Text Based Cross-Lingual Emotion Detection

1st Semester of 2024-2025

Cristian Daniel Păduraru
cristian.paduraru@s.unibuc.ro

Abstract

Many tasks from the field of Natural Language Processing have gathered a lot of popularity in recent years, mainly thanks to the rise of pre-trained Large Language Models (LLMs). One such task is Emotion Detection for languages with no labeled train set available. This project focuses on the use of Cross Lingual LLMs to solve this task, analysing the generalization capabilities of the detectors to languages with varying degrees of similarity with the ones available at train time.

1 Introduction

Emotion Detection is one of the many tasks within Natural Language Processing which have seen a rapid increase in popularity in the past few years (del Arco et al., 2024). This is tightly correlated with the advent of Large Language Models, which are trained on vast amounts of textual data and thus become capable of understanding context and linguistic cues. These are properties that have made them popular for the emotion recognition tasks (Acheampong et al., 2021), where various hints must be identified in the texts (e.g. punctuation, choice of words, phrase structre).

A more challenging setup for this task is the Cross-Lingual Emotion Detection, where annotated texts are provided in certain number of languages and the goal is to obtain a detector for new, distinct languages. In this context, previous works have tried different approaches, including the use of Cross-Lingual LLMs and data augmentations (Zhang et al., 2024), translating the annotated texts into the target languages (Hassan et al., 2022) and even zero-shot prediction using some of the most recent and largest LLMs available (Kadiyala, 2024).

In this work I chose to use Cross-Lingual LLMs and focused on analysing how the similarity of target languages with the ones available at train-time correlates with the performance of the detector. Through empirical experiments I have found that the detectors tend to generalize better to languages from the same family as at least one language used in the training phase. I also observed that depending on the target language certain emotions tend to be better recognized, even if the target language is not related with those in training set. Regarding the cross-lingual representations of the chosen LLM, I have observed that the generalization capabilities of the detectors tend to worsen when the text embeddings form language specific clusters.

2 Task and Data descriptions

The task tackled in this project is the one from SemEval 2025 Task 11[1], Track C. We have access to texts written in languages $(l_i)_{1\leq i\leq n}$ (referred to as *source languages* in the following sections) which are labeled for emotion detection and the goal is to obtain a detector for a different language l_{n+1} (referred to as *target language*). The organizers offered a large variety of source and target languages to choose from for this task. I selected English (eng), German (deu) and Spanish (esp) as my source languages and Romanian (ron), Ukrainian (ukr) and Hindi (hin) as target languages. I chose this setup with only European source languages to observe how well a detector can generalize to other European languages, whether they are from the same language family with at least one source language or not, and to a more distant Indo-European language.

The emotions to be detected are the 6 basic ones from the Ekman's model (Ekman and Friesen, 1981), namely *anger*, *disgust*, *fear*, *joy*, *sadness* and *surprise*. The source languages have between 2000 and 2800 annotated samples each, while the proportion of positive labels varies from 6 to 33%

[1] https://github.com/emotion-analysis-project/SemEval2025-task11

depending on language and emotion (with the sole exception of 58% for *fear* for the English data).

3 Approach

To make use of the available labeled texts I have decided to use multilingual Large Language Models that cover all languages $(l_i)_{1 \leq i \leq n+1}$ and map texts to a shared embedding space, irrespective of the language they are written in. The main assumption would be that, as long as similar texts written in different languages have similar representations, a classifier trained on the available labeled data should generalize to the new language, unseen at train time. In all the experiments presented in this section we use a 85%/15% train/validation split from the original train data and save the classifier weights that have the highest F_1 score on the validation set.

The first LLM tested was a LEALLA-large (Mao and Nakagawa, 2023) sentence encoder, as it is trained to produce the same embedding for sentence pairs from different languages. Using the frozen, pre-trained representations I trained linear classifiers independently for each emotion with the AdamW (Loshchilov and Hutter, 2019) optimizer, a learning rate and weight decay of 10^{-3}, a batch size of 512 and the binary cross entopy loss, for up to 50 epochs. For each emotion I trained three classifiers with different seeds and selected the one with the highest macro F_1 score on the validation set at the end. To address the label imbalance, I computed the individual losses for samples in a given batch, then averaged the losses of samples with the same label and finally averaged these two losses. This gives equal weight to both positive and negative classes, it is not more computationally expensive than using class weights and simpler to implement than balanced sampling schemes. The results of these classifiers on the dev set of the unseen languages are presented in Table 1. One thing to mention is that the English texts were not annotated for the *disgust* label, so whenever English was the only language used for training I always set the value for *disgust* as negative on all samples from the target language.

As the performance of these classifiers was rather low for both the source and target languages, I assumed that the representations did not encode enough information from the input, as the encoder is only trained with the sentence retrieval task. I thus decided to fine-tune the network, but only partially due to the low number of training samples. I selected the last 4 encoder blocks, the pooler layer and the final linear classification layers for each emotion. The classification layers were initialized with the weights from the previous experiment, resulting in an instance of LP-FT (Kumar et al., 2022). The parameters were trained for 2.000 steps with a batch size of 96, learning rate and weight decay of 10^{-4}, and validated once every 50 steps. This time I used a custom sampling scheme, always retrieving a sample for a randomly selected emotion and label (positive or negative) from the dataset. Gradient clipping was necessary to ensure converge. Table 2 contains the dev set results for this experiment.

I also tried using a much larger and more recent LLM to extract deep representations, namely the QWEN2.5 7B (Yang et al., 2025) model, which covered all the languages considered in this project. As this is a decoder only architecture I used the embedding of the last token in the sequence (truncated to a max length of 256 tokens) as the sequence embedding and I trained linear classifiers with the same setup as before. The embeddings were L_2-normalized in this case. To cross check for optimization issues, I also used the logistic regression implementation from sklearn, with the *lbfgs* and *linear* solvers. The first solver usually failed to converge, whereas the second one converged to a solution similar in performance (on the validation set) with the one from the *lbfgs* solver, but in a much longer time. Surprisingly, the results turned out worse that the linear probe on LEALLA-large embeddings. Assuming this was caused by overfitting, as the embeddings' dimension was 14 times larger, I tried to first map them to a lower-dimensional space using Principal Component Analysis and then train the linear classifiers on these representations. Unfortunately, the results presented Table 4 do not seem to support this hypothesis. The validation accuracy is decreasing for the source languages, meaning that relevant information is lost in the dimensionality reduction step. These results rather suggest that the representations are not suited for the current task.

The code is publicly available on github[2]. The main libraries used were PyTorch[3], Transformers[4] and Unsloth[5]. In order to get access to the task's

[2] https://github.com/PaduraruCristian/MachineTranslation

[3] https://pytorch.org/

[4] https://huggingface.co/docs/transformers/v4.17.0/en/index

[5] https://docs.unsloth.ai/

Source languages	Val acc.	Target language		
		ron	ukr	hin
eng	53.49	41.87	18.12	27.16
deu	46.37	45.89	20.89	26.56
esp	59.23	42.89	**26.50**	**44.43**
eng, deu, esp	52.94	**46.97**	23.58	32.68

Table 1: Results on the dev set of target languages for the linear classifiers trained on embeddings from the LEALLA-large model.

Source languages	Val acc.	Target language		
		ron	ukr	hin
eng, deu, esp	70.93	53.72	36.32	53.17

Table 2: Results of the finetuned LEALLA-large model on the dev set of target languages. The model checkpoint is selected based on the validation macro F_1 score.

data for research purpose one would have to make a request to the organizers of the competition. The total runtime for all the described experiments is estimed at less than 7 hours on Google Colab[6] with a T4 GPU, the most time consuming part being the extraction of embeddings with the QWEN2.5 (Yang et al., 2025) model (approximately half the total time).

4 Discussion

Distribution Shift In Figure 1 we can see how the macro F_1 score on the validation set is not necessarily well correlated with the score on the target languages. This is an expected behavior, as the task itself is an instance of out-of-distribution (OOD) generalization and model selection in the absence of an annotated validation set for the target distribution is known to be a hard problem (Gulrajani and Lopez-Paz, 2020).

Source languages	Val acc.	Target language		
		ron	ukr	hin
eng	46.01	36.69	14.53	19.35
deu	40.44	43.94	15.82	22.60
esp	55.58	44.10	**16.79**	21.39
eng, deu, esp	52.03	**45.21**	16.67	**23.46**
eng*	41.04	37.26	**16.85**	16.78
deu*	37.85	31.48	11.69	19.08
esp*	52.66	28.81	15.52	18.90
eng, deu, esp*	50.80	**37.35**	15.16	**21.43**

Table 3: Results on the dev set of target languages for the linear classifiers trained on embeddings from the QWEN2.5 model. The mark * indicates results for the sklearn implementation of logistic regression.

#Principal components	Val acc.	Target language		
		ron	ukr	hin
64	44.05	27.38	15.53	**23.93**
128	45.17	29.09	16.17	20.96
256	45.97	34.15	**17.64**	19.94
3584	52.03	**45.21**	16.67	23.46

Table 4: Results on the dev set of target languages for the linear classifiers trained on embeddings from the QWEN2.5 model (from all source languages) after dimensionality reduction with PCA.

To investigate this distribution shift in the embedding space I have used t-SNE (van der Maaten and Hinton, 2008) to map the embeddings to a 2-dimensional space and plotted them in figure 3 per language and LLM used. It can be seen that the embeddings produced by the QWEN model tend to make more language-specific clusters than the embeddings obtained from the LEALLA encoder, which can explain the stronger degradation of performance on the target languages.

The distribution shift between the source languages and the targets ones can also be observed in the topics that the texts cover. I manually looked at the samples from the Romanian and English datasets and concluded that the texts were most likely retrieved from news websites and social media. The Romanian texts were mostly about political debates or the recent COVID-19 pandemic, while the English ones where a lot more diverse, covering stories and experiences of people. This distinction further justifies the loss of performance when extrapolating the learned weights to the other target languages.

Per-emotion performance For a more detailed analysis of the detectors, I individually plotted in Figure 2 the F_1 score for each target language and emotion. For each language I observed a trend where certain emotions are better recognized than the others, irrespective of the source languages used for training (e.g. *joy* for Romanian, *fear* for Ukrainian and Hindi), while other emotions are always poorly detected, with respect to the other emotions.

More data is not always better Looking at the linear probe results we can see that using data from

[6] https://colab.research.google.com/

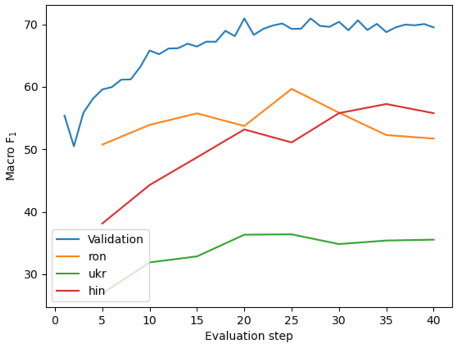

Figure 1: Macro F_1 scores on the validation set of source languages and the dev set of target languages when fine-tuning the LEALLA-large model on all the source languages.

a single language sometimes leads to better results than using all of the source languages. In generalizations setups, having access to a varied pool of distributions at train time usually leads to better performance that the case where a single distribution is available. This also involves the use of more specialized algorithms (Sagawa et al., 2020; Yao et al., 2022; Arjovsky et al., 2020) that use this additional information to obtain models which are robust to the shifts between the given distributions.

5 Limitations

The Cross-Lingual LLM is the centerpiece of the presented approach and also it's main limitation, as the structure of it's embedding space (how well the representations of texts from different languages are aligned) directly affects the generalization capabilities of the final detector, as shown by the experiments with the two distinct models. This poses a more serious problem for the low-resource languages, as they are not well represented in the pre-training data of LLMs and neither do they have large annotated datasets for the given tasks.

Based on the presented experiments, fine-tuning the LLM on the available data seems to be mandatory for greatly improving results, as the pre-trained representations are not predictive enough for all the tasks. In addition, lower level features (e.g. punctuation, typos, emojis) are oftentimes most relevant for the emotion detection task. This would thus imply that the entire LLM must be fine-tuned, which further requires more resources, both in terms of GPU-time and volume of data in order to prevent overfitting.

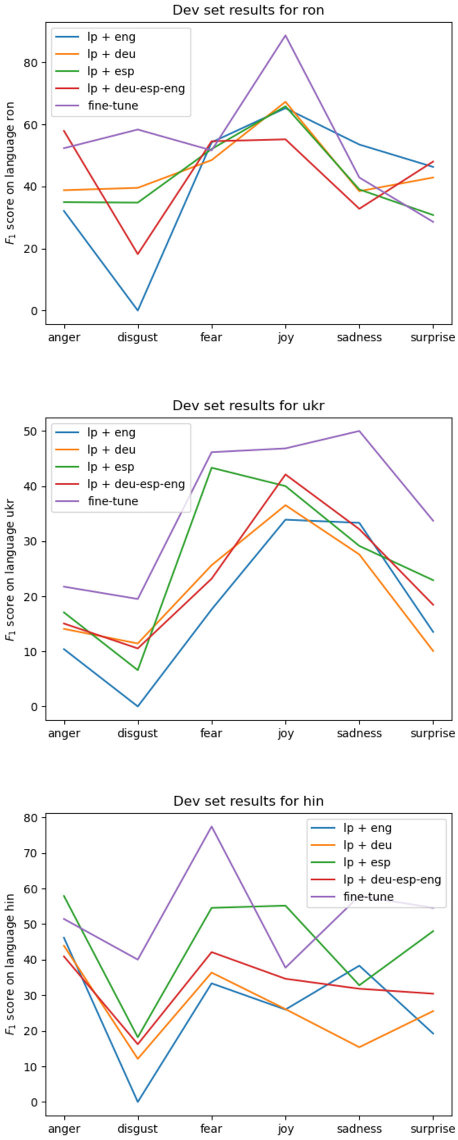

Figure 2: Per emotion F_1 score for each of the target languages on the dev set for linear probes (lp) on LEALLA embeddings and the finetuning of the encoder.

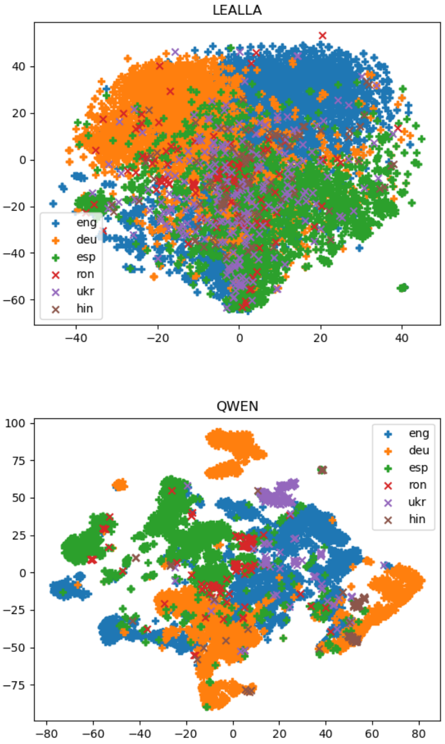

Figure 3: Scatter plot of embeddings extracted by LEALLA (top) and QWEN (bottom) based on the language of the texts. The embeddings were reduced to a 2-dimensional space using t-SNE.

6 Conclusions and Future Work

In this project I have explored the use of Cross-Lingual LLMs for the task of emotion detection under strong distributional shifts. We have seen that the generalization capabilities of the approach is heavily impacted by the similarity of the source and target languages - better results were obtained on texts written in Romanian which, together with Spanish, is part of the Romance language family, while generalization to more distance languages (Ukrainian and Hindi) has proven more difficult.

While the task and experiments were interesting and allowed me to explore properties of pre-trained representations, there is still a lot more to be investigated. Different approaches could be taken, such as translating texts from the source languages into the target one in order to escape the requirement of having a Cross-Lingual LLM, or using cross-lingual data augmentations and specialized algorithms to train detectors which are invariant to changes in the language of texts. The current project could also be improved by covering a larger number of target languages and looking for trends based on the studied language families and architectures.

References

Francisca Adoma Acheampong, Henry Nunoo-Mensah, and Wenyu Chen. 2021. Transformer models for text-based emotion detection: a review of BERT-based approaches. *Artificial Intelligence Review*, 54:5789 – 5829.

Martin Arjovsky, Léon Bottou, Ishaan Gulrajani, and David Lopez-Paz. 2020. Invariant risk minimization.

Flor Miriam Plaza del Arco, Alba Curry, Amanda Cercas Curry, and Dirk Hovy. 2024. Emotion Analysis in NLP: Trends, Gaps and Roadmap for Future Directions.

Paul Ekman and Wallace V. Friesen. 1981. *The Repertoire of Nonverbal Behavior: Categories, Origins, Usage, and Coding*, pages 57–106. De Gruyter Mouton, Berlin, Boston.

Ishaan Gulrajani and David Lopez-Paz. 2020. In Search of Lost Domain Generalization.

Sabit Hassan, Shaden Shaar, and Kareem Darwish. 2022. Cross-lingual Emotion Detection. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 6948–6958, Marseille, France. European Language Resources Association.

Ram Mohan Rao Kadiyala. 2024. Cross-lingual Emotion Detection through Large Language Models. In *Proceedings of the 14th Workshop on Computational*

Approaches to Subjectivity, Sentiment, & Social Media Analysis, pages 464–469, Bangkok, Thailand. Association for Computational Linguistics.

Ananya Kumar, Aditi Raghunathan, Robbie Jones, Tengyu Ma, and Percy Liang. 2022. Fine-tuning can distort pretrained features and underperform out-of-distribution.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled Weight Decay Regularization.

Zhuoyuan Mao and Tetsuji Nakagawa. 2023. LEALLA: Learning lightweight language-agnostic sentence embeddings with knowledge distillation. *arXiv preprint arXiv:2302.08387*.

Shiori Sagawa, Pang Wei Koh, Tatsunori B. Hashimoto, and Percy Liang. 2020. Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization.

Laurens van der Maaten and Geoffrey E. Hinton. 2008. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9:2579–2605.

Qwen: An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tianyi Tang, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. 2025. Qwen2.5 Technical Report.

Huaxiu Yao, Yu Wang, Sai Li, Linjun Zhang, Weixin Liang, James Zou, and Chelsea Finn. 2022. Improving out-of-distribution robustness via selective augmentation.

Jinghui Zhang, Yuan Zhao, Siqin Zhang, Ruijing Zhao, and Siyu Bao. 2024. Enhancing Cross-Lingual Emotion Detection with Data Augmentation and Token-Label Mapping. In *Proceedings of the 14th Workshop on Computational Approaches to Subjectivity, Sentiment, & Social Media Analysis*, pages 528–533, Bangkok, Thailand. Association for Computational Linguistics.