Optimizing Multilingual Euphemism Detection using Low-Rank Adaptation Within and Across Languages

Nicholas Hankins

nicholasjhankins@gmail.com,
hankinsn1@montclair.edu

Abstract

This short paper presents an investigation into the effectiveness of various classification methods as a submission in the Multilingual Euphemism Detection Shared Task for the Fourth Workshop on Figurative Language Processing co-located with NAACL 2024. The process utilizes pre-trained large language models combined with parameter efficient fine-tuning methods, specifically Low-Rank Adaptation (LoRA), in classifying euphemisms across four different languages - Mandarin Chinese, American English, Spanish, and Yorùbá. The study is comprised of three main components that aim to explore heuristic methods to navigate how base models can most efficiently be fine-tuned into classifiers to learn figurative language. Multilingual labeled training data was utilized to fine-tune classifiers for each language, and later combined for one large classifier, while unseen test data was finally used to evaluate the accuracy of the best performing classifiers. In addition, cross-lingual tests were conducted by applying each language's data on each of the other language's classifiers. All of the results provide insights into the potential of pre-trained base models combined with LoRA fine-tuning methods in accurately classifying euphemisms across and within different languages.

1 Introduction

In order to best understand this task, it is important to define what a euphemism is. Euphemisms are a linguistic device used to soften statements, or to make statements more polite. Some examples of a euphemism might be using the terms "between jobs" or "late" instead of "unemployed" or "dead," respectively (Lee et al. 2024). Research proves that euphemisms are a multilingual feature that exists in numerous languages (Gavidia et al. 2022). By collecting more training data and testing on unseen data, we are further able to see the extent of how state-of-the-art language modeling captures these universally figurative traits.

The ability to observe whether these elements of figurative language are taken into consideration during tasks like classification by large language models (LLM) can be speculated as a topic of increasing interest in natural language processing communities. The growing number of base models, such as XLM-RoBERTa, that can be utilized for downstream tasks like text classification, reasoning, and sequence generation is staggering and leads to further questions of how the existing methods can be tested and improved (Conneau et al. 2019). By addressing the numerous kinds of euphemistic categories, and how they can be represented multilingually, this kind of research enables a greater level of natural language understanding by embodying an ambiguous and subjective aspect of languages (Lee et al. 2024). Furthermore, by aiming to solve the problem of accurate classification of figurative language using machine learning, this task importantly measures how well a human language characteristic can be interpreted by LLMs.

2 Related Work

Prior research determined that semantic category might influence cross-lingual transfer of information (Lee et al. 2024). This insight drives the intuition for this experiment. Once the ostensibly optimal classification method is discovered, then we can perform a cross-lingual comparison to see how all other languages performed on classifiers fine-tuned for other languages. Previous work is helpful in this regard, as it enables us to have a starting point to compare and contrast base models, which were chosen heuristically. The Multilingual

Proceedings of the 4th Workshop on Figurative Language Processing (FLP), pages 8–14

Classifier predictions will be included in the results as a comparison. It is important to note that the base model for the cross-lingual experiment remained the same. In other words, all languages had context for each inference, yet the fine-tuning the classifier certainly made a difference in the results. The complete visualization of this can be seen in Figure 1.

By freezing all the parameters in the base model with the Parameter-Efficient Fine-Tuning (PEFT) method, we are able to explicitly train the new classifier on our input language datasets. This will ideally be able to increase the speed with which we fine-tune and keep majority of the base model parameters frozen (Hu et al. 2021). Using parameter efficient fine-tuning essentially allows our model to be trained on a small set of new parameters, which is why PEFT was used for this experiment. The goal is to see how well we can utilize these methods for our classification purposes. Multilingual word embeddings have been shown to also have produced positive results in text classification tasks (Plank 2017).

Through examining the problem posed in the introduction of best classifying figurative language through fine-tuning LLMs, and incorporating adjacent work that has proven successful on tasks with similar goals, we can begin to formulate an overarching methodology. Starting with different base models to explore a variety of new options, keeping in mind the limited compute resources available, we can focus our efforts in changing as little as possible from the base model in an effort to highlight the impact of the task's training and test data throughout the experiment. LoRA, specifically PEFT, allows us to do this by inspecting the given data, specifically how semantic information is transferred accordingly, in the model predictions. The aim is to emphasize how the arrangement of the model's data can affect classification predictions.

3 Experiment 1 - Choosing the Base Model

3.1 Methodology

The training data included examples with a column for the text containing the Potentially Euphemistic Term (PET), the assigned label of 1 signifying that the term is euphemistic, or 0 if not, and the PET category. After obtaining the training data, one of the primary characteristics observed was the imbalance of the data. More specifically, each language had more positive, or euphemistic, labels which indicates that the training datasets are imbalanced (See Table 1). This imbalance problem was addressed with adjusting the learning rate during hyperparameter weight setting.

	En.	Sp.	Ch.	Yo.
Euph.	1383	1143	1484	1281
Non-Euph.	569	718	521	660

Table 1: Training Data Split between Euphemistic and Non-Euphemistic Examples (English, Spanish, Chinese, Yorùbá)

It is important to note the unique counts of PET categories present in each dataset considering the impact that they might have, despite the PET category not being explicitly included in the fine-tuning process. Only the text and label columns were input into the trainer function. Nonetheless, this way, we can intuitively observe the classifier results and see the PET characteristics, such as quantity, uniqueness, and frequency in the data. The total counts by themselves do not provide much insight given that each example has one by design, however, it is helpful to know the count of unique PET categories that may be prompting the fine-tuning step with semantically relevant information embedded within the associated input text. English has 163 unique PET categories, Spanish has 147, Chinese has 110, and Yorùbá has 133.

All the classifiers were trained on T4 GPU and incorporated the models, tokenizers, and LoRA adapters via HuggingFace's platform (Wolf et al. 2020). They were all preprocessed the same way by utilizing an Autotokenizer and were set to initiate data collation for training (*YouTube-Blog* 2024). While tuning the hyperparameters, it was discovered that having all the linear target modules instantiated maximized the number of trainable parameters, as supported by prior studies into LoRA techniques (Dettmers et al. 2023). This meant that the most crucial hyperparameter was the number of LoRA adapters, thus ensuring full capability of fine-tuning performance (Dettmers et al. 2023). The number of LoRA linear layers included in the

Languages	First Test	Second Test
English	0.85045	0.84158
Spanish	0.77704	0.74321
Chinese	0.84438	0.85454
Yorùbá	0.81308	0.80423

Table 2: Maximum Validation F1 scores of the 10 epochs for both experiments. The final epoch results may be lower during inference on the test data.

PEFT model instantiation was 6 in total [1].

Moreover, all the classifiers ultimately were set to having a learning rate of 1e-5 and trained on 10 epochs in order to create consistency. It should be noted that the original metric scores of the first base model experiments had varying learning rates, which may have had an impact on the training process due to the inherent data imbalance.

The first iteration of fine-tuning used the uncased DistilBERT base model for English (Sanh et al. 2019), main branch of XLM-Roberta for Spanish and Chinese (Conneau et al. 2019), and a fine-tuned version of XLM-Roberta for Yorùbá (Adelani 2021). After that, in the second iteration, the cased multilingual DistilBERT base model was utilized for each language classifier's training due to its ability to train more quickly without forfeiting performance on predicting output labels (Sanh et al. 2019). The process included splitting the data into a training set and a test set of 80/20, respectively. This split was the same for both experiments. The metric that would be used in the shared task competition was Macro F1, so the efforts of enhancing the training process made sure to especially track those results in the trainer outputs.

At this point, the decision for which base models to incorporate in training the classifiers was made after observing changes in model performance after each epoch output. Some undesirable trends were noticed, such as overfitting in one case as suggested by an increasing validation loss in the uncased DistilBERT base model for English. This trial and error process facilitated the choice for which base model would be used later by ruling out the options that do not perform well.

3.2 Results

The cased multilingual DistilBERT base model proved to be the better option moving forward, since the difference in maximum F1 validation scores were marginal, and keeping this model as the base one allowed for consistency in creating one large multilingual classifier. The reason for this is because all four languages in this experiment were all included in that particular base model's training data. Given that the cased multilingual DistilBERT model was originally chosen as simply a new option to explore, combined with its lightweight characteristics, the decision was then confirmed to move forward with a uniform base model due to its ability to include all of the languages, an increased training time efficiency, sufficient F1 metric performance, and a confidence in the prediction labels (See Table 2 for more details). The prediction labels held great importance in seeing how the configuration of the models and the training data impacted the final results.

This importance of prediction label analysis was another significant contributing factor to abandoning the implementation of different base models for a consistent one in how some languages appeared to have exceptional F1 scores during training, yet when tested on the data, the prediction labels were incredibly wrong. For example, training Mandarin Chinese on XLM-RoBERTa proved to have high F1 and accuracy scores (using glue and mrpc), yet when the training data was tested as an inference, everything was labeled as euphemistic.

4 Experiment 2 - Multilingual Classifier and Cross-Lingual Comparison

4.1 Methodology

Since there have been positive results making a large multilingual classifier for text classification, the next step of this paper will detail how that process was completed for this shared task (Plank 2017). In an effort to maximize the F1 validation scores, the first step was concatenating the data so it would all be trained at the same time. Once it was prepared, the training pipeline remained the same. That is, LoRA was used again for its ability to keep

[1] https://github.com/nhankins/multilingual-euphs-figlang2024

most parameters the same as the base model, drawing attention to the training data in particular. The results can be found in Figure 1, or numerically in Table 4 and Figures 2-5.

Another aspect of this paper focuses on how information is transferred across languages. This portion ran concurrently with the multilingual portion to see if there was a major difference in the results when compared side-by-side. In each of these experiments, it is important to emphasize the motivations in choosing what constitutes a classifier as being better is directly related to its ability to both satisfy higher Macro F1 validation scores, but give confident label predictions on completely unseen data. This was done as a way to succeed in meeting the shared task requirements, and likewise further improve figurative language text classification.

At this point, the study requires verification that there is indeed an effect in using one language classifier with a specific dataset over another. The expectation is that the languages will output more accurate predictions on the classifier which has been fine-tuned with its own language. Therefore, the cross-lingual exercise demonstrates the results of this expectation.

4.2 Results

Analyzing the euphemistic and non-euphemistic splits from all 4 individual classifiers did not appear to yield any glaringly significant observations, yet when visualized it became easier to see overarching correlations (See Figure 1). Languages did not always appear to align more closely with the multilingual classifier predictions even on their own languages, which suggests that the greater quantity of training data plays an important role in favorable predictions. The Multilingual results were added to show contrast between the splits, noting that the multilingual classifier performed better than the individual ones. The detailed shared task final results on the test data of both methods can be found in the appendix, yet the F1 scores are as follows in Table 3.

Languages	**Individual**	**Multilingual**
English	0.57	0.64
Spanish	0.59	0.60
Chinese	0.59	0.68
Yorùbá	0.60	0.65

Table 3: Final F1 Scores for the shared task after submitting predictions. The First experiment (individual classifier) showed consistently lower values for all languages compared with the Second experiment (multilingual classifier).

	En.	Sp.	Ch.	Yo.
Euph.	687	714	794	486
Non-Euph.	509	377	432	183

Table 4: Predicted labels on Test Data using Multilingual Classifier (English, Spanish, Chinese, Yorùbá)

5 Conclusion

In conclusion, we learned that the cased multilingual DistilBERT base model proved to have a faster performance and learned more about the training data during fine-tuning. Despite not much change in the metrics output, the adherence of predicted labels to the ground truth gold standard labels was much closer.

Some concluding speculations could be that English and Spanish potentially have lower success rates in predicting whether a term is being used euphemistically or not due to less ambiguity in the instances for which they are being used. This is a curious assertion to prove in future work as the definition of what is ambiguous varies between speakers of a language. A major consideration that should also be noted, and potentially the subject of future work, is the impact that the unique number of PET categories has on the training process. English, for example, as mentioned before has the highest number, whereas Chinese has the lowest number. As mentioned previously, the data imbalance problem was addressed with learning rate adjustment, due to concerns that alternative methods, such as undersampling, might eliminate crucial semantic information. Another factor that should be noted is that Chinese and Yorùbá both needed to have truncation at inference time encoding, most likely due to BERT models using word-piece tokenization (Devlin et al. 2018). In other words, they saw more unknown words in their vocabularies, thus needing them to create more tokens and increasing the total length of the sequence for each example. Future work could explore if token length, language family, more balanced training data, or different

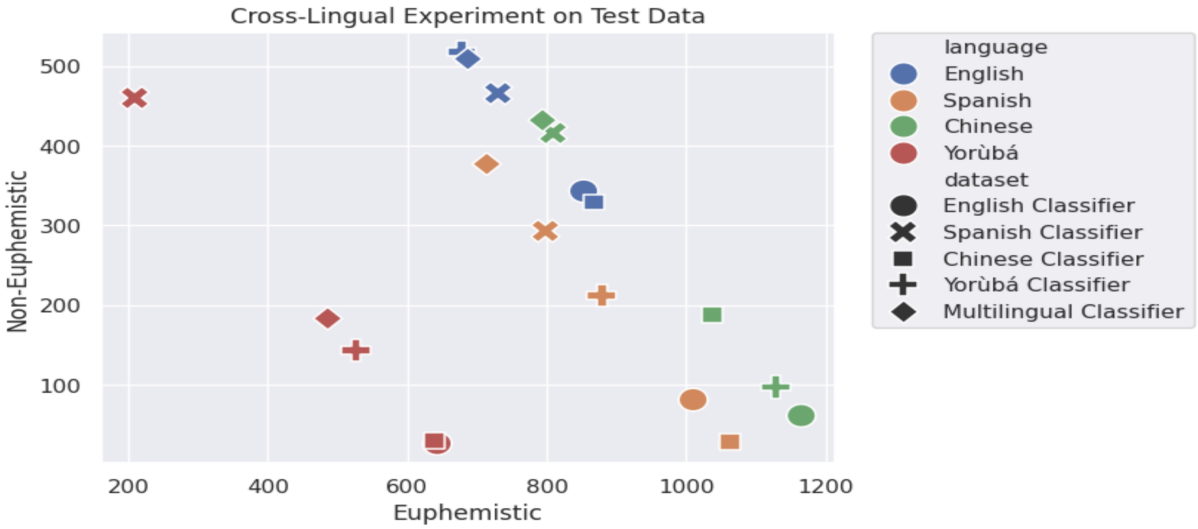

Figure 1: This chart portrays the inference results of the cross-lingual split between which sequences were labeled as euphemistic, and which were labeled as non-euphemistic. The individual language classifiers (fine-tuned only on their respective language data) are included along with the multilingual classifier to show contrast. Gold Standard labels are unknown and were not available to include in this Figure. Values can be found in Figures 2-5.

dialects play a role in greater euphemistic language understanding. The overall implications can suggest which kinds of base models could be optimal for assessing complicated linguistic devices in downstream language tasks, as well as how semantic correlation impacts deep learning throughout different languages.

6 Limitations

Please note that this paper does not account for varying dialects of all the presented languages. The only dialect of Chinese in the data is Mandarin Chinese, and the only dialect of English is American English. The Spanish and Yorùbá language data sets do, however, contain examples from different dialects. The selection of DistilBERT was partially due to the limited computatational resources of the author.

7 Ethics Statement

The author does not foresee any ethical concerns with the findings presented in this paper.

8 Acknowledgments

Special thanks goes to the Montclair State University NLP Lab for their work in organizing this shared task, answering questions concerning the limitations of this paper, as well as assembling the training and test data sets.

References

Adelani, David (2021). *Davlan/xlm-roberta-base-finetuned-yoruba · Hugging Face — huggingface.co*. https://huggingface.co/Davlan/xlm-roberta-base-finetuned-yoruba. [Accessed 03-03-2024].

Conneau, Alexis et al. (2019). "Unsupervised Cross-lingual Representation Learning at Scale". In: *CoRR* abs/1911.02116. arXiv: 1911.02116. URL: http://arxiv.org/abs/1911.02116.

Dettmers, Tim, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer (2023). *QLoRA: Efficient Finetuning of Quantized LLMs*. arXiv: 2305.14314 [cs.LG].

Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova (2018). "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding". In: *CoRR* abs/1810.04805. arXiv: 1810.04805. URL: http://arxiv.org/abs/1810.04805.

Gavidia, Martha, Patrick Lee, Anna Feldman, and Jing Peng (2022). *CATs are Fuzzy PETs: A Corpus and Analysis of Potentially Euphemistic Terms*. arXiv: 2205.02728 [cs.CL].

Hu, Edward J., Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen (2021). *LoRA: Low-Rank Adaptation of Large Language Models*. arXiv: 2106.09685 [cs.CL].

Lee, Patrick, Alain Chirino Trujillo, Diana Cuevas Plancarte, Olumide Ebenezer Ojo, Xinyi Liu, Iyanuoluwa Shode, Yuan Zhao, Jing Peng, and Anna Feldman (2024). *MEDs for PETs: Multilingual Euphemism Disambiguation for Potentially Euphemistic Terms*. arXiv: 2401.14526 [cs.CL].

Plank, Barbara (2017). *ALL-IN-1: Short Text Classification with One Model for All Languages*. arXiv: 1710.09589 [cs.CL].

Sanh, Victor, Lysandre Debut, Julien Chaumond, and Thomas Wolf (2019). "DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter". In: *ArXiv* abs/1910.01108.

Wolf, Thomas et al. (2020). *HuggingFace's Transformers: State-of-the-art Natural Language Processing*. arXiv: 1910.03771 [cs.CL].

YouTube-Blog (2024). https://github.com/ShawhinT/YouTube-Blog.

9 Appendix

A Cross-Lingual Experiment Data

Figures 2-5 are the predicted label splits from the Cross-Lingual Experiment.

English Classifier

	En.	Sp.	Ch.	Yo.
Euph.	853	1010	1165	643
Non-Euph.	343	81	61	26

Figure 2: English Classifier with Cross-Lingual Experiment

Spanish Classifier

	En.	Sp.	Ch.	Yo.
Euph.	730	798	809	209
Non-Euph.	466	293	416	460

Figure 3: Spanish Classifier with Cross-Lingual Experiment

Chinese Classifier

	En.	Sp.	Ch.	Yo.
Euph.	867	1063	1038	639
Non-Euph.	329	28	188	30

Figure 4: Chinese Classifier with Cross-Lingual Experiment

Yorùbá Classifier

	En.	Sp.	Ch.	Yo.
Euph.	678	879	1129	526
Non-Euph.	518	212	97	143

Figure 5: Yorùbá Classifier with Cross-Lingual Experiment

B Full Results from Shared Tasks

These are the detailed results output after the first and second submissions to the shared task. They include the individual classifier results, and the multilingual classifier results. As mentioned, the F1 scores were most important for this task, yet the precision and recall were included for transparency.

	En.	Sp.	Ch.	Yo.
F1	0.5736	0.5997	0.5995	0.6091
Precis.	0.6410	0.5986	0.7076	0.6537
Recall	0.6184	0.6011	0.6130	0.6104

Table 5: Detailed Results of Individual Classifiers

	En.	Sp.	Ch.	Yo.
F1	0.6446	0.6054	0.6808	0.6500
Precis.	0.6601	0.6024	0.6861	0.6716
Recall	0.6607	0.6209	0.6780	0.6457

Table 6: Detailed Results of Multilingual Classifier