When Does Language Transfer Help? Sequential Fine-Tuning for Cross-Lingual Euphemism Detection

Julia Sammartino, Libby Barak , Jing Peng, Anna Feldman
Montclair State University
New Jersey, USA
{sammartinoj1, barakl, pengj, feldmana}@montclair.edu

Abstract

Euphemisms are culturally variable and often ambiguous, posing challenges for language models, especially in low-resource settings. This paper investigates how cross-lingual transfer via sequential fine-tuning affects euphemism detection across five languages: English, Spanish, Chinese, Turkish, and Yorùbá. We compare sequential fine-tuning with monolingual and simultaneous fine-tuning using XLM-R and mBERT, analyzing how performance is shaped by language pairings, typological features, and pretraining coverage. Results show that sequential fine-tuning with a high-resource L1 improves L2 performance, especially for low-resource languages like Yorùbá and Turkish. XLM-R achieves larger gains but is more sensitive to pretraining gaps and catastrophic forgetting, while mBERT yields more stable, though lower, results. These findings highlight sequential fine-tuning as a simple yet effective strategy for improving euphemism detection in multilingual models, particularly when low-resource languages are involved.

1 Introduction

Euphemisms are used as substitutes for words or phrases that could be considered harsh, impolite, or taboo. For example, instead of overtly stating that someone died, one can instead utilize a euphemism that softens the tone: "I heard that his dad **passed on**." Due to the subjectiveness and figurative nature of euphemisms, native speakers of a language may disagree on whether a word or phrase is considered euphemistic (Gavidia et al., 2022). It is important to note that some phrases which may be used as euphemisms may also be taken at face value in certain contexts, without an underlying intended meaning (e.g. "he **passed on** the information to his boss", "I **passed on** this job offer"). Therefore, the term Potentially Euphemistic Terms (PETs) was created to reflect this ambiguity, aligning with previous research (Gavidia et al., 2022; Lee et al., 2022b). For example, the phrase 'between jobs' could be used euphemistically to mean 'unemployed', or could be taken literally to mean 'between shifts at two different jobs'. We investigate whether models can transfer euphemism knowledge across languages, and whether sequential fine-tuning—training first on a high-resource language—can improve performance in low-resource settings.

Multilingual transformer models, such as XLM-RoBERTa (XLM-R) (Conneau et al., 2019) and mBERT (Pires et al., 2019) have been used for various tasks and experiments due to their ability to capture cross-lingual representations and transfer-learning capabilities. In order to analyze the knowledge captured through this process, we performed experiments with sequential fine-tuning and compared it to monolingual baselines and paired language simultaneous fine-tuning. We investigate the cross-linguistic generalization capabilities through sequential fine-tuning, in which the model learns the same task first on one language, L1, and once it reaches its peak performance, learns the same task on a second language, L2. The model is then tested on both languages, including every pairing of English (EN), Mandarin Chinese (ZH), (Latin American and Castilian) Spanish (ES), Turkish (TR), and Yorùbá (YO).

We then compare these sequential results to the baseline monolingual score for each language, as well as simultaneous fine-tuning - learning two languages at the same time, and then testing on both.

We hypothesize that these proposed experimental settings of sequential fine-tuning will enable deeper understanding of the abilities of LLMs to learn the properties of abstract figurative language when given the chance to focus on each language in isolation. This experiment is especially important for low-resource languages, in which we have less access to rich training data and therefore depend on cross-lingual transfer to boost a model's

Proceedings of Recent Advances in Natural Language Processing,pages 1058–1065
Varna, Sep 8–10, 2025
https://doi.org/10.26615/978-954-452-098-4-122

performance. Our extensive analysis offers a new perspective on this important aspect of multilingual LLMs, and investigates the cross-lingual capabilities of XLM-R and mBERT.

Lang	Euph	Non-Euph	Total
ZH	2213 (149)	998 (56)	3211 (151)
EN	1841 (141)	1257 (85)	3098 (144)
ES	1955 (223)	997 (135)	2952 (233)
TR	1457 (67)	979 (59)	2436 (70)
YO	1689 (153)	909 (85)	2598 (157)

Table 1: Number of examples for 2025 PETs Datasets - Number of PETs for each class in parentheses. For each individual PET, there is a maximum of 40 examples of each class (euph vs. non-euph).

2 Related Work

2.1 Euphemism Detection

Recent work on euphemism detection has expanded to multilingual settings, leveraging deep learning and cross-lingual methods. Gavidia et al. (2022) introduced a PETs corpus for English, later extended to Spanish, Chinese, Yorùbá (Lee et al., 2023, 2024) and Turkish (Biyik et al., 2024).

Approaches range from lexicon-based methods (Felt and Riloff, 2020; Lee et al., 2022b) to transformer models (Zhu et al., 2021; Wang et al., 2022) to exploring various linguistic properties, e.g., vagueness (Lee et al., 2023). To address data scarcity, Kohli et al. (2022) used adversarial augmentation, and Keh et al. (2022) applied kNN-based data expansion. Shared tasks (Lee et al., 2022a; Lee and Feldman, 2024) have driven benchmarking, with ensemble models (Vitiugin and Paakki, 2024) achieving strong performance, while zero-shot evaluations Keh (2022) provide insights into cross-lingual generalization.

Multilingual work with more than two languages remains limited, though bilingual euphemism detection (Wang et al., 2022) and euphemistic abuse detection (Wiegand et al., 2023) highlight transfer challenges. We extend this by evaluating sequential fine-tuning across diverse languages, analyzing the role of dataset structure and lexical overlap in cross-lingual transfer.

2.2 Sequential Fine-Tuning

Sequential fine-tuning is an established approach to LLM experimentation, but the majority of previous work has focused on utilizing multiple tasks or 'sub-tasks' rather than exploring the cross-lingual capabilities of a model. Prior work has shown that continued pretraining on domain- or task-specific data improves downstream performance, even without labeled supervision Gururangan et al. (2020). Our study builds on this insight, applying a similar principle to the cross-lingual setting, where we use labeled data for figurative language (euphemisms) in one language as a form of task-aligned adaptation for another. One cross-lingual application focused on translation into English and then classification (Hu et al., 2024). This highlights a downside to some multilingual work with LLMs – having the model work on a dataset that has been translated into English, rather than directly interpreting the original non-English text. Our work improves upon this area by using a variety of languages **without** using the model for translation, as translation may result in a loss of underlying meanings.

Figurative language adds another layer of complexity as far as underlying meanings. Prior work has evaluated euphemism detection in multilingual or zero-shot settings Lee et al. (2023); Keh (2022), but few studies have tested sequential fine-tuning as a method for targeted cross-lingual adaptation. We address this gap by systematically evaluating whether exposure to euphemisms in a high-resource language improves detection in a low-resource language, using both mBERT and XLM-R across five typologically diverse languages.

3 Datasets

We leverage publicly available euphemism datasets that were originally published in Lee and Feldman (2024), with the addition of a Turkish dataset. Table 1 details the distribution of euphemistic and non-euphemistic examples.

Previous researchers created these datasets by first curating a list of potentially euphemistic terms, and then scraping from a variety of corpora that are listed in the following paper (Lee et al., 2023). This data is composed of extracted examples from online sources including Glowbe (English) (Davies, 2013) and curated corpora (Spanish, Chinese) (Real Academia Española, 2025; Brightmart, 2019). In the case of Yorùbá and Turkish [1], the authors utilized various sources such as news articles, religious texts, and more.

Annotations were executed by at least 3 native speakers of each language, and majority vote was utilized for the final classification. Lee et al. (2024)

[1]The Turkish dataset was used with permission from the author for a paper that is currently under review. The curation of it followed a similar schematic to previous work.

assessed inter-annotator agreement using Krippendorf's alpha on a small subset of the dataset, and found values ranging from 0.415 to 0.679 on a scale of 0 to 1. This was expected, as euphemisms can be ambiguous, even to native speakers.

4 Methodology

4.1 Model

As our research focused on the multilingual and cross-lingual learning capabilities of LLMs, we chose to experiment with two prominent multilingual models - XLM-R and mBERT.

XLM-R was pretrained on English, Chinese, Spanish, and Turkish, but not on Yorùbá. The model was trained on 2.5TB of CommonCrawl data spanning 100 languages, notably with English, Chinese, and Spanish receiving significantly higher representation, Turkish having moderate coverage, and Yorùbá absent from the pretraining corpus. XLM-R has approximately 125 million trainable parameters with 12 hidden layers with 768-dimensional hidden states (Conneau et al., 2019).

mBERT was pretrained on Wikipedia data for 104 languages - with the explicit caveat that lower resource languages have less training data overall. For our experimentation, however, all five languages are included in pretraining, a major difference from XLM-R. This model has 110 million trainable parameters and 12 hidden layers with 768-dimensional hidden states, making its size and fine-tuning capabilities comparable to XLM-R (Pires et al., 2019).

Model	EN	ES	ZH	YO	TR
XLM-R	0.821	0.768	0.878	0.809	0.790
mBERT	0.791	0.712	0.860	0.800	0.720

Table 2: Average Macro-F1s for Monolingual Fine-Tuning

4.2 Experimental Setup

Each experiment consisted of five trials with an 80-10-10 training-validation-testing split. Datasets included a 'euph_status' feature distinguishing always-euphemistic from sometimes-euphemistic PETs. To prevent memorization, always-euphemistic PETs appeared only in training or testing, ensuring the model learned euphemism use from context. We used standard hyperparameters, including a 1×10^{-5} learning rate, AdamW optimizer, and batch size of 4. Fixed learning rate experiments performed similarly or worse. Models were trained for up to 15 epochs with early stopping (patience = 5). Training on GPUs took ∼6 hours for sequential fine-tuning and ∼5 hours for simultaneous fine-tuning per language pair.

These experiments are designed to test the ability to transfer knowledge of euphemisms learned in one language to another language. To assess the directionality of transfer, all language pairs are evaluated bidirectionally (e.g., English → Yorùbá and Yorùbá →English), allowing us to analyze both symmetric and asymmetric patterns of cross-lingual adaptation.

5 Simultaneous and Sequential Fine-Tuning Results

5.1 Baseline

We maintain a consistent parameter setting for the monolingual experiments as done for sequential and simultaneous fine-tuning models. We observe a decrease in training time due to the relatively smaller size of the training data. As shown in Table 2, XLM-R consistently outperforms mBERT, possibly due to the slight difference in number of trainable parameters.

5.2 Simultaneous Fine-Tuning

Simultaneous fine-tuning involves combining two languages' datasets for training and validation, while testing on each language separately. Along with the addition of Turkish (Biyik et al., 2024), our experimentation differs from prior work by using the aforementioned euphemism status-based zero-shot setting, where datasets are shuffled without designating an L1 or L2. Table 4 reports the results.

For XLM-R, Chinese performed well across most pairs (see Table 3), likely due to strong pretraining data and corpus quality. The only minor drop occurred when paired with Turkish, but it was negligible. Turkish showed stable performance, suggesting compatibility with other languages. Yorùbá struggled in some cases, especially with English, likely due to limited pretraining and English's dominance in XLM-R. Spanish benefited from typologically similar pairs, while English showed mixed results depending on its counterpart. These findings highlight that typological similarity, dataset composition, and pretraining exposure

[1]Pretraining coverage for mBERT can be found here and for XLM-R here.

Github for Code and Data

Lang. Pair	XLM-R		mBERT	
	Lang A	Lang B	Lang A	Lang B
EN & ES	0.821 (0.821)	0.781 (0.768)	0.801 (0.791)	0.733 (0.712)
EN & ZH	0.829 (0.821)	0.885 (0.878)	0.808 (0.791)	0.852 (0.860)
EN & YO	0.829 (0.821)	0.455 (0.809)	0.789 (0.791)	0.814 (0.800)
EN & TR	0.832 (0.821)	0.817 (0.790)	0.803 (0.791)	0.759 (0.720)
ES & ZH	0.768 (0.768)	0.893 (0.878)	0.732 (0.712)	0.850 (0.860)
ES & YO	0.741 (0.768)	0.797 (0.809)	0.728 (0.712)	0.800 (0.800)
ES & TR	0.751 (0.768)	0.802 (0.790)	0.700 (0.712)	0.731 (0.720)
ZH & YO	0.882 (0.878)	0.824 (0.809)	0.855 (0.860)	0.808 (0.800)
ZH & TR	0.873 (0.878)	0.808 (0.790)	0.831 (0.860)	0.747 (0.720)
YO & TR	0.811 (0.809)	0.795 (0.790)	0.793 (0.800)	0.729 (0.720)

Table 3: Average Macro-F1s for Simultaneous Fine-Tuning. Monolingual (Baseline) scores are reported in parentheses. F1 scores outperforming the baseline are underscored.

all impact multilingual, simultaneous fine-tuning effectiveness.

mBERT's results for simultaneous fine-tuning were not as prominently different from a language's corresponding baseline as seen in the results for XLM-R.

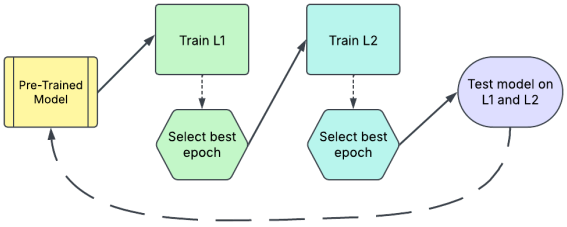

Figure 1: Model Archetype

5.3 Sequential Fine-Tuning

Sequential fine-tuning starts with an off-the-shelf model, which is first fine-tuned on a source language (L1) and then on a target language (L2). The best epoch (based on validation macro F1) on L1 is then fine-tuned on L2. Once the model reaches its highest validation F1 for L2, it is tested on both languages. This process is repeated with a new model for each trial split, and final F1 scores are averaged. The full setup is illustrated in Figure 1, and is performed for both XLM-R and mBERT.

5.3.1 When Does Sequential Fine-Tuning Help L2 Performance?

Sequential fine-tuning for the majority of experiments with both models reported higher scores for the L2, supporting our original hypothesis that prior knowledge of euphemisms in L1 aids understanding in L2. This setup shows more distinct differences than simultaneous fine-tuning, where scores remained closer to the monolingual baseline.

Examples from Table 4 (XLM-R):

- **EN** → **YO**: 0.812 vs. YO baseline 0.809
- **ES** → **YO**: 0.830 vs. 0.809
- **EN** → **TR**: 0.801 vs. TR baseline 0.790
- **YO** → **ZH**: 0.900 vs. ZH baseline 0.878

These gains are particularly notable in low-resource target languages like Yorùbá and Turkish, reinforcing the benefit of high-resource L1 transfer.

5.3.2 When Does Sequential Fine-Tuning Hurt L1 Performance? (Catastrophic Forgetting)

In most cases, L1 performance drops after L2 training—particularly in XLM-R, with some cases resulting in severe performance degradation indicating catastrophic forgetting. This is evident in:

- **YO** → **EN**: 0.490 vs. YO baseline 0.809
- **YO** → **ZH**: 0.432 vs. YO baseline 0.809

These considerable drops are not observed in mBERT, likely due to balanced pretraining coverage for all five languages. XLM-R lacks pretraining exposure to Yorùbá, which leads to shallow integration that is easily overwritten. Agreement metrics support this interpretation: Cohen's Kappa between YO monolingual and YO → ZH is 0.145, while YO monolingual and ZH → YO yields 0.667.

5.3.3 What Roles Do Pretraining, Typology, and Data Play?

Pretraining coverage significantly impacts transfer success. XLM-R achieves stronger L2 gains but

Train		Test - XLM-R		Test - mBERT	
L1	**L2**	**L1**	**L2**	**L1**	**L2**
ES	EN	0.733 (0.768)	0.824 (0.821)	0.702 (0.712)	0.799 (0.791)
ZH	EN	0.791 (0.878)	0.830 (0.821)	0.809 (0.860)	0.812 (0.791)
YO	EN	0.490 (0.809)	0.800 (0.821)	0.678 (0.800)	0.785 (0.791)
TR	EN	0.732 (0.790)	0.835 (0.821)	0.660 (0.720)	0.791 (0.791)
EN	ES	0.780 (0.821)	0.761 (0.768)	0.745 (0.791)	0.738 (0.712)
ZH	ES	0.843 (0.878)	0.746 (0.768)	0.843 (0.860)	0.722 (0.712)
YO	ES	0.709 (0.809)	0.746 (0.768)	0.770 (0.800)	0.717 (0.712)
TR	ES	0.676 (0.790)	0.764 (0.768)	0.622 (0.720)	0.690 (0.712)
EN	ZH	0.797 (0.821)	0.876 (0.878)	0.783 (0.791)	0.868 (0.860)
ES	ZH	0.743 (0.768)	0.876 (0.878)	0.727 (0.712)	0.885 (0.860)
YO	ZH	0.432 (0.809)	0.900 (0.878)	0.701 (0.800)	0.854 (0.860)
TR	ZH	0.704 (0.790)	0.857 (0.878)	0.676 (0.720)	0.858 (0.860)
EN	YO	0.761 (0.821)	0.812 (0.809)	0.735 (0.791)	0.817 (0.800)
ES	YO	0.661 (0.768)	0.830 (0.809)	0.734 (0.712)	0.801 (0.800)
ZH	YO	0.837 (0.878)	0.798 (0.809)	0.827 (0.860)	0.809 (0.800)
TR	YO	0.727 (0.790)	0.824 (0.809)	0.703 (0.720)	0.816 (0.800)
EN	TR	0.767 (0.821)	0.801 (0.790)	0.765 (0.791)	0.780 (0.720)
ES	TR	0.644 (0.768)	0.777 (0.790)	0.662 (0.712)	0.741 (0.720)
ZH	TR	0.692 (0.878)	0.792 (0.790)	0.727 (0.860)	0.758 (0.720)
YO	TR	0.674 (0.809)	0.776 (0.790)	0.760 (0.800)	0.742 (0.720)

Table 4: Comparison of Average Macro-F1s for Sequential Fine-Tuning vs. Monolingual Baseline. Parentheses contain monolingual F1 for reference. F1-scores that outperform the baseline are underscored. Scores highlighted in blue are where L2 performs better than L1, those highlighted in yellow are where L1 outperforms.

suffers more from volatility, while mBERT shows steadier though lower performance. Yorùbá performs well as L2 but poorly as L1, which tracks with its absence from XLM-R's pretraining corpus.

Typological similarity alone does not explain results. For instance, EN → TR and ES → YO show strong gains despite language distance, whereas EN → ES does not yield consistent improvement.

These findings suggest that dataset characteristics and pretraining exposure are more influential than typological features in euphemism detection transfer.

5.3.4 Summary of Key Results

Table 5 highlights the top-performing configurations for each language in both models. For XLM-R, the strongest gain occurs in the YO → ZH setting (0.900), outperforming the ZH monolingual baseline (0.878).

In contrast, mBERT produces more balanced results across languages, with no extreme gains but consistent improvements when English is used as the source language. While mBERT avoids catastrophic forgetting due to its more uniform pretraining coverage, XLM-R achieves higher absolute L2 performance, especially for low-resource L2s. These results underscore that sequential fine-tuning is a lightweight and effective strategy for improving euphemism detection across typologically diverse and unevenly resourced languages.

Model	Lang.	Pair	Type	F1
XLM-R	EN	TR → EN	Seq.	0.835
XLM-R	ES	ES & ZH	Sim.	0.768*
XLM-R	ZH	YO → ZH	Seq.	0.9
XLM-R	YO	ES → YO	Seq.	0.830
XLM-R	TR	EN & TR	Sim.	0.817
mBERT	EN	ZH → EN	Seq.	0.812
mBERT	ES	EN → ES	Seq.	0.738
mBERT	ZH	ES → ZH	Seq.	0.885
mBERT	YO	EN → YO	Seq.	0.817
mBERT	TR	EN → TR	Seq.	0.780

Table 5: Highest F1 scores for models in each of the languages over the two fine-tuning setups: sequential (Seq.) and simultaneous (Sim.). * indicates score matches baseline performance.

6 Conclusion and Future Work

This paper explored whether euphemism detection can benefit from cross-lingual transfer, specifically through sequential fine-tuning. We evaluated XLM-R and mBERT across five typologically and resource-diverse languages: English, Spanish, Chinese, Turkish, and Yorùbá.

Our findings show that sequential fine-tuning with a high-resource language improves L2 euphemism detection, especially for low-resource languages like Yorùbá and Turkish. XLM-R achieves larger gains, but is more sensitive to catastrophic forgetting and pretraining gaps. mBERT, by contrast, shows more stable performance across language pairs, albeit with smaller improvements.

Interestingly, the success of transfer was not predicted by typological similarity. Instead, per-

formance was shaped more by dataset structure and pretraining exposure. Strong results for Yorùbá→Chinese and English→Turkish demonstrate that meaningful transfer can occur even between distant languages.

Overall, these results highlight sequential fine-tuning as a lightweight and effective adaptation strategy for figurative language tasks and extends previous studies by introducing cross-lingual transfer investigations in relation to a challenging task. In future work, we plan to explore few-shot sequential fine-tuning, hybrid multilingual-sequential setups, and extensions to languages with non-Latin scripts and richer morphology.

Future work could explore cyclical fine-tuning or interleaved exposure to counteract forgetting, and longer L2 training where L1 outperforms. Testing whether Yorùbá's weaker performance extends to other low-resource languages could reveal if sequential fine-tuning serves as implicit pretraining.

Evaluating larger multilingual models (e.g., mT5, GPT-4, Mistral) may enhance cross-lingual euphemism detection, particularly for low-resource languages. Expanding to morphologically rich and non-Latin scripts could uncover new challenges, while discourse-level modeling may improve context sensitivity. This study shows that prior exposure to euphemisms in L1 enhances cross-lingual transfer, but effectiveness depends on pretraining data, dataset structure, and linguistic differences. Sequential fine-tuning provides a scalable strategy for improving LLM's ability to detect figurative language in low-resource settings, thus contributing to the development of more effective multilingual NLP models.

Limitations

Our study has several limitations in cross-lingual euphemism detection. Dataset imbalance affects comparability, as Spanish contains significantly more PETs than other languages, which may skew model performance. XLM-R's pretraining bias favors English, Spanish, and Chinese, while Turkish has moderate coverage, and Yorùbá has none, contributing to its weaker performance in some settings. Furthermore, most of the datasets were skewed towards the 1's (i.e. euphemistic contexts), with the Chinese dataset and the Spanish dataset having nearly 2/3 of their instances labeled as Euphemistic.

Catastrophic forgetting occurred in sequential fine-tuning with XLM-R, where L1 performance dropped after exposure to L2, particularly in YO → EN and YO → ZH, indicating interference in euphemism learning. Typology did not strictly predict transfer success – some distant pairs (e.g., EN → TR, ES → YO) showed gains, while structurally similar languages (e.g., English → Spanish) did not, suggesting dataset complexity and euphemism structures play a larger role.

Computational constraints may have impacted results, but we did not systematically test training duration and batch sizes with larger models. Due to efficiency constraints, we were only able to perform 5 trials on each pair for sequential fine-tuning - although the model still sees the entirety of our datasets, it does not receive as much variability in regards to the random shuffles.

Generalizability remains uncertain, as all studied languages use relatively simple scripts, with the exception of Chinese, which uses a logographic script, leaving open questions about languages with complex morphology or non-Latin scripts. Finally, euphemism detection is inherently subjective, meaning dataset inconsistencies and cultural variation may introduce noise.

Ethics Statement

This study acknowledges the cultural and linguistic variations in euphemism detection, as meanings shift across contexts. The data used in this work was made publicly available by the authors of Lee and Feldman (2024) and is used in accordance with their original intent.

Our dataset includes euphemisms related to sensitive topics like death, illness, and socio-political issues, and may include vulgar language. We are not policing language – our goal is to enhance cross-lingual understanding of euphemistic language. The data used does not contain personally identifying information.

Our research includes low-resource languages like Yorùbá, which often lack strong NLP infrastructure. By working with these languages, we aim to support more inclusive language technologies.

Acknowledgments

Thanks to Patrick Lee, Hasan Biyik, and Whitney Poh for help with annotation, experiments, and feedback. This material is based upon work supported by the National Science Foundation under Grants #2226006 and #2428506.

References

Hasan Biyik, Patrick Lee, and Anna Feldman. 2024. Turkish delights: a dataset on Turkish euphemisms. In *Proceedings of the First Workshop on Natural Language Processing for Turkic Languages (SIGTURK 2024)*, pages 71–80, Bangkok, Thailand and Online. Association for Computational Linguistics.

Brightmart. 2019. NLP Chinese Corpus: Release version 1.0. Accessed via Zenodo.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Unsupervised cross-lingual representation learning at scale. *CoRR*, abs/1911.02116.

Mark Davies. 2013. Corpus of global web-based english: 1.9 billion words from speakers in 20 countries (glowbe). https://corpus.byu.edu/glowbe/. Accessed: 2025-05-25.

Christian Felt and Ellen Riloff. 2020. Recognizing euphemisms and dysphemisms using sentiment analysis. In *Proceedings of the Second Workshop on Figurative Language Processing*, pages 136–145, Online. Association for Computational Linguistics.

Martha Gavidia, Patrick Lee, Anna Feldman, and Jing Peng. 2022. CATs are fuzzy PETs: A corpus and analysis of potentially euphemistic terms. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 2658–2671, Marseille, France. European Language Resources Association.

Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A. Smith. 2020. Don't stop pretraining: Adapt language models to domains and tasks. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8342–8360, Online. Association for Computational Linguistics.

Hanxu Hu, Simon Yu, Pinzhen Chen, and Edoardo M. Ponti. 2024. Fine-tuning large language models with sequential instructions. *Preprint*, arXiv:2403.07794.

Sedrick Scott Keh. 2022. Exploring Euphemism Detection in Few-Shot and Zero-Shot Settings. In *Proceedings of the 3rd Workshop on Figurative Language Processing (FLP)*, pages 167–172, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Sedrick Scott Keh, Rohit Bharadwaj, Emmy Liu, Simone Tedeschi, Varun Gangal, and Roberto Navigli. 2022. EUREKA: EUphemism Recognition Enhanced through Knn-based methods and Augmentation. In *Proceedings of the 3rd Workshop on Figurative Language Processing (FLP)*, pages 111–117, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Guneet Singh Kohli, Prabsimran Kaur, and Jatin Bedi. 2022. Adversarial Perturbations Augmented Language Models for Euphemism Identification. In *Proceedings of the 3rd Workshop on Figurative Language Processing*, pages 154–159. Association for Computational Linguistics.

Patrick Lee, Alain Chirino Trujillo, Diana Cuevas Plancarte, Olumide Ojo, Xinyi Liu, Iyanuoluwa Shode, Yuan Zhao, Anna Feldman, and Jing Peng. 2024. MEDs for PETs: Multilingual euphemism disambiguation for potentially euphemistic terms. In *Findings of the Association for Computational Linguistics: EACL 2024*, pages 875–881, St. Julian's, Malta. Association for Computational Linguistics.

Patrick Lee and Anna Feldman. 2024. Report on the Multilingual Euphemism Detection Task. In *Proceedings of the 4th Workshop on Figurative Language Processing*, pages 110–114. Association for Computational Linguistics.

Patrick Lee, Anna Feldman, and Jing Peng. 2022a. A report on the euphemisms detection shared task. In *Proceedings of the 3rd Workshop on Figurative Language Processing (FLP)*, pages 184–190, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Patrick Lee, Martha Gavidia, Anna Feldman, and Jing Peng. 2022b. Searching for PETs: Using distributional and sentiment-based methods to find potentially euphemistic terms. In *Proceedings of the Second Workshop on Understanding Implicit and Underspecified Language*, pages 22–32, Seattle, USA. Association for Computational Linguistics.

Patrick Lee, Iyanuoluwa Shode, Alain Trujillo, Yuan Zhao, Olumide Ojo, Diana Plancarte, Anna Feldman, and Jing Peng. 2023. FEED PETs: Further experimentation and expansion on the disambiguation of potentially euphemistic terms. In *Proceedings of the 12th Joint Conference on Lexical and Computational Semantics (*SEM 2023)*, pages 437–448, Toronto, Canada. Association for Computational Linguistics.

Telmo Pires, Eva Schlinger, and Dan Garrette. 2019. How multilingual is multilingual BERT? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4996–5001, Florence, Italy. Association for Computational Linguistics.

Real Academia Española. 2025. CORPES XXI: Corpus del Español del Siglo XXI. Accessed: 2025-05-25.

Fedor Vitiugin and Henna Paakki. 2024. Ensemble-Based Multilingual Euphemism Detection: A Behavior-Guided Approach. In *Proceedings of the 4th Workshop on Figurative Language Processing*, pages 73–78. Association for Computational Linguistics.

Yuting Wang, Yiyi Liu, Ruqing Zhang, Yixing Fan, and Jiafeng Guo. 2022. Euphemism Detection by Transformers and Relational Graph Attention Network.

In *Proceedings of the 3rd Workshop on Figurative Language Processing*, pages 79–83. Association for Computational Linguistics.

Michael Wiegand, Jana Kampfmeier, Elisabeth Eder, and Josef Ruppenhofer. 2023. Euphemistic Abuse – A New Dataset and Classification Experiments for Implicitly Abusive Language. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 16280–16297. Association for Computational Linguistics.

Wanzheng Zhu, Hongyu Gong, Rohan Bansal, Zachary Weinberg, Nicolas Christin, Giulia Fanti, and Suma Bhat. 2021. Self-supervised euphemism detection and identification for content moderation. *CoRR*, abs/2103.16808.