UncleLM at SemEval-2025 Task 11: RAG-Based Few-Shot Learning and Fine-Tuned Encoders for Multilingual Emotion Detection

Mobin Barfi Sajjad Mehrpeyma Nasser Mozayani

Iran University of Science and Technology

{m_barfi, sajjad_mehrpeyma}@comp.iust.ac.ir, mozayani@iust.ac.ir

Abstract

This paper introduces our approach for SemEval-2025 Task 11: Bridging the Gap in Text-Based Emotion Detection. We investigate a diverse set of methodologies, including fine-tuning encoder-based models and employing prompt engineering with large language models (LLMs) augmented by retrieval-augmented generation (RAG). Our system is evaluated across multiple languages, with a particular focus on low-resource languages, to assess the robustness and adaptability of these techniques. The findings provide valuable insights into enhancing emotion detection in multilingual and resource-constrained settings. The code and implementation details are publicly available at GitHub.

1 Introduction

Emotion detection from text has emerged as a fundamental task in Natural Language Processing (NLP), enabling advancements in domains such as sentiment analysis, affective computing, and human-computer interaction (Mohammad and Kiritchenko, 2018; Cambria, 2016). Despite progress, challenges persist, especially in multilingual settings where emotional expressions vary across languages and cultures (Öhman et al., 2020). Detecting nuanced, fine-grained emotional states is further complicated by the inherent context-dependence and subtlety of emotions.

This paper is motivated by our participation in the SemEval-2025 Task 11 (Muhammad et al., 2025b). This shared task spans three sub-tracks: Track A (Multi-label Emotion Classification), Track B (Emotion Intensity Prediction), and Track C (Cross-Lingual Emotion Detection). Each track poses distinct challenges, ranging from the classification of overlapping emotional states in a single text snippet to the estimation of emotion intensity across scales, and even transferring emotion detection across languages.

To address these challenges, we adopt state-of-the-art transformer-based architectures. Specifically, we fine-tune RoBERTa-large (Liu et al., 2019) for English tasks and XLM-RoBERTa-large (Conneau, 2019) for multilingual and cross-lingual settings. For Track A, our approach incorporates fine-tuning these models with a strong focus on threshold optimization to balance precision and recall. Additionally, we employ back-translation as a data augmentation technique for English models to improve their robustness (Sennrich et al., 2015; Edunov et al., 2018).

In Track B, we leverage the capabilities of GPT-4o-mini through few-shot learning, a paradigm particularly well-suited for estimating emotion intensity (Brown et al., 2020; Zhao et al., 2021). Relevant training examples are retrieved using cosine similarity of text embeddings, derived via OpenAI's embedding models. By retrieving semantically similar examples, the model is guided toward making more contextually appropriate predictions for varying emotion intensities. Notably, we introduce tailored prompting and iterative refinement of responses, which significantly improve the model's capacity to handle complex emotional expressions (Reynolds and McDonell, 2021).

We further experimented with a hybrid approach in Track B, combining outputs from GPT-4o-mini with predictions from an independently trained Multi-Layer Perceptron (MLP) model using OpenAI embeddings. While the hybrid model achieved notable gains in detecting "surprise", it fell short of surpassing the few-shot GPT-based predictions in other categories.

For Track C, our work extends the multilingual and cross-lingual capabilities of XLM-RoBERTa-large. Leveraging insights from prior work (Conneau, 2019), we exploit shared linguistic representations in related languages to enhance cross-lingual transfer learning. Datasets like XED (Öhman et al., 2020) and GoEmotions (Demszky et al.,

Proceedings of the The 19th International Workshop on Semantic Evaluation (SemEval-2025), pages 1563–1569

	ary	chn	deu	eng	esp	hau	hin	mar	ptmz	ron	rus	tat	ukr	amh	arq	ptbr
Train	1,608	2,642	2,603	2,768	1,996	2,145	2,556	2,415	1,546	1,241	2,679	1,000	2,234	2,556	1,608	2,226
Dev	267	200	200	116	184	356	100	100	257	123	199	200	249	100	100	200
Test	812	2,642	2,604	2,767	1,695	1,080	1,010	1,000	776	1,119	1,000	1,000	1,119	1,010	1,000	2,226

Table 1: Statistics for selected languages from the BRIGHTER dataset. Each row reflects the number of text samples available for training, development (dev), and testing in specific languages.

2020) underscore the importance of high-quality annotated resources for training robust emotion detection models, and we align our methodology with these principles.

2 Datasets

This study utilizes the **BRIGHTER dataset** (Muhammad et al., 2025a), a multilingual corpus for emotion recognition spanning 28 diverse languages. The dataset addresses a significant gap in the availability of annotated emotional data, particularly for low-resource languages, and is constructed from a range of textual sources, including social media posts, news articles, personal narratives, and literary texts.

In tasks related to Ethiopian languages, particularly Amharic, we incorporated insights from prior work on multi-label emotion classification (Belay et al., 2025), using knowledge derived from the EthioEmo dataset (Belay et al., 2025). Each data instance in the dataset is annotated for one or more of the six core emotions: *anger*, *disgust*, *fear*, *joy*, *sadness*, and *surprise*. A *neutral* label indicates the complete absence of these emotions, where all emotion intensities are set to zero.

The BRIGHTER dataset enhances the conventional multi-label emotion detection task by providing an emotion intensity scale ranging from 0 to 3 for each emotion, thereby enabling a more granular analysis of emotional nuances. This capability is pivotal for fine-grained emotion detection and cross-lingual tasks.

To further enhance model robustness for English, we augmented the BRIGHTER dataset with samples from **GoEmotions** (Demszky et al., 2020), a widely-used resource for fine-grained English emotion annotation. The augmentation enriched the training data and provided broader context for modeling diverse emotional expressions.

3 Methods

Our approach combines fine-tuned transformer models, few-shot learning with large language models (LLMs), and threshold optimization for multi-label classification. To handle the multilingual aspect, we fine-tuned **XLM-RoBERTa-large** (Conneau, 2019) and **RoBERTa-large** (Liu et al., 2019) for Track A, applying back-translation as a data augmentation technique for English (Sennrich et al., 2015). We also optimized classification thresholds to improve model calibration across all languages.

For emotion intensity estimation in Track B, we utilized a Retrieval-Augmented Generation (RAG) framework (Lewis et al., 2020) to enhance few-shot learning with GPT-4o-mini. Specifically, we retrieved the most relevant training examples using embedding-based cosine similarity search and incorporated them directly into the prompts. By including these contextually relevant examples, the model benefited from enhanced in-context learning, allowing it to better generalize across different intensity levels (Brown et al., 2020).

3.1 Data Preparation and Augmentation

To mitigate class imbalance within the English dataset, we augmented it with additional samples drawn from the GoEmotions dataset(Demszky et al., 2020). As GoEmotions encompasses a broader spectrum of emotion labels than those defined in this task, we filtered the dataset to retain only the six target emotion categories: *anger, disgust, fear, joy, sadness*, and *surprise*. Instances annotated as *neutral* were excluded due to their limited emotional content. Following this preprocessing step, we selectively incorporated records to achieve a balanced distribution across all emotion classes. This process resulted in a dataset comprising 6,195 samples for English.

In addition to balancing, we employed back-translation as a data augmentation technique, applied exclusively to the English samples. Each sentence was translated into German and then back to English using the Deep-Translator library with the Google Translate API. Back-translation introduced natural linguistic variation while preserving the underlying emotional content of the text. No additional filtering was applied to the backtranslated outputs. The augmentation procedure yielded

a final dataset of 12,330 samples for English.

3.2 Fine-Tuning Encoder-Based Models

For Tracks A and C, we employed a full fine-tuning strategy on encoder-based language models to perform multi-label emotion classification across multiple languages. Specifically, RoBERTa-large was utilized for English emotion classification, while XLM-R was selected for multilingual emotion classification. Additionally, back-translation was explored as a data augmentation approach; however, due to resource constraints, it was primarily implemented for English. The choice of these models stems from their robust contextual representation capabilities, which are particularly effective in capturing the intricate nuances required for emotion classification tasks (Devlin et al., 2018; Liu et al., 2019; Conneau, 2019).

3.2.1 RoBERTa-large for English

We conducted fine-tuning of the RoBERTa-large model for the classification of five emotion labels in English. The model was trained in a multi-label classification setting, where each emotion label was independently predicted using a sigmoid activation function, thereby allowing the assignment of multiple emotions to a single input text. Hyperparameters such as the learning rate, batch size, and number of epochs were selected after testing several configurations to optimize performance, ensuring effective gradient updates during training with the AdamW optimizer.

3.2.2 XLM-R for Multilingual Emotion Detection

For multilingual emotion classification, we performed fine-tuning of XLM-R (Conneau, 2019), a multilingual encoder-based model pretrained on a diverse array of languages, including many low-resource ones. XLM-R was selected due to its ability to capture shared linguistic features across different languages, making it highly effective for tasks such as those in Track C, which require the transfer of emotion classification knowledge from one language to another. This aligns with prior studies demonstrating that multilingual pretraining enables models to leverage language-independent representations, resulting in enhanced cross-lingual generalization, particularly when fine-tuned on linguistically similar source languages (Conneau, 2019; Lim et al., 2024). The model was optimized using AdamW with a learning rate and batch size chosen after testing various configurations to achieve optimal performance.

Source Language	Inference Language
Romanian (ron)	Spanish (esp)
Romanian (ron)	German (deu)
Russian (rus)	Tatar (tat)
Hindi (hin)	Marathi (mar)
Hindi (hin)	Russian (rus)
Marathi (mar)	Hindi (hin)

Table 2: Source and inference language mapping for Track C.

3.3 Multi-Label Classification and Threshold Optimization

The task of emotion classification is modeled as a multi-label classification problem, where a single text instance may simultaneously express multiple emotions. Instead of employing a softmax activation function, we adopt a sigmoid activation function to independently estimate the probability of each emotion.

To address the challenge of class imbalance, we perform emotion-specific threshold optimization. Thresholds are tuned by maximizing the macro F1-score on the validation set through a grid search over values ranging from 0.1 to 0.9 in increments of 0.01. These optimized thresholds are subsequently applied to the predicted probabilities to generate the final classification labels.

3.4 Few-Shot Learning and Structured Prompt Engineering

To address the task of multilingual emotion detection and classification, we employed a sophisticated few-shot learning paradigm powered by GPT-4.0-mini, an advanced large language model (LLM). This approach leveraged prompt engineering to systematically guide the model's predictions for emotion detection, accounting for the subtle, multi-label, and multi-intensity nature of the task.

3.4.1 Role-Based Prompting for Multilingual Emotion Detection

We employed a structured "Role-Based Prompting" methodology to instruct the LLM for multi-label emotion classification across all emotional dimensions, utilizing a four-point intensity scale (0–3). The prompt explicitly framed the task as an analysis of the *perceived emotions of the speaker*, emphasiz-

ing linguistic markers that most observers would associate with the speaker's emotional state.

Key design elements of the prompts included:

- Emphasis on the co-existence of multiple emotions at varying intensities, reflecting the nuanced, multi-label nature of emotion classification (Demszky et al., 2020).
- A statistical framework for label distributions. We observed that the model exhibited biases in its baseline performance, and we manually calibrated the probability distributions for each emotion class. For underrepresented emotions, such as Joy and Surprise, we adjusted the distributions to better align with real-world data distributions. When the original distribution consisted of 60% for label 0 and 40% for label 1, we discovered that adjusting this distribution to 90% for label 0 and 10% for label 1 led to improved model performance.
- A strict output format to ensure consistency: "Joy: [0-3], Fear: [0-3], Anger: [0-3], Sadness: [0-3], Surprise: [0-3], Disgust: [0-3]."
- Integration of a Retrieval-Augmented Generation (RAG) framework (Lewis et al., 2020). Relevant few-shot examples were dynamically retrieved based on cosine similarity of OpenAI Ada embeddings, providing the LLM with contextually aligned examples for enhanced in-context learning. This retrieval process was pivotal in guiding the model toward more contextually appropriate predictions.

3.4.2 Iterative Re-Prompting for Enhanced Robustness

To improve the reliability of the LLM's outputs, we introduced an "Iterative Re-Prompting" technique. After the initial prediction for a given text snippet, the model was re-prompted with an additional instruction: *"Are you sure? Analyze deeper."* This iterative mechanism encouraged the model to re-evaluate its initial predictions, particularly for instances with subtle or ambiguous emotional cues. Re-prompting approaches have previously been shown to enhance model robustness by refining responses in iterative loops.

Empirical results demonstrated that this iterative querying improved classification consistency, especially for challenging samples. The method was particularly beneficial in cases where the initial prediction lacked confidence or exhibited biases toward dominant emotions.

3.4.3 Temperature-Tuned Multiple Prompting for Diversity

We experimented with a "Temperature-Tuned Multiple Prompting" strategy to introduce diversity into the predictions. This involved generating outputs from five separate prompts with a higher temperature setting (*temperature = 0.7*) and averaging the results to mitigate prediction biases. While this approach introduced controlled variability in predictions, the aggregated performance did not exceed that of the single-prompt, low-temperature setting (*temperature = 0*). Consequently, the re-prompting method was adopted as the primary mechanism, given its superior consistency and accuracy.

3.4.4 Hybrid Framework with MLP for Specific Emotions

To address notable limitations observed in the LLM's performance for specific emotion classes, such as *Surprise*, we incorporated a hybrid framework. While the LLM exhibited strong performance across most emotional dimensions, it struggled with *Surprise*, as evidenced by lower evaluation metrics. To mitigate this, we trained a Multi-Layer Perceptron (MLP) classifier using OpenAI Ada embeddings.

The MLP model demonstrated remarkable effectiveness in handling the *Surprise* dimension, likely due to its ability to leverage consistent embedding representations for this class. The final hybrid system integrated outputs from both models: the LLM predictions were retained for all emotions except *Surprise*, which was handled exclusively by the MLP classifier. Specifically, the MLP consists of two hidden layers with 128 and 64 units respectively, each followed by ReLU activation and dropout, and a final linear layer projecting to the six emotion classes. This hybrid approach aligns with recent research advocating for the combination of LLMs and traditional classifiers to achieve task-specific enhancements.

4 Results and Analysis

4.1 Track A: Multi-Label Classification

We evaluated the performance of RoBERTa-large and XLM-RoBERTa-large on multiple languages. The results of our experiments, including macro F1-scores and emotion-specific scores, are presented

in Table 3. The evaluation metrics employed in this task, including those for multi-label classification, adhere to the definitions provided in the task description paper (Muhammad et al., 2025b). Our findings suggest that optimized thresholds significantly outperformed the default 0.5 threshold by effectively balancing precision and recall, particularly for underrepresented emotions like *surprise*.

4.2 Track B: Emotion Intensity Estimation

For Track B, we used GPT-4o-mini in a few-shot learning setup. The performance was evaluated using Pearson correlation between the predicted and gold intensity labels. The results of our experiments, including correlation scores across different emotions, are presented in Table 4. While the model performed well overall, it struggled with the surprise emotion, which consistently showed weaker results compared to other emotions.

4.3 Track C: Cross-Lingual Emotion Detection

For Track C, we fine-tuned XLM-RoBERTa-large on a source language and evaluated it on a target language. The evaluation metric used was macro F1-score. The results of our experiments, including performance scores across target language, are presented in Table 5. The best performance was achieved with linguistically related source-target pairs, such as Romanian → Spanish and Hindi → Marathi, which has been shown to improve performance in cross-lingual tasks (Conneau, 2019).

4.4 Analysis

The methods used in this study, including fine-tuned transformer models, Retrieval-Augmented Generation (RAG), and threshold optimization, were effective for emotion detection but also revealed key challenges and areas for improvement.

Threshold optimization in multi-label emotion classification played a crucial role in balancing precision and recall, particularly for underrepresented emotions like Surprise. This method helped mitigate class imbalance. However, emotion distributions in training data impacted model performance, emphasizing the need for dynamic thresholding to handle imbalances, especially for languages with skewed emotion distributions.

In emotion intensity estimation, GPT-4o-mini was used in a few-shot learning setup, enhanced by RAG-based retrieval. This allowed the model to improve predictions by retrieving relevant examples through embedding-based similarity search. While this approach worked well overall, it faced difficulties with emotions like Surprise. Further refinement of the retrieval process, especially for emotions with subtle markers, could improve accuracy.

For Track C, fine-tuning XLM-RoBERTa-large across linguistically related language pairs showed that shared linguistic features improve performance in cross-lingual emotion detection. However, the linguistic distance between some languages posed challenges. This highlights the need for techniques that can handle greater linguistic diversity and improve cross-lingual transferability.

A common challenge across all tracks was the model's struggle with emotions that are less represented in the training data. While common emotions like joy and anger performed well, rare emotions like Surprise were more difficult to detect, indicating the need for more diverse datasets to capture a broader range of emotional expressions.

While LLMs are highly capable at nuanced emotion analysis, careful alignment of their predictions to the statistical and linguistic realities of multilingual datasets is essential. Additionally, our hybrid framework highlights the benefits of combining fine-tuned classical ML models with advanced LLM-based pipelines to improve performance in specialized emotion detection tasks.

4.5 Conclusion

This work presents a framework for emotion detection, combining fine-tuned transformer models with Retrieval-Augmented Generation (RAG) techniques. We demonstrated the effectiveness of multilingual fine-tuning and threshold optimization to improve emotion classification and handle class imbalance.

The results highlight the importance of linguistic relatedness for cross-lingual emotion detection, with fine-tuning on related languages enhancing transferability. RAG proved valuable in retrieving relevant examples for more accurate intensity predictions. This approach sets the stage for future improvements in cross-lingual transfer and emotion detection for underrepresented emotions.

References

Tadesse Destaw Belay, Israel Abebe Azime, Abinew Ali Ayele, Grigori Sidorov, Dietrich Klakow, Philip

Slusallek, Olga Kolesnikova, and Seid Muhie Yimam. 2025. Evaluating the capabilities of large language models for multi-label emotion understanding. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 3523–3540, Abu Dhabi, UAE. Association for Computational Linguistics.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS '20, Red Hook, NY, USA. Curran Associates Inc.

Erik Cambria. 2016. Affective computing and sentiment analysis. *IEEE Intelligent Systems*, 31(2):102–107.

A Conneau. 2019. Unsupervised cross-lingual representation learning at scale. *arXiv preprint arXiv:1911.02116*.

Dorottya Demszky, Dana Movshovitz-Attias, Jeongwoo Ko, Alan Cowen, Gaurav Nemade, and Sujith Ravi. 2020. Goemotions: A dataset of fine-grained emotions. *arXiv preprint arXiv:2005.00547*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805.

Sergey Edunov, Myle Ott, Michael Auli, and David Grangier. 2018. Understanding back-translation at scale. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 489–500, Brussels, Belgium. Association for Computational Linguistics.

Patrick S. H. Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. Retrieval-augmented generation for knowledge-intensive NLP tasks. *CoRR*, abs/2005.11401.

Seong Hoon Lim, Taejun Yun, Jinhyeon Kim, Jihun Choi, and Taeuk Kim. 2024. Analysis of multi-source language training in cross-lingual transfer. *arXiv preprint arXiv:2402.13562*.

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. *arXiv preprint arXiv:1907.11692*.

Saif Mohammad and Svetlana Kiritchenko. 2018. Understanding emotions: A dataset of tweets to study interactions between affect categories. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

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.

Emily Öhman, Marc Pàmies, Kaisla Kajava, and Jörg Tiedemann. 2020. Xed: A multilingual dataset for sentiment analysis and emotion detection. *arXiv preprint arXiv:2011.01612*.

Laria Reynolds and Kyle McDonell. 2021. Prompt programming for large language models: Beyond the few-shot paradigm. *CoRR*, abs/2102.07350.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2015. Improving neural machine translation models with monolingual data. *arXiv preprint arXiv:1511.06709*.

Tony Z. Zhao, Eric Wallace, Shi Feng, Dan Klein, and Sameer Singh. 2021. Calibrate before use: Improving few-shot performance of language models. *CoRR*, abs/2102.09690.

Language	Anger (%)	Disgust (%)	Fear (%)	Joy (%)	Sadness (%)	Surprise (%)	Macro F1 (%)	Micro F1 (%)
deu	74.13	68.95	39.62	68.39	63.16	35.11	**58.23**	**64.80**
eng	66.67	-	81.52	76.05	74.60	73.90	**74.55**	**76.48**
esp	68.67	75.51	78.28	84.73	78.75	72.68	**76.44**	**76.56**
hin	78.80	81.22	90.91	88.89	85.38	88.14	**85.56**	**85.55**
mar	74.72	81.25	83.85	76.92	80.70	82.78	**80.04**	**79.84**
rus	86.56	86.78	93.27	90.77	81.45	83.84	**87.11**	**87.13**
ary	53.16	49.38	45.00	70.00	60.06	45.99	**53.93**	**55.20**
chn	82.87	47.86	46.62	85.69	56.82	48.92	**61.46**	**71.11**
hau	31.60	28.68	25.54	27.42	46.77	30.66	**31.78**	**32.60**
ptmz	30.88	27.59	51.22	54.47	63.19	36.92	**44.05**	**50.87**
ron	62.10	71.32	85.22	96.04	74.96	57.40	**74.51**	**74.18**

Table 3: Track A results for multi-label classification across multiple languages (F1 score in percentage).

Language	Anger (%)	Disgust (%)	Fear (%)	Joy (%)	Sadness (%)	Surprise (%)	Avg Pearson r (%)
amh	39.36	39.91	27.69	59.99	56.24	16.33	**39.92**
deu	74.61	67.83	52.43	77.14	70.68	42.41	**64.18**
eng	76.57	-	79.88	81.80	76.71	64.21	**75.83**
esp	72.39	48.15	79.16	80.52	79.46	68.32	**71.33**
ptbr	67.59	29.31	56.56	76.14	72.17	42.65	**57.40**
rus	89.03	87.93	83.89	84.31	81.21	79.71	**84.35**
arq	57.41	35.76	53.59	64.41	50.36	41.44	**50.50**
chn	71.44	42.33	41.26	79.44	57.59	31.53	**53.93**
hau	57.16	76.10	64.16	64.40	67.69	48.74	**63.04**

Table 4: Track B results for emotion intensity prediction across multiple languages (pearson correlation in percentage)

Language	Anger (%)	Disgust (%)	Fear (%)	Joy (%)	Sadness (%)	Surprise (%)	Macro F1 (%)	Micro F1 (%)
spa	61.50	68.14	65.10	64.40	56.67	49.34	**60.86**	**61.30**
hin	58.91	55.56	80.00	85.35	72.16	73.65	**70.94**	**72.41**
mar	74.85	73.74	85.41	76.41	74.77	84.62	**78.30**	**77.97**
rus	68.12	70.83	83.33	61.06	59.48	59.33	**67.03**	**66.86**
tat	40.12	11.94	03.17	53.22	49.56	62.73	**36.79**	**45.92**
ukr	38.77	46.24	72.53	65.29	57.20	55.38	**55.90**	**59.69**

Table 5: Track C results for cross-lingual emotion detection across multiple languages (F1 score in percentage).