# Report on the Multilingual Euphemism Detection Task

**Patrick Lee** and **Anna Feldman**
Montclair State University
New Jersey, USA
{leep,feldmana}@montclair.edu

## Abstract

This paper presents the Multilingual Euphemism Detection Shared Task for the Fourth Workshop on Figurative Language Processing (FigLang 2024) held in conjunction with NAACL 2024. Participants were invited to attempt the euphemism detection task on four different languages (American English, global Spanish, Yorùbá, and Mandarin Chinese): given input text containing a potentially euphemistic term (PET), determine if its use is euphemistic or not. We present the expanded datasets used for the shared task, summarize each team's methods and findings, and analyze potential implications for future research.

## 1 Introduction

Euphemisms are a linguistic device used to soften or neutralize language that may otherwise be harsh or awkward to state directly (e.g., "between jobs" instead of "unemployed","late" instead of "dead", "collateral damage" instead of "war-related civilian deaths"). By acting as alternative words or phrases, euphemisms are used in everyday language to maintain politeness, mitigate discomfort, or conceal the truth. While they are culturally-dependent, the need to discuss sensitive topics in a non-offensive way is universal, suggesting similarities in the way euphemisms are used across languages and cultures.

Terms which may be used euphemistically sometimes require context to determine a euphemistic usage:

*Asked to choose between jobs and the environment, a majority – at least in our warped, first-past-the-post system – will pick jobs.* (non-euphemistic)

*This summer, the budding talent agent was between jobs and free to babysit pretty much any time.* (euphemistic)

In this shared task, participants were invited to develop approaches and models to disambiguate texts (in multiple languages) as either euphemistic or not. The previous iteration of this task resulted in numerous insights from participating teams, but featured only an English dataset (Lee et al., 2022a). By providing a multilingual iteration, we hoped to extend these findings to other languages and employ transfer learning to uncover possible cross-lingual patterns (Shode et al., 2023). This paper is structured as follows: Section 2 describes related work, Section 3 describes the additional data collected for the competition[1] and the task setting, Section 4 summarizes the participants' methods and results, and Section 5 analyzes common findings and the future directions they suggest.

## 2 Related Work

Magu and Luo (2018) and Felt and Riloff (2020) explored word embeddings and sentiment analysis, respectively, for detecting euphemisms. Zhu and Bhat (2021) and subsequent works such as (Lee et al., 2022a) and Lee et al. (2023) advanced this research using BERT and other transformers for euphemism detection and disambiguation. Keh (2022) focused on classifying previously unseen euphemistic phrases. Gavidia et al. (2022) built a corpus of potentially euphemistic terms (PETs) influencing further studies (Lee et al., 2022b,a, 2023). Most recently, (Lee et al., 2024) demonstrated the effectiveness of XLM-RoBERTa in multilingual euphemism disambiguation, showing superior performance of multilingual over monolingual models and enabling zero-shot learning across languages (refer to Table 1 for the average macro-F1 scores from multilingual and cross-lingual experiments).

[1]The final datasets, as well as the specific train-test split used for the competition, are available at https://github.com/pl464/euph-detection-datasets/tree/main/EACL_2024



*Proceedings of the 4th Workshop on Figurative Language Processing (FLP)*, pages 110–114

Table 1: Average Macro-F1s for multi- and cross-lingual experiments. ZH=Mandarin Chinese, EN=American English, ES=Global Spanish, and YO=Yorùbá

| TrainTest | ZH | EN | ES | YO |
|---|---|---|---|---|
| **Baseline** | 0.426 | 0.416 | 0.381 | 0.394 |
| **ZH** | **0.879** | 0.653 | 0.535 | 0.300 |
| **EN** | 0.607 | **0.765** | 0.567 | 0.381 |
| **ES** | 0.613 | 0.639 | **0.752** | 0.384 |
| **YO** | 0.417 | 0.407 | 0.383 | **0.790** |
| **ZH+EN** | 0.897 | 0.804 | 0.508 | 0.397 |
| **EN+ES** | 0.650 | 0.781 | 0.764 | 0.416 |
| **ES+YO** | 0.605 | 0.630 | 0.758 | 0.794 |
| **ZH+ES** | 0.884 | 0.670 | 0.764 | 0.377 |
| **EN+YO** | 0.616 | 0.772 | 0.602 | 0.802 |
| **ZH+YO** | 0.881 | 0.646 | 0.585 | 0.795 |
| **ZH+EN+ES** | 0.898 | 0.805 | 0.775 | 0.389 |
| **EN+ES+YO** | 0.647 | 0.783 | 0.772 | 0.791 |
| **ZH+EN+YO** | 0.899 | 0.801 | 0.555 | 0.794 |
| **ZH+ES+YO** | 0.885 | 0.664 | 0.778 | 0.778 |
| **All** | 0.895 | 0.792 | 0.776 | 0.793 |

## 3 Task Setting

### 3.1 Multilingual Datasets

The training data used in this competition were the labelled datasets in American English (EN), Spanish (ES), Yorùbá (YO), and Mandarin Chinese (ZH) constructed and described by Lee et al. (2023). Source texts were collected from a variety of sources that comprised primarily of online articles and webpages (though the Spanish and Yorùbá datasets included other sources, such as transcribed texts and social media posts). Each instance contained up to 3 sentences and contained a potentially euphemistic term (PET). These texts were also human-annotated with labels indicating either a euphemistic (1) or non-euphemistic (0) usage of the PET. Special tokens were placed before and after the PET in each instance, which we standardize for the shared task as "**[PET_BOUNDARY]**". Additionally, as euphemisms can be language-specific, data for each language were collected separately (i.e. are not translations of each other) and differed in PET and label distributions.

Since these datasets were already publicly available, we collected additional data in each of the four languages to comprise the test sets. The data were from the same source corpora as the training data and were annotated by 2-3 native speakers in each language. The final distribution of examples in the training and test set can be found in Table 2. Note that the goal was not only to provide unseen examples for the shared task, but also to contribute additional data for multilingual euphemism detection in general; therefore, test sets sometimes contained entirely new PETs, but to varying extents across languages as shown in Table 3. Prior work has shown that when new PETs are introduced at test time, models have a more difficult time correctly classifying them (Keh, 2022). As a result of this and other differences between the datasets, classification metrics among the participants should not be compared across languages, but "within" languages.

| Lang | Train | | Test | |
|---|---|---|---|---|
| | 1s | 0s | 1s | 0s |
| EN | 1339 | 563 | 502 | 694 |
| ES | 1146 | 715 | 809 | 282 |
| YO | 1270 | 659 | 419 | 250 |
| ZH | 1469 | 516 | 744 | 482 |

Table 2: Number of examples per label in train and test. "1s" refers to euphemistic examples, and "0s" refers to non-euphemistic examples.

| Lang | Number of PETs | | |
|---|---|---|---|
| | Train | Test | Overlap |
| EN | 121 | 67 | 44 |
| ES | 148 | 85 | 0 |
| YO | 133 | 28 | 4 |
| ZH | 110 | 48 | 7 |

Table 3: Number of PETs/overlap between train and test

### 3.2 Task Description

The shared task was hosted as a competition on Codabench[2]. During the development phase, participants were provided with datasets in all four languages. During the test phase, participants were provided a test set for each language and had the option of submitting predictions for one to four of them for scoring. However, all teams ultimately chose to submit predictions for all four. The metric for comparison was Macro-F1, and the submissions were ranked using the average Macro-F1 across all four languages, weighted equally.

| # | User | EN | ES | YO | ZH | AVG | Title of Paper |
|---|---|---|---|---|---|---|---|
| 1 | amri228 | 0.83 | 0.60 | 0.72 | 0.78 | 0.73 | Can GPT4 Detect Euphemisms across Multiple Languages? (Firsich and Rios, 2024) |
| 2 | vitiugin | 0.74 | 0.67 | 0.63 | 0.71 | 0.69 | Ensemble-based Multilingual Euphemism Detection: a Behavior-Guided Approach (Vitiugin and Paaki, 2024) |
| 3 | nhankins | 0.65 | 0.61 | 0.65 | 0.68 | 0.65 | Optimizing Multilingual Euphemism Detection using Low-Rank Adaptation Within and Across Languages (Hankins, 2024) |
| 4 | Baseline | 0.30 | 0.43 | 0.39 | 0.38 | 0.37 | – |

Table 4: Results of submitted systems to the Multilingual Euphemism Detection Task

## 4 Participants and Results

In all, there were 3 teams that participated in the task and also submitted descriptions of their systems. A summary of their performances are in Table 4, along with a majority class baseline. In this section, we briefly describe each team's approach and results.

### 4.1 GPT-4 in Zero-Shot and Few-Shot Settings

Firsich and Rios (2024) submitted the highest-scoring approach (based on averaged F1 across all four languages), which explored zero-shot and few-shot prompts with GPT4 for the task. Their zero-shot setting consisted of instructions and then the task prompt, optionally accompanied by "Context", or a description of what euphemisms are. Their few-shot setting consisted of the above, plus $k$ examples of euphemistic and non-euphemistic instances with labels. On the development set, they confirm that the highest setting of $k$=8 yields the highest scores by a significant margin over $k$=2, which is also significantly better than $k$=0. Moreover, providing few-shot examples that contained the same PET as in the task prompt was always better. This is an intuitive result, and it seems the model is able to better leverage more directly related examples to do a better job of disambiguating PET usages. Additionally, providing the "context" of what euphemisms are boosted performance significantly for the zero-shot setting (e.g. for Yorùbá, 0.400 $\rightarrow$ 0.610).

On the shared task's test set, they scored the highest in all categories except Spanish. Performances on all languages except English dropped significantly from the best setting in the development set (ES: 0.761 $\rightarrow$ 0.598, YO: 0.872 $\rightarrow$ 0.723, ZH: 0.858 $\rightarrow$ 0.776). This likely correlates with the degree of "PET overlap" (see Table 3) for which English is very high, Spanish and Yorùbá very low, and Chinese in-between.

### 4.2 Behavior-Guided, Ensemble-based Approach

Vitiugin and Paaki (2024) develop an approach using an ensemble of multilingual transformers (XLM-RoBERTa-large, or XLM-R), each fine-tuned on either the euphemism detection task or one of several "behavior-related" tasks (sarcasm and irony detection, sexism detection, racism detection, and sentiment classification) that are potentially related to general euphemism understanding. The authors cite multiple works in which training on such tasks, as well as ensembling, have been shown to improve performance on figurative language tasks. Unlike the previous system, they train and test on data from all four languages at once.

They found the best approach on the development set to be a Random Forest ensemble of 6 models: all 4 behavior-related fine-tuned models, and 2 trained on the euphemism detection task, one of which with PETs removed from the text, and the other as normal. This decision may have stemmed from the observation that PETs are unevenly distributed in the dataset, and the model should learn to classify based on context. While their reported performance on the development set was very high (F1 = 0.95), it was much lower on the test set (average F1 across the four languages = 0.69), though they yielded the highest score on Spanish in the competition (F1 = 0.67). This suggests some kind of significant overfitting, perhaps in regards to PETs, though no connection to "PET overlap" can be made, as their validation perfor-

[2] https://www.codabench.org/competitions/1959/

mance was not reported for each of the languages separately.

### 4.3 Optimizing and Low-Rank Adaptation Approach

Hankins (2024) experiment with multiple multilingual transformer models with a focus on efficient methods. On the development data, they find that fine-tuning multilingual DistilBERT (base, cased) with Low-Rank Adaptation (LoRA) yields comparable performances to using XLM-R (F1 ∼0.74-0.85), while being much lighter and faster to train. However, as with the other teams' approaches, the performance on the test set was much lower all around (∼0.61-0.68). This may suggest that, while more parameter-efficient approaches work well when tested on PETs seen during training, a larger number of parameters may be needed for capturing the nuances associated with unseen PETs.

## 5 Discussion and Future Work

Here, we discuss some common themes among the participants' approaches and suggest related directions for future work.

### 5.1 PETs Matter

There are many indications that the distribution of PETs in the data seems to matter to a large extent. Not only are the test score degradations correlated with PET overlaps in each language, but each language's relative score also seems correlated with the overall number of PETs in the dataset (e.g., Spanish had the most unique PETs total, 233, and generally performed the worst; English had the least, 144, and performed the best). Furthermore, the degree of difficulty may vary by PET, as well. In addition to the varying label distributions per PET (e.g. ten 1's and ten 0's for one PET, but thirty 1's and no 0's of another), the complexity of some PETs may also differ. Firsich and Rios (2024) noted the examples of the English PET "disabled" and Chinese PET "环卫工人", which intuitively seemed difficult to classify and in fact required relatively many examples of the same PET to improve performance.

All in all, it seems that this task is inherently tied to the varying types of PETs present. It is suggested that future work should pay special attention to this aspect, perhaps experimenting with different ranges, amounts, or linguistic qualities of PETs.

### 5.2 Analyzing Model Predictions

As mentioned in the previous section, Firsich and Rios (2024) observed that PETs may have different "classification difficulties" by looking past the classification metrics and at actual predictions. Hankins (2024) additionally report the distributions of predictions made by models trained on different languages. While they found, somewhat unsurprisingly, that test performance on language X is highest with a model trained on data from all four languages (i.e. is trained four times as much data), it makes significantly different predictions than a model only trained on language X, particularly for Chinese and English. This suggests that training on multiple languages results in significantly different learned representations of languages for this task. Overall, it is suggested to analyze prediction distributions and error analyses to further understand model behavior.

### 5.3 Linguistically Related Knowledge

Euphemism detection may involve many different forms of pragmatic knowledge - politeness, offensiveness, directness, conciseness, sentiment, sensitivity, etc. One way to leverage this intuition computationally is to explicitly teach models these tasks, as explored by Vitiugin and Paaki (2024), or include them as part of model inputs. The validation scores from Firsich and Rios (2024) show that including a definition of euphemisms in prompts benefits GPT4 in the zero-shot setting almost as much as providing randomized (i.e. not having the same PET) few-shot examples. Additionally, models trained on euphemism detection may also implicitly encode this knowledge, and perhaps differently across languages. These are all potential findings for future computational work to uncover.

## 6 Conclusion

We present the results of the Multilingual Euphemism Detection Shared Task. Participants' systems scored well above the baselines, but well below their reported validation metrics. Taken together, these results invite further work into using LLMs, ensembling/related tasks, and efficient models, which showed proficiency across languages, but leave much room for improvement. From a synthesis of the teams' findings, we also suggest that future work explore the impact of PETs, model behavior beyond performance metrics, and connections with related linguistic tasks.

## Limitations

The primary limitations of the work include inconsistent performance across languages, particularly in non-English languages due to varying degrees of potentially euphemistic term overlap and limited model robustness in handling diverse linguistic data.

## Ethics Statement

The authors foresee no ethical concerns with the work presented in this paper.

## Acknowledgements

We would like to thank Alain Chirino Trujillo, Diana Cuevas Plancarte, Iyanuoluwa Shode, Julia Sammartino, Olumide Ebenezer Ojo, Thomas Hicks, Xinyi Liu, and Yuan Zhao for for all the help with collecting and annotating datasets and preparing them for the shared task.

This material is based upon work supported by the National Science Foundation under the Grant number 2226006.

## References

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.

Todd Firsich and Anthony Rios. 2024. Can gpt4 detect euphemisms across multiple languages? In *In Proceedings of the 4th Workshop on Figurative Language Processing co-located with NAACL 2024, Mexico City, June, 2024. To appear.*

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.

Nicholas Hankins. 2024. Optimizing multilingual euphemism detection using low-rank adaptation within and across languages. In *In Proceedings of the 4th Workshop on Figurative Language Processing co-located with NAACL 2024, Mexico City, June, 2024. To appear.*

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.

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, 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.

Rijul Magu and Jiebo Luo. 2018. Determining code words in euphemistic hate speech using word embedding networks. In *Proceedings of the 2nd workshop on abusive language online (ALW2)*, pages 93–100.

Iyanuoluwa Shode, David Ifeoluwa Adelani, Jing Peng, and Anna Feldman. 2023. Nollysenti: Leveraging transfer learning and machine translation for nigerian movie sentiment classification. pages 986–998.

Fedor Vitiugin and Henna Paaki. 2024. Ensemble-based multilingual euphemism detection: a behavior-guided approach. In *In Proceedings of the 4th Workshop on Figurative Language Processing co-located with NAACL 2024, Mexico City, June, 2024. To appear.*

Wanzheng Zhu and Suma Bhat. 2021. Euphemistic phrase detection by masked language model. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 163–168, Punta Cana, Dominican Republic. Association for Computational Linguistics.