Team UBD at SemEval-2025 Task 11: Balancing Class and Task Importance for Emotion Detection

Cristian Daniel Păduraru
University of Bucharest, Romania
cristian.paduraru@s.unibuc.ro

Abstract

This article presents the systems used by Team UBD in Task 11 of SemEval-2025. We participated in all three sub-tasks, namely Emotion Detection, Emotion Intensity Estimation and Cross-Lingual Emotion Detection. In our solutions we make use of publicly available Language Models (LMs) already fine-tuned for the Emotion Detection task, as well as open-sourced models for Neural Machine Translation (NMT). We robustly adapt the existing LMs to the new data distribution, balance the importance of all emotions and classes and also use a custom sampling scheme. We present fine-grained results in all sub-tasks and analyze multiple possible sources for errors for the Cross-Lingual Emotion Detection sub-task.

1 Introduction

In this project, we address the three sub-tasks (tracks) of Emotion Detection (ED), Emotion Intensity Estimation (EIE), and Cross-Lingual Emotion Detection (CL-ED) from SemEval-2025 Task11 (Muhammad et al., 2025b). While the organizers provided a dataset with samples from 28 different languages (Muhammad et al., 2025a), we only focused on English, German and Spanish for the first two sub-tasks and Romanian, Portuguese, Ukrainian, Russian, Hindi and Indonesian in the last one.

Our solutions mainly rely on language-specific encoders that have already been fine-tuned for the emotion detection task and robustly adapt them to the new data distribution. To bridge the gap between languages in the last task we use an open-source NMT system to translate the test sets of other languages, and also experiment with cross-lingual LMs (XLMs).

We find that the emotion-specific performance of our systems correlates well with the frequency of positive examples in the first two sub-tasks. This indicates that data scarcity can still be a problem, even when it is addressed through common means. For the CL-ED sub-task we find multiple factors that lead to degraded performance on new languages, some of which are system-specific, while others are common to both of them. Notably, the relatedness of languages does not correlate well with cross-lingual performance in our experiments.

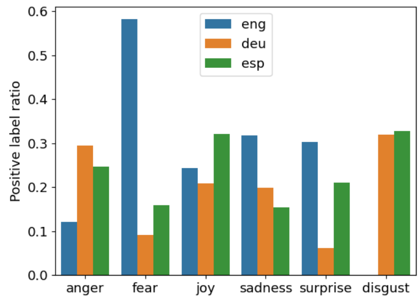

Figure 1: The ratio of positive labels in the train set of the three languages addressed in the ED sub-task.

The implementation of our solutions will be published on github[1].

2 Background

Related Work The task of Emotion Detection has found diverse applications (del Arco et al., 2024), such as analyzing social interactions, monitoring mental well-being (Chiruzzo et al., 2024; Paduraru and Anghelina, 2024), highlighting mental health concerns or understanding people's emotions during stressful events (Sosea et al., 2022). While earlier works have tried to use a mix of low-level features and more abstract ones, obtained from Deep Neural Networks (Khanpour and Caragea, 2018), the Transformers architecture (Vaswani et al., 2017) has become the default architecture for this task (Acheampong et al., 2021) in recent years, with people using even specialized

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

Proceedings of the The 19th International Workshop on Semantic Evaluation (SemEval-2025), pages 874–884
July 31 - August 1, 2025

pre-training objectives (Sosea and Caragea, 2021) to improve results in this downstream task.

For the Cross-Lingual variant of the task, multiple solutions have been proposed. Some notable ones are: using multilingual encoders and training detectors on language agnostic representations of the texts (Alejo et al., 2020; Zhang et al., 2024; Hassan et al., 2022), distilling monolingual detectors into cross-lingual models (Wang et al., 2024), translating texts into a language where annotated training data is available (Alejo et al., 2020; Hassan et al., 2022) and even using Large Language Models as zero-shot detectors (Kadiyala, 2024).

Dataset We work on the dataset provided by Muhammad et al. (2025a), which contains texts from 28 different languages, annotated with the 6 emotions from Ekman's model (Ekman and Friesen, 1981).

3 System Overview

In our solutions we use language-specific encoders, based on the Transformer (Vaswani et al., 2017) architecture, that have already been fine-tuned for the emotion detection task, in order to extract deep representations of the textual samples. We then add a linear classification layer to the encoders and train them with dynamic weights to balance the importance of each class and emotion.

3.1 Track A: ED

Linear Probing By keeping the encoders frozen we can individually train the classifiers for each emotion, as their optimization is completely independent from one another. To address the imbalance between positive and negative classes we computed the individual loss of each sample, averaged the losses of samples from the same class, and finally averaged the losses for the positive and negative classes. After training a linear classifier we also adjust its detection thresholds by iterating through a range of values and selecting the one that leads to the highest dev set F_1 score.

Fine-tuning In order to robustly fine-tune the encoders, we follow Kumar et al. (2022) and initialize the linear classification layer with the one previously trained on the frozen encoder embeddings. We then jointly update both this layer and the encoder's parameters, balancing the classes in the same manner as before. As the detectors for all emotions are simultaneously trained in this case, we further balance the importance of each one by averaging the emotion specific losses.

Besides this balancing, we also implemented a custom sampling scheme to make it unlikely for a batch to have no positive examples for an emotion. At each step, we uniformly select a random emotion and label and then retrieve a sample from the dataset with the selected label on that emotion.

The fine-tuning process does not ensure that the classification layers are optimal for each emotion with respect to the current encoder parameters. We thus decided to keep the fine-tuned encoder and remake the classification layers for each emotion individually, with the same procedure previously presented. We provide in the Appendix (Tab. 5) the F_1 scores on the dev set for the fine-tuned detectors and the second linear probes for comparison.

3.2 Track B: EIE

As the intensity levels for the emotions were discrete, we modeled this sub-task as a multi-label classification problem. We trained linear probes on text embeddings from the same encoders used in Track A (the frozen ones and their fine-tuned variants) with the class-balanced loss described in Sec. 3.1.

3.3 Track C: CL-ED

For this sub-task, we chose to translate the test sets of multiple languages into Spanish, using an NMT system from the NLLB (NLLB et al., 2022) family of LMs. We then use a classifier trained in Task A for Spanish in order to detect the emotions in these translated texts.

In the development period of the task we have also experimented with classifiers based on cross-lingual LMs. The classifiers were trained using only texts written in the three languages addressed in Track A, and then applied on texts from other languages of this sub-task.

4 Experimental Setup

In all sub-tasks we use only the data provided by the task organizers, without applying data augmentations or any pre-processing before tokenizing the texts. We used the data splits provided by the organizers (train/dev/test), with a single exception in the experiments based on cross-lingual models, where 15% of the train data is used for validation and the dev split is used for testing. All classifiers were trained using the

Language	Emotion						Macro F_1
	anger	fear	joy	sadness	surprise	disgust	
English	48.08	77.26	68.97	68.69	62.35	-	65.07
German	63.79	32.63	63.59	52.54	19.18	64.96	49.44
Spanish	74.15	75.00	74.87	76.36	70.8	79.36	75.09

Table 1: Test set F_1 scores in Track A of the linear probes trained on frozen embeddings.

Language	Emotion						Macro F_1	Official Ranking
	anger	fear	joy	sadness	surprise	disgust		
English	55.51	79.82	68.97*	68.69*	67.06	-	68.01	58/74
German	73.62	37.31	67.47	52.54*	29.76	64.96*	54.27	32/44
Spanish	75.09	75.00*	74.87*	78.74	72.16	81.28	76.19	25/44

Table 2: Test set F_1 scores in Track A, mixed between the linear probes trained on embeddings from the frozen encoders (marked with *) and those on embeddings from the fine-tuned encoders. These are the results of our final submission in Track A.

AdamW (Loshchilov and Hutter, 2019) optimizer implemented in Pytorch (Ansel et al., 2024) and the pre-trained encoders were downloaded from HuggingFace[2]. The following encoders were used in Tracks A&B for each language:
English: SamLowe/roberta-base-go_emotions[3]
German: visegradmedia-emotion/Emotion_RoBERTa_german6_v7[4]
Spanish: pysentimiento/robertuito-emotion-analysis[5] (del Arco et al., 2020; Pérez et al., 2021; Pérez et al., 2022)

4.1 Track A

Linear Probing We used the hidden state of the CLS token from the last transformer block as the sequence representation, ignoring the pooler layer if the encoder happened to have one. The linear probes were trained for 50 epochs with the binary cross entropy loss, a learning rate and weight decay of 1e-3, a batch size of 512, and a cosine annealing learning rate schedule with a minimum learning rate of 1e-5. The linear probes are trained individually for each emotion with three different random seeds and the final weights are selected based on the dev set F_1 score. The detection threshold on the logits is adapted for each emotion by iterating through values in the [-2, 2] interval with a step of 0.1, computing the F_1 score on the dev set at each threshold, and selecting the best one.

Fine-tuning Due to the low number of examples we only adjust the parameters of the last two transformer blocks and the final classification layer. The weights are tuned with the binary cross entropy loss, a learning rate of 1e-5, weight decay of 1e-2, and batch size of 256. To ensure the stability of training, we also clipped the gradients to a maximum global value of 3. The weights are trained for up to 500 steps and evaluated on the dev set every 25 steps (due to the uniform sampling the concept of *epoch* is no longer well-defined).

4.2 Track B

In this sub-task we trained a linear classifier for each emotion with the cross-entropy loss and the same hyper-parameters from Track A's Linear Probing setup. Due to the lower number of samples for higher intensity levels, we increased the batch size to 1024 and the number of epochs to 150, to make up for the reduced number of steps per epoch.

4.3 Track C

We translated the texts into Spanish using the distilled NLLB-1.3B (NLLB et al., 2022) model, always producing up to 200 tokens. English could not be used as a target language for translation because it lacks a classifier for the *disgust* emotion, while for German the results in Track A were worse compared to the other two languages. We didn't apply any post-translation processing on the texts to ensure that the LM did not start hallucinating or went in a loop, repeating the same token at output.

[2] https://huggingface.co/
[3] https://huggingface.co/SamLowe/roberta-base-go_emotions
[4] https://huggingface.co/visegradmedia-emotion/Emotion_RoBERTa_german6_v7
[5] https://huggingface.co/pysentimiento/robertuito-emotion-analysis

Language	Fine-tuned Encoder	Emotion						Avg	Official Ranking
		anger	fear	joy	sadness	surprise	disgust		
English	✘	0.452	0.616	0.654	**0.612**	0.454	-	0.558	-
	✓	**0.584**	**0.648**	**0.692**	0.556	**0.585**	-	**0.613**	-
German	✘	0.427	0.127	0.559	0.512	0.166	0.480	0.378	-
	✓	**0.592**	**0.339**	**0.648**	**0.516**	**0.314**	**0.527**	**0.489**	19/24
Spanish	✘	0.649	0.714	0.649	0.697	0.649	0.691	0.675	-
	✓	**0.679**	**0.721**	**0.706**	**0.745**	**0.672**	**0.712**	**0.706**	13/26

Table 3: Pearson Correlation on the test set of Track B (maximum value is 1). The final submission contained the predictions made with fine-tuned encoders only for the German and Spanish languages (gray background).

Target Language	Emotion						Macro F_1	Official Ranking
	anger	fear	joy	sadness	surprise	disgust		
Spanish	75.09	75.00	74.87	78.74	72.16	81.28	76.19	-
Romanian	40.74	**72.27**	**78.93**	34.29	27.83	**51.15**	50.87	11/13
Portuguese (ptbr)	**60.99**	35.38	52.94	45.67	35.53	12.59	40.52	9/11
Ukrainian	26.46	54.55	41.99	51.11	37.41	20.22	38.63	10/15
Russian	54.07	66.67	49.93	46.51	54.98	48.08	53.37	11/14
Hindi	**60.92**	62.86	57.28	**62.54**	**69.46**	47.51	**60.09**	10/14
Indonesian	29.06	27.42	69.61	40.69	34.34	45.05	41.03	11/15

Table 4: Test set F_1 scores in Track C, obtained by the linear probes trained on fine-tuned embedding for Spanish texts in Track A. The results from Track A on Spanish are also added for comparison. The results on the other six languages correspond to our final submission in Track C.

For the experiments with cross-lingual LMs we have performed linear probing on embeddings extracted with the LEALLA-large (Mao and Nakagawa, 2023) and QWEN2.5-7B (Yang et al., 2025) models and also fine-tuned the LEALLA model. The complete setup is detailed in Appx. A.

5 Results

Track A The results of the linear probes on the test set of the competition are presented in Table 1. We observe that the detectors trained on Spanish texts are the only ones to consistently perform well on all emotions. For texts written in German, the detectors for *fear* and *surprise* lack in performance when compared to the detectors for other emotions. In the case of English texts, the detector for *anger* is the only one that is well below the average performance level, while *fear* is highly above it. This pattern is correlated with the frequency of positive labels in the provided train sets (see Fig. 1) for English and German. For detectors trained on Spanish texts however, we notice that this correlation does not hold anymore. The correlations on English and German data are still maintained even after fine-tuning (see Table 2). We also notice that each emotion attains different levels of improvements in the second linear probing (Table 5 in the Appendix), but this is not correlated with the initial performance of the fine-tuned detectors.

The frequency of positive samples alone is not a good indicator for the final performance. While *joy* and *sadness* have similar frequencies in German data, there is a 10% gap in F_1 score between them in the linear probing scenario (Table 1). Also, the *disgust* label has similar frequency in German and Spanish data, but the difference in F_1 score is close to 15%. We believe that this is where the inherent task difficulty and quality of the encoders used are most likely to make the difference.

For certain emotions, the initial linear probes performed better than those trained on the fine-tuned encoders. Thus, we decided to select for each emotion the test set predictions from the linear probe that had the highest F_1 score on the dev set. The results of this **final submission** are presented in detail in Table 2.

Track B The test set results for the previously described experiments in this sub-task are presented in Table 3. Using the fine-tuned encoder resulted in increased performance in almost every case (the only exception is the *sadness* label for the English data). The largest improvements are in the German

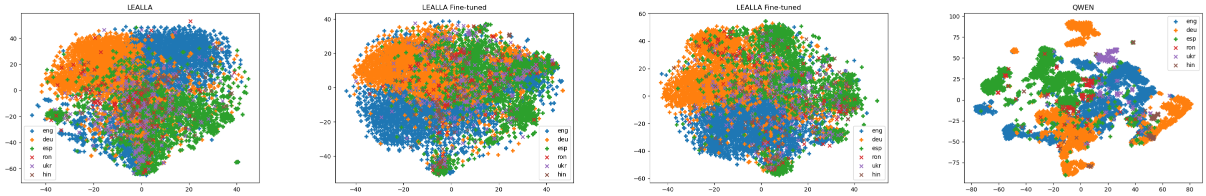

Figure 2: t-SNE visualizations of language specific embeddings extracted with LEALLA-large pre-trained (left), 4 blocks fine-tuned (middle left), 8 blocks fine-tuned (middle right) and QWEN2.5-7B (right) from the train set of Track A for English, German and Spanish, and from the dev set of Track C for Romanian, Ukrainian and Hindi.

data, which also had the worst performance without fine-tuning. The improvement on each emotion is also not uniform - the highest gains are on the emotions that initially had the lowest scores. We consider this to be thanks to the balancing of emotions from the previous sub-task, which helped the encoder attend more to the ones with higher loss, but without disregarding the others, so that their performance did not degrade through fine-tuning. The results of the **final submission** are marked in Table 3 with a gray background.

Track C The results of our system for this Track are presented in Table 4, along with the results for Spanish from Track A, only for comparison. While a decrease in overall performance (Macro F_1) was expected, we note that this decrease is not uniform across the individual emotions and languages. Certain detectors transfer well only to a single language, losing less than 5% in terms of F_1, but most of the times the decrease is well above 10%. Surprisingly, the detector for *joy* achieved better performance on texts translated from Romanian than on the texts from Track A, originally written in Spanish. We also noticed that the drop in performance is not necessarily correlated with the similarity of Spanish and the target languages. For example, the performance on Hindi texts is the highest on average, surpassing the Romance languages considered (Romanian and Portuguese).

In the described framework we have identified multiple sources of errors. The first one is the quality of the translations - we validated that in certain cases the NMT system started repeating a single token multiple times. Nonetheless, as almost all languages have at least one emotion with an F_1 higher than 60% (Ukrainian is the sole exception), we expect this type of errors to be limited. Another possibility is for low-level cues for the perceived emotions (e.g. punctuation) to be lost in the process or for the choice of words to be unusual due to translationese (Rabinovich et al., 2017). We expect these problems to be correlated to the data distribution used for training the NMT model.

A general source of errors is the distributional shift between languages, regarding the topics they covered. We manually determined that texts in Romanian seemed to be mainly scraped from news websites and covered topics like politics and the recent COVID-19 pandemic, whereas the texts in English contained mostly short texts that were likely written on social media websites. These particularities might bias the detectors towards detecting certain topics, not the emotions themselves, resulting in degraded performance in other contexts.

Through our experiments in the dev period with Cross-Lingual models we have noticed that they also suffer a great performance degradation on new languages (consult Appx. A for the results). We provide in Fig. 2 a t-SNE visualization of text embeddings extracted with the LEALLA-large and QWEN2.5-7B models. We observed that the data tends to form language-specific clusters, which we assume to be the main reason for the poor generalization of the trained detectors to new languages. This embeddings space structure can be caused both by the topic changes between languages, and the language specificity of the embeddings.

In order to quantify the impact of the two factors mentioned above one would require high quality translations pairs, as well as topic annotated samples. We leave this detailed study as future work and only investigate in Appendix A the text pairs translated with the NLLB model in Track C. While no clear conclusion can be reached, we reason based on the observed evidence that severe translations errors are surely present.

6 Conclusions

In this work we have presented our systems and results for the three tracks of Task 11 from SemEval-2025. For the ED task we observed that properly balancing the classes and emotions in the fine-tuning of LMs leads to consistent performance improvements. In the EIE one we have shown that fine-tuning with the simple detection objective from before can greatly increase performance in this task, especially for the under-performing emotions. Lastly, in the CL-ED task we have tested two types of systems, one relying on NMT and one on cross-lingual LMs. We presented the specific and common issues of both system types, proposing future research directions for quantifying the impact of these error sources.

Acknowledgments

The authors would like to thank Sergiu Nisioi and Ana-Sabina Uban for the provided references and resources, and also for reviewing an earlier version of this article. We also thank the anonymous reviewers for their feedback and suggestions for improving the article.

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.

Irean Navas Alejo, Toni Badia, and Jeremy Barnes. 2020. Cross-lingual Emotion Intensity Prediction. *Preprint*, arXiv:2004.04103.

Jason Ansel, Edward Yang, Horace He, Natalia Gimelshein, Animesh Jain, Michael Voznesensky, Bin Bao, Peter Bell, David Berard, Evgeni Burovski, Geeta Chauhan, Anjali Chourdia, Will Constable, Alban Desmaison, Zachary DeVito, Elias Ellison, Will Feng, Jiong Gong, Michael Gschwind, Brian Hirsh, Sherlock Huang, Kshiteej Kalambarkar, Laurent Kirsch, Michael Lazos, Mario Lezcano, Yanbo Liang, Jason Liang, Yinghai Lu, CK Luk, Bert Maher, Yunjie Pan, Christian Puhrsch, Matthias Reso, Mark Saroufim, Marcos Yukio Siraichi, Helen Suk, Michael Suo, Phil Tillet, Eikan Wang, Xiaodong Wang, William Wen, Shunting Zhang, Xu Zhao, Keren Zhou, Richard Zou, Ajit Mathews, Gregory Chanan, Peng Wu, and Soumith Chintala. 2024. PyTorch 2: Faster Machine Learning Through Dynamic Python Bytecode Transformation and Graph Compilation. In *29th ACM International Conference on Architectural Support for Programming Languages and Operating Systems, Volume 2 (ASPLOS '24)*. ACM.

Luis Chiruzzo, Salud María Jiménez-Zafra, and Francisco Rangel. 2024. Overview of IberLEF 2024: Natural Language Processing Challenges for Spanish and other Iberian Languages. In *IberLEF@SEPLN*.

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. *Preprint*, arXiv:2403.01222.

Flor Miriam Plaza del Arco, Carlo Strapparava, L Alfonso Urena Lopez, and M Teresa Martín-Valdivia. 2020. EmoEvent: A multilingual emotion corpus based on different events. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 1492–1498.

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.

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.

Hamed Khanpour and Cornelia Caragea. 2018. Fine-Grained Emotion Detection in Health-Related Online Posts. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1160–1166, Brussels, Belgium. 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. *Preprint*, arXiv:2202.10054.

Weixin Liang, Yuhui Zhang, Yongchan Kwon, Serena Yeung, and James Zou. 2022. Mind the Gap: Understanding the Modality Gap in Multi-modal Contrastive Representation Learning. *Preprint*, arXiv:2203.02053.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled Weight Decay Regularization. *Preprint*, arXiv:1711.05101.

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

Michael McCloskey and Neal J. Cohen. 1989. Catastrophic Interference in Connectionist Networks: The Sequential Learning Problem.

Shamsuddeen Hassan Muhammad, Nedjma Ousidhoum, Idris Abdulmumin, Jan Philip Wahle, Terry Ruas, Meriem Beloucif, Christine de Kock, Nirmal Surange, Daniela Teodorescu, Ibrahim Said Ahmad, David Ifeoluwa Adelani, Alham Fikri

Aji, Felermino D. M. A. Ali, Ilseyar Alimova, Vladimir Araujo, Nikolay Babakov, Naomi Baes, Ana-Maria Bucur, Andiswa Bukula, Guanqun Cao, Rodrigo Tufino Cardenas, Rendi Chevi, Chiamaka Ijeoma Chukwuneke, Alexandra Ciobotaru, Daryna Dementieva, Murja Sani Gadanya, Robert Geislinger, Bela Gipp, Oumaima Hourrane, Oana Ignat, Falalu Ibrahim Lawan, Rooweither Mabuya, Rahmad Mahendra, Vukosi Marivate, Andrew Piper, Alexander Panchenko, Charles Henrique Porto Ferreira, Vitaly Protasov, Samuel Rutunda, Manish Shrivastava, Aura Cristina Udrea, Lilian Diana Awuor Wanzare, Sophie Wu, Florian Valentin Wunderlich, Hanif Muhammad Zhafran, Tianhui Zhang, Yi Zhou, and Saif M. Mohammad. 2025a. BRIGHTER: BRIdging the Gap in Human-Annotated Textual Emotion Recognition Datasets for 28 Languages. *Preprint*, arXiv:2502.11926.

Shamsuddeen Hassan Muhammad, Nedjma Ousidhoum, Idris Abdulmumin, Seid Muhie Yimam, Jan Philip Wahle, Terry Ruas, Meriem Beloucif, Christine De Kock, Tadesse Destaw Belay, Ibrahim Said Ahmad, Nirmal Surange, Daniela Teodorescu, David Ifeoluwa Adelani, Alham Fikri Aji, Felermino Ali, Vladimir Araujo, Abinew Ali Ayele, Oana Ignat, Alexander Panchenko, Yi Zhou, and Saif M. Mohammad. 2025b. SemEval task 11: Bridging the gap in text-based emotion detection. In *Proceedings of the 19th International Workshop on Semantic Evaluation (SemEval-2025)*, Vienna, Austria. Association for Computational Linguistics.

team NLLB, Marta Ruiz Costa-jussà, James Cross, Onur cCelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Alison Youngblood, Bapi Akula, Loïc Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, John Hoffman, Semarley Jarrett, Kaushik Ram Sadagopan, Dirk Rowe, Shannon L. Spruit, C. Tran, Pierre Yves Andrews, Necip Fazil Ayan, Shruti Bhosale, Sergey Edunov, Angela Fan, Cynthia Gao, Vedanuj Goswami, Francisco Guzm'an, Philipp Koehn, Alexandre Mourachko, Christophe Ropers, Safiyyah Saleem, Holger Schwenk, and Jeff Wang. 2022. No Language Left Behind: Scaling Human-Centered Machine Translation. *ArXiv*, abs/2207.04672.

Cristian Daniel Paduraru and Ion Marian Anghelina. 2024. Early Risk Detection for Mental Health Disorders: UnibucAI at MentalRiskES 2024. In *IberLEF@SEPLN*.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Juan Manuel Pérez, Damián Ariel Furman, Laura Alonso Alemany, and Franco M. Luque. 2022. RoBERTuito: a pre-trained language model for social media text in Spanish. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 7235–7243, Marseille, France. European Language Resources Association.

Juan Manuel Pérez, Juan Carlos Giudici, and Franco Luque. 2021. pysentimiento: A Python Toolkit for Sentiment Analysis and SocialNLP tasks. *Preprint*, arXiv:2106.09462.

Ella Rabinovich, Noam Ordan, and Shuly Wintner. 2017. Found in translation: Reconstructing phylogenetic language trees from translations. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 530–540, Vancouver, Canada. Association for Computational Linguistics.

Tiberiu Sosea and Cornelia Caragea. 2021. eMLM: A New Pre-training Objective for Emotion Related Tasks. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 286–293, Online. Association for Computational Linguistics.

Tiberiu Sosea, Chau Pham, Alexander Tekle, Cornelia Caragea, and Junyi Jessy Li. 2022. Emotion analysis and detection during COVID-19. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 6938–6947, Marseille, France. European Language Resources Association.

Ashish Vaswani, Noam M. Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is All you Need. In *Neural Information Processing Systems*.

Yuqi Wang, Zimu Wang, Nijia Han, Wei Wang, Qi Chen, Haiyang Zhang, Yushan Pan, and Anh Nguyen. 2024. Knowledge Distillation from Monolingual to Multilingual Models for Intelligent and Interpretable Multilingual Emotion Detection. In *Proceedings of the 14th Workshop on Computational Approaches to Subjectivity, Sentiment, & Social Media Analysis*, pages 470–475, Bangkok, Thailand. Association for Computational Linguistics.

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. *Preprint*, arXiv:2412.15115.

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.

A Cross-Lingual Models

t-SNE visualization Regarding the t-SNE plots from Figure 2, we want to highlight that the tanh activation of the pooler layer in the LEALLA-large model is likely the main reasons why all the embeddings are clumped together more tightly than the embeddings of the QWEN model.

Setup In our experiments with cross-lingual LMs from the development period, we used the output of the pooler layer for the LEALLA encoder as sequence representation, while for the QWEN LM we used the last hidden state of the last token. The QWEN embeddings were extracted from a 4-bit quantized version of the model, using the Unsloth[6] library, and L_2 normalized. The linear classifiers were trained with the same methodology and hyper-parameters as before, but without further adjusting the detection thresholds.

LEALLA encoder The results of the linear probes trained on embeddings from the LEALLA-large model are presented in Table 6. We also fine-tuned the LEALLA encoder on all 3 languages addressed in Track A, following the recipe presented in the main article (see Table 7 for the results). The performance degradation on new languages is higher than 10% in most of the cases. Another observation is that training the linear probes only on embeddings from Spanish texts leads to better cross-lingual performance on Ukrainian and Hindi than training on all 3 languages from Track A.

We also observe that the cross-lingual performance improvements from fine-tuning are not uniform across the new languages. For Romanian the macro F_1 score either improves by less than 1%, or decreases by almost 2%, while on Ukrainian we observe and improvement of 5-7% and 13-15% for Hindi.

QWEN embeddings The results of the linear probes trained on QWEN embeddings are presented listed in Table 8. We also trained logistic regression models using the implementation in sklearn (Pedregosa et al., 2011) with the *lbfgs* and *linear* solvers. The results of these models (see Table 8) were actually worse than the results with LEALLA embeddings. We initially assumed that overfitting was the most likely cause, as the dimensionality of the embeddings was 14 times larger. To address this, we used PCA to reduce the dimensionality of the embeddings and then retrained the linear classifiers with the Pytorch implementation. We present in Table 9 the results with increasing number of principal components. As the validation Macro F_1 score keeps increasing with the number of components we conclude that the dimensionality reduction actually removes useful information from the embeddings. We also note that for Ukrainian and Hindi the best results are obtained with fewer components, not with the original embeddings, meaning that the dimensionality reduction also removed some information that was damaging for cross-lingual transferability.

Similarity of translated text pairs We provide in Figure 3 a set of t-SNE plots for LEALLA-large embeddings of the test set from Track C for the 6 addressed languages, both in its initial form and the version translated into Spanish with the NLLB model. We observe that the translated variants are more spread out than the originals, but they remain centered in the same region as the initial embeddings.

In Figure 4 we present histograms for the cosine similarity of original and translated text pairs, encoded with the pre-trained LEALLA model. The low cosine values can indicate both translation errors and language specificity of embeddings. While it is not clear based on these figures what the main source of errors is in the cross-lingual setting of ED, we believe that very low cosine values (less than 0.2) are most likely caused by severe translation errors. We assumed this based on the tendency of deep networks to restrict their outputs to a narrow cone (Liang et al., 2022). Thus, embeddings that largely deviate from this cone are most likely extract from nonsensical inputs, which are outside the distribution of texts used for training the encoder.

As for the generally poor performance of the models trained from this encoder (including the in-ditribution setting), we assume that this is caused by the data used for pre-training, which may not contain emotion-showcasing samples. The pre-training objective itself is more oriented towards matching information, not emotions, thus the em-

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

Language	Classifier	Emotion						Macro F_1
		anger	fear	joy	sadness	surprise	disgust	
English	Fine-tuned	68.75	79.03	73.33	76.32	71.88	-	73.86
	Fine-tune + LP	**72.73**	**80.6**	**73.68**	**80.56**	**73.68**	-	**76.25**
German	Fine-tune	79.7	42.11	67.39	61.11	36.84	66.17	58.89
	Fine-tune + LP	**80.0**	**50.00**	**74.36**	**62.96**	**40.74**	**68.96**	**62.84**
Spanish	Fine-tune	72.73	83.58	77.59	81.36	68.42	85.04	78.12
	Fine-tune + LP	**76.32**	**86.96**	**79.66**	**82.76**	**71.43**	**86.61**	**80.62**

Table 5: Dev set F_1 scores in track A for the fine-tuned classifiers and the second linear probes.

Source language	Validation macro F_1	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 6: Results on the dev set of target languages for the linear probes trained on embeddings from the LEALLA-large model.

#Transformer Blocks	Val. F_1	Target language		
		ron	ukr	hin
4	61.80	**47.60**	28.44	45.67
8	**62.64**	45.20	**30.22**	**47.05**

Table 7: Results of the finetuned LEALLA-large model on the dev set of target languages, based on the number of fine-tuned Transformer blocks.

Source languages	Validation macro F_1	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 8: 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.

beddings are unlikely to capture emotions-related features. While fine-tuning can help address this issues, a complete one would fair better than the partial fine-tune that we have done. Even in this case, one would have to take measures to prevent the occurrence of catastrophic forgetting (McCloskey and Cohen, 1989), making sure that the encoder remains language agnostic.

B Language Families Covered

We listed in Table 10 the 9 languages addressed in this paper and the language family that they are part of.

#Principal components	Val F_1	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 9: 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.

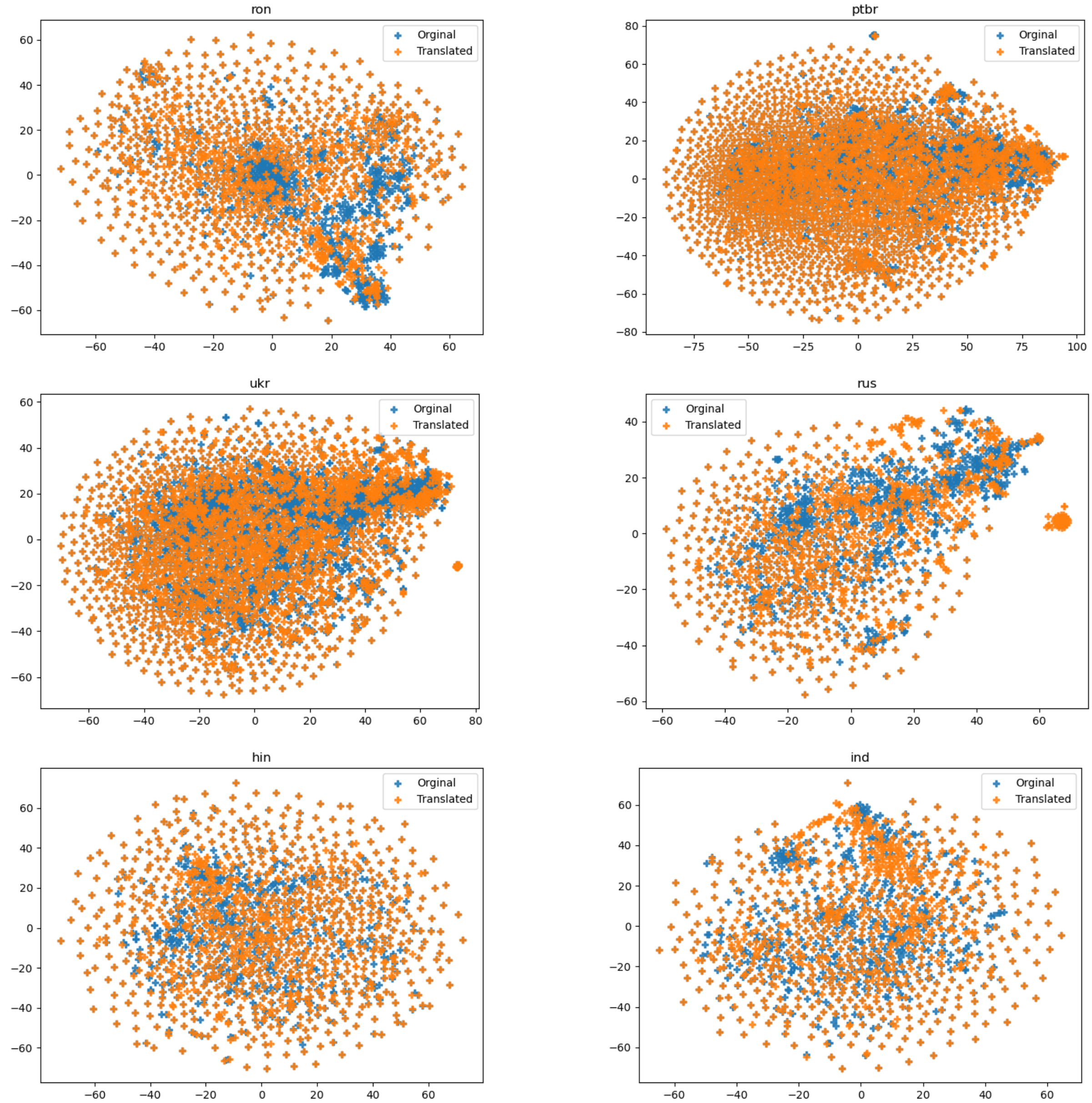

Figure 3: t-SNE plot of LEALLA-large embeddings for the test set of Track C, both in the original form and their Spanish translations done with the distilled NLLB-1.3B.

Language	Family
English	Indo-European; Germanic
German	Indo-European; Germanic
Spanish	Indo-European; Romance
Romanian	Indo-European; Romance
Portuguese (ptbr)	Indo-European; Romance
Ukrainian	Indo-European; Balto Slavic
Russian	Indo-European; Balto Slavic
Hindi	Indo-European; Indo-Iranian
Indonesian	Austronesian; Malayo-Polynesian

Table 10: Languages addressed in this work and the Language Families that they are part of.

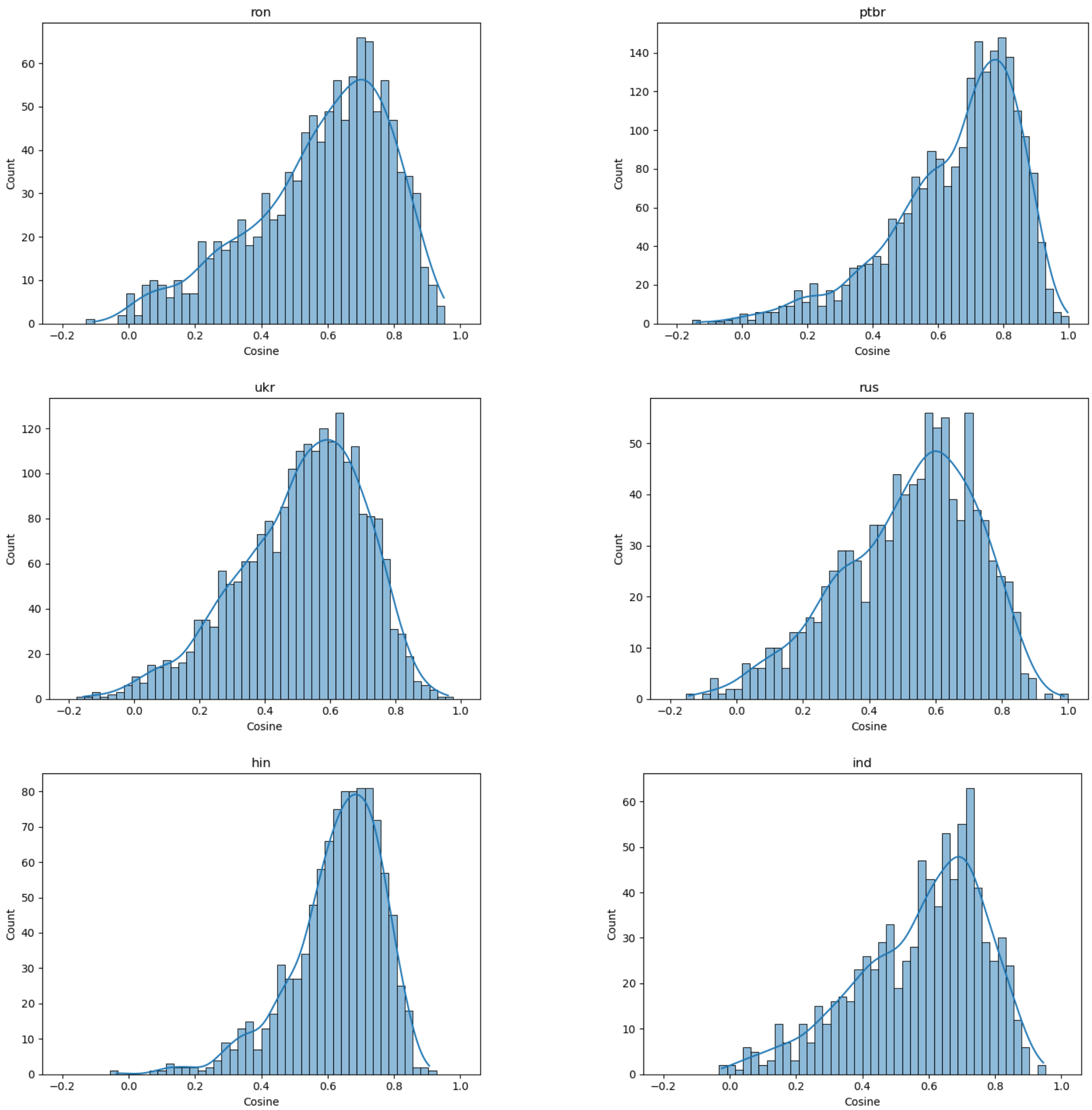

Figure 4: Cosine similarity for LEALLA-large embeddings of pairs of original and translated texts from the test set of Track C.