Cross-Cultural Transfer Learning for Chinese Offensive Language Detection

Li Zhou[1][2], **Laura Cabello**[2], **Yong Cao**[2][3], **Daniel Hershcovich**[2]
[1]University of Electronic Science and Technology of China
[2]Department of Computer Science, University of Copenhagen
[3]Huazhong University of Science and Technology
li_zhou@std.uestc.edu.cn, {lcp,dh}@di.ku.dk, yongcao_epic@hust.edu.cn

Abstract

Detecting offensive language is a challenging task. Generalizing across different cultures and languages becomes even more challenging: besides lexical, syntactic and semantic differences, pragmatic aspects such as cultural norms and sensitivities, which are particularly relevant in this context, vary greatly. In this paper, we target Chinese offensive language detection and aim to investigate the impact of transfer learning using offensive language detection data from different cultural backgrounds, specifically Korean and English. We find that culture-specific biases in what is considered offensive negatively impact the transferability of language models (LMs) and that LMs trained on diverse cultural data are sensitive to different features in Chinese offensive language detection. In a few-shot learning scenario, however, our study shows promising prospects for non-English offensive language detection with limited resources. Our findings highlight the importance of cross-cultural transfer learning in improving offensive language detection and promoting inclusive digital spaces.

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

1 Introduction

The proliferation of offensive language and hate speech in online platforms, especially on social media, has significantly increased in recent years (Zampieri et al., 2019, 2020; Gao et al., 2020). There is a fine line between offensive language and hate speech as few universal definitions exist (Davidson et al., 2017). Therefore, hate speech can be classified as a subtype of offensive language. In this paper, we do not differentiate them in detail, and instead, refer to the task of offensive language detection (OLD).

Despite numerous breakthroughs in the development of NLP methods for OLD (Liu et al., 2022; Rusert et al., 2022), some significant obstacles remain unsolved (Vidgen et al., 2019), including the shortage of data resources for research purposes and bias in human annotation. Since most of the available approaches and resources for OLD are designed for English (Arango Monnar et al., 2022), the resulting trained models operate within a monocultural background that caters to English speakers.[1] However, Schmidt and Wiegand (2017) believe that OLD has strong cultural implications, unlike other NLP tasks, because an utterance's offensiveness can vary based on an individual's cultural background.

People with different backgrounds react to inputs differently and communicate differently, so their tolerance for the presence of offensive terms, e.g., slur, may differ, as well as what is altogether considered offensive (Jay and Janschewitz, 2008). Cultural differences have been explored in humor perception (Jiang et al., 2019), swearing reception (Pavesi and Zamora, 2022), translation in semantic inconsistencies (Sperber et al., 1994) and honorifics expression (Song, 2015; Liu and Kobayashi, 2022). Even in less obvious cases, however, they bear meaningful significance on how to pose and solve NLP tasks, as cultures differ with respect to style, values, common ground and topics of interest (Hershcovich et al., 2022).

Therefore, we argue that there is a need for addressing cross-cultural aspects in offensive language detection. Although culture is intricate and challenging to define clearly, language still remains as one of the most straightforward manifestation of culture. While recent work (Ringel et al., 2019; Ranasinghe and Zampieri, 2021) has demonstrated the effectiveness of cross-lingual transfer learning

[1] Importantly, "culture" is multifaceted and complex. When referring to English speakers, we assume that there are general unique features that characterize them, but of course there is enormous diversity within speakers of the same language. As a first step towards the analysis of cross-cultural OLD, we restrict ourselves to the level of language categories.

Proceedings of the First Workshop on Cross-Cultural Considerations in NLP (C3NLP), pages 8–15
May 5, 2023 ©2023 Association for Computational Linguistics

Dataset	Language	Train	Dev	Test
COLD	Chinese	25726 (12723:13003=0.98)	6431 (3211:3220=1.00)	5323 (2107:3216=0.66)
KOLD	Korean	24257 (12190:12067=1.01)	8086 (4076:4010=1.02)	8086 (4044:4022=1.01)
HatEn	English	9000 (3782:5217=0.72)	1000 (427:573=0.75)	3000 (2343:657=3.57)
Region		8449	2104	2087
Gender		6579	1657	1551
Race		10698	2670	1685

Table 1: Datasets statistics (**top**) and topic distributions of COLD (**bottom**). Particularly, statistics of offensive and non-offensive data and the ratio between them are indicated in **parentheses**.

in the text classification and offensive Language (hate speech) detection, they don't consider the impact of cultural background differences (e.g., Eastern and Western culture). In this paper, we take a step forward in this direction and explore the influence of offensive content from diverse cultural background on OLD, focusing on evaluation in Chinese.

Our contributions are as follows: 1) We explore the impact of transfer learning using offensive language data from different cultural backgrounds on Chinese offensive language detection (§3). 2) We find cultural differences in offensive language are expressed in the text topics, and that LMs are sensitive to these differences, learning culture-specific biases that negatively impact their transfer ability (§4). 3) We find that in the few-shot scenario, even with very limited Chinese examples, the model quickly adapts to the target culture.

2 Related work

Offensive language detection. Although most of the research on OLD has focused on English (Fortuna and Nunes, 2018), there exist datasets in multiple languages: Chinese (Deng et al., 2022), Korean (Jeong et al., 2022), Danish (Sigurbergsson and Derczynski, 2020), Bengali (Das et al., 2022), and Nepali (Niraula et al., 2021), to name a few. However, language models commonly rely on prior distributions from training data, that reflects a discourse that is temporally and culturally situated (Ghosh et al., 2021). In a comprehensive analysis of geographically-related content and its influence on performance disparities of offensive language detection models, Lwowski et al. (2022) find that current models do not generalize across locations. Sap et al. (2022) call for contextualizing offensive (toxicity) labels in social variables as determining what is toxic is subjective, and annotator beliefs can be reflected in the data collected.

Cross-lingual transfer learning. Cross-lingual transfer appears as a potential solution to the issue of language-specific resource scarcity (Lamprinidis et al., 2021). Nozza (2021) demonstrates the limits of cross-lingual zero-shot transfer for hate speech detection in English, Italian and Spanish. The benefits of few-shot learning is evident in works from Stappen et al. (2020) and Röttger et al. (2022), who confirmed the effectiveness of few-shot learning for the task of hate speech detection in under-resourced languages. Ringel et al. (2019) harness cross-cultural differences for English formality and sarcasm detection based on German and Japanese, respectively. Litvak et al. (2022) show that, in the context of OLD, knowledge transfer is not bidirectional and efficient transfer learning holds from Arabic to Hebrew in terms of recall.

3 Method

3.1 Datasets

To explore the influence of different cultural backgrounds on Chinese OLD, the most straightforward approach is to adopt OLD datasets whose context and annotation process reflect diverse cultural backgrounds. We first select COLD (Deng et al., 2022), a Chinese benchmark dataset covering the topics of racial, gender, and regional bias as our test dataset. We then select two other datasets that will be used in different training scenarios (see § 3.2): KOLD (Jeong et al., 2022), a Korean dataset suited for OLD covering topics such as race, gender, political affiliation and religion; and HatEn, the

English subset of HatEval (Basile et al., 2019) composed of tweets which tends to capture a Western cultural background. Table 1 reports the statistics of the three datasets and the topic distributions of COLD. Notably, the three languages come from three different language families, making linguistic similarities between them less likely to be a factor in effective transfer learning between the datasets.

3.2 Learning settings

We explore different learning settings by utilizing **intra-cultural** and **cross-cultural** training sets during fine-tuning. For the intra-cultural setting, we only use COLD as the training set, which ensures cultural consistency in the training and testing process. In the cross-cultural setting, we further set up two ways: 1) *zero-shot*: only use KOLD or HatEn as the training set, which makes the fine-tuning process of LMs come from completely different cultural backgrounds; 2) *mix-training few-shot*: mix COLD with another language (KOLD or HatEn) as the final training set, which introduces cultural interference and makes the acquisition of the target culture more challenging. For convenience, we use $\mathcal{D}\left[X\right]$ to represent the detector with X as training set. Since the datasets are in different languages, we apply multilingual LMs in these experiments.

Translated data setting. As an additional control experiment, to avoid the difference from the language itself, we also translate COLD and KOLD into English with *googletrans*[2] and conduct experiments with *English* PLMs under the same settings.

4 Experiments

Implementation. In our experiments, we only evaluate on COLD and try different training settings with COLD, KOLD and HatEn. In particular, because the data volume of HatEn is relatively small, we use all of its data as the training set. The actual training set of three datasets has offensive data to non-offensive data ratios of 0.98, 1.01, and 1.02 (refer to Table 1). In the cross-cultural zero-shot setting, we also randomly sample 13,000 examples[3] from the Korean training set to ensure the consistency of the training data sizes with HatEn. For the multilingual LMs, we choose mBERT$_{\texttt{base}}$ (Devlin et al., 2019), XLM-R$_{\texttt{base}}$ and XLM-R$_{\texttt{large}}$ (Conneau et al., 2020). In the translated data setting, we apply the English models BERT$_{\texttt{base}}$ (Devlin et al., 2019), RoBERTa$_{\texttt{base}}$ and RoBERTa$_{\texttt{large}}$ (Liu et al., 2019).

Model	Train Set	Test F1	Test ACC
mBERT$_{\texttt{base}}$	COLD	77.90±0.25	80.86±0.26
	CO+KO	78.23±0.05*	81.16±0.19
	CO+HE	78.19±0.18*	81.07±0.10
	KOLD	49.27±4.04**	67.85±0.70**
	KOLD†	50.34±3.49**	69.47±0.71**
	HatEn	35.96±3.95**	63.54±0.54**
XLM-R$_{\texttt{base}}$	COLD	78.77±0.27	81.51±0.20
	CO+KO	78.90±0.10	81.78±0.15*
	CO+HE	78.96±0.15	81.66±0.18
	KOLD	58.13±1.78**	72.14±0.67**
	KOLD†	60.86±1.44**	72.93±0.37**
	HatEn	29.84±2.07**	63.36±0.90**
XLM-R$_{\texttt{large}}$	COLD	79.09±0.24	81.87±0.16
	CO+KO	79.76±0.19**	82.45±0.19**
	CO+HE	79.43±0.22*	82.16±0.26**
	KOLD	63.48±1.63**	74.45±0.34**
	KOLD†	61.71±2.37**	74.09±0.80**
	HatEn	28.94±2.50**	63.76±0.40**

Table 2: Overall results on COLD test set. † marks KOLD training set is the same size as HatEn. CO, KO and HE are short for COLD, KOLD and HatEn respectively. By conducting Paired Student's t-test, $*$ = differs significantly from intra-cultural at $p < 0.05$, $**$ = significant difference at $p < 0.01$.

Our models are optimized with a learning rate of $5e-5$. We fine-tune each model for 100 epochs using early-stopping with a patience of 5, and run 5 times with different random seeds for each setting.

Overall results. The experimental results on COLD test set are shown in Table 2.[4] Compared to the intra-cultural setting, we find that: 1) In the cross-cultural few-shot scenario, the performance differences between $\mathcal{D}\left[\text{COLD}\right]$ and $\mathcal{D}\left[\text{CO}+\text{KO}\right]$, $\mathcal{D}\left[\text{COLD}\right]$ and $\mathcal{D}\left[\text{CO}+\text{HE}\right]$ are both very small (less than one point at the maximum), which implies that with sufficient knowledge of the Chinese target culture, the intervention of other cultures does not diminish the ability to detect Chinese offensive language, but has a slight contribution. 2) In the cross-cultural zero-shot scenario, the detection ability of $\mathcal{D}\left[\text{KOLD}\right]$ and $\mathcal{D}\left[\text{HatEn}\right]$ get worse. In particular, the former is slightly better than the latter. This implies that it is easier to detect Chinese offensive language in Korean cultural background compared to a Western cultural background.

[2] https://pypi.org/project/googletrans/

[3] The ratio of offensive data to non-offensive data is 0.96.

[4] We only report the test set score, because only the test set of COLD is annotated manually, and the training and dev sets are labeled semi-automatically.

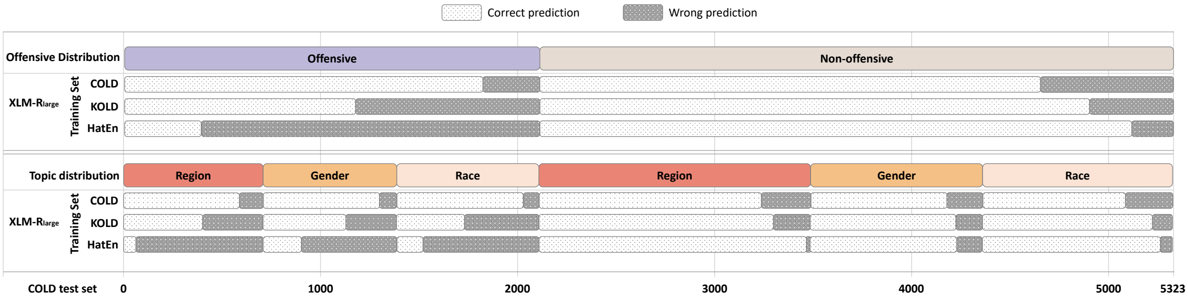

Figure 1: A fine-grained view of the distribution of offensive detection results based on XLM-R$_{\text{large}}$. For reference, the colored part represent the distribution of related data in COLD test set. The model learns culture-specific biases—e.g., when training on English, it tends not to classify region-related text as offensive.

To better understand the detection ability of Chinese offensive language with different cultural backgrounds, we look closer at offensive detection results for the intra-cultural and cross-cultural zero-shot settings. Figure 1 shows the distribution of the data and the predictions from our best performing model XLM-R$_{\text{large}}$. First, $\mathcal{D}$ [COLD], which is in the same cultural background as the test set, has the best ability to detect offense. $\mathcal{D}$ [HatEn] is the worst detector, with less than 50% accuracy for offensive data. Because of this, it can be highly accurate in non-offensive data. This is why $\mathcal{D}$ [HatEn] gets a spurious high accuracy on the test set but a very low F1 score (Table 2). However, it is noteworthy that the HatEn-trained model requires more severe language to be labeled as offensive,[5] so some instances that should be classified as offensive, may not be considered hate speech and will not be classified as such. Moreover, for specific-topic offensive language detection, the performance of each detector is also different, with $\mathcal{D}$ [HatEn] performing the worst in the regional topic.

Translated results. For the experiments of the translated version of the Chinese and Korean datasets into English. The experimental results are shown in Table 3, showing similar trends to the results in Table 2. This demonstrates that the results hold for cross-cultural transfer and are not simply due to linguistic similarities.

Few-shot learning. While the diverse cultural backgrounds of Korean and English may not enable precise detection of Chinese offensive language in a zero-shot scenario, it is not detrimental when integrated into the target culture in a few-shot scenario. Therefore, when mixing heterogeneous cultural background knowledge, is it necessary to provide sufficient target cultural background knowledge? To investigate this problem, we conduct an analytical experiment under a few-shot setting by incorporating different scales of COLD data into the training set. Figure 2 displays experimental results indicating that the correlation between the ability to detect offensive language and target cultural knowledge follows a pattern similar to that of an increasing logarithmic function. This implies that offensive language detection performance improves rapidly with limited target cultural knowledge acquisition, but gradually slows down as the

Model	**Train Set**	**Test F1**	**Test ACC**
BERT$_{\text{base}}$	COLD	77.59±0.41	80.67±0.37
	CO+KO	77.86±0.19*	80.90±0.20
	CO+HE	77.50±0.17*	80.47±0.18
	KOLD	61.84±1.46**	71.26±0.34**
	KOLD†	61.64±1.06**	71.21±0.27**
	HatEn	21.20±1.36**	61.53±0.21**
RoBERTa$_{\text{base}}$	COLD	77.89±0.46	81.01±0.40
	CO+KO	78.25±0.40	81.35±0.37*
	CO+HE	78.08±0.34	81.12±0.25
	KOLD	63.85±1.12**	73.60±0.43**
	KOLD†	63.47±0.84**	73.21±0.25**
	HatEn	26.09±2.82**	62.81±0.36**
RoBERTa$_{\text{large}}$	COLD	78.22±0.40	81.24±0.33
	CO+KO	78.74±0.21**	81.70±0.15**
	CO+HE	78.24±0.30*	81.17±0.25**
	KOLD	65.56±1.16**	73.70±0.49**
	KOLD†	64.39±1.60**	73.71±0.37**
	HatEn	26.69±1.38**	63.20±0.44**

Table 3: The experimental results on the COLD test set, with all training and testing data translated to English. † marks KOLD training set is the same size as HatEn. By conducting Paired Student's t-test, $*$ = differs significantly from intra-cultural at $p < 0.05$, $**$ = significant difference at $p < 0.01$.

[5] This could be a reason to treat Hate Speech Detection as a separate task, contrary to our simplified view here.

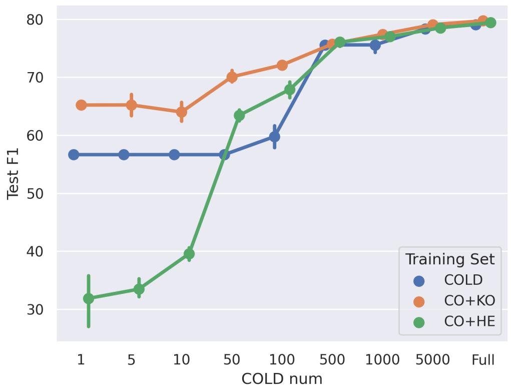

Figure 2: The experimental results (F1) in few-shot setting based on XLM-R$_{\texttt{large}}$, evaluated on the COLD (Chinese) test set. Performance improves rapidly with training examples from the target culture. Pre-training on KOLD (Korean) provides a better starting point, while pre-training on HatEn (English) is detrimental.

amount of target knowledge increases. Specifically, when the training focuses on COLD within the range of 1 to 50, $\mathcal{D}$ [COLD] possesses limited knowledge of the training concentration, and its detection capability stems primarily from the pre-training model itself. At this stage, HatEn has a clearly negative effect, while KOLD has a positive effect. Within the range of 50 to 500, both HatEn and KOLD have an obvious positive effect, while for COLD data scales greater than 500, the effect is still present but less pronounced. These findings offer promising opportunities for low-resource offensive language detection systems.

Case study. To provide an intuitive explanation of cultural differences, we use semantic similarity retrieval (Reimers and Gurevych, 2019) to find the most similar cases from KOLD to COLD with the similarity threshold set to 0.7. As depicted in Table 4, sentences with similar topics and semantics (e.g. racial discrimination, politics) hold different labels among languages, suggesting the presence of cultural distinctions in offensive language detection and highlighting the significant obstacles for few-shot learning. Thus, we emphasize the necessity of greater cultural adaptation models that can integrate diverse cultural knowledge.

Chinese	Korean	Labels
黑人反对歧视黑人有啥错? *What is wrong with blacks against discrimination against blacks?*	흑인 대통령도 나온 미국, 이제 인종차별은 사라졌다? *America with a black president, now racism has disappeared?*	0 / 1
中国哪有那么容易搞到毒品? *How can it be so easy to get drugs in China.*	중국에서 범은 잡히면 뭐가 잘릴까.. *What will happen if a criminal is caught in China?*	0 / 1

Table 4: Cases with reversed labels through semantic vector retrieval were listed, suggesting the existence of cultural differences across languages. Non-offensive and offensive cases are labeled as 0 and 1.

5 Conclusion

Our study highlights the challenges of detecting offensive language across different cultures and languages. We show that transfer learning using data from diverse cultural backgrounds have different negative effects on the transferability of language models due to culture-specific biases. However, our findings also indicate promising prospects for improving offensive language detection in promoting inclusive digital spaces, particularly in a few-shot learning scenario. We call for more research on cross-cultural offensive language detection, which is important to deploy effective moderation strategies for social media platforms, improving cross-cultural communication, and reducing harmful online behavior.

Limitations

Our study explores the impact of transfer learning on offensive language detection using data from different cultural backgrounds. However, treating HatEn as representative of “Western cultural backgroun” is too vague, as it ignores the cultural differences between American and British cultures. Moreover, “culture” is multifaceted and complex, and there is enormous diversity among speakers of the same language. To focus on language categories, we limit our analysis to a first step towards cross-cultural offensive language detection.

Ethics Statement

The datasets used in this study are publicly available, and we strictly follow the ethical implications of previous research related to the data sources. It is important to note that the content of these datasets does not represent our opinions or views.

Acknowledgments

Thanks to the anonymous reviewers for their helpful feedback. The authors gratefully acknowledge financial support from China Scholarship Council. (CSC No. 202206070002 and No. 202206160052).

References

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

Valerio Basile, Cristina Bosco, Elisabetta Fersini, Debora Nozza, Viviana Patti, Francisco Manuel Rangel Pardo, Paolo Rosso, and Manuela Sanguinetti. 2019. SemEval-2019 task 5: Multilingual detection of hate speech against immigrants and women in Twitter. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 54–63, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Mithun Das, Somnath Banerjee, Punyajoy Saha, and Animesh Mukherjee. 2022. Hate speech and offensive language detection in Bengali. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 286–296, Online only. Association for Computational Linguistics.

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

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

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Paula Fortuna and Sérgio Nunes. 2018. A survey on automatic detection of hate speech in text. *ACM Comput. Surv.*, 51(4).

Zhiwei Gao, Shuntaro Yada, Shoko Wakamiya, and Eiji Aramaki. 2020. Offensive language detection on video live streaming chat. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 1936–1940, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Sayan Ghosh, Dylan Baker, David Jurgens, and Vinodkumar Prabhakaran. 2021. Detecting cross-geographic biases in toxicity modeling on social media.

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.

Timothy Jay and Kristin Janschewitz. 2008. The pragmatics of swearing. 4(2):267–288.

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.

Tonglin Jiang, Hao Li, and Yubo Hou. 2019. Cultural differences in humor perception, usage, and implications. *Frontiers in Psychology*, 10.

Sotiris Lamprinidis, Federico Bianchi, Daniel Hardt, and Dirk Hovy. 2021. Universal joy a data set and results for classifying emotions across languages. In *Proceedings of the Eleventh Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 62–75, Online. Association for Computational Linguistics.

Marina Litvak, Natalia Vanetik, Chaya Liebeskind, Omar Hmdia, and Rizek Abu Madeghem. 2022. Offensive language detection in Hebrew: can other languages help? In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 3715–3723, Marseille, France. European Language Resources Association.

Jiexi Liu, Dehan Kong, Longtao Huang, Dinghui Mao, and Hui Xue. 2022. Multiple instance learning for offensive language detection. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 7387–7396, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Muxuan Liu and Ichiro Kobayashi. 2022. Construction and validation of a Japanese honorific corpus based on systemic functional linguistics. In *Proceedings of the Workshop on Dataset Creation for Lower-Resourced Languages within the 13th Language Resources and Evaluation Conference*, pages 19–26, Marseille, France. European Language Resources Association.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Brandon Lwowski, Paul Rad, and Anthony Rios. 2022. Measuring geographic performance disparities of offensive language classifiers. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 6600–6616, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Nobal B. Niraula, Saurab Dulal, and Diwa Koirala. 2021. Offensive language detection in Nepali social media. In *Proceedings of the 5th Workshop on Online Abuse and Harms (WOAH 2021)*, pages 67–75, Online. Association for Computational Linguistics.

Debora Nozza. 2021. Exposing the limits of zero-shot cross-lingual hate speech detection. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 907–914, Online. Association for Computational Linguistics.

Maria Pavesi and Pablo Zamora. 2022. The reception of swearing in film dubbing: a cross-cultural case study. *Perspectives*, 30(3):382–398.

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.

Dor Ringel, Gal Lavee, Ido Guy, and Kira Radinsky. 2019. Cross-cultural transfer learning for text classification. 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 3873–3883, Hong Kong, China. Association for Computational Linguistics.

Paul Röttger, Debora Nozza, Federico Bianchi, and Dirk Hovy. 2022. Data-efficient strategies for expanding hate speech detection into under-resourced languages. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 5674–5691, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Jonathan Rusert, Zubair Shafiq, and Padmini Srinivasan. 2022. On the robustness of offensive language classifiers. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7424–7438, Dublin, Ireland. Association for Computational Linguistics.

Maarten Sap, Swabha Swayamdipta, Laura Vianna, Xuhui Zhou, Yejin Choi, and Noah A. Smith. 2022. Annotators with attitudes: How annotator beliefs and identities bias toxic language detection. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5884–5906, Seattle, United States. Association for Computational Linguistics.

Anna Schmidt and Michael Wiegand. 2017. A survey on hate speech detection using natural language processing. In *Proceedings of the Fifth International Workshop on Natural Language Processing for Social Media*, pages 1–10, Valencia, Spain. 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.

Sanghoun Song. 2015. Representing honorifics via individual constraints. In *Proceedings of the Grammar Engineering Across Frameworks (GEAF) 2015 Workshop*, pages 57–64, Beijing, China. Association for Computational Linguistics.

Ami D. Sperber, Robert F. Devellis, and Brian Boehlecke. 1994. Cross-cultural translation: Methodology and validation. *Journal of Cross-Cultural Psychology*, 25(4):501–524.

Lukas Stappen, Fabian Brunn, and Björn Schuller. 2020. Cross-lingual zero- and few-shot hate speech detection utilising frozen transformer language models and axel. *ArXiv*, abs/2004.13850.

Bertie Vidgen, Alex Harris, Dong Nguyen, Rebekah Tromble, Scott Hale, and Helen Margetts. 2019. Challenges and frontiers in abusive content detection. In *Proceedings of the Third Workshop on Abusive Language Online*, pages 80–93, Florence, Italy. Association for Computational Linguistics.

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.