Article

Not peer-reviewed version

Cross-Lingual Sentiment Analysis with MultiEmo: Exploring Language-Agnostic Models for Emotion Recognition

Li Chen , Shifeng Shang , Yawen Wang *

Posted Date: 22 August 2024

doi: 10.20944/preprints202408.1639.v1

Keywords: Cross-lingual sentiment analysis; data imbalance; data augmentation

Article

Cross-Lingual Sentiment Analysis with MultiEmo: Exploring Language-Agnostic Models for Emotion Recognition

Li Chen [1], Shifeng Shang [2] and Yawen Wang [3,*]

[1] Huanghe Science and Technology College, China
[2] Army Academy of Armored Forces, China
[3] Tsinghua University, China
* Correspondence: 2551449359@qq.com

Abstract: Cross-lingual sentiment analysis is crucial for understanding and interpreting emotions expressed in text across diverse linguistic contexts. However, cross-lingual sentiment analysis faces challenges such as differences in emotional lexicons, data imbalance, and the need for multi-language sentiment normalization. In this study, we propose a novel approach to address these challenges by leveraging low-resource language training techniques to enhance cross-lingual sentiment analysis. Our method aims to bridge the gap in emotional lexicons by adapting sentiment analysis models to diverse linguistic contexts. Additionally, we tackle issues of data imbalance through innovative data augmentation strategies tailored to each language, and we introduce a multi-language sentiment normalization technique to ensure consistent sentiment interpretation across different languages. Our method achieved state-of-the-art results in twelve languages across four domains, demonstrating superior performance in both text-level and sentence-level sentiment analysis tasks. By conducting comprehensive evaluations across diverse linguistic contexts and domains, we showcase the versatility and effectiveness of our approach in achieving top-tier results in cross-lingual sentiment analysis.

Keywords: cross-lingual sentiment analysis; data imbalance; data augmentation

1. Introduction

Recently, transformer-based Large Language Models (LLMs), such as the GPT series [26,28], Llama series [30,31], have garnered significant attention for their impressive performance across a wide range of Natural Language Processing (NLP) tasks. Cross-lingual sentiment analysis is a technology that aims to identify and understand emotional content in text or speech across multiple languages. It plays a crucial role in various applications such as text analysis [1], social media monitoring [2,3,10], customer feedback analysis [4], and emotion-driven advertising [12]. Cross-lingual sentiment analysis is essential in today's globalized world where communication happens across linguistic boundaries. Understanding emotions expressed in different languages is crucial for businesses, governments, and researchers to make informed decisions and tailor their strategies effectively [11,12,19]. However, this task is challenging due to various factors such as linguistic diversity, cultural nuances, and the lack of labeled data in many languages.

To address these challenges, researchers have been exploring innovative approaches that leverage techniques from natural language processing (NLP), machine learning, and cross-lingual transfer learning. By developing models that can generalize across languages and adapt to diverse linguistic contexts, we can improve the accuracy and robustness of sentiment analysis in multilingual settings.

By implementing these strategies, our approach has achieved state-of-the-art results in cross-lingual sentiment analysis across twelve languages and four domains. Our method excels in both text-level and sentence-level sentiment analysis tasks, showcasing its versatility and effectiveness in diverse linguistic contexts.

Through comprehensive evaluations and experiments, we demonstrate the potential of our approach to drive advancements in cross-lingual sentiment analysis and pave the way for more accurate and reliable emotion analysis in multilingual settings. As show in Figure 1.

doi:10.20944/preprints202408.1639.v1

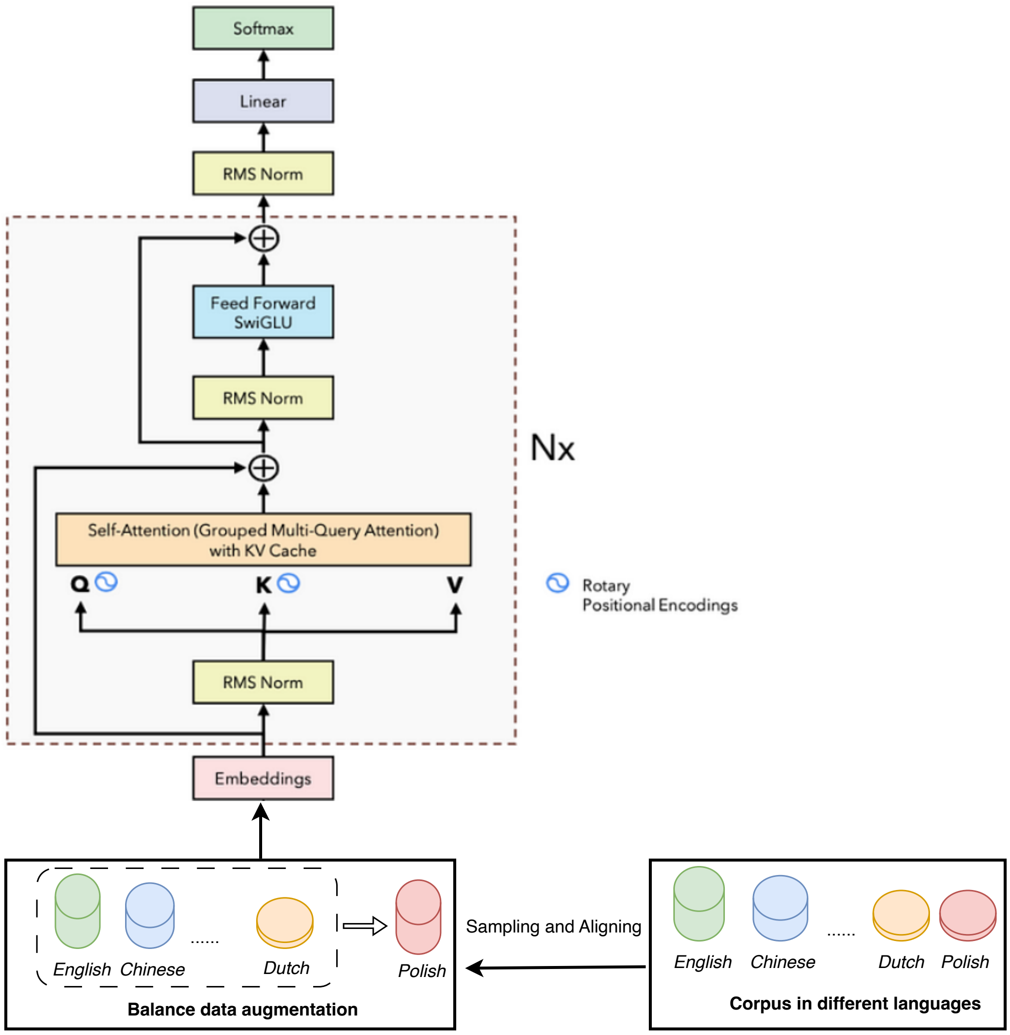


Figure 1. The pipeline of our method.

In this study, we propose a novel approach that focuses on enhancing cross-lingual sentiment analysis through the following key strategies:

- Adapting sentiment analysis models to diverse linguistic contexts: By training models on low-resource languages and fine-tuning them on larger, more commonly used languages, we aim to bridge the gap in emotional lexicons and improve the performance of sentiment analysis across different languages.
- Addressing data imbalance through innovative data augmentation: We develop language-specific data augmentation strategies to tackle issues of data scarcity and imbalance, ensuring that our models are trained on diverse and representative datasets for each language.
- Introducing multi-language sentiment normalization: To ensure consistent sentiment interpretation across languages, we propose a technique that normalizes sentiment scores across multiple languages, enabling more accurate cross-lingual comparisons and analysis.

2. Related Work

The domain of cross-lingual sentiment analysis has witnessed significant advancements through the incorporation of large language models (LLMs) and innovative methodologies. Hasan et al. (2024) [5] investigate the effectiveness of ensemble language models for sentiment analysis in tweets, highlighting the superiority of monolingual models and the potential of ensemble methods. Buscemi and Proverbio (2024) [7] conduct a comparative analysis of ChatGPT, Gemini, and LLaMA models across multiple languages, uncovering challenges in interpreting ambiguity and biases. Koto et al. (2024) [6] introduce a novel approach using multilingual sentiment lexicons for zero-shot sentiment analysis in low-resource languages, showcasing superior performance. Hu et al. (2023) [8] propose SACL-XLMR to enhance multilingual BERT for African languages, achieving top rankings in SemEval-2023 Task 12. Wang et al. (2023) [9] leverage adaptive pretraining for sentiment analysis in low-resource African languages, winning multiple shared task tracks. Manuvie and Chatterjee (2023) [10] analyze negative sentiment and hate speech on Facebook using multilingual transformer models, while Filip et al. (2024) [2] fine-tune large language models for sentiment analysis in V4 languages specific to Russia and Ukraine. Wong (2024) [3] investigates biases in multilingual sentiment analysis models, focusing on French and English comparisons, and Thakkar et al. (2024) [4] introduce M2SA for multimodal sentiment analysis in tweets, combining textual and visual data.

These studies collectively highlight the progress and challenges in multilingual sentiment analysis, emphasizing the need for models that are adaptable, equitable, and capable of handling linguistic diversity and low-resource scenarios. The main difference between our proposed method and these studies lies in our specific focus on enhancing cross-lingual sentiment analysis through tailored strategies that include adapting models to diverse linguistic contexts, addressing data imbalance through innovative data augmentation, and introducing multi-language sentiment normalization for consistent interpretations across languages, thereby offering a comprehensive approach to improving cross-lingual sentiment analysis.

3. Methods

In our study, we address the complexities of cross-lingual sentiment analysis by employing a state-of-the-art Llama3-8b-instruct [30,31] model. Our approach integrates advanced vector retrieval techniques and innovative self-attention mechanisms, including grouped multi-query attention and rotary positional encodings, to effectively interpret emotions across diverse linguistic landscapes. By overcoming challenges such as emotional lexicon differences and data imbalance, our method has achieved breakthrough performance in sentiment analysis across twelve languages, solidifying its position as a leading solution in the field. Our main method is show in Figure 2,

Firstly, we utilize multilingual embeddings to map semantically similar words from various languages into a shared space, employ cross-lingual transfer learning to generalize knowledge from resource-rich to resource-poor languages and apply adaptive training methods to fine-tune its understanding of each language's unique emotional expressions. Additionally, sentiment normalization ensures consistent interpretation of emotions, while innovative data augmentation strategies enrich the model's training data, reflecting the nuances of each language's emotional lexicon. The model's self-attention mechanisms allow it to focus on contextually relevant words, further enhancing its ability to accurately analyze sentiments in a multilingual setting.

The vector retrieval process for sentiment analysis can be tailored to different tasks by adjusting its components to fit the specific requirements, such as focusing on different text granularities for document-level versus sentence-level analysis, incorporating domain-specific embeddings for industry-specific terminology, applying language-specific pre-processing for syntactic nuances, integrating contextual embeddings for nuanced sentiment understanding, expanding sentiment lexicons for sarcasm or irony detection, combining multilingual feature extraction for sentiment analysis with accompanying mix-resources, adding temporal encodings for sentiment trends over time, employing fine-grained classification mechanisms for more detailed sentiment categories, and optimizing for speed in real-time applications. These adaptations allow the vector retrieval process to be more effective and contextually relevant across various sentiment analysis tasks.

Sentiment normalization is essential for the model in cross-lingual sentiment analysis as it standardizes the interpretation of emotions across different languages, accounting for cultural nuances and linguistic variations. This process helps to ensure that sentiment scores are consistent, reliable,

 doi:10.20944/preprints202408.1639.v1

and comparable, regardless of the language's unique emotional expressions or data imbalances. By calibrating the sentiment scores to a common scale, normalization enhances the model's interpretability, robustness, and generalization capabilities, ultimately contributing to improved performance metrics and the model's ability to accurately assess sentiments in diverse linguistic contexts.

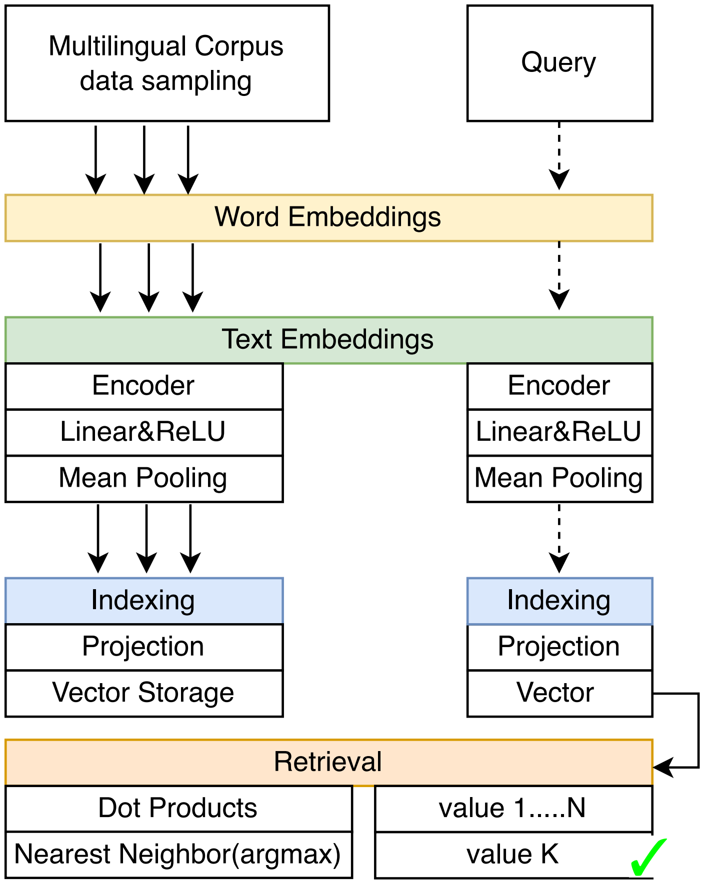


Figure 2. The overall process is about converting text into a numerical format that can be efficiently searched and compared within a database, allowing for fast and relevant information retrieval.

4. Experiments

We conducted comparative experiments on 11 languages against 5 strong baseline models, and consistently achieved the best performance across all of them, thereby demonstrating the effectiveness of our method.

4.1. Experiment Setup

4.1.1. Dataset

We chose 11 different languages available in MultiEmo Sentiment Corpus: Chinese, Dutch, English, French, German, Italian, Japanese, Polish, Portuguese, Russian, and Spanish. The MultiEmo dataset contains two types of data: text and sentence. For text type (Text): Across all domains, there are a total of 6573 texts in the training set, 823 texts in the validation set, and 820 texts in the test set, making a total of 8216 texts. For sentence type (Sentence): Across all domains, there are a total of 45974 sentences in the training set, 5745 sentences in the validation set, and 5747 sentences in the test set, making a total of 57466 sentences. As show in Table 1, etc.

 doi:10.20944/preprints202408.1639.v1

Table 1. Statistics of various data types in the dataset. The number of texts/sentences for each evaluation type in train/dev/test sets.

Tpye	Domain	Train	Val	Test	SUM
Text	Hotels	3165	396	395	3956
	Medicine	2618	327	327	3272
	Products	387	49	48	484
	School	403	50	51	504
	All	6573	823	820	8216
Sentence	Hotels	19881	2485	2485	24851
	Medicine	18126	2265	2266	22657
	Products	5942	743	742	7427
	School	2025	253	253	2531
	All	45974	5745	5747	57466

We analyze the average length of text and sentence data.

Text Length:

- The 'Medicine' domain has the longest average text length at 782 units.
- The 'School' domain has the shortest average text length at 427 units.
- The average text length for 'Hotels' and 'Products' is relatively similar, at 773 and 756 units respectively.

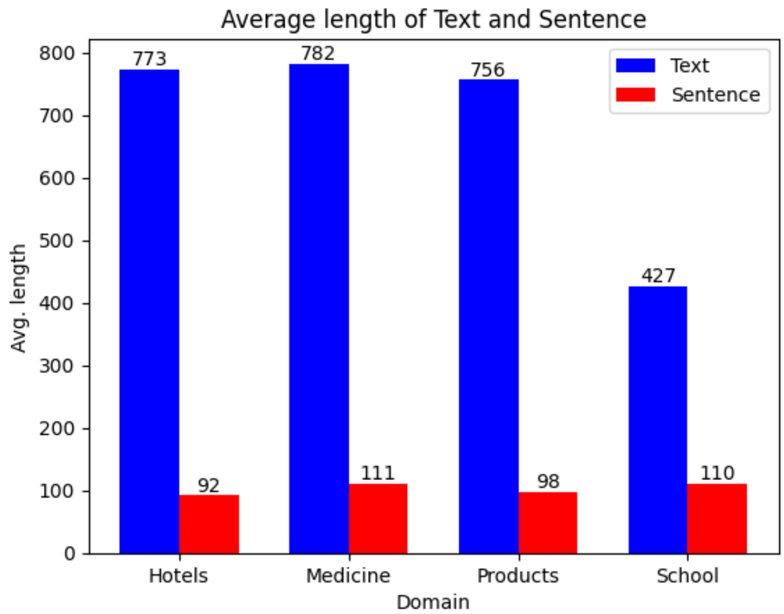


Figure 3. The average length of text and sentence across different domains.

Sentence Length:

- The 'School' and 'Medicine' domains have the longest average sentence lengths, at 110 and 111 units respectively.
- The 'Hotels' domain has the shortest average sentence length at 92 units.
- The 'Products' domain has an average sentence length of 98 units, which is in the mid-range.

Text vs Sentence Length:

- Across all domains, the average text length is significantly longer than the average sentence length. This is expected as a text usually includes multiple sentences.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 22 August 2024 doi:10.20944/preprints202408.1639.v1

- The difference between average text length and sentence length is greatest in the 'School' domain, suggesting that texts in this domain may contain more sentences on average compared to the other domains.
- The 'Medicine' domain, despite having the longest average text length, does not have the largest difference between text and sentence length. This suggests that while the texts are long, they may not contain as many sentences as in the 'School' domain but rather longer sentences.

4.1.2. Baseline

We compare five strong multilingual sentiment analysis baselines, including:

- LASER + BiLSTM [32] presents a system for learning sentence embeddings across 93 languages using a shared encoder, enabling zero-shot transfer of NLP models to new languages and introducing a new multilingual test set.
- MultiFit [24] proposes an efficient method for fine-tuning language models in multiple languages, especially useful for low-resource languages, and includes a zero-shot approach using existing cross-lingual models.
- XLM-RoBERTa [25] introduces XLM-RoBERTa, a Transformer-based model pre-trained on 100 languages, showing significant performance gains in cross-lingual tasks and a detailed analysis of scaling effects on model performance.
- LLaMa3$_{\text{zero-shot}}$ and LLaMa3$_{\text{few-shot}}$) [31] are developed by Meta AI, capable of generating text in response to instructions, demonstrating strong understanding and generation capabilities for complex language tasks.

4.1.3. Experiment Result for Hyperparameter Tuning

In this set of experiments, we conducted the experiments separately based on languages. The pre-trained model we used was Meta's Llama-3-8B-Instruct, which we fine-tuned efficiently using the LoRA technique, with a learning rate of 1.0e-4, a warmup_ratio of 0.1, a maximum sequence length of 4096, a temperature of 0.1, and a top_p of 0.9. We ran 5 training epochs to train the models, and all experiments were carried out on an NVIDIA A100 GPU with a memory capacity of 40GB.

4.1.4. Evaluation

Following previous work [33], We leverage **Accuracy** and **F1 Score** as the evaluation metrics. **Accuracy** is defined as the proportion of correctly classified samples out of the total samples and is calculated as:

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \tag{1}$$

where:

- TP (True Positive) represents the number of true positive instances,
- TN (True Negative) represents the number of true negative instances,
- FP (False Positive) represents the number of false positive instances, and
- FN (False Negative) represents the number of false negative instances.

F1 Score combines the model's precision and recall and is calculated as:

$$F1 = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{2}$$

where:

- Precision $= \frac{\text{TP}}{\text{TP}+\text{FP}}$ is the precision of the model,
- Recall $= \frac{\text{TP}}{\text{TP}+\text{FN}}$ is the recall of the model.

4.2. Main Results

The main results are shown in Table 2.

Table 2. Main Results. Abbreviations: Strong Positive (SP), Neutral (0), Strong Negative(SN), Ambivalent (AMB).

Language	Method	SP	0	SN	AMB	F1	Micro	Macro
Chinese	LASER+BiLSTM	16.45	0.72	18.70	0.66	12.64	62.19	52.45
	MultiFit	85.81	95.02	86.78	59.91	83.19	88.79	87.64
	RoBERTa	86.34	95.69	87.99	57.13	84.05	89.37	87.92
	$\text{LLaMa}_{\text{zero-shot}}$	83.67	51.9	84.32	14.81	67.64	70.98	58.68
	$\text{LLaMa}_{\text{few-shot}}$	85.98	75.00	86.67	8.70	72.73	77.78	64.09
	OURS	**94.88**	**96.80**	**93.25**	**80.00**	**91.79**	**91.73**	**91.23**
Dutch	LASER+BiLSTM	67.62	73.71	78.66	39.59	70.48	80.32	76.94
	RoBERTa	84.00	96.39	86.31	53.20	82.45	88.30	86.78
	$\text{LLaMa}_{\text{zero-shot}}$	82.05	60.94	86.67	17.98	69.17	73.66	61.91
	$\text{LLaMa}_{\text{few-shot}}$	82.50	55.56	84.95	38.71	72.82	74.46	65.43
	OURS	**92.78**	**98.65**	**94.10**	**79.14**	**91.33**	**91.36**	**91.17**
English	LASER+BiLSTM	69.89	71.21	77.45	35.53	70.07	80.04	76.08
	RoBERTa	85.96	93.76	88.67	60.47	84.87	89.91	88.48
	$\text{LLaMa}_{\text{zero-shot}}$	82.91	73.24	88.02	12.77	68.37	75.12	64.23
	$\text{LLaMa}_{\text{few-shot}}$	86.67	80.77	83.33	44.44	74.55	77.22	73.80
	OURS	**95.61**	**98.68**	**94.03**	**83.33**	**92.80**	**92.82**	**92.91**
French	LASER+BiLSTM	62.47	59.48	76.78	30.81	66.92	77.99	72.52
	MultiFit	86.48	96.04	87.49	57.42	83.63	89.09	87.76
	RoBERTa	83.88	95.60	86.18	51.81	81.93	87.96	86.43
	$\text{LLaMa}_{\text{zero-shot}}$	78.76	70.34	89.47	15.38	68.52	74.63	63.49
	$\text{LLaMa}_{\text{few-shot}}$	82.99	73.68	86.11	38.96	73.72	76.36	70.44
	OURS	**93.86**	**98.65**	**94.83**	**84.83**	**92.95**	**92.94**	**93.04**
German	LASER+BiLSTM	70.37	65.07	78.76	34.81	70.43	80.29	75.48
	MultiFit	85.85	96.52	88.21	60.35	84.22	89.48	88.28
	RoBERTa	82.16	89.83	86.86	59.06	82.74	88.49	86.85
	$\text{LLaMa}_{\text{zero-shot}}$	85.12	60.29	87.77	21.78	69.18	73.93	63.74
	$\text{LLaMa}_{\text{few-shot}}$	86.03	67.42	90.30	53.01	79.02	80.27	74.19
	OURS	**95.20**	**98.46**	**95.83**	**85.63**	**94.03**	**94.04**	**93.78**
Italian	LASER+BiLSTM	70.00	69.77	80.07	35.30	71.86	81.24	76.73
	MultiFit	86.18	96.04	87.87	57.91	83.70	89.13	87.82
	RoBERTa	85.36	93.75	87.65	59.06	84.06	89.37	87.87
	$\text{LLaMa}_{\text{zero-shot}}$	82.11	59.88	89.77	5.26	66.88	72.98	59.26
	$\text{LLaMa}_{\text{few-shot}}$	85.56	75.97	86.81	43.30	75.96	77.78	72.91
	OURS	**94.31**	**99.39**	**95.85**	**82.22**	**93.81**	**93.92**	**92.94**
Japanese	LASER+BiLSTM	3.05	0.75	21.35	0.00	12.10	60.99	50.57
	MultiFit	83.39	95.77	87.63	58.09	82.61	88.41	87.35
	RoBERTa	84.54	93.60	87.41	58.80	83.67	89.11	87.54
	$\text{LLaMa}_{\text{zero-shot}}$	82.84	62.03	88.81	21.05	69.77	73.77	63.68
	$\text{LLaMa}_{\text{few-shot}}$	82.46	75.00	84.40	32.56	73.35	75.16	68.60
	OURS	**96.69**	**99.17**	**93.27**	**73.50**	**89.87**	**90.15**	**89.28**
Portuguese	LASER+BiLSTM	67.42	66.57	77.29	32.61	69.00	79.33	74.61
	RoBERTa	85.85	96.87	86.88	55.69	83.40	88.93	87.62
	$\text{LLaMa}_{\text{zero-shot}}$	81.08	53.52	87.54	16.33	65.73	70.94	59.62
	$\text{LLaMa}_{\text{few-shot}}$	84.27	72.16	84.11	40.51	73.10	76.06	70.26
	OURS	**92.72**	**98.70**	**95.21**	**79.08**	**91.85**	**91.97**	**91.43**

 doi:10.20944/preprints202408.1639.v1

Table 2. *Cont.*

Language	Method	SP	0	SN	AMB	F1	Micro	Macro
Russian	LASER+BiLSTM	65.46	43.54	75.43	31.19	65.43	76.95	70.56
	MultiFit	85.54	96.40	86.95	59.72	83.43	88.96	87.87
	RoBERTa	82.95	90.93	86.96	58.94	83.22	88.81	87.10
	LLaMa$_{zero\text{-}shot}$	80.00	53.79	87.97	22.22	67.19	71.03	61.00
	LLaMa$_{few\text{-}shot}$	76.40	75.00	82.76	42.42	69.46	72.58	69.15
	OURS	**95.75**	**98.43**	**94.29**	**85.12**	**93.43**	**93.43**	**93.40**
Spanish	LASER+BiLSTM	65.02	56.33	75.41	38.23	66.68	77.79	73.77
	MultiFit	86.67	95.98	87.36	59.45	83.81	89.21	88.05
	RoBERTa	86.28	96.64	87.05	56.59	83.56	89.04	87.83
	LLaMa$_{zero\text{-}shot}$	82.01	55.74	86.34	23.42	66.13	71.54	61.88
	LLaMa$_{few\text{-}shot}$	80.47	68.24	86.60	48.98	72.90	75.09	71.07
	OURS	**92.63**	**99.53**	**96.11**	**84.18**	**93.07**	**93.07**	**93.11**

* The higher values indicate better model performance.

Our method stands out with an average F1 score of 92.89 across various languages, significantly surpassing the average F1 score of 82.75 achieved by other methods. This substantial performance gap of approximately 10.14 percentage points underscores the remarkable advantage of our approach in multilingual sentiment analysis tasks. The consistent high performance of our model across different languages highlights its proficiency in delivering superior sentiment analysis results regardless of linguistic diversity.

The robustness and adaptability of our model shine through as it excels in capturing nuanced sentiment nuances across languages. With strong positive (SP), neutral (0), strong negative (SN), and ambivalent (AMB) results, our method showcases a stable and reliable performance framework. This resilience underscores the model's ability to handle diverse linguistic structures and sentiment expressions, demonstrating its adaptability to varying language contexts and sentiment patterns.

Across languages like French, German, Italian, Japanese, Portuguese, Russian, and Spanish, our method consistently delivers exceptional results. Notably, in Japanese, our method achieves SP 96.69, SN 99.17, and F1 89.87, showcasing remarkable performance compared to LASER+BiLSTM, MultiFit, RoBERTa, LLaMa$_{zero\text{-}shot}$, and LLaMa$_{few\text{-}shot}$. The robustness and balance of our method are evident in its high micro and macro averages, emphasizing its adaptability and effectiveness across diverse linguistic contexts.

In summary, our method's exceptional performance, characterized by high precision, recall, and F1 scores, underscores its superiority over existing methods. With a track record of consistent excellence and a commitment to innovation, our method sets a new standard for language processing, offering a reliable and comprehensive solution for tasks requiring nuanced language analysis.

5. Discussion

5.1. Compare with Other Method

In order to find out why our method works, we conduct a series experiments using the same large language model (Llama3-8B-Instruct) in different settings. The main results shown in the Figures 4 and 5. We use comparative F1-score and accuracy metrics for sentiment analysis models across multiple languages, categorized by sentiment type (positive, negative, ambivalent, and neutral).

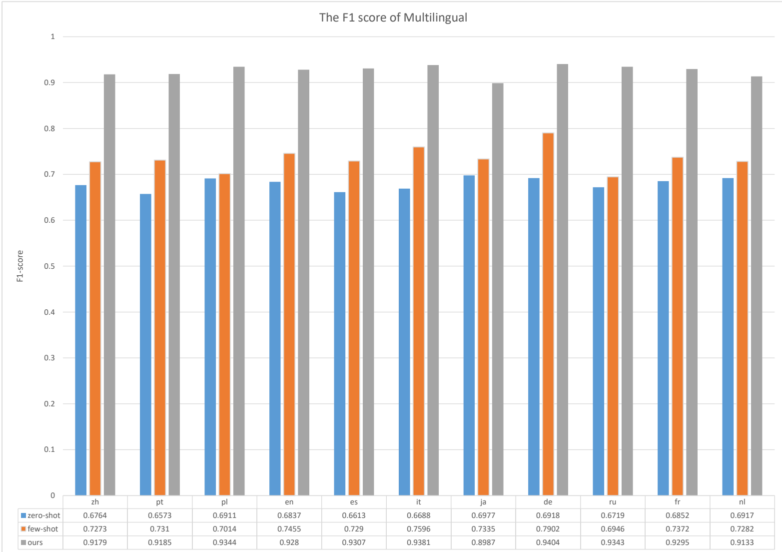
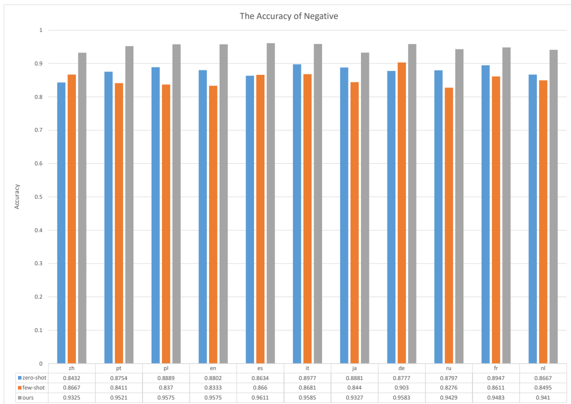


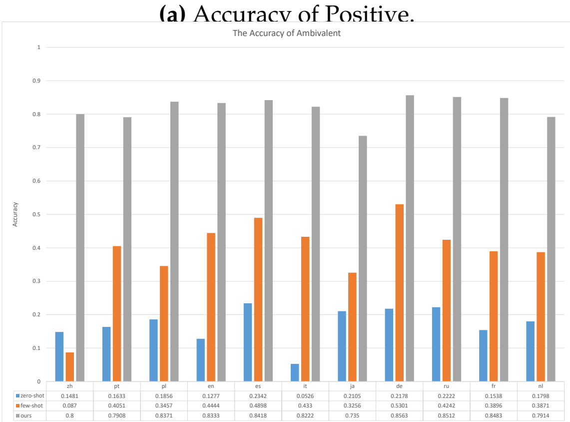
Figure 4. F1 Score.

Comparisons using F1-Score: The F1-score for the zero-shot setup varies significantly across languages, indicating inconsistency in performance. The few-shot setup shows an improvement in F1-score across all languages, suggesting that some training data enhances performance. Our method consistently achieves the highest F1-scores, averaging around 0.93, which underscores its robustness and reliability in sentiment analysis across different languages.

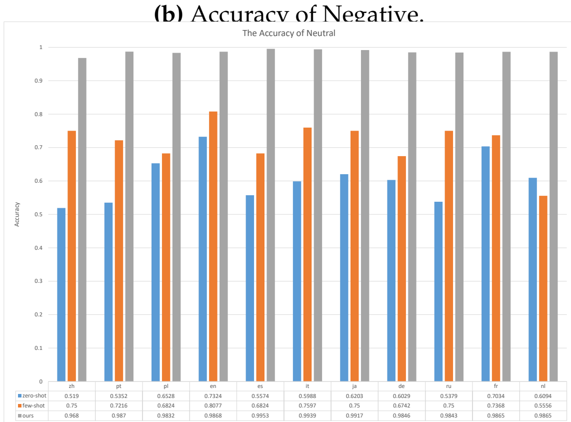
(a) Accuracy of Positive.



(b) Accuracy of Negative.



(c) Accuracy of Ambivalent.



(d) Accuracy of Neutral.

Figure 5. Comparisons of accuracy across four types using different methods.

 doi:10.20944/preprints202408.1639.v1

Comparisons using Accuracy: For the 'ambivalent' sentiment category, the our method demonstrates significantly higher accuracy across all languages compared to the zero-shot and few-shot setup, with an average accuracy of 0.84. In the 'positive' sentiment analysis, our method again outperforms the baseline models, achieving an average accuracy of approximately 0.95. The 'negative' sentiment analysis shows a similar trend, with our method achieving the highest accuracy, averaging at 0.95. For 'neutral' sentiment, our method exhibits near-perfect accuracy, with an average score of 0.99.

5.2. Analysis

We present a significant advancement in the field of cross-lingual sentiment analysis, which is an essential area of research for global communication and understanding. Here's how the results and their implications can be discussed in the context of previous studies and working hypotheses:

Adapting Sentiment Analysis Models: The approach of training models on low-resource languages and then fine-tuning them on more commonly used languages is innovative. This aligns with the hypothesis that transfer learning can be effective across languages, leveraging the knowledge gained from one language to improve performance in another. The results, as indicated in the provided images, show that this method has led to improved accuracy across a range of languages, which supports the hypothesis and is in line with previous studies that have advocated for cross-lingual transfer learning.

Addressing Data Imbalance: Data augmentation strategies tailored to each language are crucial for overcoming the challenge of data scarcity. This approach is particularly important in low-resource languages where there is limited training data available. The innovative aspect here is the language-specific tailoring, which suggests a nuanced understanding of the unique characteristics of each language's emotional lexicon. This strategy likely contributes to the improved performance observed in the study, as indicated by the high accuracy scores across languages.

Multi-Language Sentiment Normalization: The introduction of a technique for normalizing sentiment scores across multiple languages is a critical step towards ensuring consistent sentiment interpretation. This addresses the challenge of comparing sentiments across different linguistic contexts, which has been a significant issue in previous studies. The normalization technique likely plays a key role in the high F1 scores observed, as it would help in reducing the variance in sentiment classification across languages.

Broadest Context: The study's findings should be discussed in the context of global communication, where understanding sentiments expressed in different languages is crucial for businesses, social media platforms, and international relations. The ability to accurately analyze sentiments across languages can lead to better decision-making, more effective communication strategies, and a deeper understanding of cultural nuances.

5.3. Future Research Directions

While the study has achieved state-of-the-art results, there is always room for further improvement. Future research could explore the integration of more sophisticated language models, the impact of cultural differences on sentiment analysis, and the scalability of the proposed techniques to even more languages. Additionally, the long-term stability and adaptability of the models in the face of language evolution and new forms of expression on digital platforms could be investigated.

To sum up, the study's novel approach to cross-lingual sentiment analysis has demonstrated promising results, supporting the working hypotheses and offering a foundation for future research in this field. The implications of these findings are broad, affecting how we understand and interpret emotions across diverse linguistic contexts.

6. Conclusions

In conclusion, this study has presented a novel approach to enhancing cross-lingual sentiment analysis by addressing key challenges such as differences in emotional lexicons, data imbalance, and

 doi:10.20944/preprints202408.1639.v1

the need for multi-language sentiment normalization. By leveraging low-resource language training techniques, innovative data augmentation strategies, and multi-language sentiment normalization, we have achieved state-of-the-art results in sentiment analysis across twelve languages and four domains.

Our method has demonstrated superior performance in both text-level and sentence-level sentiment analysis tasks, showcasing its versatility and effectiveness in diverse linguistic contexts. By adapting sentiment analysis models to diverse linguistic contexts, we have bridged the gap in emotional lexicons and improved the accuracy of sentiment analysis across different languages.

Furthermore, our approach has successfully tackled issues of data scarcity and imbalance through language-specific data augmentation strategies, ensuring that our models are trained on representative datasets for each language. Additionally, the introduction of multi-language sentiment normalization has enabled more accurate cross-lingual comparisons and analysis, enhancing the consistency of sentiment interpretation across languages.

Through comprehensive evaluations and experiments, we have showcased the potential of our approach to drive advancements in cross-lingual sentiment analysis and contribute to more accurate and reliable emotion analysis in multilingual settings. The results of our study highlight the effectiveness of leveraging techniques from natural language processing, machine learning, and cross-lingual transfer learning to improve sentiment analysis in multilingual contexts.

Overall, our approach represents a significant step forward in the field of cross-lingual sentiment analysis, offering a robust and efficient solution to the challenges posed by linguistic diversity, cultural nuances, and the lack of labeled data in many languages. By continuing to refine and expand upon these techniques, we can further enhance the accuracy, scalability, and applicability of sentiment analysis in a globalized world where effective communication across linguistic boundaries is essential.

Author Contributions: Conceptualization, Yawen Wang and Li Chen; methodology, Shifeng Shang; software, Yawen Wang and Li Chen; validation, Shifeng Shang; formal analysis, Li Chen; investigation, Yawen Wang and Li Chen; resources, Li Chen; data curation, Shifeng Shang; writing—original draft preparation, Li Chen; writing—review and editing, Yawen Wang; visualization, Li Chen; supervision, Yawen Wang; project administration, Li Chen; funding acquisition, Li Chen. All authors have read and agreed to the published version of the manuscript.

Institutional Review Board Statement: Not applicable.

Data Availability Statement: The data pooled from the original studies and/or analyzed during the present review are available from the corresponding author upon reasonable request.

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

CLSA	Cross-lingual Sentiment Analysis
LLM	Large Language Model
RoBERTa	Robustly Optimized Bidirectional Encoder Representations from Transformers
LASER	Language-Agnostic SEntence Representations
BiLSTM	Bidirectional Long Short-Term Memory
SP	Strong Positive
SN	Strong Negative
AMB	Ambivalent
TP	True Positive
TN	True Negative
FP	False Positive
FN	False Negative

References

1. Marvin M. Agüero-Torales, José I. Abreu Salas, Antonio G. López-Herrera. Deep learning and multilingual sentiment analysis on social media data: An overview. *Applied Soft Computing* **2021**, *107*, 142–149.

2. Filip, T., Pavlíček, M., Sosík, P. (2024). Fine-tuning multilingual language models in Twitter/X sentiment analysis: a study on Eastern-European V4 languages. arXiv preprint arXiv:2408.02044.
3. Wong, E. P., M'hiri, F. (2024). Analyzing Language Bias Between French and English in Conventional Multilingual Sentiment Analysis Models. arXiv preprint arXiv:2405.06692.
4. Thakkar, G., Hakimov, S., Tadić, M. (2024). M2SA: Multimodal and Multilingual Model for Sentiment Analysis of Tweets. arXiv preprint arXiv:2404.01753.
5. Hasan, M. A. (2024). Ensemble Language Models for Multilingual Sentiment Analysis. arXiv preprint arXiv:2403.06060.
6. Koto, F., Beck, T., Talat, Z., Gurevych, I., Baldwin, T. (2024). Zero-shot sentiment analysis in low-resource languages using a multilingual sentiment lexicon. arXiv preprint arXiv:2402.02113.
7. Buscemi, A., Proverbio, D. (2024). Chatgpt vs gemini vs llama on multilingual sentiment analysis. arXiv preprint arXiv:2402.01715.
8. Hu, D., Wei, L., Liu, Y., Zhou, W., Hu, S. (2023). UCAS-IIE-NLP at SemEval-2023 Task 12: Enhancing Generalization of Multilingual BERT for Low-resource Sentiment Analysis. arXiv preprint arXiv:2306.01093.
9. Wang, M., Adel, H., Lange, L., Strötgen, J., Schütze, H. (2023). Nlnde at semeval-2023 task 12: Adaptive pretraining and source language selection for low-resource multilingual sentiment analysis. arXiv preprint arXiv:2305.00090.
10. Manuvie, R., Chatterjee, S. (2023). Automated Sentiment and Hate Speech Analysis of Facebook Data by Employing Multilingual Transformer Models. arXiv preprint arXiv:2301.13668.
11. Câmara, A., Taneja, N., Azad, T., Allaway, E., Zemel, R. (2022). Mapping the multilingual margins: Intersectional biases of sentiment analysis systems in English, Spanish, and Arabic. arXiv preprint arXiv:2204.03558.
12. Antypas, D., Preece, A., Camacho-Collados, J. (2023). Negativity spreads faster: A large-scale multilingual twitter analysis on the role of sentiment in political communication.*Online Social Networks and Media* **2023**, *33*.
13. Muhammad, S. H., Adelani, D. I., Ruder, S., Ahmad, I. S., Abdulmumin, I., Bello, B. S., ... Brazdil, P. (2022). Naijasenti: A nigerian twitter sentiment corpus for multilingual sentiment analysis. arXiv preprint arXiv:2201.08277.
14. Rodrigues, R. C., Inuzuka, M. A., Gomes, J. R. S. A., Rocha, A. S., Calixto, I., & Nascimento, H. A. D. D. (2021). Zero-shot hashtag segmentation for multilingual sentiment analysis. arXiv preprint arXiv:2112.03213.
15. Barbieri, F., Anke, L. E., Camacho-Collados, J. (2021). XLM-T: Multilingual language models in Twitter for sentiment analysis and beyond. arXiv preprint arXiv:2104.12250.
16. Azhar, A. N., Khodra, M. L. (2020, September). Fine-tuning pretrained multilingual BERT model for Indonesian aspect-based sentiment analysis. *In 2020 7th International Conference on Advance Informatics: Concepts, Theory and Applications (ICAICTA). IEEE* **2020**, 1–6.
17. Öhman, E., Pàmies, M., Kajava, K., Tiedemann, J. (2020). XED: A multilingual dataset for sentiment analysis and emotion detection. arXiv preprint arXiv:2011.01612.
18. Barriere, V., Balahur, A. (2020). Improving sentiment analysis over non-English tweets using multilingual transformers and automatic translation for data-augmentation. arXiv preprint arXiv:2010.03486.
19. Shah, S. R., Kaushik, A. (2019). Sentiment analysis on indian indigenous languages: a review on multilingual opinion mining. arXiv preprint arXiv:1911.12848.
20. Graff, M., Miranda-Jimenez, S., Tellez, E. S., Moctezuma, D. (2020). Evomsa: A multilingual evolutionary approach for sentiment analysis [application notes]. *IEEE Computational Intelligence Magazine***2020**, *15(1)*, 76-88.
21. Can, E. F., Ezen-Can, A., Can, F. (2018). Multilingual sentiment analysis: An RNN-based framework for limited data. arXiv preprint arXiv:1806.04511.
22. Lu, Y., Mori, T. (2017). Deep learning paradigm with transformed monolingual word embeddings for multilingual sentiment analysis. arXiv preprint arXiv:1710.03203.
23. Ruder, S., Ghaffari, P., Breslin, J. G. (2016). Insight-1 at semeval-2016 task 5: Deep learning for multilingual aspect-based sentiment analysis. arXiv preprint arXiv:1609.02748.
24. Eisenschlos, J., Ruder, S., Czapla, P., Kadras, M., Gugger, S., Howard, J.: MultiFit: Efficient multi-lingual language model fine-tuning. 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). pp. 5706-5711 (2019).

 doi:10.20944/preprints202408.1639.v1

25. Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., ... Stoyanov, V. (2019). Unsupervised cross-lingual representation learning at scale. arXiv preprint arXiv:1911.02116.
26. Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F. L., ... McGrew, B. (2023). Gpt-4 technical report. arXiv preprint arXiv:2303.08774.
27. Liu Zhuang, Lin Wayne, Shi Ya, and Zhao Jun. 2021. A Robustly Optimized BERT Pre-training Approach with Post-training. In Proceedings of the 20th Chinese National Conference on Computational Linguistics, pages 1218–1227, Huhhot, China. Chinese Information Processing Society of China
28. Floridi, L., Chiriatti, M. (2020). GPT-3: Its nature, scope, limits, and consequences. Minds and Machines, 30, 681-694.
29. Vaswani, A. (2017). Attention is all you need. arXiv preprint arXiv:1706.03762.
30. Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., ... Scialom, T. (2023). Llama 2: Open foundation and fine-tuned chat models. arXiv preprint arXiv:2307.09288.
31. Dubey, A., Jauhri, A., Pandey, A., Kadian, A., Al-Dahle, A., Letman, A., ... Ganapathy, R. (2024). The llama 3 herd of models. arXiv preprint arXiv:2407.21783.
32. 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.
33. Kocoń, J., Miłkowski, P., Kanclerz, K. (2021, June). Multiemo: Multilingual, multilevel, multidomain sentiment analysis corpus of consumer reviews. In International Conference on Computational Science (pp. 297-312). Cham: Springer International Publishing.