“Is Hate Lost in Translation?”: Evaluation of Multilingual LGBTQIA+ Hate Speech Detection

Fai Leui Chan, Duke Nguyen, Aditya Joshi
University of New South Wales, Sydney, Australia
aditya.joshi@unsw.edu.au

Abstract

This paper explores the challenges of detecting LGBTQIA+ hate speech of large language models across multiple languages, including English, Italian, Chinese and (code-mixed) English-Tamil, examining the impact of machine translation and whether the nuances of hate speech are preserved across translation. We examine the hate speech detection ability of zero-shot and fine-tuned GPT. Our findings indicate that: (1) English has the highest performance and the code-mixing scenario of English-Tamil being the lowest, (2) fine-tuning improves performance consistently across languages whilst translation yields mixed results. Through simple experimentation with original text and machine-translated text for hate speech detection along with a qualitative error analysis, this paper sheds light on the socio-cultural nuances and complexities of languages that may not be captured by automatic translation.

Warning: The paper contains examples of multilingual hate speech towards LGBTQIA+ community because of the nature of the work.

1 Introduction

LGBTQIA+ individuals are particularly vulnerable to hate speech due to their sexual orientation and gender identity. They are frequently subject to harassment, discrimination, violence due to their identity (Chakravarthi et al., 2024). Therefore, many social media platforms have implemented hate speech detection as part of content sanitation on their platforms to create safer online environments. As social media platforms become increasingly diverse with people coming from different linguistic backgrounds, we investigate if hate speech detection is sustained across different languages, translations, and code-mixing environments. In other words, is hate speech detection “lost in translation”[1]?

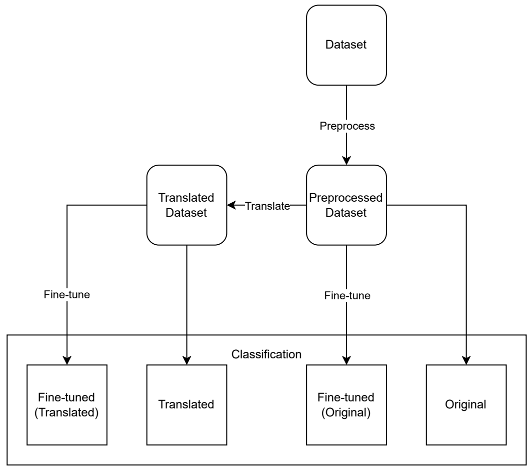

Figure 1: Evaluation methodology of machine translation-based hate speech detection.

The approach of using machine translation to translate the test data into English and running inference using an English-only model has long been studied (Pikuliak et al., 2021). This method may be better for complex tasks that require common sense or real-world knowledge, as it benefits from the use of a stronger English-only model (Artetxe et al., 2023), which may be useful for the complex task of hate speech detection.

Therefore, we ask the question: “How does hate speech detection perform for original text and translated text?” We do so for the case of hate speech towards LGBTQIA+ people. While it is intuitive that machine translation will not preserve all semantics, our experiments with zero-shot and fine-tuned GPT show that it particularly holds true for hate speech detection. Our error analysis sheds light on the nature of errors to highlight ‘what’ is lost in translation.

[1] As part of a discussion on his poem “Stopping by Woods on a Snowy Evening”, Robert Frost famously remarked “You’ve often heard me say – perhaps too often – that poetry is what is lost in translation. It is also what is lost in interpretation.” (Untermeyer, 1964, p. 18)

Proceedings of The 22nd Annual Workshop of the Australasian Language Technology Association, pages 146–152
December 2-4, 2024 ©2024 Association for Computational Linguistics

Language	Source	Total Samples	Non-Homotransphobic	Homotransphobic
English	(McGiff and Nikolov, 2024)	1,277	656 (51.4%)	621 (48.6%)
Italian	(Nozza et al., 2023)	5,000	2,992 (59.8%)	2,008 (40.2%)
Chinese	(Lu et al., 2023)	2011	1247 (62.0%)	764 (38.0%)
English-Tamil	(Chakravarthi et al., 2021)	6033	5384 (89.2%)	649 (10.8%)

Table 1: Comparison of datasets. % in the Non-Homotransphobic and Homotransphobic columns refer to the proportion of each class relative to the total samples in each dataset, with each row summing to 100%.

2 Methodology

Our methodology is as shown in Figure 1. We utilise labeled datasets in English, Italian, Chinese, and English-Tamil (code-mixed[2]), each focusing on LGBTQIA+-specific hate speech. Our preprocessing involves removing excess spaces and invalid characters.

We translate non-English datasets (Italian, Chinese, and English-Tamil (code-mixed) into English via the chosen LLM (large language model) in zero-shot setting using the following user prompt 'Translate this sentence into English: 'text". This forms our **Translated Dataset**.

We then perform zero-shot classification using the chosen LLM to the detect homotransphobia[3], with 1 referring to homotransphobic content and 0 referring to non-homotransphobic content. We use the following system prompt "You are an AI assistant that classifies text as either homotransphobic (1) or not homotransphobic (0). Respond with only 0 or 1.", and the user prompt being "Classify the following text: 'text". This is applied on both the **Preprocessed Dataset** and the **Translated Dataset**. This gives us classification results for **Original** and **Translated** respectively.

We then perform fine-tuning on the LLM via the OpenAI API[4] using the **Preprocessed Dataset** and **Translated Dataset** using the same prompts as what was used for the earlier round of classification. We then get the classification results for **Fine-tuned (Original)** and **Fine-tuned (Translated)** respectively.

Finally, we perform comparative analysis between the classification results from four models **Original**, **Fine-tuned (Original)**, **Translated**, and **Fine-tuned (Translated)** and evaluate the impact of translation on the final effectiveness of the model and measure the performance improvement, if any, achieved through fine-tuning.

3 Experiment Setup

The LLM which we use for our experiments is the gpt-3.5-turbo model[5], a chat-bot based on the GPT-3.5 language model developed by OpenAI. This model is optimised for prompt-based usage but performs equally well for traditional NLP tasks (Das et al., 2024).

The datasets which we employ are shown in Table 1 with a train-validation-test split of 60:20:20. It is noted that the datasets display varying degrees of imbalance which could affect model performance across languages. While the English and Italian datasets are fairly balanced, and the Chinese dataset shows a moderate imbalance, the English-Tamil dataset exhibits severe imbalance, with only 10.8% of samples being homotransphobic, broadly referred to as hate speech towards the LGBTQIA+ community.

The downstream task is hate speech detection, and is evaluated using the following metrics: F1 score, precision, recall, and Cohen-Kappa agreement. In particular, Cohen's Kappa is used to measure the agreement between the predicted labels and the true label. It is chosen as it is a good measure of intra-rater reliability, while correcting for times when the raters may agree by chance (Cohen, 1960). F1 score, precision, and recall are weighted to account for class imbalances.

4 Results

4.1 Quantitative Evaluation

Table 2 compares the performance of gpt3.5-turbo on original text versus translated text across different languages. English yields the highest F1-score

[2]Code-mixing indicates the use of vocabulary from multiple languages. The English-Tamil (code-mixed) dataset employed in this paper are remarks written in mostly Roman character employing Tamil vocabulary with either Tamil or English grammar (Chakravarthi et al., 2021).

[3]'Homotransphobic' is used as an umbrella term to indicate hate speech towards the LGBTQIA+ community

[4]https://platform.openai.com/docs/guides/fine-tuning

[5]https://platform.openai.com/docs/models

Language	Condition	F1	P	R	K	Δ F	Δ K
English	Original	0.7952	0.7082	0.9066	0.5488	-	-
	Fine-tuned	0.8689	0.8833	0.8548	0.7486	+0.0737	+0.1998
Italian	Original	0.5990	0.4514	0.8899	0.1414	-	-
	Translated	0.5355	0.4424	0.6783	0.0960	-0.0635	-0.0454
	Fine-tuned (Original)	0.8375	0.8417	0.8333	0.7292	+0.2385	+0.5878
	Fine-tuned (Translated)	0.7417	0.7371	0.7463	0.5662	+0.2062	+0.4702
Chinese	Original	0.7464	0.7493	0.7435	0.5878	-	-
	Translated	0.6839	0.7099	0.6597	0.2463	-0.0625	-0.3415
	Fine-tuned (Original)	0.8146	0.8255	0.8039	0.7030	+0.0682	+0.1152
	Fine-tuned (Translated)	0.7661	0.7958	0.7386	0.6308	+0.0822	+0.3845
English- Tamil	Original	0.3619	0.2843	0.4977	0.1998	-	-
	Translated	0.3202	0.3511	0.2943	0.2463	-0.0417	+0.0465
	Fine-tuned (Original)	0.5391	0.6200	0.4769	0.2452	+0.1772	+0.0454
	Fine-tuned (Translated)	0.4037	0.5000	0.3385	0.3469	+0.0835	+0.1006

Table 2: Performance Metrics (F1: F1-score, P: Precision, R: Recall: K: Cohen's Kappa) and Changes Across Languages and Conditions. All scores are weighted. Δ columns represent the changes in F1-score and Cohen's Kappa between different conditions: Fine-tuned (Original → Fine-tuned), Translated (Original → Translated), and Fine-tuned (Translated → Fine-tuned).

(0.7952), followed by Chinese (0.7464), Italian (0.5990), and English-Tamil (0.3619). The strong performance in Chinese suggests good generalisation to non-Latin scripts after translation, while the low score for English-Tamil highlights challenges with code-mixed content (Doğruöz et al., 2021).

We also evaluate whether applying the subsequent transformation process degrades or improves the performance. Translating non-English content to English produces mixed results. English-Tamil sees a slight improvement in Cohen's Kappa (+0.0465) despite a decrease in F1-score (-0.0417), which suggests translating and classifying may improve model performance in code-mixed languages (Gautam et al., 2021). Italian shows marginal decreases in both metrics. Chinese experiences the most significant performance drop (F1: -0.0625, Kappa: -0.3415), suggesting substantial loss of context during translation. These findings indicate that in general, translation decreases the effectiveness of hate speech detection. However, the degree of reduction is language-dependent.

Fine-tuning consistently improves performance across all languages, with the most substantial gains in Italian (ΔF1: +0.2385, ΔKappa: +0.5878) and English-Tamil (ΔF1: +0.1772). Even English and Chinese, which have strong baseline performances, see notable improvements. Fine-tuning on translated text also shows benefits, though generally not as substantial as fine-tuning on original text, with Chinese being an exception.

4.2 Qualitative Analysis

We now show qualitative analysis of how hate is 'lost in translation' as shown in the previous section. This is visible in the case of slang and culturally specific references. We request the assistance of native speakers of Italian, Tamil, and Chinese to identify prominent translation errors for the misclassified case as shown in Table 3.

Table 3 indicates that most of LGBTQ terminologies, derogatory language involving LGBTQ people, and sometimes even non-LGBTQ slang words (in the case of English-Tamil) are mistranslated across the three languages (Italian, English-Tamil, and Chinese). This suggests that while the translation models may handle standard language adequately, they struggle with specialised or subcultural terms, which are often crucial in detecting hate speech. Despite these challenges, the model shows some strengths, such as correctly identifying some LGBTQIA+-related slang like "BL", "CP", and "腐女" in translations.

The qualitative analysis reveals significant challenges in translating and detecting LGBTQIA+ hate speech across languages, particularly with slang, implicit hate, and cultural-specific expressions. While the model shows promise in some areas, there is a clear need for more nuanced, language-specific approaches to improve accuracy in multilingual hate speech detection.

Language	Original Text	GPT Translation	Human Translation	T	P	Analysis
Chinese	你怎么骂基佬我也不生气我也不是男的我还有男朋友	I'm not angry even if you scold me, I'm not a man, I have a boyfriend.	I won't be offended if you call me gay. I'm not a man. I have a boyfriend.	0	1	Mistranslation of "基佬" (brash slang for gay people).
Chinese	男同没皮没脸要借别人的好脸	Men who have no shame want to borrow other people's good reputation.	Gay men who have no shame want to borrow other people's good reputation.	1	0	"男同" (abbreviation for gay men) was missed in translation.
Chinese	妇男，快12点了，你上飞机没有?	Woman, it's almost 12 o'clock, have you boarded the plane?	House-husband, it's almost 12 o'clock, have you boarded the plane?	1	0	Mistranslation of "妇男" (house-husband) as woman.
English-Tamil	Yogesh Waran Hey nee 9 taane	Yogesh, you are a 9 out of 10	Yogesh Waran, aren't you trans?	1	0	Mistranslation of "9" (double meaning word for the transgender community).
English-Tamil	Serupa alla ADI naramunda	"Similar to ADI, naramunda"	Will beat you with slipper, fool	1	0	Translation is completely wrong. "Serupu" means "slipper/footwear" and "naramunda" is pejorative term meaning "fool".
English-Tamil	Serpala adikanum......enga ponanulm ithuka tholla thanga mudila ...	I need to talk to Serpala...I can't handle this on my own...	I will hit with slipper, I cant bear this trouble.	1	0	Mistranslation of Serpala, which means "with slipper/footwear" with an informal and aggressive connotation, as a name.
Italian	@user_ab @user_abcde @user_abcdef @user_a Sono tutti innamorati del busone	@user_ab @user_abcde @user_abcdef @user_a They are all in love with the big bus	They're all in love with anal sex	1	0	Mistranslation of "Busone" (Northern Italian slang for anal sex, used for mocking homosexuals) as "big bus" (-one affix is used to indicate something that is big)
Italian	MADONNA SUCCHIACAZZI E PUTTANA #bestemmie #world	MADONNA SUCKS DICK AND WHORE #blasphemy #world	GOD FUCKING DAMMIT	0	1	Mistranslation of "Madonna Puttana" which is a common rage expression in Italian involving saints but is not homotransphobic
Italian	Non ho letto nulla ma nel dubbio Sala ricchione	I haven't read anything but just in case Sala is a rich man	I haven't read anything but just in case Sala is a faggot	1	0	Mistranslation of "ricchione" (Southern Italian derogatory slang for homosexual people) as "rich" (which is "ricco")

Table 3: Qualitative error analysis of misclassified examples for the **Zero-shot on Translated**. Each sample is given the 'Original Text', the 'GPT Translation', and the 'Human Translation'. 'T' stands for 'Truth' and 'P' stands for 'Prediction'. 'Truth' and 'Prediction' values are either 0 (non-homotransphobic) or 1 (homotransphobic). 'Analysis' are comments on the translation error.

5 Related Work

Despite broad interest in hate speech detection, research specifically addressing LGBTQIA+ communities remain limited. Challenges to create a generalised hate speech model for various targets have been reported in particular(Nozza et al., 2023). Shared tasks have been particularly important for hate speech detection towards LGBTQIA+ community. The LT-EDI@EACL series (2022-2024) focuses on the identification of homophobia, transphobia, and nonanti-LGBTQIA+ content in Tamil, English, and code-mixed English-Tamil (Chakravarthi et al., 2022, 2023, 2024). The shared task has expanded to include various languages to look at homotransphobia in a multilingual context. There have also been other shared tasks on the topic, focusing on various languages. Examples include HOMO-MEX2023@IberLEF which focuses on hate speech detection towards the Mexican Spanish-Speaking LGBTQIA+ population (Bel-Enguix et al., 2023; Tash et al., 2023). In a similar vein, HODI is a shared task for the automatic detection of homotransphobia in Italian presented at EVALITA 2023 (Nozza et al., 2023). Beyond shared tasks, some research has employed Transformer-based models like BERT and XLM-RoBERTa to identify transphobic and homophobic insults in social media comments (Manikandan et al., 2022). Benchmarks such as WinoQueer (Felkner et al., 2023) provide pairs of sentences to measure anti-LGBTQIA+ bias in language models. To the best of our knowledge, this is the first hate speech detection comparison centered around machine translation. The datasets we use are reported in past work.

6 Conclusion

This study provides valuable insights into the effectiveness of LLM in hate speech detection in diverse linguistic settings involving LGBTQIA+ communities. We compare the ability of zero-shot and fine-tuned GPT for hate speech detection of multilingual text in the original language and translated versions to English. Our insights were: (1) hate speech detection via LLM is in general effective (including in non-Latin script settings), however LLMs perform significantly worse when dealing with code-mixed languages; (2) hate speech detection via LLM can be improved simply via fine-tuning, although the degree of improvement is language-dependent; (3) translation is ineffective in transferring nuanced ideas and show visible degradation on hate speech detection performance.

To the best of our knowledge, this is the first work in hate speech detection with machine translation as our anchor. While the technique itself is simplistic, our research demonstrates the complexity of hate speech detection, especially for LGBTQIA+ communities in multilingual contexts and the need for continued research in this area. By advancing our understanding of multilingual hate speech detection, we can work towards creating safer, more inclusive online spaces for LGBTQIA+ individuals across different linguistic communities.

Limitations and Future Work

We now discuss limitation and future work. First of all, large language models have shown to exhibit bias towards LGBTQIA+ communities (Sosto and Barrón-Cedeño, 2024; Felkner et al., 2023), and there may exist potential biases in the training data and model itself.

Secondly, the cascaded approach of using gpt3.5-turbo for both translation and classification makes the process vulnerable to errors from both stages and may introduce biases or errors that are difficult to isolate (Unanue et al., 2023). Future work could benefit from variations to the translation and classification process in order to study the influence of each component on the final evaluation.

In addition, the use of GPT is prompt-dependent. The quality of the prompt can significantly impact the quality and accuracy of the model's outputs (Li et al., 2024). Our works have not analyzed the effects of insignificant prompt variation on the model's performance on selected tasks. Furthermore, we have also used English prompts for non-English datasets. Future work can experiment with prompts in the language that corresponds to each dataset.

Moreover, there is a lack of context beyond single sentences in our analysis. Providing more contextual information could lead to a more robust understanding of the cultural context and lead to better results. This could be done via adding slang words and their translations in the prompt.

Additionally, we have not analyzed if there was any correlation between the translation quality and the performance on the downstream tasks. In addition, whilst English, Italian, and Chinese are high-resource languages, Tamil is much more low-resourced and this could have contributed to the low performance of English-Tamil. Future work could include an LLM that has been trained more intensively on Tamil.

Lastly, it would be highly beneficial to compare gpt3.5-turbo with other large language models and specialised hate speech detection systems to benchmark its effectiveness.

Acknowledgment

This paper is the outcome of a Taste-of-Research scholarship awarded to Fai Leui Chan. The scholarship was funded by Google's exploreCSR grant.

We would like to acknowledge Marsha Mariya Kappan and Steffano Mezza from the School of Computer Science and Engineering at the University of New South Wales for providing qualitative error analysis of English-Tamil and Italian classification examples.

References

Mikel Artetxe, Vedanuj Goswami, Shruti Bhosale, Angela Fan, and Luke Zettlemoyer. 2023. Revisiting machine translation for cross-lingual classification. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 6489–6499, Singapore. Association for Computational Linguistics.

Gemma Bel-Enguix, Helena Gómez-Adorno, Gerardo Sierra, Juan Vásquez, Scott Thomas Andersen, and Sergio Ojeda-Trueba. 2023. Overview of HOMO-MEX at iberlef 2023: Hate speech detection in online messages directed towards the mexican spanish speaking LGBTQ+ population. *Proces. del Leng. Natural*, 71:361–370.

Bharathi Raja Chakravarthi, Prasanna Kumaresan, Ruba Priyadharshini, Paul Buitelaar, Asha Hegde, Hosahalli Shashirekha, Saranya Rajiakodi, Miguel Ángel García, Salud María Jiménez-Zafra, José García-Díaz, Rafael Valencia-García, Kishore Ponnusamy, Poorvi Shetty, and Daniel García-Baena. 2024. Overview of third shared task on homophobia and transphobia detection in social media comments. In *Proceedings of the Fourth Workshop on Language Technology for Equality, Diversity, Inclusion*, pages 124–132, St. Julian's, Malta. Association for Computational Linguistics.

Bharathi Raja Chakravarthi, Rahul Ponnusamy, Malliga S, Paul Buitelaar, Miguel Ángel García-Cumbreras, Salud María Jimenez-Zafra, Jose Antonio Garcia-Diaz, Rafael Valencia-Garcia, and Nitesh Jindal. 2023. Overview of second shared task on homophobia and transphobia detection in social media comments. In *Proceedings of the Third Workshop on Language Technology for Equality, Diversity and Inclusion*, pages 38–46, Varna, Bulgaria. INCOMA Ltd., Shoumen, Bulgaria.

Bharathi Raja Chakravarthi, Ruba Priyadharshini, Thenmozhi Durairaj, John McCrae, Paul Buitelaar, Prasanna Kumaresan, and Rahul Ponnusamy. 2022. Overview of the shared task on homophobia and transphobia detection in social media comments. In *Proceedings of the Second Workshop on Language Technology for Equality, Diversity and Inclusion*, pages 369–377, Dublin, Ireland. Association for Computational Linguistics.

Bharathi Raja Chakravarthi, Ruba Priyadharshini, Rahul Ponnusamy, Prasanna Kumar Kumaresan, Kayalvizhi Sampath, Durairaj Thenmozhi, Sathiyaraj Thangasamy, Rajendran Nallathambi, and John Phillip McCrae. 2021. Dataset for identification of homophobia and transophobia in multilingual youtube comments. *Preprint*, arXiv:2109.00227.

Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1):37–46.

Mithun Das, Saurabh Kumar Pandey, and Animesh Mukherjee. 2024. Evaluating ChatGPT against functionality tests for hate speech detection. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 6370–6380, Torino, Italia. ELRA and ICCL.

A. Seza Doğruöz, Sunayana Sitaram, Barbara E. Bullock, and Almeida Jacqueline Toribio. 2021. A survey of code-switching: Linguistic and social perspectives for language technologies. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1654–1666, Online. Association for Computational Linguistics.

Virginia Felkner, Ho-Chun Herbert Chang, Eugene Jang, and Jonathan May. 2023. Winoqueer: A community-in-the-loop benchmark for anti-lgbtq+ bias in large language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9126–9140.

Devansh Gautam, Kshitij Gupta, and Manish Shrivastava. 2021. Translate and classify: Improving sequence level classification for English-Hindi code-mixed data. In *Proceedings of the Fifth Workshop on Computational Approaches to Linguistic Code-Switching*, pages 15–25, Online. Association for Computational Linguistics.

Lingyao Li, Lizhou Fan, Shubham Atreja, and Libby Hemphill. 2024. "hot" chatgpt: The promise of chatgpt in detecting and discriminating hateful, offensive, and toxic comments on social media. *ACM Trans. Web*, 18(2).

Junyu Lu, Bo Xu, Xiaokun Zhang, Changrong Min, Liang Yang, and Hongfei Lin. 2023. Facilitating fine-grained detection of chinese toxic language: Hierarchical taxonomy, resources, and benchmarks. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 16235–16250. Association for Computational Linguistics.

Deepalakshmi Manikandan, Malliga Subramanian, and Kogilavani Shanmugavadivel. 2022. A system for detecting abusive contents against LGBT community using deep learning based transformer models. In *Working Notes of FIRE 2022 - Forum for Information Retrieval Evaluation, Kolkata, India, December 9-13, 2022*, volume 3395 of *CEUR Workshop Proceedings*, pages 106–116. CEUR-WS.org.

Josh McGiff and Nikola S Nikolov. 2024. Bridging the gap in online hate speech detection: a comparative analysis of bert and traditional models for homophobic content identification on x/twitter. *arXiv preprint arXiv:2405.09221*.

Debora Nozza, Alessandra Teresa Cignarella, Greta Damo, Tommaso Caselli, and Viviana Patti. 2023. Hodi at evalita 2023: Overview of the first shared task on homotransphobia detection in italian. In *Proceedings of the Eighth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2023)*, CEUR Workshop Proceedings. CEUR Workshop Proceedings (CEUR-WS.org). Publisher Copyright: © 2023 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).; 8th Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop, EVALITA 2023 ; Conference date: 07-09-2023 Through 08-09-2023.

Matúš Pikuliak, Marián Šimko, and Mária Bieliková. 2021. Cross-lingual learning for text processing: A survey. *Expert Systems with Applications*, 165:113765.

Mae Sosto and Alberto Barrón-Cedeño. 2024. Queerbench: Quantifying discrimination in language models toward queer identities. *Preprint*, arXiv:2406.12399.

Moein Shahiki Tash, Jesús Armenta-Segura, Zahra Ahani, Olga Kolesnikova, Grigori Sidorov, and Alexander F. Gelbukh. 2023. LIDOMA at homomex2023@iberlef: Hate speech detection towards the mexican spanish-speaking LGBT+ population. the importance of preprocessing before using bert-based models. In *Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2023) co-located with the Conference of the Spanish Society for Natural Language Processing (SEPLN 2023), Jaén, Spain, September 26, 2023*, volume 3496 of *CEUR Workshop Proceedings*. CEUR-WS.org.

Inigo Jauregi Unanue, Gholamreza Haffari, and Massimo Piccardi. 2023. T3L: Translate-and-test transfer learning for cross-lingual text classification. *Transactions of the Association for Computational Linguistics*, 11:1147–1161.

L. Untermeyer. 1964. *Robert Frost: a Backward Look*.
Reference Department, Library of Congress.