toxicity. This would help assess the stability and adaptability of transfer learning approaches across both cultural and temporal dimensions.

Conclusion

This study explored the role of cultural relatedness in cross-lingual transfer learning for toxic speech detection across 14 languages. Our results show that cultural alignment between train and test languages improves model performance, particularly for culturally and socially embedded toxic content. Transfer learning proves especially beneficial for low-resource languages when culturally similar training data is available. However, even in the absence of cultural relatedness, certain types of toxic expressions—such as overt insults or emotionally extreme language—can still be reliably transferred. Furthermore, our complementary experiments (RQ3) reveal that training on large-scale integration of all languages benefits high-resource languages but risks weakening culturally specific cues in low-resource ones, while grouping culturally similar languages consistently yields the strongest and most balanced gains. These findings point to cultural relatedness as a valuable consideration in designing cross-lingual models.

Limitations

While this study provides insights into the cultural dimensions of cross-lingual transfer learning, several limitations should be noted. First, our definition of cultural relatedness is based on geographic proximity, which may not fully capture the complexity of cultural relationships. More granular cultural variables—such as religion, historical events, or media exposure—were not explicitly modeled. Second, the datasets used vary in size, source, and collection period, which may introduce confounding factors in performance comparisons.

Ethics Statement

This study follows ethical principles, emphasizing fairness, privacy, and harm prevention in toxic con- tent detection. Methods aim to minimize bias and ensure transparency and accountability. All data were handled in line with ethical guidelines and data protection laws.

References

Amirhossein Abaskohi, Sara Baruni, Mostafa Masoudi, Nesa Abbasi, Mohammad Hadi Babalou, Ali Edalat, Sepehr Kamahi, Samin Mahdizadeh Sani, Nikoo Naghavian, Danial Namazifard, Pouya Sadeghi, and Yadollah Yaghoobzadeh. 2024. Benchmarking large language models for Persian: A preliminary study focusing on ChatGPT. In *Proceedings of the 2024 Joint*

International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024), pages 2189–2203, Torino, Italia. ELRA and ICCL.

Sindhu Abro, Sarang Shaikh, Zahid Hussain Khand, Ali Zafar, Sajid Khan, and Ghulam Mujtaba. 2020. Automatic hate speech detection using machine learning: A comparative study. *International Journal of Advanced Computer Science and Applications*, 11(8).

Ife Adebara and Muhammad Abdul-Mageed. 2021. Improving similar language translation with transfer learning. In *Proceedings of the Sixth Conference on Machine Translation*, pages 273–278, Online. Association for Computational Linguistics.

Imatitikua D Aiyanyo, Hamman Samuel, and Heuiseok Lim. 2020. A systematic review of defensive and offensive cybersecurity with machine learning. *Applied Sciences*, 10(17):5811.

Nuha Albadi, Maram Kurdi, and Shivakant Mishra. 2018. Are they our brothers? analysis and detection of religious hate speech in the arabic twittersphere. In *2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, pages 69–76.

Doris Chinedu Asogwa, Chiamaka Ijeoma Chukwuneke, CC Ngene, and GN Anigbogu. 2022. Hate speech classification usingSVM and Naive Bayes. *arXiv preprint arXiv:2204.07057*.

Girma Bade, Olga Kolesnikova, Grigori Sidorov, and José Oropeza. 2024. Social media hate and offensive speech detection using machine learning method. In *Proceedings of the Fourth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*, pages 240–244, St. Julian's, Malta. Association for Computational Linguistics.

Pinkesh Badjatiya, Shashank Gupta, Manish Gupta, and Vasudeva Varma. 2017. Deep learning for hate speech detection in tweets. In *Proceedings of the 26th international conference on World Wide Web companion*, pages 759–760.

Sina Bagheri Nezhad, Ameeta Agrawal, and Rhitabrat Pokharel. 2025. Beyond data quantity: Key factors driving performance in multilingual language models. In *Proceedings of the First Workshop on Language Models for Low-Resource Languages*, pages 225–239, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Imene Bensalem, Meryem Mout, and Paolo Rosso. 2023. Offensive language detection in Arabizi. In *Proceedings of ArabicNLP 2023*, pages 423–434.

Mohit Bhardwaj, Md. Shad Akhtar, Asif Ekbal, Amitava Das, and Tanmoy Chakraborty. 2020. Hostility detection dataset in hindi. *CoRR*, abs/2011.03588.

Zahra Bokaei, Walid Magdy, and Bonnie Webber. 2025. Culture matters in toxic language detection in Persian. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9290–9304, Vienna, Austria. Association for Computational Linguistics.

Amparo Elizabeth Cano Basave, Yulan He, Kang Liu, and Jun Zhao. 2013. A weakly supervised Bayesian model for violence detection in social media. In *Proceedings of the Sixth International Joint Conference on Natural Language Processing*, pages 109–117, Nagoya, Japan. Asian Federation of Natural Language Processing.

Rui Cao, Roy Ka-Wei Lee, and Tuan-Anh Hoang. 2020. Deephate: Hate speech detection via multi-faceted text representations. In *Proceedings of the 12th ACM Conference on Web Science*, pages 11–20.

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.

Ke-Li Chiu, Annie Collins, and Rohan Alexander. 2021. Detecting hate speech with GPT-3. *arXiv preprint arXiv:2103.12407*.

Jan Christian Blaise Cruz, Julianne Agatha Tan, and Charibeth Cheng. 2020. Localization of fake news detection via multitask transfer learning. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 2596–2604, Marseille, France. European Language Resources Association.

Thomas Davidson, Dana Warmsley, Michael Macy, and Ingmar Weber. 2017. Automated hate speech detection and the problem of offensive language. In *Proceedings of the international AAAI conference on web and social media*, volume 11, pages 512–515.

Zahra Delbari, Nafise Sadat Moosavi, and Mohammad Taher Pilehvar. 2024. Spanning the spectrum of hatred detection: a Persian multi-label hate speech dataset with annotator rationales. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 17889–17897.

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.

Emna Fsih, Sameh Kchaou, Rahma Boujelbane, and Lamia Hadrich-Belguith. 2022. Benchmarking transfer learning approaches for sentiment analysis of Arabic dialect. In *Proceedings of the Seventh Arabic Natural Language Processing Workshop (WANLP)*, pages 431–435, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Aditya Gaydhani, Vikrant Doma, Shrikant Kendre, and Laxmi Bhagwat. 2018. Detecting hate speech and offensive language on twitter using machine learning: An n-gram and tfidf based approach. *arXiv preprint arXiv:1809.08651*.

Preni Golazizian, Behnam Sabeti, Seyed Arad Ashrafi Asli, Zahra Majdabadi, Omid Momenzadeh, and Reza Fahmi. 2020. Irony detection in Persian language: A transfer learning approach using emoji prediction. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 2839–2845, Marseille, France. European Language Resources Association.

Xiaoyu Guo and Susan Gauch. 2024. Using sarcasm to improve cyberbullying detection. In *Proceedings of the Fourth Workshop on Threat, Aggression & Cyberbullying @ LREC-COLING-2024*, pages 52–59, Torino, Italia. ELRA and ICCL.

Vikram Gupta, Sumegh Roychowdhury, Mithun Das, Somnath Banerjee, Punyajoy Saha, Binny Mathew, Animesh Mukherjee, et al. 2022. Multilingual abusive comment detection at scale for Indic languages. *Advances in Neural Information Processing Systems*, 35:26176–26191.

Nhat Hoang, Xuan Long Do, Duc Anh Do, Duc Anh Vu, and Anh Tuan Luu. 2024. ToXCL: A unified framework for toxic speech detection and explanation. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 6460–6472, Mexico City, Mexico. Association for Computational Linguistics.

Fatemah Husain, Hana Al-Ostad, and Halima Omar. 2022. A weak supervised transfer learning approach for sentiment analysis to the Kuwaiti dialect. In *Proceedings of the Seventh Arabic Natural Language Processing Workshop (WANLP)*, pages 161–173, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Muhammad Okky Ibrohim and Indra Budi. 2019. Multi-label hate speech and abusive language detection in Indonesian Twitter. In *Proceedings of the Third Workshop on Abusive Language Online*, pages 46–57, Florence, Italy. Association for Computational Linguistics.

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.

Raghav Kapoor, Yaman Kumar, Kshitij Rajput, Rajiv Ratn Shah, Ponnurangam Kumaraguru, and Roger Zimmermann. 2019. Mind your language: Abuse and offense detection for code-switched languages. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 9951–9952.

Md Tawkat Islam Khondaker, Abdul Waheed, El Moatez Billah Nagoudi, and Muhammad Abdul-Mageed. 2023. GPTAraEval: A comprehensive evaluation of ChatGPT on Arabic NLP. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 220–247, Singapore. Association for Computational Linguistics.

Yunsu Kim, Petre Petrov, Pavel Petrushkov, Shahram Khadivi, and Hermann Ney. 2019. Pivot-based transfer learning for neural machine translation between non-English languages. 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 866–876, Hong Kong, China. Association for Computational Linguistics.

Ankit Kumar, Richa Sharma, and Punam Bedi. 2024. Towards optimal NLP solutions: Analyzing GPT and LLaMA-2 models across model scale, dataset size, and task diversity. *Engineering, Technology & Applied Science Research*, 14(3):14219–14224.

Jean Lee, Taejun Lim, Heejun Lee, Bogeun Jo, Yangsok Kim, Heegeun Yoon, and Soyeon Caren Han. 2022. K-MHaS: A multi-label hate speech detection dataset in Korean online news comment. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 3530–3538, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Krishanu Maity, Shaubhik Bhattacharya, Sriparna Saha, and Manjeevan Seera. 2023. A deep learning framework for the detection of malay hate speech. *IEEE Access*, 11:79542–79552.

Jitendra Singh Malik, Hezhe Qiao, Guansong Pang, and Anton van den Hengel. 2024. Deep learning for hate speech detection: a comparative study. *International Journal of Data Science and Analytics*, pages 1–16.

Thomas Mandl, Sandip Modha, Prasenjit Majumder, Daksh Patel, Mohana Dave, Chintak Mandlia, and Aditya Patel. 2019. Overview of the hasoc track at fire 2019: Hate speech and offensive content identification in Indo-European languages. In *Proceedings of the 11th Annual Meeting of the Forum for Information Retrieval Evaluation*, FIRE '19, page 14–17, New York, NY, USA. Association for Computing Machinery.

Jihyung Moon, Won Ik Cho, and Junbum Lee. 2020. BEEP! Korean corpus of online news comments for toxic speech detection. In *Proceedings of the Eighth International Workshop on Natural Language Processing for Social Media*, pages 25–31, Online. Association for Computational Linguistics.

Wajid Hassan Moosa and Najiba. 2022. Multilingual hate speech dataset. `https://www.kaggle.com/datasets/wajidhassanmoosa/`

multilingual-hatespeech-dataset. Open Data Commons Attribution License (ODC-By) v1.0.

Hamdy Mubarak, Hend Al-Khalifa, and Abdulmohsen Al-Thubaity. 2022. Overview of OSACT5 shared task on Arabic offensive language and hate speech detection. In *Proceedinsg of the 5th Workshop on Open-Source Arabic Corpora and Processing Tools with Shared Tasks on Qur'an QA and Fine-Grained Hate Speech Detection*, pages 162–166, Marseille, France. European Language Resources Association.

Hamdy Mubarak, Ammar Rashed, Kareem Darwish, Younes Samih, and Ahmed Abdelali. 2021. Arabic offensive language on Twitter: Analysis and experiments. In *Proceedings of the Sixth Arabic Natural Language Processing Workshop*, pages 126–135, Kyiv, Ukraine (Virtual). Association for Computational Linguistics.

Nanlir Sallau Mullah and Wan Mohd Nazmee Wan Zainon. 2021. Advances in machine learning algorithms for hate speech detection in social media: a review. *IEEE Access*, 9:88364–88376.

Thanh Thi Nguyen, Campbell Wilson, and Janis Dalins. 2023. Fine-tuning LLAMA 2 large language models for detecting online sexual predatory chats and abusive texts. *arXiv preprint arXiv:2308.14683*.

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.

Siddhesh Pawar, Junyeong Park, Jiho Jin, Faiz Ghifari Haznitrama, Junho Myung, Alice Oh, Isabelle Augenstein, et al. 2024. Survey of cultural awareness in language models: Text and beyond. *arXiv preprint arXiv:2411.00860*.

Zesis Pitenis, Marcos Zampieri, and Tharindu Ranasinghe. 2020. Offensive language identification in Greek. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 5113–5119, Marseille, France. European Language Resources Association.

Julian Risch, Robin Ruff, and Ralf Krestel. 2020. Offensive language detection explained. In *Proceedings of the Second Workshop on Trolling, Aggression and Cyberbullying*, pages 137–143, Marseille, France. European Language Resources Association (ELRA).

Hammad Rizwan, Muhammad Haroon Shakeel, and Asim Karim. 2020. Hate-speech and offensive language detection in Roman Urdu. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2512–2522, Online. Association for Computational Linguistics.

Nauros Romim, Mosahed Ahmed, Hriteshwar Talukder, and Md Saiful Islam. 2020. Hate speech detection in the bengali language: A dataset and its baseline evaluation. *Preprint*, arXiv:2012.09686.

Pradeep Kumar Roy, Asis Kumar Tripathy, Tapan Kumar Das, and Xiao-Zhi Gao. 2020. A framework for hate speech detection using deep convolutional neural network. *IEEE Access*, 8:204951–204962.

Hind Saleh, Areej Alhothali, and Kawthar Moria. 2023. Detection of hate speech using BERT and hate speech word embedding with deep model. *Applied Artificial Intelligence*, 37(1):2166719.

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.

Gudbjartur Ingi Sigurbergsson and Leon Derczynski. 2020. Offensive language and hate speech detection for Danish. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 3498–3508, Marseille, France. European Language Resources Association.

Manuel Tonneau, Diyi Liu, Samuel Fraiberger, Ralph Schroeder, Scott Hale, and Paul Röttger. 2024. From languages to geographies: Towards evaluating cultural bias in hate speech datasets. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Cynthia Van Hee, Els Lefever, Ben Verhoeven, Julie Mennes, Bart Desmet, Guy De Pauw, Walter Daelemans, and Veronique Hoste. 2015. Detection and fine-grained classification of cyberbullying events. In *Proceedings of the International Conference Recent Advances in Natural Language Processing*, pages 672–680, Hissar, Bulgaria. INCOMA Ltd. Shoumen, BULGARIA.

Kanishk Verma, Tijana Milosevic, Keith Cortis, and Brian Davis. 2022. Benchmarking language models for cyberbullying identification and classification from social-media texts. In *Proceedings of the First Workshop on Language Technology and Resources for a Fair, Inclusive, and Safe Society within the 13th Language Resources and Evaluation Conference*, pages 26–31, Marseille, France. European Language Resources Association.

Kunze Wang, Dong Lu, Caren Han, Siqu Long, and Josiah Poon. 2020. Detect all abuse! toward universal abusive language detection models. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6366–6376, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Michael Wiegand, Melanie Siegel, and Josef Ruppenhofer. 2018. Overview of the GermEval 2018 shared task on the identification of offensive language.

Marcos Zampieri, Shervin Malmasi, Preslav Nakov, Sara Rosenthal, Noura Farra, and Ritesh Kumar. 2019. SemEval-2019 task 6: Identifying and categorizing offensive language in social media (OffensEval). In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 75–86, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

Marcos Zampieri, Preslav Nakov, Sara Rosenthal, Pepa Atanasova, Georgi Karadzhov, Hamdy Mubarak, Leon Derczynski, Zeses Pitenis, and Çağrı Çöltekin. 2020. SemEval-2020 task 12: Multilingual offensive language identification in social media (OffensEval 2020). In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 1425–1447, Barcelona (online). International Committee for Computational Linguistics.

Marcos Zampieri, Sara Rosenthal, Preslav Nakov, Alphaeus Dmonte, and Tharindu Ranasinghe. 2023. Offenseval 2023: Offensive language identification in the age of large language models. *Natural Language Engineering*, 29(6):1416–1435.

Ziqi Zhang, David Robinson, and Jonathan Tepper. 2018. Detecting hate speech on twitter using a convolution-gru based deep neural network. In *The Semantic Web: 15th International Conference, ESWC 2018, Heraklion, Crete, Greece, June 3–7, 2018, Proceedings 15*, pages 745–760. Springer.

Li Zhou, Laura Cabello, Yong Cao, and Daniel Hershcovich. 2023. Cross-cultural transfer learning for Chinese offensive language detection. In *Proceedings of the First Workshop on Cross-Cultural Considerations in NLP (C3NLP)*, pages 8–15, Dubrovnik, Croatia. Association for Computational Linguistics.

Yanling Zhou, Yanyan Yang, Han Liu, Xiufeng Liu, and Nick Savage. 2020. Deep learning based fusion approach for hate speech detection. *IEEE Access*, 8:128923–128929.

Steven Zimmerman, Udo Kruschwitz, and Chris Fox. 2018. Improving hate speech detection with deep learning ensembles. In *Proceedings of the eleventh international conference on language resources and evaluation (LREC 2018)*.

Barret Zoph, Deniz Yuret, Jonathan May, and Kevin Knight. 2016. Transfer learning for low-resource neural machine translation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1568–1575, Austin, Texas. Association for Computational Linguistics.