Large Language Models for Cross-lingual Emotion Detection

Anonymous ACL submission

Abstract

This paper presents a detailed system description of our entry for the WASSA 2024 Task 2, focused on cross-lingual emotion detection. We utilized a combination of large language models (LLMs) and their ensembles to effectively understand and categorize emotions across different languages. Our approach not only outperformed other submissions with a large margin, but also demonstrated the strength of integrating multiple models to enhance performance. Additionally, We conducted a thorough comparison of the benefits and limitations of each model used. An error analysis is included along with suggested areas for future improvement. This paper aims to offer a clear and comprehensive understanding of advanced techniques in emotion detection, making it accessible even to those new to the field.

1 Introduction

Emotion detection in texts across different languages is a challenging yet crucial task, especially in the context of global digital communication. The ability to accurately identify emotions in text, regardless of the language, can significantly enhance interactions in various domains such as customer service, social media monitoring, and mental health assessments. This paper introduces our approach to cross-lingual emotion detection, which was recently recognized as the top submission in WASSA 2024 Task 2(Maladry et al., 2024). Our system leveraged the capabilities of several open source and proprietary Large Language Models (LLMs), including GPT-4(OpenAI, 2024) and Claude-Opus(Anthropic, 2024) in a zero-shot configuration, as well as LLAMA-3-8B(Touvron et al., 2023), Gemma-7B(GemmaTeam, 2024), and Mistral-v2-7B(Jiang et al., 2023), which were fine-tuned. To assess the robustness and efficiency of these models, we conducted tests in both 4-bit and 16-bit precision. This varied precision testing helps in understanding the trade-offs between computational efficiency and model performance. Additionally, we compared the performance of our models against the best submission(Patkar et al., 2023) on a similar monolingual task from the previous years' shared tasks. Furthermore, we experimented with enhancing model performance by incorporating additional training data from previous editions of the shared task, specifically WASSA 2023(Barriere et al., 2023) and WASSA 2022(Barriere et al., 2022) EMO task datasets.

2 Dataset

The dataset consisted of texts belonging to one the 5 languages - Dutch, English, French, Russian and Spanish annotated as one of the 6 classes - Anger, Fear, Love, Joy, Neutral and Sadness. The distribution of languages and each class in each of the datasets can be seen in Table 1 and Table 2.

Class ↓	Train	Dev	Test
Anger	1028	129	614
Fear	143	14	77
Joy	1293	102	433
Love	579	40	190
Neutral	1397	157	916
Sadness	560	58	270
Total	5000	500	2500

Table 1: Class distribution in each dataset split

Class ↓	Train	Dev	Test
English	5000	100	500
French	-	100	500
Dutch	-	100	500
Russian	-	100	500
Spanish	-	100	500
Total	5000	500	2500

Table 2: Language distribution in each dataset split

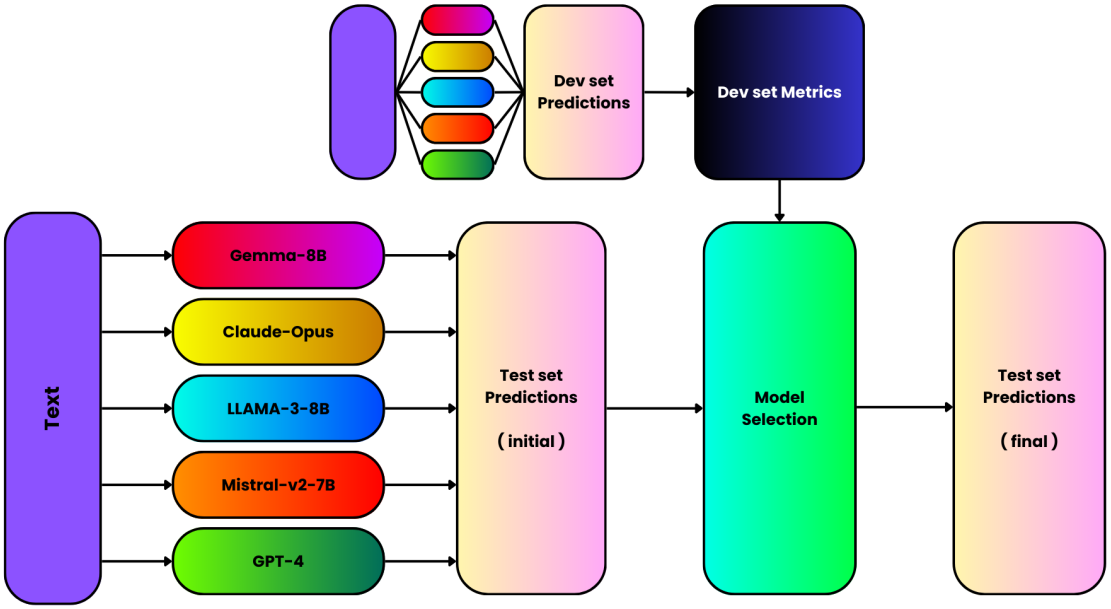

Figure 1: System Overview : Ensembles of LLMs

3 System Description

The non-proprietary LLMs were fine-tuned over just the training dataset over 5 epochs with a learning rate of 0.0002 and weight decay of 0.01. The proprietary systems were tested with various prompt over the development set and the best performing prompt was used to make predictions over the test set. Additionally the previous year's benchmark was also tested alongside by replacing RoBERTa(Liu et al., 2019) with Xlm-RoBERTa(Conneau et al., 2020). Additionally other ensembles like majority vote, model selection based on features were also were also tested. The Code and Models are available over the GitHub repository[1] and Huggingface[2] [3] [4]. Additional information is available here[5]. The primary metric was weighted F1 score, additionally Precision and Recall have also been observed.

3.1 Results Comparison

The results using each of the models on the development set by fine-tuning over 3 epochs on the training set can be seen in Table 3. Other approaches like data augmentation using previous years' emotion detection datasets of the current classes, translating dev and test sets to English before making predictions did not improve the metrics. No preprocessing steps have been used. The metrics on the Test set can be seen in Table 4.

Model ↓	Description	F1
GPT-4	Zero-shot	**0.5616**
Claude-Opus	Zero-shot	0.5581
LLaMa-3-8B	Fine-tuned 3 epochs	0.5474
Mistral-v2-7B	Fine-tuned 3 epochs	0.5466
Gemma-8B	Fine-tuned 3 epochs	0.5300
Xlm-R	10 epochs + SWA	0.5392

Table 3: Performance of each model on Dev set

3.2 Error Analysis

Each of the models had its own advantages and drawbacks likely due to the differences in the training data used by each of the models. The performance of each of the models was observed separately on each of the languages over the development set, this can be seen in Table 5. It can be seen that certain models performed better on some of the languages. This led to the conclusion that selecting appropriate model based on language of the text to be classified might yield better results.

[1]Code Used : https://github.com/1024-m/ACL-2024-WASSA-EXALT

[2]The finetuned LLAMA Model : https://huggingface.co/1024m/EXALT-1A-LLAMA3-5A-16bit

[3]The fine-tuned Mistral Model : https://huggingface.co/1024m/EXALT-1A-MISTRAL-5A-16bit

[4]The finetuned GEMMA Model : https://huggingface.co/1024m/EXALT-1A-GEMMA-5A-16bit

[5]More Info : https://rkadiyala.com/papers

Model ↓	Description	F1 score
llama-3-8b	fine-tuned , 5 epochs	0.5931
llama-3-8b	fine-tuned , test data translated , 5 epochs	0.5701
gemma-8b	fine-tuned , 5 epochs	0.5450
mistral-v2-7b	fine-tuned , 5 epochs	0.5915
gpt-4	few-shot : one sample of each class	0.5918
claude-opus	zero-shot	0.5257
ensemble	model selection based on weighted-f1 scores , 5 epochs each	0.5810
ensemble	model selection based on macro-f1 scores , 5 epochs each	0.5977
ensemble	model selection based on micro-f1 scores , 5 epochs each	0.5725
ensemble	majority vote or model selection based on macro-f1 , 5 epochs each	**0.6295**

Table 4: Performance of each models / approaches on Test set

Language	Metric	GPT-4	GEMMA	Claude-Opus	Mistral-v2	LLAMA-3
English	Micro F1	0.610	0.650	0.580	**0.680**	0.610
English	Macro F1	0.443	**0.594**	0.470	0.590	0.481
English	Weighted F1	0.582	0.655	0.563	**0.671**	0.587
Russian	Micro F1	**0.620**	0.550	0.570	0.590	0.610
Russian	Macro F1	**0.506**	0.425	0.454	0.434	0.457
Russian	Weighted F1	**0.633**	0.574	0.584	0.603	0.627
Spanish	Micro F1	0.670	0.700	0.630	0.740	**0.770**
Spanish	Macro F1	0.521	0.552	0.597	0.659	**0.687**
Spanish	Weighted F1	0.676	0.725	0.666	0.751	**0.779**
French	Micro F1	0.590	0.610	0.610	**0.630**	**0.630**
French	Macro F1	0.509	0.533	0.499	**0.549**	0.522
French	Weighted F1	0.579	**0.607**	0.596	0.596	0.589
Dutch	Micro F1	**0.670**	0.550	0.650	0.620	0.660
Dutch	Macro F1	**0.540**	0.394	**0.540**	0.413	0.533
Dutch	Weighted F1	**0.657**	0.576	0.636	0.610	0.642

Table 5: Performance of each model on Dev set : by each Language and Metric

3.3 Our System

Each of the model's predictions whenever are not among the target labels were retried with a paraphrased prompt till the output matches one of the target labels for classification as seen in Figure 2. Several approaches of using ensembles based on majority voting, model selection based on macro F1, micro F1 and the weighted F1 scores were tested. The best performing system uses a majority voting criteria from the 5 models used. In cases where consensus is not achieved i.e no clear majority, the output of the model with highest weighted F1 score was chosen as the final label.

3.4 Possible Extensions

As seen in Table 5, each of the models had their own advantages and disadvantages with varying performances on each language. It is likely that adding more models and features into model selection like text length or using different models for binary classification tasks as to whether if a text belongs to certain class. This can be seen in Table 6 displaying varying effectiveness of each model in predicting each emotion. A viable approach would be predicting each emotion as a binary task and then using other methods in cases where none or more than one class ends up as true. The fine-tuned LLMs were loaded in 4bit precision and later fine tuned using LoRA(Hu et al., 2021) and tested in both 4bit precision and 16bit precision versions. The drop in performance in 4bit overall was minimal, However in many cases the predictions in 4bit ended up as correct while 16bit were incorrect. Another viable approach is to pick the top 2 likely class labels for each of the texts' predictions and using other methods to classify more effectively.

Class ↓	GPT-4	GEMMA-8B	Claude-Opus	Mistral-v2-7B	LLAMA-3-8B
Anger	**0.75**	0.69	0.71	0.72	0.74
Fear	0.26	0.27	**0.42**	0.30	0.40
Joy	0.62	0.59	0.61	**0.67**	0.65
Love	0.40	0.44	0.33	**0.46**	0.34
Neutral	0.69	0.73	0.67	0.74	**0.75**
Sadness	0.42	**0.47**	0.45	0.39	0.40

Table 6: Performance of each model class wise : class F1 scores on Dev set

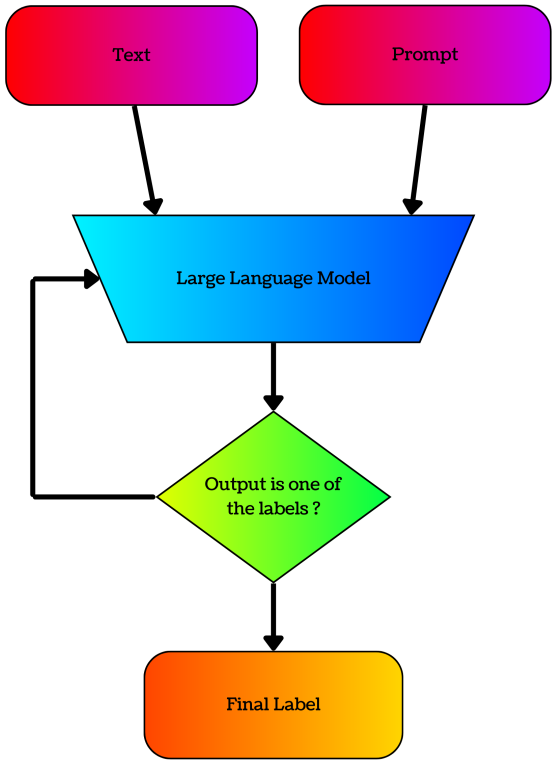

Figure 2: System Overview : Each LLM's predictions

4 Conclusion

It can be seen from Table 4 that ensemble models have achieved a significantly better result over direct approaches. This was 3 percent better that rest of the participants. However not all approaches have been tested due to limit on number of submissions. As seen in Table 5, It can also be observed that from Table 6 that a similar trend was observed in using different models for each emotion detection too might aid in improving the performance further. As seen in Table 4 and Table 3, further training is likely to improve the results as the dev set results of fine-tuned models were lower that the proprietary models when trained on 3 epochs, but when the same models were further tuned over 2 more epochs, they performed better than proprietary models. Most of the errors when using proprietary models were with the neutral class texts being classified incorrectly or other classes being classified as being neutral. While the fine-tuned models were able to learn to be able to distinguish texts as neutral or some other class in a better way as seen in Table 6. The classes with lesser data samples as shown in Table 1 had significantly worse compared to other classes as seen in Table 6. Techniques like Stochastic weight averaging (SWA)(Izmailov et al., 2019) in this case only led to a minor improvement and techniques like augmentation using other datasets did not improve performance. It is likely that adding sufficient data for all classes can make the current proposed system better as enough correlation can be seen in training data amount from Table 1 and average performance of the discussed models on each of the classes from Table 6. The current proposed approach can be extended to other languages by testing performance on a small sample of that language to decide the extent of reliability of each model in making predictions over texts of that language. In case of using proprietary models the same prompt used for all texts, it is worth testing different prompts for texts of each language due to varying features of each language where one class might to higher number of false positives than other in a different language. Approaches like removal of stop words did not improve the performance. While using ensembles, texts completely in one language performed better that the texts where a portion of the text is in English and rest in a different language. These texts led to higher frequency of errors. The performance of proprietary models was a bit better on these kind of texts compared to the rest of the models tested probably due to larger model size and more code-mixed data in training. Other information like the specific prompts used on each of the LLMs, Prompt format for the fine-tuned LLMs used and other relevant plots are available in Appendix A.

Limitations

Due to computational resource limitations, the models used (non-proprietary) were loaded in 4bit precision before being fine-tuned. It is likely that with higher precision usage of the models can yield better results. The models used (non-proprietary) were of the 7B or 8B variants. It is likely that higher variants may yield better results. The approaches might not be extendable to all languages as not all languages' data were covered in the training data of the LLMs used in the current proposed system. The current set of classes (6) are different from other set of classes used in other similar benchmarks, where one of the other classes can be close or similar to one of the current classes and has a possibility of skewing the results to a small extent if included.

References

Anthropic. 2024. Claude-opus technical report.

Valentin Barriere, João Sedoc, Shabnam Tafreshi, and Salvatore Giorgi. 2023. Findings of WASSA 2023 shared task on empathy, emotion and personality detection in conversation and reactions to news articles. In *Proceedings of the 13th Workshop on Computational Approaches to Subjectivity, Sentiment, & Social Media Analysis*, pages 511–525, Toronto, Canada. Association for Computational Linguistics.

Valentin Barriere, Shabnam Tafreshi, João Sedoc, and Sawsan Alqahtani. 2022. WASSA 2022 shared task: Predicting empathy, emotion and personality in reaction to news stories. In *Proceedings of the 12th Workshop on Computational Approaches to Subjectivity, Sentiment & Social Media Analysis*, pages 214–227, Dublin, Ireland. Association for Computational Linguistics.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale.

GemmaTeam. 2024. Gemma: Open models based on gemini research and technology.

Edward J. Hu, 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.

Pavel Izmailov, Dmitrii Podoprikhin, Timur Garipov, Dmitry Vetrov, and Andrew Gordon Wilson. 2019. Averaging weights leads to wider optima and better generalization.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach.

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*.

OpenAI. 2024. Gpt-4 technical report.

Aditya Patkar, Suraj Chandrashekhar, and Ram Mohan Rao Kadiyala. 2023. AdityaPatkar at WASSA 2023 empathy, emotion, and personality shared task: RoBERTa-based emotion classification of essays, improving performance on imbalanced data. In *Proceedings of the 13th Workshop on Computational Approaches to Subjectivity, Sentiment, & Social Media Analysis*, Toronto, Canada. Association for Computational Linguistics.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. Llama: Open and efficient foundation language models.

A Appendix