Enhancing Cross-Lingual Emotion Detection with Data Augmentation and Token-Label Mapping

Jinghui Zhang[1] and **Yuan Zhao** [1] and **Siqin Zhang**[1] and **Ruijing Zhao**[1] and **Siyu Bao**[1]
[1]China Telecom Cloud Technology Co., Ltd
{zhangjh33,zhaoyuan1,zhangsq20,zhaorj1, baosy}@chinatelecom.cn

Abstract

Cross-lingual emotion detection faces challenges such as imbalanced label distribution, data scarcity, cultural and linguistic differences, figurative language, and the opaqueness of pretrained language models. This paper presents our approach to the EXALT shared task at WASSA 2024, focusing on emotion transferability across languages and trigger word identification. We employ data augmentation techniques, including back-translation and synonym replacement, to address data scarcity and imbalance issues in the emotion detection subtask. For the emotion trigger identification subtask, we utilize token and label mapping to capture emotional information at the subword level. Our system achieves competitive performance, ranking 13th, 1st, and 2nd in the Emotion Detection, Binary Trigger Word Detection, and Numerical Trigger Word Detection tasks.

1 Introduction

Emotion detection in text has attracted significant attention in recent years due to its diverse applications (Nandwani and Verma, 2021). The growing presence of multilingual and code-mixed content on social media has emphasized the need for cross-lingual emotion detection systems (Balahur and Turchi, 2014; Dashtipour et al., 2016).

However, developing accurate and interpretable cross-lingual emotion detection models presents several challenges, including the limited availability of labeled data in many languages (Xue et al., 2020), cultural and linguistic variations in emotion expression (Hareli et al., 2015), the use of figurative language and sarcasm in social media (Bouazizi and Ohtsuki, 2019; Reyes et al., 2012), and the lack of transparency in large pre-trained language models (PLMs) (Hase and Bansal, 2020; Feder et al., 2021).

To tackle these challenges and foster research on interpretable cross-lingual emotion detection, the EXALT shared task[1] was organized at WASSA 2024. EXALT focuses on the transferability of emotion information across languages and the identification of emotion triggers, encouraging the development of interpretable and explainable emotion detection systems.

This paper presents our approach to the EXALT shared task. The XLM-RoBERTa-XL (Goyal et al., 2021) model with encoder-only architecture is selected to better capture the emotion information contained in the context. For the emotion detection sub-task, we employ data augmentation techniques, such as back-translation and synonym replacement, to address the imbalanced and insufficient training data. For the emotion trigger identification subtask, we utilize token and label mapping to align subword-level predictions with word-level labels.

2 Related Work

2.1 Emotion Detection

Emotion detection aims to identify the emotional states expressed in text, such as joy, sadness, anger, and fear (Acheampong et al., 2020). Compared to sentiment analysis, which focuses on the overall polarity of text (positive, negative, or neutral), emotion detection provides a more fine-grained understanding of the affective information conveyed in text.

Traditional approaches to emotion detection relied on rule-based methods and machine learning algorithms, such as Naïve Bayes, Support Vector Machines (SVM), and decision trees (Alswaidan and Menai, 2020).Recently, deep learning models, including Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), and Transformers, have achieved state-of-the-art performance on emotion detection tasks by learning rich feature representations from text (Birjali et al., 2021). However, the success of these models relies

[1] https://lt3.ugent.be/exalt/

Proceedings of the 14th Workshop on Computational Approaches to Subjectivity, Sentiment, & Social Media Analysis, pages 528–533
August 15, 2024 ©2024 Association for Computational Linguistics

heavily on large annotated datasets, which are time-consuming and expensive to collect, especially for low-resource languages and domains. Researchers are exploring techniques such as transfer learning (Hazarika et al., 2021), few-shot learning, and data augmentation to address this challenge and improve emotion detection performance in low-resource settings.

2.2 Data Augmentation

Data augmentation techniques have been widely adopted to address the issue of data scarcity in various natural language processing tasks, aiming to generate synthetic training data and improve model performance (Pellicer et al., 2023). These methods operate at different levels of granularity, such as word-level, phrase-level, and document-level.

At the word level, techniques like EDA (Wei and Zou, 2019) generate new examples by manipulating words or embeddings. Phrase-level augmentation, e.g., PPDB Augmenter (Ganitkevitch et al., 2013; Pavlick et al., 2015), uses paraphrase databases to replace phrases with semantic equivalents, adding linguistic diversity. Document-level techniques include back-translation (Mallinson et al., 2017), paraphrasing with models like T5 (Raffel et al., 2020) and BART (Lewis et al., 2019; Dopierre et al., 2021), and text generation using language models such as Llama (Touvron et al., 2023) and GPT-4 (Achiam et al., 2023).

2.3 Cross-lingual Transfer Learning

Cross-lingual transfer learning has emerged as a promising approach to overcome the data scarcity issue in low-resource languages. Methods such as machine translation (Demirtas and Pechenizkiy, 2013) and multilingual PLMs (Lewis et al., 2019; Xue et al., 2020) have been used to transfer knowledge from high-resource to low-resource languages.

However, these methods often fail to capture language-specific nuances in emotion expression and may suffer from translation errors. Adversarial training (Chen et al., 2018) and contrastive learning (Lin et al., 2023) have been proposed to learn language-invariant representations, but they often require parallel data or emotion lexicons.

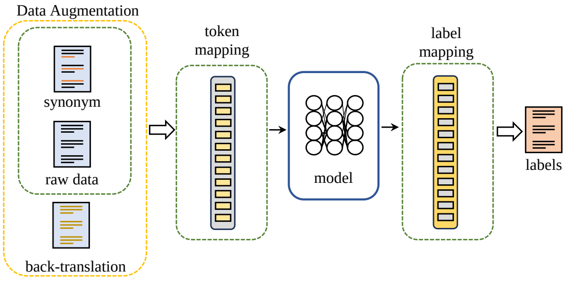

Figure 1: The overview of our work. The yellow and green dashed boxes represent the methods used in Task 1 and Tasks 2/3, respectively.

2.4 Interpretability in Natural Language Processing

Interpretability has gained significant attention in natural language processing, particularly with the increasing complexity of deep learning models (Danilevsky et al., 2020). Various approaches have been proposed to provide explanations for model predictions, such as attention mechanisms (Bibal et al., 2022), saliency maps (Wallace et al., 2020), and post-hoc explanations (Madsen et al., 2022).

However, most existing methods focus on providing explanations at the input feature level and may not offer fine-grained interpretability at the token level. In the context of emotion detection, identifying the specific words or phrases that trigger the predicted emotions is crucial for understanding the model's behavior.

3 Methods

This section illustrates our approach and main work, as shown in Figure 1. We focus on describing our data augmentation and token and label mapping methods. The details of the shared task can be found in this work (Maladry et al., 2024).

Briefly, Task 1 is sequence-classification task, where there are 6 possible emotion classes in the datasets. Task 2 and 3 are token-classification tasks that are focused on explaining which words are used to express the emotion. The training dataset contains only English data, while the test dataset contains five languages, i.e. Dutch, Russian, Spanish, English, and French.

3.1 Data Augmentation

The dataset for Task 1 includes 5,000 training samples, 500 development samples, and 2,500 test samples. For Tasks 2 and 3, the dataset consists of 3,000 training samples, 300 development samples, and 832 test samples.

Label	Raw	Back-translation	Synonym
Anger	1,028	192	1,028
Fear	143	576	143
Joy	1,293	239	1,290
Love	579	144	578
Neutral	1,397	259	1,396
Sadness	560	658	560

Table 1: Statistics of the labels distribution on the training set for Task 1.

The original data presents the characteristics of imbalanced label distribution, as shown in Table 1. For example, the proportion of Fear class is less than 3%. In order to mitigate this characteristic and enhance the diversity of the data to improve the generalisation of the model, we apply two data augmentation methods as detailed as follows. Other data augmentation methods would break sentence syntax, such as random deletion, or change emotion information, such as text generation with PLMs, and are therefore not considered in this work.

Back-translation: The original training samples are translated into other languages, in this case Dutch, French, Spanish and Russian, and then back into English to generate new instances. We adopted a stratified sampling approach to mitigate the imbalance in label distribution. Notably, back-translation may alter the position of trigger words, causing misalignment with their corresponding labels.

Synonym replacement: Words are randomly selected in the original samples and then replaced with their synonyms. In this paper, we implemented synonym replacement using the NLTK WordNet corpus (Loper and Bird, 2002). Notably, synonym replacement may alter the number of words, which lead to misalignment of the augmented data with the original labels.

We use DeepL to complete the back-translation. In Task 1, 2,068 new instances are added via back-translation, while 4,995 instances are added via synonym replacement. The details are shown in Table 1. In Task 2 and 3, 2,291 instances are added through synonym replacement.

3.2 Token and Label Mapping

The mapping between each subword and its corresponding input word is recorded when the tokenizer processes the input data. The label of the subword is assigned to be the same as the label of its corresponding word. During prediction, the label of each word is set to the maximum value among the predictions of all its corresponding subwords. Figure 2 illustrates the configuration of token and label mapping.

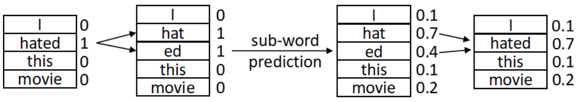

Figure 2: Configuration of token and label mapping.

By applying labels at the subword level, the emotional information of the text is captured more fine-grained. During the prediction phase, taking the maximum value of the subword predictions as the label of the corresponding word effectively integrates the emotional information at the subword level. Moreover, this method is easy to implement and compatible with various PLMs.

For Task 3, we map the output of the classifier layer of the fine-tuned language model to word-level numerical results. For Task 2, we compute the softmax of the classifier layer output, then map it into word-level numerical probabilities and finally convert it to a binary result using a threshold.

We have shared the main scripts of this paper on Github[2] for other researchers.

4 Experiments

We conducted experiments on NPU training machines, equipped with 8 Ascend 910B 64G NPUs, to compare the performance of multiple PLMs and investigate the impact of hyperparameter settings, data augmentation strategies on cross-lingual emotion detection.

4.1 Comparison of Pre-trained Language Models

Although many current PLMs adopt the decoder-only architecture, this architecture has limited support for natural language understanding tasks. In our tasks, we consider the information in the subsequent text to be equally important as the preceding text. Therefore, we compared multiple encoder-only models pre-trained on Dutch, Russian, Spanish, English, and French corpora to better understand the contextual content and cross-lingual emotion information.

We fine-tuned these models on the original data without data augmentation and tried different parameter freezing strategies. Figure 3 demonstrates that that larger models achieved better cross-lingual

[2] https://github.com/QQJellyy/CTcloud-EXALT-WASSA2024

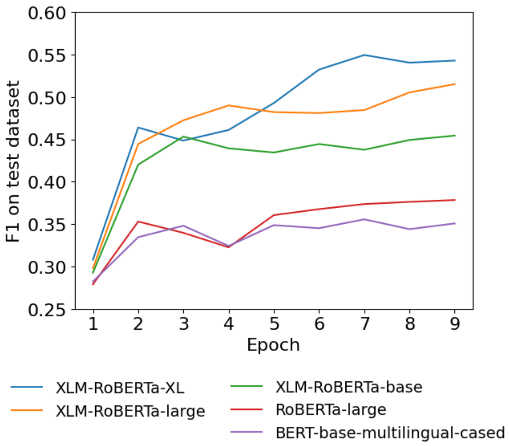

Figure 3: F1 on test dataset of each PLM for Task 1. All parameters of each PLM are fine-tuned using a learning rate of 4e-5 and a batch size of 128.

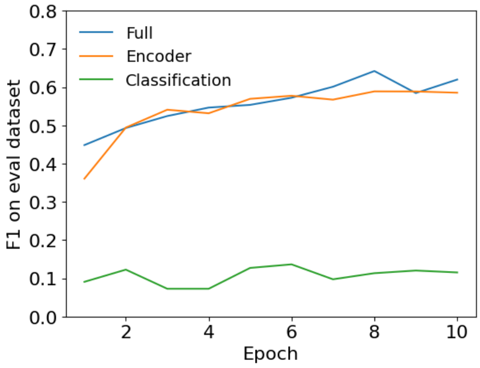

Figure 4: F1 of XLM-RoBERTa-large on eval dataset of each parameter freezing strategies for Task 1. Full, Encoder and Classification represent parameters of all layers, parameters of encoder layers and parameters of output layer are trainable, respectively.

emotion detection ability. In the meanwhile, Figure 4 suggests that the PLMs lack emotion detection capabilities, as the results by training only on the classification layer are significantly lower than the other training strategies.

Training on the English dataset significantly improves emotion detection abilities on five languages, indicating that emotion information has been effectively transferred across languages. We explain this phenomenon with a task analogy theory, e.g., 'Queen = King + (Woman-Man)' (Ethayarajh et al., 2018), where cross-lingual capabilities are inherent in PLMs, and emotion detection capability is attained through fine-tuning. The fine-tuned models integrate these capabilities and exhibit cross-lingual emotion detection abilities.

4.2 Hyperparameter Optimisation

Rank	Learning Rate	F1
1	2e-6	0.3221
2	5e-6	0.4676
3	1e-5	0.4821
4	2e-5	0.5284
5	3e-5	0.5318
6	4e-5	**0.5705**
7	5e-5	0.4676

Table 2: F1 of XLM-RoBERTa-large on eval dataset for Task 1 after being fine-tuned without augmented dataset for 10 epochs. The batch size is 128.

Task	Augmentation	F1
Task1	Null	0.5242
Task1	Synonym	0.5420
Task1	Translation	0.5393
Task1	Synonym + Translation	**0.5432**
Task2	Null	0.6042
Task2	Synonym	**0.6158**
Task3	Null	0.6833
Task3	Synonym	**0.6972**

Table 3: F1 of XLM-RoBERTa-XL on test dataset with different data augmentation. All parameters of each scenario are fine-tuned using a learning rate of 4e-5 and a batch size of 64.

We conduct hyperparameter optimisation on XLM-RoBERTa-large and employ an orthogonal approach to optimize each hyperparameter. As shown in Table 2, a learning rate of 4e-5 achieves the best classification results for Task 1. Learning rates that are too large or too small can be less effective. Similarly, we find that a batch size of 64 yields the best detection results.

For Task 1, we train the XLM-RoBERTa-XL model on the augmented data using the optimal hyperparameters described above. For Tasks 2 and 3, we use the same hyperparameters and further train the model saved from Task 1. When making predictions, we use the default value of 0.1 as the threshold for identifying trigger words.

4.3 Data Augmentation

Table 3 demonstrates that both synonym replacement and back-translation improve the model's performance. Synonym replacement enriches the training corpus, while back-translation implicitly introduces information about the target test language without altering the language of the training data.

5 Conclusion

In this paper, we address the label imbalance issue and enhanced data diversity in the training data by employing data augmentation techniques, including back-translation and synonym replacement. The augmented data is used to fine-tune the XLM-RoBERTa-XL model, achieving competitive results in all three tasks: 13th, 1st, and 2nd places in Task 1, Task 2, and Task 3, respectively. These results demonstrate the effectiveness of our methods for the transferability of emotion information across languages and the identification of emotion triggers.

The reason why we choose XLM-RoBERTa-XL model is that the encoder architecture is able to capture the context of the data, making it well-suited for token-level tasks. Furthermore, we select PLMs that have been trained on five languages to ensure that the models have the transferability of emotion information across languages, potentially contributing to the improved performance of our proposed system.

6 Limitations

In this paper, we assume that different models behave similarly, e.g., the optimal hyperparameters of XLM-RoBERTa-XL and XLM-RoBERTa-large are similar. However, this may not be the case in practice.

We use the default value of 0.1 as the threshold for identifying the trigger word in Task 2 and apply two data augmentation techniques, i.e. back-translation and synonym replacement. In the future, we can explore the impact of different classification threshold on model performance and try other data augmentation methods. Moreover, we can also go deep into model enhancing techniques on cross-lingual emotion detection task, such as task analogy and model fusion.

References

Francisca Adoma Acheampong, Chen Wenyu, and Henry Nunoo-Mensah. 2020. Text-based emotion detection: Advances, challenges, and opportunities. *Engineering Reports*, 2(7):e12189.

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Nourah Alswaidan and Mohamed El Bachir Menai. 2020. A survey of state-of-the-art approaches for emotion recognition in text. *Knowledge and Information Systems*, 62(8):2937–2987.

Alexandra Balahur and Marco Turchi. 2014. Comparative experiments using supervised learning and machine translation for multilingual sentiment analysis. *Computer Speech & Language*, 28(1):56–75.

Adrien Bibal, Rémi Cardon, David Alfter, Rodrigo Wilkens, Xiaoou Wang, Thomas François, and Patrick Watrin. 2022. Is attention explanation? an introduction to the debate. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3889–3900.

Marouane Birjali, Mohammed Kasri, and Abderrahim Beni-Hssane. 2021. A comprehensive survey on sentiment analysis: Approaches, challenges and trends. *Knowledge-Based Systems*, 226:107134.

Mondher Bouazizi and Tomoaki Ohtsuki. 2019. Multi-class sentiment analysis on twitter: Classification performance and challenges. *Big Data Mining and Analytics*, 2(3):181–194.

Xilun Chen, Yu Sun, Ben Athiwaratkun, Claire Cardie, and Kilian Weinberger. 2018. Adversarial deep averaging networks for cross-lingual sentiment classification. *Transactions of the Association for Computational Linguistics*, 6:557–570.

Marina Danilevsky, Kun Qian, Ranit Aharonov, Yannis Katsis, Ban Kawas, and Prithviraj Sen. 2020. A survey of the state of explainable ai for natural language processing. *arXiv preprint arXiv:2010.00711*.

Kia Dashtipour, Soujanya Poria, Amir Hussain, Erik Cambria, Ahmad YA Hawalah, Alexander Gelbukh, and Qiang Zhou. 2016. Multilingual sentiment analysis: state of the art and independent comparison of techniques. *Cognitive computation*, 8:757–771.

Erkin Demirtas and Mykola Pechenizkiy. 2013. Cross-lingual polarity detection with machine translation. In *Proceedings of the Second International Workshop on Issues of Sentiment Discovery and Opinion Mining*, pages 1–8.

Thomas Dopierre, Christophe Gravier, and Wilfried Logerais. 2021. Protaugment: Intent detection meta-learning through unsupervised diverse paraphrasing. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 2454–2466. Association for Computational Linguistics.

Kawin Ethayarajh, David Duvenaud, and Graeme Hirst. 2018. Towards understanding linear word analogies. *arXiv preprint arXiv:1810.04882*.

Amir Feder, Nadav Oved, Uri Shalit, and Roi Reichart. 2021. Causalm: Causal model explanation through counterfactual language models. *Computational Linguistics*, 47(2):333–386.

Juri Ganitkevitch, Benjamin Van Durme, and Chris Callison-Burch. 2013. Ppdb: The paraphrase database. In *Proceedings of the 2013 conference of the north american chapter of the association for computational linguistics: Human language technologies*, pages 758–764.

Naman Goyal, Jingfei Du, Myle Ott, Giri Anantharaman, and Alexis Conneau. 2021. Larger-scale transformers for multilingual masked language modeling. *arXiv preprint arXiv:2105.00572*.

Shlomo Hareli, Konstantinos Kafetsios, and Ursula Hess. 2015. A cross-cultural study on emotion expression and the learning of social norms. *Frontiers in psychology*, 6:152022.

Peter Hase and Mohit Bansal. 2020. Evaluating explainable ai: Which algorithmic explanations help users predict model behavior? *arXiv preprint arXiv:2005.01831*.

Devamanyu Hazarika, Soujanya Poria, Roger Zimmermann, and Rada Mihalcea. 2021. Conversational transfer learning for emotion recognition. *Information Fusion*, 65:1–12.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. 2019. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461*.

Nankai Lin, Yingwen Fu, Xiaotian Lin, Dong Zhou, Aimin Yang, and Shengyi Jiang. 2023. Cl-xabsa: Contrastive learning for cross-lingual aspect-based sentiment analysis. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*.

Edward Loper and Steven Bird. 2002. Nltk: The natural language toolkit. *arXiv preprint cs/0205028*.

Andreas Madsen, Siva Reddy, and Sarath Chandar. 2022. Post-hoc interpretability for neural nlp: A survey. *ACM Computing Surveys*, 55(8):1–42.

Aaron Maladry, Pranaydeep Singh, and Els Lefever. 2024. Findings of the wassa 2024 exalt shared task on explainability for cross-lingual emotion in tweets. In *Proceedings of the 14th Workshop of on Computational Approaches to Subjectivity, Sentiment & Social Media Analysis@ACL 2024*, Bangkok, Thailand.

Jonathan Mallinson, Rico Sennrich, and Mirella Lapata. 2017. Paraphrasing revisited with neural machine translation. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 881–893.

Pansy Nandwani and Rupali Verma. 2021. A review on sentiment analysis and emotion detection from text. *Social network analysis and mining*, 11(1):81.

Ellie Pavlick, Pushpendre Rastogi, Juri Ganitkevitch, Benjamin Van Durme, and Chris Callison-Burch. 2015. Ppdb 2.0: Better paraphrase ranking, fine-grained entailment relations, word embeddings, and style classification. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 425–430.

Lucas Francisco Amaral Orosco Pellicer, Taynan Maier Ferreira, and Anna Helena Reali Costa. 2023. Data augmentation techniques in natural language processing. *Applied Soft Computing*, 132:109803.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research*, 21(140):1–67.

Antonio Reyes, Paolo Rosso, and Davide Buscaldi. 2012. From humor recognition to irony detection: The figurative language of social media. *Data & Knowledge Engineering*, 74:1–12.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Eric Wallace, Matt Gardner, and Sameer Singh. 2020. Interpreting predictions of nlp models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Tutorial Abstracts*, pages 20–23.

Jason Wei and Kai Zou. 2019. Eda: Easy data augmentation techniques for boosting performance on text classification tasks. *arXiv preprint arXiv:1901.11196*.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2020. mt5: A massively multilingual pre-trained text-to-text transformer. *arXiv preprint arXiv:2010.11934*.