Ensemble-based Multilingual Euphemism Detection: a Behavior-Guided Approach

Fedor Vitiugin and **Henna Paakki**
Department of Computer Science, Aalto University, Finland
{fedor.vitiugin, henna.paakki}@aalto.fi

Abstract

This paper describes the system submitted by our team to the Multilingual Euphemism Detection Shared Task for the Fourth Workshop on Figurative Language Processing (FigLang 2024). We propose a novel model for multilingual euphemism detection, combining contextual and behavior-related features. The system classifies texts that potentially contain euphemistic terms with an ensemble classifier based on outputs from behavior-related fine-tuned models. Our results show that, for this kind of task, our model outperforms baselines and state-of-the-art euphemism detection methods. As for the leader-board, our classification model achieved a macro averaged F1 score of 69%, reaching the third place.

1 Introduction

Euphemism, as defined by the Oxford English Dictionary, is the substitution of mild or indirect expressions for harsh or blunt ones when referring to unpleasant topics. The American Heritage Dictionary of the English Language similarly defines euphemism as replacing harsh or offensive terms with milder, indirect ones.

This paper explores the task of detecting euphemisms across multiple languages. Euphemism is a linguistic strategy employed to soften the impact of direct or uncomfortable language, such as using ‘collateral damage” instead of “war-related civilian deaths”. Euphemisms are commonly employed to maintain politeness, ease discomfort, or veil harsh realities in everyday communication. Despite cultural differences in their usage, the universal need to discuss sensitive topics without causing offense suggests commonalities in how euphemisms are applied across languages and cultures. This study investigates how multilingual models can leverage these similarities in processing euphemisms.

Our work is part of a Shared Task for the Fourth Workshop on Figurative Language Processing (FigLang 2024) and focuses on the euphemism disambiguation task, in which potentially euphemistic termss (PETs) are classified as euphemistic or not in a given context in four languages (Chinese, English, Spanish, and Yorùbá). This set of languages helps to encompass a diverse range of linguistic and cultural backgrounds (Lee et al.).

Our approach achieved the third-best score in the multilingual euphemism detection shared task. This paper describes our model [1] participating in the task.

2 Related Work

In this section, we explore related work about figurative language detection and euphemism detection in particular, utilization of behavior-related models for detecting specific types of content, and use of ensemble learning for combining different approaches for text classification.

2.1 Euphemism Detection

Euphemism allows writers to address taboo topics indirectly, facilitating better cross-cultural communication. Consequently, there’s a growing interest in computational methods for detecting euphemisms within Natural Language Processing (NLP) (Lee et al., 2022; Gavidia et al., 2022; Lee et al., 2023).

Recent work demonstrates semantic lexicon induction and the development of sentiment analysis methods could help to detect of euphemisms by investigating their connection with sentiment analysis. The study suggests analyzing affective polarity and connotation within sentence contexts yields better results than directly labeling phrases (Felt and Riloff, 2020).

[1] Our code is available at https://github.com/vitiugin/med

Proceedings of the 4th Workshop on Figurative Language Processing (FLP), pages 73–78
June 21, 2024 ©2024 Association for Computational Linguistics

Pre-trained transformer models are extensively employed in various NLP-related tasks including euphemism detection through task-specific fine-tuning (Tiwari and Parde, 2022), in combination with relational graph attention network (Wang et al., 2022), with adversarial augmentation technique (Kohli et al., 2022). Additionally, the utilization of clustering algorithms to provide additional signals of PETs similarity improves performance of pre-trained model in ensemble methods (Keh et al., 2022).

Leveraging of prompt tuning pre-trained language models is another direction in euphemism detection. Use of RoBERTa as the pre-trained language model and creation of suitable templates and verbalizers could be effectively used (Maimaitituoheti et al., 2022).

Large Language Modelss (LLMs) have been the subject of exploration regarding their multilingual and cross-lingual transfer capabilities in prior studies (Lee et al.). Multilingual LLMs extensively leverage data from multiple languages, acquiring both complementary and reinforcing information (Choenni et al., 2023). Transfer learning from out-of-language data within a particular domain yielded superior results compared to utilizing same-language data from a different domain (Shode et al., 2023).

2.2 Behavior-Related Fine-Tuning for Euphemism Detection

Since euphemisms are established social speaking and behaving norms, ways of thinking as well as outlook of value, it is essential to study their application. Euphemism exists in all aspects of English in great numbers and is categorized into eight types (Li-Na, 2015): *death, aging and disease* ("passed away", "passed", "departed"), *disability and handicap* ("mentally challenged", "special needs", "full-figured"), *education* ("slow student", "peer homework"), *marriage and pregnancy* ("renovate", "unwedding", "tie the knot"), *military* ("collateral damage", "neutralizing", "involvement"), *profession* ("sanitation engineer", "comfort woman"), *politics* ("the deprived", "economic downturn"), *profanity* ("private parts", "choke the chicken").

Utilizing models to detect sociopolitical threads can enhance euphemism detection performance according to the provided classification. Behavior-related fine-tuning (Ruder, 2021) involves teaching models relevant capabilities for excelling in a target task, necessitating an understanding of diverse human behavioral patterns in language (Founta et al., 2019; Zhang et al., 2023). This process involves fine-tuning the model on related tasks to acquire practical behaviors (Vitiugin and Purohit, 2024), contrasting with adaptive fine-tuning. Behavioral fine-tuning, particularly with labeled data, has proven effective in teaching models various linguistic features such as named entities (Broscheit, 2020), paraphrasing (Arase and Tsujii, 2019), syntax (Glavaš and Vulić, 2021), answer sentence selection (Garg et al., 2020), and question answering (Khashabi et al., 2020). A recent study emphasized the importance of a diverse task selection for optimal transfer performance, based on fine-tuning a model on nearly 50 labeled datasets in a massively multitask environment (Aghajanyan et al., 2021).

2.3 Ensemble Learning

Ensemble multifeatured deep learning is a powerful method to improve model generalization and performance, which has been used effectively in figurative language detection. Combining ensemble outputs can boost metaphor detection performance (Brooks and Youssef, 2020). Additionally, utilizing an Adaptive Boosting classifier with Decision Tree as a base estimator shows promise in predicting sarcasm probabilities (Lemmens et al., 2020).

By combining the strengths of multiple models and features, ensemble multifeatured deep learning models have demonstrated improved performance and adaptability in diverse problem settings. While these models have such challenges as model interpretability, computational complexity, ensemble model selection, adversarial robustness, and personalized and federated learning (Abimannan et al., 2023).

3 Model Architecture

The model's architecture is presented in Figure 1 and includes two main steps: fine-tuning for behavior-related downstream tasks and ensemble method for classification.

First, we fine-tuned the multilingual transformer-based model (XLM-RoBERTa (Conneau et al., 2019)) for classifying contextual texts (without PETs) and classifying PETs separately. Based on review of related work, we fine-tuned the same pre-trained language model for the several behav-

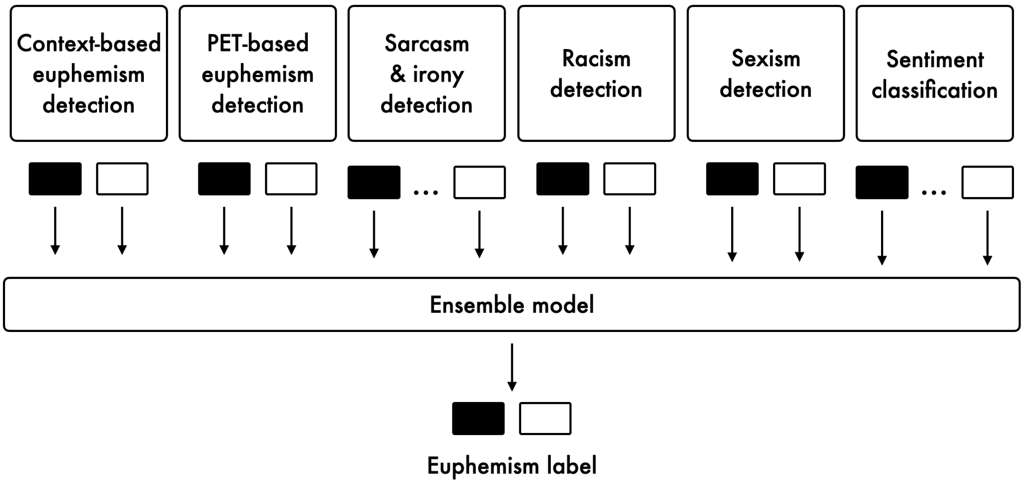

Figure 1: Model architecture

ioral tasks: detection of sarcasm and irony (Ling and Klinger, 2016), sexism, racism (Albright, 2021), and sentiment classification (Passionate-NLP, 2021). After fine-tuning, we had 6 fine-tuned models with the same architecture, and tokenizers.

Second, our final model used the ensemble learning method for classification, which received logits from described models as features. During the developing step, we tested several ensemble models including: Adaptive Boosting, Extra Trees, Gradient Boosting, and Random Forest.

Finally, we used the best performing ensemble learning method to train model for detection euphemisms in four languages.

4 Experiment

For the shared task, we made only multilingual experiments, i.e. training and developing datasets contain entities in all four presented languages.

4.1 Dataset

The dataset for the experiment includes texts in four languages: Mandarin Chinese (ZH), American English (EN), Spanish (ES), and Yorùbá (YO) (Lee et al., 2023). The dataset for each language contains texts, PETs, and labels (euphemistic or non-euphemistic). Dataset statics is presented in Table 1. For each test run, we use 80-10-10 split to create training, validation, and test sets.

4.2 Implementation Details

We maintain the same number of layers in each model – 24 layers for XML-RoBERTa (Conneau et al., 2019). During fine-tuning, we used the same hyperparameters and number of frozen layers (detected for task-related fine-tuning by grid search.) For LLMs' fine-tuning, we used $0.5 * 10^{-5}$ learning rate, 10 epochs. The number of frozen layers for each model were detected by grid search. The models were trained on NVIDIA A100-SXM4 with 40Gb GPU RAM.

Table 1: Experiment dataset statistics

language	euphemistic	non-euphemistic	total
Chinese (ZH)	1484	521	2005
English (EN)	1383	569	1952
Spanish (ES)	1143	718	1861
Yorùbá (YO)	1281	660	1941

Table 2: Comparison of ensemble learning methods for classification. 10-fold CV for multilingual data.

scheme	ACC	AUC	F1
Adaptive Boosting	96.06	95.38	95.13
Extra Trees	96.01	95.32	95.06
Gradient Boosting	**96.10**	95.39	94.75
Random Forest	**96.10**	**95.42**	**95.27**

4.3 Baselines and Compared Methods

To compare our proposed method for multilingual euphemism detection problem, we construct baseline scheme using deep learning model that use LASER embeddings (Artetxe and Schwenk, 2019) as input features. Additionally, we also compare our method in combination with varied sets of behavior-related models. The full list of schemes includes:

- **[LSTM_text&PET]** – method uses combines pre-trained LASER embeddings of text and PET, which are passed as input to a

Table 3: Comparison of baseline schemes and proposed approach. 10-fold CV for multilingual data.

scheme	ACC	AUC	F1
LSTM_text&PET	79.52 ± 0.5	79.66 ± 0.4	88.30 ± 0.9
RoBERTa_text&PET	91.29 ± 0.7	90.42 ± 0.9	90.25 ± 1.1
RoBERTa_text&PET&sexism	95.84 ± 0.8	95.13 ± 1.0	94.92 ± 0.9
RoBERTa_text&PET&racism	95.79 ± 0.7	95.07 ± 0.9	94.90 ± 1.1
RoBERTa_text&PET&social	95.82 ± 0.7	95.11 ± 0.9	94.87 ± 1.1
RoBERTa_text&PET&social&sarcasm	96.02 ± 0.7	95.23 ± 0.9	94.94 ± 1.1
RoBERTa_text&PET&social&sentiment	96.03 ± 0.7	95.35 ± 0.8	95.09 ± 1.1
RoBERTa_text&PET&all	$\mathbf{96.10 \pm 0.7}$	$\mathbf{95.42 \pm 0.9}$	$\mathbf{95.27 \pm 1.1}$

Long Short-Term Memory (LSTM) Network model (Vitiugin and Barnabo, 2021);

- **[RoBERTa_text&PET]** – method uses logits of fine-tuned RoBERTa for euphemism detection in text and PET;
- **[RoBERTa_text&PET&sexism]** – method uses logits of fine-tuned RoBERTa for euphemism detection in text and PET, as well as logits of the model for sexism detection;
- **[RoBERTa_text&PET&racism]** – method uses logits of fine-tuned RoBERTa for euphemism detection in text and PET, as well as logits of the model for racism detection;
- **[RoBERTa_text&PET&social]** – method uses logits of fine-tuned RoBERTa for euphemism detection in text and PET, as well as logits from models for sexism and racism detection;
- **[RoBERTa_text&PET&social&sarcasm]** – method uses logits of fine-tuned RoBERTa for euphemism detection in text and PET, as well as logits from models for sexism, racism, and sarcasm detection;
- **[RoBERTa_text&PET&social&sentiment]** – method uses logits of fine-tuned RoBERTa for euphemism detection in text and PET, as well as logits from models for sexism and racism detection and from sentiment classification model;
- **[RoBERTa_text&PET&all]** – method uses logits of fine-tuned RoBERTa for euphemism detection in text and PET, as well as logits from all behaviour-related models.

4.4 Results

First, we compare several ensemble methods applying for the euphemism detection task. In this experiment we use outputs from all fine-tuned models and all ensemle methods' parameters were optimized by applying Greed Search. Table 2 demonstrates that the Random Forest classifier reaches the highest results. While Adaptive Boosting, Extra Trees, and Gradient Boosting perform less effective, 10-fold cross-validation demonstrates that the difference between the performance of different models is insignificant (p-value ≥ 0.05). As a result of this experiment, we chose the Random Forest model for combining outputs of fine-tuned models.

Comparison of baseline and proposed models on training data provided by organizers of the shared task demonstrates high performance of ensemble learning method with behavior-related models. Use of logits from all fine-tuned models shows the best performance. Even use of logits from the only one behaviour-related model significantly improves results (p-value ≤ 0.05) comparing to combination of logits provided only by contextual and PET models. While our experiments didn't show significant improvement of performance between models used outputs from one behaviour-related model and outputs from all behaviour related models (p-value ≈ 0.4). The full results of schemes comparison are presented in Table 3.

4.5 Shared Task Results

During the test phase of the shared task, we employed our most effective model, *RoBERTa_text&PET&all*. However, its performance significantly declined compared to the development phase, achieving a macro-averaged F1 score of 69%. This highlights the model's reliance on contextual familiarity, particularly as the test data incorporates numerous new PETs. Notably, English and Chinese languages exhibited better performance overall, aligning with trends observed in similar methods. Noteworthy, our model excelled with the Spanish dataset. For detailed results, please refer to Table 4.

Table 4: Shared task results for test dataset provided by organizers.

Language	P	R	F1
English	75.29	75.57	73.90
Spanish	68.78	66.56	67.43
Yorùbá	65.53	62.77	63.06
Chinese	71.10	82.00	70.44

5 Conclusion

We have described a method for multilingual euphemism detection. This method is based on behaviour-related fine-tuning of transformer model for combining their logits in ensemble learning. Experiments with four different languages demonstrate that our approach could reach high performance in the task.

5.1 Limitations

In the work, we used only English datasets for behavior-related fine-tuning. The use of datasets in other languages could show different results.

5.2 Future Work

One of the directions of future research is exploration of grammatical features of euphemisms. Grammatical methods, such as past tense and passive voice, create psychological distance and politeness. Extracting these types of features from the text could enhance multilingual euphemism detection.

6 Acknowledgements

This work is supported by the Trust-M research project, a partnership between Aalto University, University of Helsinki, Tampere University, and the City of Espoo, funded in-part by a grant from the Strategic Research Council (SRC) in Finland. The research is also supported in-part by a grant from the Helsingin Sanomat Foundation for the project AI-infused Disinformation in Media Communications (DiME).

References

Satheesh Abimannan, El-Sayed M El-Alfy, Yue-Shan Chang, Shahid Hussain, Saurabh Shukla, and Dhivyadharsini Satheesh. 2023. Ensemble multi-featured deep learning models and applications: A survey. *IEEE Access*.

Armen Aghajanyan, Anchit Gupta, Akshat Shrivastava, Xilun Chen, Luke Zettlemoyer, and Sonal Gupta. 2021. Muppet: Massive multi-task representations with pre-finetuning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 5799–5811.

Munki Albright. 2021. Suspicious tweets dataset. `https://www.kaggle.com/datasets/munkialbright/classified-tweets`.

Yuki Arase and Jun'ichi Tsujii. 2019. Transfer fine-tuning: A BERT case study. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5393–5404, Hong Kong, China. Association for Computational Linguistics.

Mikel Artetxe and Holger Schwenk. 2019. Massively multilingual sentence embeddings for zero-shot cross-lingual transfer and beyond. *Transactions of the Association for Computational Linguistics*, 7:597–610.

Jennifer Brooks and Abdou Youssef. 2020. Metaphor detection using ensembles of bidirectional recurrent neural networks. In *Proceedings of the Second Workshop on Figurative Language Processing*, pages 244–249.

Samuel Broscheit. 2020. Investigating entity knowledge in bert with simple neural end-to-end entity linking. *arXiv preprint arXiv:2003.05473*.

Rochelle Choenni, Dan Garrette, and Ekaterina Shutova. 2023. How do languages influence each other? studying cross-lingual data sharing during llm fine-tuning. *arXiv preprint arXiv:2305.13286*.

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. *arXiv preprint arXiv:1911.02116*.

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.

Antigoni Maria Founta, Despoina Chatzakou, Nicolas Kourtellis, Jeremy Blackburn, Athena Vakali, and Ilias Leontiadis. 2019. A unified deep learning architecture for abuse detection. In *Proceedings of the 10th ACM conference on web science*, pages 105–114.

Siddhant Garg, Thuy Vu, and Alessandro Moschitti. 2020. Tanda: Transfer and adapt pre-trained transformer models for answer sentence selection. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 7780–7788.

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.

Goran Glavaš and Ivan Vulić. 2021. Is supervised syntactic parsing beneficial for language understanding? an empirical investigation.

Sedrick Scott Keh, Rohit K Bharadwaj, Emmy Liu, Simone Tedeschi, Varun Gangal, and Roberto Navigli. 2022. Eureka: Euphemism recognition enhanced through knn-based methods and augmentation. *arXiv preprint arXiv:2210.12846*.

Daniel Khashabi, Sewon Min, Tushar Khot, Ashish Sabharwal, Oyvind Tafjord, Peter Clark, and Hannaneh Hajishirzi. 2020. UNIFIEDQA: Crossing format boundaries with a single QA system. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1896–1907, Online. Association for Computational Linguistics.

Guneet 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 (FLP)*, pages 154–159.

Patrick Lee, Anna Feldman, and Jing Peng. 2022. A report on the euphemisms detection shared task. *arXiv preprint arXiv:2211.13327*.

Patrick Lee, Iyanuoluwa Shode, Alain Chirino Trujillo, Yuan Zhao, Olumide Ebenezer Ojo, Diana Cuervas Plancarte, Anna Feldman, and Jing Peng. 2023. Feed pets: Further experimentation and expansion on the disambiguation of potentially euphemistic terms. *arXiv preprint arXiv:2306.00217*.

Patrick Lee, Alain Chirino Trujillo, Diana Cuevas Plancarte, Olumide Ebenezer Ojo, Xinyi Liu, Iyanuoluwa Shode, Yuan Zhao, Jing Peng, and Anna Feldman. Meds for pets: Multilingual euphemism disambiguation for potentially euphemistic terms.

Jens Lemmens, Ben Burtenshaw, Ehsan Lotfi, Ilia Markov, and Walter Daelemans. 2020. Sarcasm detection using an ensemble approach. In *proceedings of the second workshop on figurative language processing*, pages 264–269.

Zhou Li-Na. 2015. Euphemism in modern american english. *Sino-US English Teaching*, 12(4):265–270.

Jennifer Ling and Roman Klinger. 2016. An empirical, quantitative analysis of the differences between sarcasm and irony. In *The Semantic Web: ESWC 2016 Satellite Events, Heraklion, Crete, Greece, May 29–June 2, 2016, Revised Selected Papers 13*, pages 203–216. Springer.

Abulimiti Maimaitituoheti, Yang Yong, and Fan Xiaochao. 2022. A prompt based approach for euphemism detection. In *Proceedings of the 3rd Workshop on Figurative Language Processing (FLP)*, pages 8–12.

Passionate-NLP. 2021. Twitter sentiment analysis dataset. `https://www.kaggle.com/datasets/jp797498e/twitter-entity-sentiment-analysis`.

Sebastian Ruder. 2021. Recent advances in language model fine-tuning. `http://ruder.io/recent-advances-lm-fine-tuning`,.

Iyanuoluwa Shode, David Ifeoluwa Adelani, Jing Peng, and Anna Feldman. 2023. Nollysenti: Leveraging transfer learning and machine translation for nigerian movie sentiment classification. *arXiv preprint arXiv:2305.10971*.

Devika Tiwari and Natalie Parde. 2022. An exploration of linguistically-driven and transfer learning methods for euphemism detection. In *Proceedings of the 3rd Workshop on Figurative Language Processing (FLP)*, pages 131–136.

Fedor Vitiugin and Giorgio Barnabo. 2021. Emotion detection for spanish by combining laser embeddings, topic information, and offense features.

Fedor Vitiugin and Hemant Purohit. 2024. Multilingual serviceability model for detecting and ranking help requests on social media during disasters. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 18.

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 (FLP)*, pages 79–83.

Dong Zhang, Wenwen Li, Baozhuang Niu, and Chong Wu. 2023. A deep learning approach for detecting fake reviewers: Exploiting reviewing behavior and textual information. *Decision Support Systems*, 166:113911.