Cultural Diversity Enhances Offensive Language Detection in Multilingual Models

Anonymous ACL submission

Abstract

The proliferation of offensive online content across diverse languages necessitates culturally-aware NLP solutions. While Cross-Lingual Transfer Learning (CLTL) shows promise in other NLP tasks, its application to offensive language detection overlooks crucial cultural nuances in how offensiveness is perceived. This work investigates the effectiveness of CLTL for offensive language detection, considering both linguistic and cultural factors. Specifically, we investigated transfer learning across 105 language pairs, and uncovered several key findings. Firstly, training exclusively on English data impedes performance in certain target languages. Secondly, linguistic proximity between languages does not have a significant impact on transferability. Lastly, there is a significant correlation between cultural distance and performance. Importantly, for each unit increase of cultural distance, there was an increase of 0.3 in the AUC. These findings emphasize the limitations of English-centric approaches and highlight the need to integrate cultural context into NLP solutions for offensive language detection.

1 Introduction

In recent years, the escalating prevalence of offensive language on prominent social media platforms such as Facebook and Twitter has emerged as a significant and pressing concern. The landscape of online discourse has been further complicated with the introduction of content generated by language models (United States Senate Committee on the Judiciary, Jan 31st, 2024; Atlantic-Council, 2023). Within the NLP community, extensive research efforts have been dedicated to developing resources and methodologies for detecting offensive content (See Yin and Zubiaga, 2021, for a review). Initial endeavors were predominantly concentrated on monolingual settings, with the majority of the research focusing on the English language (Vidgen and Derczynski, 2020). However, recently, the trajectory of research has shifted towards addressing the challenge of offensive language detection in other languages or in multilingual settings (Al-Hassan and Al-Dossari, 2019). This shift, however, is hindered by the constrained availability of labeled data and the considerable variability in what constitutes offensive language across diverse cultures and languages (Röttger et al., 2022b).

In numerous NLP tasks, Cross-Lingual Transfer Learning (CLTL) has emerged as a promising avenue for addressing challenges related to data scarcity. CLTL leverages domain knowledge from high-resource languages to benefit low-resource languages. However, the application of many CLTL methods to offensive language detection has proven less successful (Nozza, 2021). The intricate linguistic structures and cultural variations across languages pose significant challenges for CLTL (Jiang and Zubiaga, 2024). Davani et al. (2023) emphasize the pivotal role of cultural and psychological factors in determining what is deemed offensive. Despite this recognition, a considerable portion of recent studies overlook the significance of cultural context and advocate a one-size-fits-all solution, using English data to enhance the performance of offensive language classifiers in low-resource languages (Röttger et al., 2022a). Consequently, as demonstrated in recent findings by Lee et al. (2023), hate speech classifiers are culturally insensitive.

In this study, we systematically investigate the influence of linguistic and cultural similarities on the cross-lingual transferability of hate speech and offensive language detection. Contrary to previous suggestions, we observe that training on English corpora before delving into offensive language detection in a different target language leads to diminished performance in certain cases (section 4). Furthermore, we find that including culturally diverse datasets in the first stage of CLTL significantly improves the performance of target languages in low-resource settings (section 5).

Based on our findings, we advocate for CLTL methods that leverage cultural diversity. Our results suggest that the model's exposure to culturally diverse datasets not only broadens the model's cultural repertoire but also increases its ability to precisely identify offensive content across different languages. Our detailed analysis of cross-lingual transfer learning across 15 languages, and 105 language pairs, aims to disentangle the respective roles of linguistic and cultural similarities between datasets on cross-lingual transferability among them. This work underscores the necessity of moving beyond English-centric approaches and integrating cultural context into NLP solutions for offensive language detection.

2 Background

2.1 Cross lingual Transfer Learning

The primary objective in CLTL for offensive language detection is to leverage knowledge from a language with existing resources (i.e., the auxiliary language) to enhance the effectiveness of offensive language detection in a language with limited resources (i.e., the target language). Various methods have been proposed for CLTL of offensive language detection. These approaches can be broadly categorized as instance transfer, feature transfer, and parameter transfer (Jiang and Zubiaga, 2024).

Instance transfer involves approaches that transfer either the labels (e.g., via label projection) or the text (e.g., via translation) to the new language. Translation approaches, however, may be prone to errors, possibly neglecting cultural nuances and resulting in translations inconsistent with the original language (Das et al., 2022). Feature transfer methods focus on using latent representations of texts (e.g., multilingual embeddings) to transfer knowledge from the source to the target language. However, Nozza (2021) demonstrated that multilingual embeddings exhibit poor generalization across languages when lacking training data in the target language. Finally, parameter transfer approaches use the parameters of a model trained on an auxiliary language to enhance performance on the target language. An essential element in parameter transfer approaches is the choice of target and auxiliary languages. Since cultural factors can influence language use, connotations, and perceptions of offensiveness, it becomes crucial to systematically investigate their impact on CLTL approaches.

2.2 Culture, Language, and Offensiveness

Culture broadly encompasses a range of "good-enough" solutions that each society has developed to address survival problems (Oyserman, 2011), often operationalized as causally distributed patterns of mental representations across a population (Atran et al., 2005). Cultural solutions manifest in a diverse array of beliefs, values, norms, and practices (Boyd and Richerson, 2005).

One of the dimensions of cultural differences is *individualism vs. collectivism* (Triandis, 2018). Individualistic cultures emphasize values of autonomy, distinction, and the pursuit of uniqueness. In contrast, collectivistic cultures prioritize unity, conformity, communal harmony, and mutual responsibility (Oyserman, 2017; Markus and Kitayama, 2010). A critical domain where individualistic and collectivistic cultures diverge is in perceptions of offensiveness, including the nature of offenses, the intensity of emotional reactions they provoke, and views on suitable retribution (Maitner et al., 2017). Collectivistic cultures perceive offenses against communal entities such as national symbols, religious beliefs, or family honor as grave threats to social unity (Kim et al., 2008). Conversely, in individualistic cultures, offenses against an individual's achievements, professional reputation (Günsoy et al., 2023), or personal identity, like gender or sexual orientation, are taken with equal gravity.

The individualism vs. collectivism difference, while providing valuable insights into the cultural psychology of offense, fails to account for other dimensions of cultural differences such as a society's tolerance for norm violations, known as the *tightness–looseness* dimension (Gelfand et al., 2011), which influences how people perceive and react to offensive language.

In recent years, cultural psychologists have introduced a new comprehensive index for quantifying cultural differences, known as the *WEIRDness* score (Muthukrishna et al., 2020). "WEIRD", in this context, stands for "Western, Educated, Industrialized, Rich, and Democratic" (Henrich et al., 2010). This index is a composite score derived from several measures of cultural differences, including Hofstede's (Hofstede, 2001) cultural dimensions (which encompass, among others, individualism-collectivism scores), the tightness–looseness, dimension, Schwartz's values (Schwartz, 2006), and a range of other psychological and behavioral measures. The WEIRDness

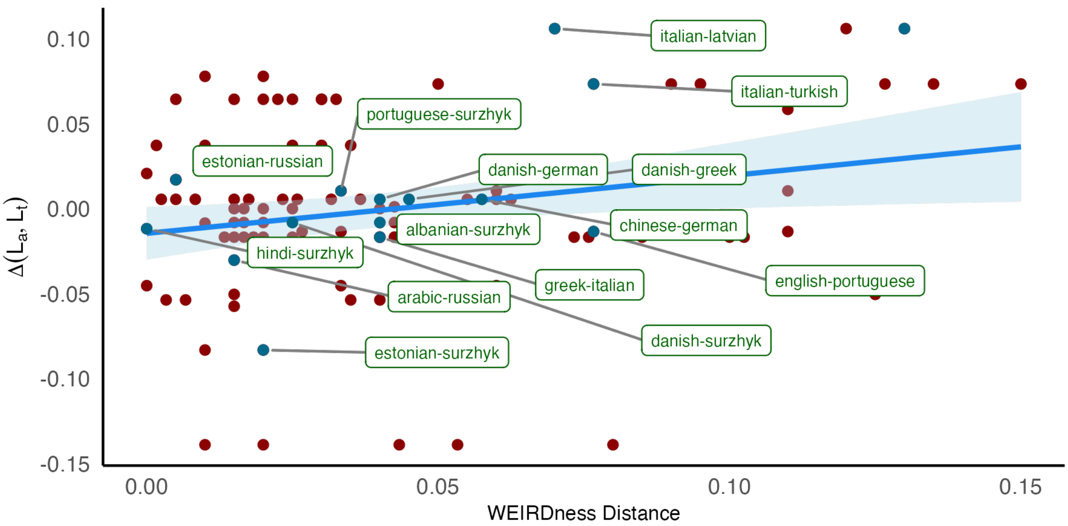

Figure 1: The relationship between cultural differences, as measured by the WEIRDness distance between L_a and L_t, and $\Delta(L_a, L_t)$. The regression line, derived from Equation 1, indicates that the WEIREDness distance predicts CLTL performance gains ($\beta = 0.3$, $p = 0.02$).

score is a quantitative measure designed to assess the cultural distance of a country to the U.S., which is considered a quintessential WEIRD nation (For a more in-depth discussion refer to Section B). Countries that align closely with the characteristics of the U.S. are deemed more WEIRD, while those diverging from the US traits are labeled as more non-WEIRD. Past cross-cultural evidence documents how WEIRDness can reliably predict a multitude of psychological variances across nations such as differences in moral values and the perception and interpretation of hate content among different populations (Henrich et al., 2010; Atari et al., 2023). Previous studies have indicated that individuals from WEIRD countries tend to classify fewer items as offensive, particularly when China is excluded from the analysis (Davani et al., 2023).

Linguistic similarity is another crucial factor in understanding cross-lingual transferability of offensive language detection due to its potential impact on the effectiveness of multilingual models. Languages vary not only in vocabulary but also in syntax, semantics, and phonetics, and various approaches have been proposed to quantify the similarity between languages (ten Thije and Zeevaert, 2007; Maedche et al., 2002; Gomaa et al., 2013). To measure linguistic similarity, we adopt a data-driven approach for language comparison, emphasizing the identification of cognates through computational analysis of phonetic data, especially consonants (eLinguistics C., 2020). This method applies phonological rules to systematically identify potential cognates. An advanced scoring system evaluates the similarity between languages at multiple levels, from phonetics to broader structures. Finally, statistical analysis of cognate scores ensures the validity and reliability of the language-relatedness findings, distinguishing true linguistic connections from coincidental similarities. For a comparative analysis between available indices, and the rationale behind our choice of linguistic similarity, see Appendix C.

3 Experimental Setup

Our goal is to investigate how linguistic and cultural differences affect cross-lingual transferability of offensive language detection. Let M_θ denote a pretrained multilingual language model M parameterized by θ and let L_a and L_t denote auxiliary and target languages, respectively. Let f_{L_t} and f_{L_a} denote the offensive language detection models that were initialized with M_θ and have only been trained on data from the target and auxiliary languages. Furthermore, let $f_{L_a \to L_t}$ denote the cross-lingual transfer model that has two training stages: In the first stage, M_θ has been trained on the auxiliary language to get f_{L_a}. Then in the second stage, f_{L_a} has been fine-tuned on data from the target language. The overall goal in CLTL is to maximize the performance gains resulting from the first stage of training formally defined as

$$\Delta(L_a \to L_t) = \mathrm{AUC}(f_{L_a \to L_t}) - \mathrm{AUC}(f_{L_a})$$

where AUC(.) is used to denote the area under the

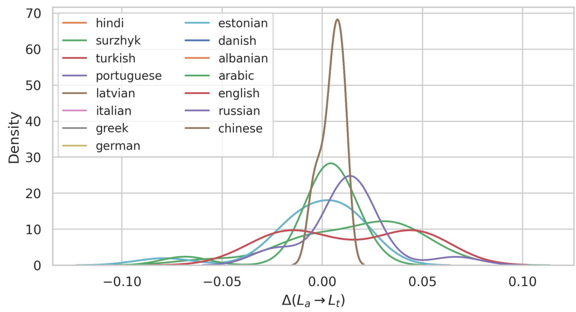

Figure 2: Distribution of $\Delta(L_a \rightarrow L_t)$ by auxiliary language L_a. Most languages exhibit both positive and negative impacts on CLTL, underscoring the significance of considering cultural factors when choosing L_a.

operating characteristic curve of a model on the test set form L_t. We use $\Delta(L_a, L_t)$ to denote the average of $\Delta(L_a \rightarrow L_t)$ and $\Delta(L_t \rightarrow L_a)$. In Section 4, we assess if English (or any auxiliary language) universally guarantees positive transfer ($\Delta(L_a \rightarrow L_t) > 0$). Subsequently, in Section 5 to quantify how cultural and linguistic differences between the L_a and L_t influence $\Delta(L_a, L_t)$, we rely on the following regression:

$$\begin{aligned}\Delta(L_a, L_t) = \beta_0 + \beta_1 \times \Delta_{\text{WEIRDness}}(L_a, L_t) \\ + \beta_2 \times \Delta_{\text{Language}}(L_a, L_t) + \epsilon\end{aligned} \quad (1)$$

where $\Delta_{\text{WEIRDness}}(L_a, L_t)$ denotes the difference in WEIRDness score of L_a and L_t (Muthukrishna et al., 2020), and $\Delta_{\text{Language}}(L_a, L_t)$ denotes the linguistic distance (eLinguistics C., 2020).

We conduct our experiments on 15 languages, namely, Albanian, Danish, English, Estonian, German, Greek, Italian, Latvian, Portuguese, Russian, Turkish, Surzhyk, Chinese, Hindi, and Arabic. More information on the datasets used in this work can be found in Appendix D and Table 1. We split each dataset into a 80/10/10 train, dev, and test split. To control for the differences in dataset size in different languages, we take a stratified sample of a fixed number of instances ($n = 1000$) from each language for the training set. Given that the language model needs to be able to handle data from multiple languages we used XLM-r (Conneau et al., 2020) and trained all model parameters for 10 epochs.

4 English Data Does Not Always Help

Recognizing the pivotal role of cultural factors in shaping perceptions of offensiveness, we reevaluate the one-size-fits-all approach proposed by previous researchers (Röttger et al., 2022a) on a diverse set of languages and cultural backgrounds. Specifically, we test the assumption that employing English as the auxiliary language consistently enhances the performance of the target language (L_t). Our empirical investigation reveals that contrary to this assumption, using English as the auxiliary language results in performance degradation ($\Delta(\text{English} \rightarrow L_t) < 0$) in 40% of the cases. Specifically, we observe diminished performance for Russian, Portuguese, Hindi, Estonian, Latvian, and Italian (Appendix E). As shown in Figure 2 most languages exhibit diverse effects, encompassing both positive and negative impacts on CLTL. This analysis underscores the limitations of English-centric approaches, highlighting the potential of considering cultural factors in offensive language detection.

5 Cultural Diversity Improves Models

To quantify the impact of cultural and linguistic factors on CLTL gains, we conducted a linear regression analysis predicting $\Delta(L_a, L_t)$ based on language difference and WEIRDness difference (see Equation 1). We find evidence that WEIRDness difference significantly predicts CLTL performance gains ($\beta = 0.3$, $p = 0.02$) even after controlling for linguistic factors. Specifically, for each unit increase in the WEIRDness difference, there is an expected increase of 0.3 units in $\Delta(L_a, L_t)$. However, language similarity was not a significant predictor ($p = 0.21$) of $\Delta(L_a, L_t)$. In the model the assumptions of linearity, independence, and normality were met, with a residual standard error of 0.05. Our results imply that training models across languages from diverse cultural contexts could serve as a potential solution to building culturally sensitive models capable of capturing a more accurate reflection of cultural nuances.

6 Conclusion

This study underscores the crucial role of cultural diversity in cross-lingual approaches to offensive language detection. We conducted a systematic examination of the influence of both cultural and linguistic factors on cross-lingual transferability across 15 languages. Interestingly, we find that linguistic proximity does not impact transferability. However, transfer significantly improves when using culturally diverse language pairs. This emphasizes the importance of cultural context in offensive language detection and exposes the shortcomings of relying on English-centric approaches.

7 Limitations

Our study is constrained by the specific languages and datasets chosen for our analysis. We leave further verification of our analysis in different languages and datasets for future work. The language models utilized in our study introduce limitations. Different language models may yield distinct results due to variations in architecture, training data, and underlying algorithms. Consequently, the findings should be interpreted within the context of the chosen language models. The study is based on data available up February 2024. Changes in language usage, cultural trends, or advancements in language models beyond this date are not considered. Consequently, our findings may not reflect the most current linguistic landscape or the latest developments in natural language processing. The accuracy and reliability of our study are contingent upon the quality and availability of the selected datasets. Issues such as data biases, incompleteness, or inaccuracies within the datasets may impact the robustness of our conclusions. Even though our study highlights the significance of incorporating cultural diversity in CLTL for offensive language detection, we do not endorse an approach that disregards universal ethical standards. Recognizing that certain expressions of hate, such as calls for genocide, are universally unacceptable based on the Declaration of Human Rights, our findings advocate for a balanced perspective that respects cultural nuances while upholding global ethics. Acknowledging these limitations is crucial for a nuanced interpretation of our study's findings and encourages future research to address these constraints for a more comprehensive understanding of the broader linguistic landscape.

References

Areej Al-Hassan and Hmood Al-Dossari. 2019. Detection of hate speech in social networks: a survey on multilingual corpus. In *6th international conference on computer science and information technology*, volume 10, pages 10–5121.

Bohdan Andrusyak, Mykhailo Rimel, and Roman Kern. 2018. Detection of abusive speech for mixed sociolects of russian and ukrainian languages. In *The 12th Workshop on Recent Advances in Slavonic Natural Languages Processing, RASLAN 2018, Karlova Studanka, Czech Republic, December 7-9, 2018*, pages 77–84. Tribun EU.

Dennis Assenmacher, Marco Niemann, Kilian Müller, Moritz Seiler, Dennis Riehle, Heike Trautmann, and Heike Trautmann. 2021. Rp-mod & rp-crowd: Moderator- and crowd-annotated german news comment datasets. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, volume 1. Curran.

Mohammad Atari, Jonathan Haidt, Jesse Graham, Sena Koleva, Sean T Stevens, and Morteza Dehghani. 2023. Morality beyond the weird: How the nomological network of morality varies across cultures. *Journal of Personality and Social Psychology*.

Atlantic-Council. 2023. Scaling trust on the web. Technical report, Atlantic Council.

Scott Atran, Douglas L Medin, and Norbert O Ross. 2005. The cultural mind: environmental decision making and cultural modeling within and across populations. *Psychological review*, 112(4):744.

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

Cristina Bosco, Felice Dell'Orletta, Fabio Poletto, Manuela Sanguinetti, and Maurizio Tesconi. 2018. Overview of the evalita 2018 hate speech detection task. In *EVALITA@CLiC-it*.

Robert Boyd and Peter J Richerson. 2005. *The origin and evolution of cultures*. Oxford University Press.

Luigi Luca Cavalli-Sforza, Paolo Menozzi, and Alberto Piazza. 1994. *The history and geography of human genes*. Princeton university press.

Çağrı Çöltekin. 2020. A corpus of Turkish offensive language on social media. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 6174–6184, Marseille, France. European Language Resources Association.

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, and Animesh Mukherjee. 2022. Data bootstrapping approaches to improve low resource abusive language detection for indic languages. In *Proceedings of the 33rd ACM Conference on Hypertext and Social Media*, pages 32–42.

Aida Davani, Mark Díaz, Dylan Baker, and Vinodkumar Prabhakaran. 2023. Disentangling perceptions of offensiveness: Cultural and moral correlates. *arXiv preprint arXiv:2312.06861*.

Ona de Gibert, Naiara Perez, Aitor García-Pablos, and Montse Cuadros. 2018. Hate Speech Dataset from a White Supremacy Forum. In *Proceedings of the*

2nd Workshop on Abusive Language Online (ALW2)*, pages 11–20, Brussels, Belgium. Association for Computational Linguistics.

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.

eLinguistics C. 2020. Quantifying the genetic proximity between languages. Retrieved on February 1, 2024.

Michele J Gelfand, Jana L Raver, Lisa Nishii, Lisa M Leslie, Janetta Lun, Beng Chong Lim, Lili Duan, Assaf Almaliach, Soon Ang, Jakobina Arnadottir, et al. 2011. Differences between tight and loose cultures: A 33-nation study. *science*, 332(6033):1100–1104.

Wael H Gomaa, Aly A Fahmy, et al. 2013. A survey of text similarity approaches. *international journal of Computer Applications*, 68(13):13–18.

A Gorbunova. 2022. GitHub - alla-g/toxicity-detection-thesis: Code and data for my thesis "Automatic toxic comment detection in social media for Russian" — github.com. `https://github.com/alla-g/toxicity-detection-thesis/tree/main`. [Accessed 31-01-2024].

Ceren Günsoy, Susan E Cross, Vanessa A Castillo, Ayse K Uskul, S Arzu Wasti, Phia S Salter, Pelin Gul, Adrienne Carter-Sowell, Afşar Yegin, Betul Altunsu, et al. 2023. Goal derailment and goal persistence in response to honor threats. *Journal of Cross-Cultural Psychology*, 54(3):365–384.

Joseph Henrich, Steven J Heine, and Ara Norenzayan. 2010. The weirdest people in the world? *Behavioral and brain sciences*, 33(2-3):61–83.

Geert Hofstede. 2001. *Culture's consequences: Comparing values, behaviors, institutions and organizations across nations*. sage.

Ronald Inglehart, Miguel Basanez, Jaime Diez-Medrano, Loek Halman, and Ruud Luijkx. 2000. World values surveys and european values surveys, 1981-1984, 1990-1993, and 1995-1997. *Ann Arbor-Michigan, Institute for Social Research, ICPSR version*.

Aiqi Jiang and Arkaitz Zubiaga. 2024. Cross-lingual offensive language detection: A systematic review of datasets, transfer approaches and challenges. *arXiv preprint arXiv:2401.09244*.

Tae-Yeol Kim, Debra L Shapiro, Karl Aquino, Vivien KG Lim, and Rebecca J Bennett. 2008. Workplace offense and victims' reactions: the effects of victim-offender (dis) similarity, offense-type, and cultural differences. *Journal of Organizational Behavior: The International Journal of Industrial, Occupational and Organizational Psychology and Behavior*, 29(3):415–433.

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

João A. Leite, Diego F. Silva, Kalina Bontcheva, and Carolina Scarton. 2020. Toxic language detection in social media for brazilian portuguese: New dataset and multilingual analysis.

Ilya Loshchilov and Frank Hutter. 2018. Decoupled weight decay regularization. In *International Conference on Learning Representations*.

Alexander Maedche, Viktor Pekar, and Steffen Staab. 2002. Ontology learning part one—on discovering taxonomic relations from the web. In *Web intelligence*, pages 301–319. Springer.

Angela T Maitner, Diane M Mackie, Janet VT Pauketat, and Eliot R Smith. 2017. The impact of culture and identity on emotional reactions to insults. *Journal of Cross-Cultural Psychology*, 48(6):892–913.

Hazel Rose Markus and Shinobu Kitayama. 2010. Cultures and selves: A cycle of mutual constitution. *Perspectives on psychological science*, 5(4):420–430.

Hala Mulki, Hatem Haddad, Chedi Bechikh Ali, and Halima Alshabani. 2019. L-hsab: A levantine twitter dataset for hate speech and abusive language. In *Proceedings of the Third Workshop on Abusive Language Online*, pages 111–118.

Michael Muthukrishna, Adrian V Bell, Joseph Henrich, Cameron M Curtin, Alexander Gedranovich, Jason McInerney, and Braden Thue. 2020. Beyond western, educated, industrial, rich, and democratic (weird) psychology: Measuring and mapping scales of cultural and psychological distance. *Psychological science*, 31(6):678–701.

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.

Erida Nurce, Jorgel Keci, and Leon Derczynski. 2021. Detecting abusive albanian. *arXiv preprint arXiv:2107.13592*.

Daphna Oyserman. 2011. Culture as situated cognition: Cultural mindsets, cultural fluency, and meaning making. *European review of social psychology*, 22(1):164–214.

Daphna Oyserman. 2017. Culture three ways: Culture and subcultures within countries. *Annual review of psychology*, 68:435–463.

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.

Senja Pollak, Marko Robnik-Šikonja, Matthew Purver, Michele Boggia, Ravi Shekhar, Marko Pranjić, Salla Salmela, Ivar Krustok, Tarmo Paju, Carl-Gustav Linden, Leo Leppänen, Elaine Zosa, Matej Ulčar, Linda Freienthal, Silver Traat, Luis Adrián Cabrera-Diego, Matej Martinc, Nada Lavrač, Blaž Škrlj, Martin Žnidaršič, Andraž Pelicon, Boshko Koloski, Vid Podpečan, Janez Kranjc, Shane Sheehan, Emanuela Boros, Jose G. Moreno, Antoine Doucet, and Hannu Toivonen. 2021. EMBEDDIA tools, datasets and challenges: Resources and hackathon contributions. In *Proceedings of the EACL Hackashop on News Media Content Analysis and Automated Report Generation*, pages 99–109, Online. Association for Computational Linguistics.

Paul Röttger, Debora Nozza, Federico Bianchi, and Dirk Hovy. 2022a. 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.

Paul Röttger, Haitham Seelawi, Debora Nozza, Zeerak Talat, and Bertie Vidgen. 2022b. Multilingual HateCheck: Functional tests for multilingual hate speech detection models. In *Proceedings of the Sixth Workshop on Online Abuse and Harms (WOAH)*, pages 154–169, Seattle, Washington (Hybrid). Association for Computational Linguistics.

Shalom Schwartz. 2006. A theory of cultural value orientations: Explication and applications. *Comparative sociology*, 5(2-3):137–182.

Ravi Shekhar, Marko Pranjić, Senja Pollak, Andraž Pelicon, and Matthew Purver. 2020. Automating news comment moderation with limited resources: Benchmarking in croatian and estonian. *Journal for Language Technology and Computational Linguistics*, 34(1):49–79.

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.

Jan D ten Thije and Ludger Zeevaert. 2007. *Receptive multilingualism: Linguistic analyses, language policies and didactic concepts*, volume 6. John Benjamins Publishing.

Harry C Triandis. 2018. *Individualism and collectivism*. Routledge.

United States Senate Committee on the Judiciary. Jan 31st, 2024. Hearings to examine big tech and the online child sexual exploitation crisis. 118th Congress (2023-2024), Presiding: Chair Durbin, G50 Dirksen Senate Office Building, Washington, D.C.

Bertie Vidgen and Leon Derczynski. 2020. Directions in abusive language training data, a systematic review: Garbage in, garbage out. *Plos one*, 15(12):e0243300.

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

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.

A Hardware and Implementation Details

All the experiments were conducted on an NVIDIA RTX A6000 with 48GB RAM. The entire experiment takes around 9 hours on a single GPU. We used a learning rate of 1e-4. For optimization, we used Adamw (Loshchilov and Hutter, 2018) using a L_2 regularization of 0.01.

B Measuring WEIRDness

Using the fixation index (FST), Muthukrishna et al., 2020 quantified variations in cultural beliefs and behaviors across societies. Initially used in genetics for assessing differentiation among subpopulations, FST has been adapted to cultural psychology (Cavalli-Sforza et al., 1994), serving to measure the deviation of cultural traits and assign a numerical value to cultural distances. The study significantly leveraged data from the World Values Survey (WVS) (Inglehart et al., 2000), a global initiative exploring the evolution of people's values and beliefs. Through WVS, (Muthukrishna et al., 2020) investigated the diverse responses of individuals from various societies to a broad set of queries about values and beliefs.

For each language, we assigned a WEIRDness score specific to the country from which the language's corpus data was sourced. For instance, the corpus for Arabic was derived from tweets originating in Lebanon; therefore, we applied the WEIRDness score specific to Lebanon for this dataset. However, for the Greek and Portuguese datasets, we adapted our approach due to the unavailability of specific WEIRDness scores for Greece and Portugal. Instead, we selected the WEIRDness scores of geographically proximal countries: Macedonia's score was used for the Greek dataset, and Spain's score was applied to the Portuguese dataset.

C Linguistic Similarity

Traditional indices like the Levenshtein distance (ten Thije and Zeevaert, 2007), Jaccard similarity (Maedche et al., 2002), and Cosine similarity (Gomaa et al., 2013) have significantly contributed to various linguistic applications, providing broad insights into text and content analysis. The Levenshtein distance is noted for its effectiveness in spelling correction and character-level analysis, Jaccard similarity in identifying word set overlaps for document comparisons, and Cosine similarity in gauging thematic content based on word frequency for information retrieval tasks.

However, our research, which delves into the nuanced detection of offensive language across languages, demands a linguistic analysis that captures more than what these traditional metrics offer. Our used index enhances these foundational indices by incorporating advanced phonological, syntactic, and semantic analyses. This is crucial for comprehensively understanding the intricacies of offensive language within various linguistic and cultural contexts.

Phonological sensitivity, a pivotal feature of this index, is instrumental in discerning subtle pronunciation or intonation differences that can significantly alter the meanings or connotations of words or phrases. For instance, homophones or words with similar sounds might have different meanings based on slight pronunciation nuances. Furthermore, the meaning or offensiveness of a word or phrase can change dramatically with intonation, such as in sarcasm or culturally specific jokes. Also, the same word can have different connotations across dialects or cultures based on pronunciation variations. The index's proficiency in analyzing these phonological aspects enhances the accuracy of offensive content detection in diverse linguistic landscapes.

Additionally, the index's capacity for syntactic and semantic analysis ensures a deep understanding of sentence structures and the contextual meaning of phrases. This surpasses the capabilities of traditional indices and is particularly beneficial for interpreting idiomatic expressions, colloquial language, and context-dependent language use. For example, the index can accurately interpret idiomatic expressions that may carry meanings not directly inferable from the individual words and are often deeply embedded in cultural contexts. It can also discern contextual nuances, enabling more accurate detection and interpretation of offensive content that varies dramatically with context.

D Datasets

Here we review all the datasets used in this work. It is essential to emphasize that all mentioned datasets are publicly available and have been specifically curated to facilitate research on hate speech and offensive language detection, which is aligned with our use case in this work.

D.1 Albanian

(Nurce et al., 2021) contains 11,874 posts collected from Instagram and YouTube. Four annotators have annotated the posts using hierarchical annotation proposed in (Zampieri et al., 2019). In this annotation three subtasks are defined as distinguishing between: 1) offensive and non-offensive, 2) targeted or untargeted offense, 3) individual, group, or other targets. In this study we use data from subtask 1.

D.2 Danish

Sigurbergsson and Derczynski (2020) consists of 800 Facebook posts and 2,800 Reddit posts and their respective comments. Annotation is done based on subtask of (Zampieri et al., 2019) and one binary label indicating offensiveness is provided.

D.3 English

de Gibert et al. (2018) introduced a dataset of 10,568 sentences sourced from 22 sub-forums of `Stormfront.org`, covering the period from 2002 to 2017. Each sentence is categorized based on whether it fulfills three criteria: a) deliberate attack, b) directed towards a specific group of people, and c) motivated by aspects of the group's identity.

D.4 Estonian

(Shekhar et al., 2020) contains 31.5M comments on news articles from Eesti Ekspress and labels to determine why deleted comments were considered inappropriate. The eight defined labels are as follows: 1) Disallowed content, 2) Threats, 3) Hate Speech, 4) Obscenity, 5) Deception and trolling, 6) Vulgarity, 7) Language, and 8) abuse. We take a comment as offensive if any of the aforementioned categories are present.

D.5 German

Assenmacher et al. (2021) contains 85,000 comments from the German newspaper Rheinische Post and the moderator's binary decision of abusiveness. The data is further annotated using the following fine-grained categories: 1) sexism, 2) racism, 3) threats, 4) insults, 5) profane, 6) meta/organizational, and 7) advertisement. In this work we aggregate the first five labels and create a new label for offensiveness.

D.6 Greek

Pitenis et al. (2020) introduce the Offensive Greek Tweet Dataset (OGTD) containing 4,779 tweets collected between May and June 2019. (Zampieri et al., 2019) guidelines and schema for subtask a is used and each tweet is labeled as offensive or not-offensive.

D.7 Italian

(Bosco et al., 2018) consists of 17,567 comments on 99 Facebook posts and 6,928 tweets. The task defined on these two datasets is a binary classification for detecting hate speech.

D.8 Latvian

Pollak et al. (2021) provide EMBEDDIA, a set of tools, datasets, and challenges for European languages. One of their datasets is 12M comments on Latvian news from Ekspress media group collected from 2015 to 2019. The labels indicate whether the comment was deleted or not from the website. Similar to Estonian, comments are often in Russian as well.

D.9 Portuguese

(Leite et al., 2020) contains 21K tweets collected from July to August 2019. The data is annotated for hate speech detection. Six fine-grained labels are also provided to indicate the type of hate speech. These labels include 1) LGBTQ+ phobia, 2) Insult, 3) Xenophobia, 4) Misogyny, 5) Obscene, and 6) Racism. In this work we aggregate all labels and create a new label for offensiveness.

D.10 Russian

(Gorbunova, 2022) contains 3,000 comments Russian social network VKontakte and was collected to evaluate existing classifiers on distorted words. Two binary labels are assigned to each comment to indicate toxicity and distortion.

D.11 Turkish

(Çöltekin, 2020) contains 40,000 tweets collected from March 2018 to September 2019 with a gap of two weeks during November 2018. The tweets are then labeled using subtask a of the hierarchical labeling introduced in (Zampieri et al., 2019).

D.12 Surzhyk

(Andrusyak et al., 2018) contains 2,000 YouTube comments in Surzhyk which is spoken in Russia and Ukraine. A binary label is then assigned to each comment to indicate if the comments is abusive or not.

D.13 Chinese

(Deng et al., 2022) consists of 37,480 posts from Zhiho and Weibo social media platforms. The data is annotated using a binary label to indicate offensiveness and a categorical label named topic that takes values of race, gender, and region. The topic label shows what topic the offender targeted.

D.14 Hindi

Bhardwaj et al. (2020) provide 8,200 posts collected from Twitter, Facebook, and WhatsApp. The posts are then categorized into five categories: 1) fake, 2) hate, 3) offense, 4) defame, and 5) non-hostile.

D.15 Arabic

The dataset provided by Mulki et al. (2019) consists of 6,000 tweets collected from March 2018 to February 2019. Each tweet has been assigned to one of the three categories: 1) Normal, 2) Hate, and 3) Abusive. We treat the tweets in the normal category as non-offensive and assign an offensive label to Tweets in the hate and abusive categories.

E Detailed Results

L_a	L_t	$\Delta(L_a, L_t)$
hindi	surzhyk	0.0080
hindi	turkish	0.0466
hindi	portuguese	-0.0134
hindi	latvian	-0.0220
hindi	italian	-0.0134
hindi	greek	0.0466
hindi	german	0.0466
hindi	estonian	-0.0220
hindi	danish	0.0466
hindi	albanian	0.0466
hindi	arabic	0.0138
hindi	english	-0.0134
hindi	russian	-0.0295
hindi	chinese	0.0466
surzhyk	hindi	-0.0092
surzhyk	turkish	0.0354
surzhyk	portuguese	-0.0092
surzhyk	latvian	0.0140
surzhyk	italian	-0.0092
surzhyk	greek	0.0354
surzhyk	german	0.0354
surzhyk	estonian	0.0140
surzhyk	danish	0.0354
surzhyk	albanian	0.0354
surzhyk	arabic	0.0600
surzhyk	english	-0.0092
surzhyk	russian	-0.0562
surzhyk	chinese	0.0354
turkish	hindi	0.0101
turkish	surzhyk	0.0111
turkish	portuguese	0.0101
turkish	latvian	-0.0001
turkish	italian	0.0101
turkish	greek	0.0055
turkish	german	0.0055
turkish	estonian	-0.0001
turkish	danish	0.0055
turkish	albanian	0.0055
turkish	arabic	-0.0043
turkish	english	0.0101
turkish	russian	-0.0051
turkish	chinese	0.0055
portuguese	hindi	-0.0134
portuguese	surzhyk	0.0080
portuguese	turkish	0.0466
portuguese	latvian	-0.0220
portuguese	italian	-0.0134
portuguese	greek	0.0466
portuguese	german	0.0466
portuguese	estonian	-0.0220
portuguese	danish	0.0466
portuguese	albanian	0.0466
portuguese	arabic	0.0138
portuguese	english	-0.0134
portuguese	russian	-0.0295
portuguese	chinese	0.0466
latvian	hindi	-0.0109
latvian	surzhyk	-0.0204
latvian	turkish	0.0108
latvian	portuguese	-0.0109
latvian	italian	-0.0109
latvian	greek	0.0108
latvian	german	0.0108
latvian	estonian	-0.0085
latvian	danish	0.0108
latvian	albanian	0.0108
latvian	arabic	0.0200
latvian	english	-0.0109
latvian	russian	-0.0789
latvian	chinese	0.0108
italian	hindi	-0.0134
italian	surzhyk	0.0080
italian	turkish	0.0466
italian	portuguese	-0.0134
italian	latvian	-0.0220
italian	greek	0.0466
italian	german	0.0466
italian	estonian	-0.0220
italian	danish	0.0466
italian	albanian	0.0466
italian	arabic	0.0138
italian	english	-0.0134
italian	russian	-0.0295
italian	chinese	0.0466
greek	hindi	0.0101
greek	surzhyk	0.0111
greek	turkish	0.0055
greek	portuguese	0.0101
greek	latvian	-0.0001
greek	italian	0.0101
greek	german	0.0055
greek	estonian	-0.0001
greek	danish	0.0055
greek	albanian	0.0055
greek	arabic	-0.0043
greek	english	0.0101
greek	russian	-0.0051
greek	chinese	0.0055
german	hindi	0.0101
german	surzhyk	0.0111

L_a	L_t	$\Delta(L_a, L_t)$
german	turkish	0.0055
german	portuguese	0.0101
german	latvian	-0.0001
german	italian	0.0101
german	greek	0.0055
german	estonian	-0.0001
german	danish	0.0055
german	albanian	0.0055
german	arabic	-0.0043
german	english	0.0101
german	russian	-0.0051
german	chinese	0.0055
estonian	hindi	-0.0109
estonian	surzhyk	-0.0204
estonian	turkish	0.0108
estonian	portuguese	-0.0109
estonian	latvian	-0.0085
estonian	italian	-0.0109
estonian	greek	0.0108
estonian	german	0.0108
estonian	danish	0.0108
estonian	albanian	0.0108
estonian	arabic	0.0200
estonian	english	-0.0109
estonian	russian	-0.0789
estonian	chinese	0.0108
danish	hindi	0.0101
danish	surzhyk	0.0111
danish	turkish	0.0055
danish	portuguese	0.0101
danish	latvian	-0.0001
danish	italian	0.0101
danish	greek	0.0055
danish	german	0.0055
danish	estonian	-0.0001
danish	albanian	0.0055
danish	arabic	-0.0043
danish	english	0.0101
danish	russian	-0.0051
danish	chinese	0.0055
albanian	hindi	0.0101
albanian	surzhyk	0.0111
albanian	turkish	0.0055
albanian	portuguese	0.0101
albanian	latvian	-0.0001
albanian	italian	0.0101
albanian	greek	0.0055
albanian	german	0.0055
albanian	estonian	-0.0001
albanian	danish	0.0055
albanian	arabic	-0.0043
albanian	english	0.0101

L_a	L_t	$\Delta(L_a, L_t)$
albanian	russian	-0.0051
albanian	chinese	0.0055
arabic	hindi	-0.0006
arabic	surzhyk	0.0006
arabic	turkish	0.0054
arabic	portuguese	-0.0006
arabic	latvian	0.0157
arabic	italian	-0.0006
arabic	greek	0.0054
arabic	german	0.0054
arabic	estonian	0.0157
arabic	danish	0.0054
arabic	albanian	0.0054
arabic	english	-0.0006
arabic	russian	-0.0684
arabic	chinese	0.0054
english	hindi	-0.0134
english	surzhyk	0.0080
english	turkish	0.0466
english	portuguese	-0.0134
english	latvian	-0.0220
english	italian	-0.0134
english	greek	0.0466
english	german	0.0466
english	estonian	-0.0220
english	danish	0.0466
english	albanian	0.0466
english	arabic	0.0138
english	russian	-0.0295
english	chinese	0.0466
russian	hindi	0.0120
russian	surzhyk	0.0244
russian	turkish	0.0141
russian	portuguese	0.0120
russian	latvian	-0.0229
russian	italian	0.0120
russian	greek	0.0141
russian	german	0.0141
russian	estonian	-0.0229
russian	danish	0.0141
russian	albanian	0.0141
russian	arabic	0.0667
russian	english	0.0120
russian	chinese	0.0141
chinese	hindi	0.0101
chinese	surzhyk	0.0111
chinese	turkish	0.0055
chinese	portuguese	0.0101
chinese	latvian	-0.0001
chinese	italian	0.0101
chinese	greek	0.0055
chinese	german	0.0055

L_a	L_t	$\Delta(L_a, L_t)$
chinese	estonian	-0.0001
chinese	danish	0.0055
chinese	albanian	0.0055
chinese	arabic	-0.0043
chinese	english	0.0101
chinese	russian	-0.0051

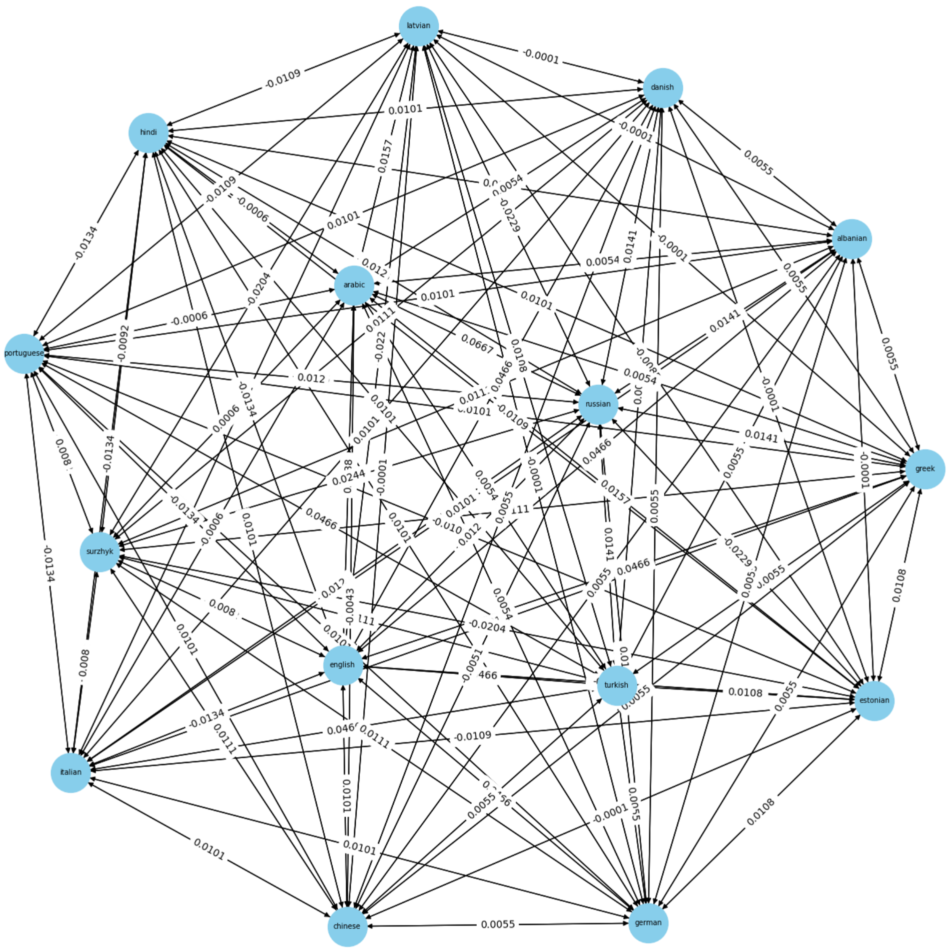

Figure 3: The CLTL performance change between 105 language pairs.

Language	Dataset	Text Source	Task/Label	Total #/Positive #
Albanian	(Nurce et al., 2021)	Instagram & YouTube	Offensive Language/subtask_a	1,568/11,874 (13.20%)
Danish	(Sigurbergsson and Derczynski, 2020)	Facebook & Reddit	Offensive Language/label	384/2,960(12.97%)
English	(de Gibert et al., 2018)	Stormfront	Offensive Language and Hate Speech	1119/9916 (11.29%)
Estonian	(Shekhar et al., 2020)	News Comments	Deleted Comment/infringed_on_rule	126,386/1.5M (8.02%)
German	(Assenmacher et al., 2021)	News Comments	Offensive Language/aggregated labels	23044/85,000 (27.11%)
Greek	(Pitenis et al., 2020)	Twitter	Offensive Language/subtask_a	2,486/8,743 (28.43%)
Italian	(Bosco et al., 2018)	Facebook & Twitter	Hate Speech/hate	2,764/6,000/ (46.06%)
Latvian	(Pollak et al., 2021)	News Comments	Deleted Comment/is_enabled	485,679/3,379,490 (14.37%)
Portuguese	(Leite et al., 2020)	Tweeter	Hate Speech/aggregated labels	9,255/21,000 (44.07%)
Russian	(Gorbunova, 2022)	Vkontakte Social Network	Toxicity/toxicity	456/2,400 (19.00%)
Turkish	(Çöltekin, 2020)	Twitter	Offensive Language/subtask_a	6,046/31,277 (19.33%)
Surzhyk	(Andrusyak et al., 2018)	YouTube	Abusive Language/abusive	654/2,000 (32.70%)
Chinese	(Deng et al., 2022)	Zhiho & Sina Weibo	Offensive Language/label	12,723/25,726 (49.45%)
Hindi	Bhardwaj et al. (2020)	Twitter & Facebook & WhatsApp	Hate Speech/Labels Set	2,678/5,728 (46.75%)
Arabic	(Mulki et al., 2019)	Twitter	Hate Speech/Class	1,791/4,676 (38.30%)

Table 1: Source, task, statistics, and reference of datasets used in this work.