FiRC-NLP at SemEval-2025 Task 11: To Prompt or to Fine-Tune? Approaches for Multilingual Emotion Classification

Wondimagegnhue Tufa[†◇], **Fadi Hassan**[†⋆], **Evgenii Migaev**[⋆], and **Yalei Fu**[⋆]

[◇]Faculty of Humanities, Vrije Universiteit Amsterdam
[⋆]Huawei Technologies Finland Research Center
{w.t.tufa}@vu.nl
{firstname.lastname}@huawei.com

Abstract

In this paper, we present our system developed for participation in SemEval-2025 Task 11: Bridging the Gap in Text-Based Emotion Detection. We compare three approaches for multilingual, multi-label emotion classification: fine-tuning a single XLM-R model, ensembling multiple XLM-R models, and a prompt-based method using a large language model. We evaluate these approaches across a diverse set of languages, ranging from high-resource to low-resource settings. Our experiments show that fine-tuning encoder models consistently outperforms the prompt-based approach for most languages. Additionally, we compare the performance of monolingual and multilingual models, observing mixed results: while multilingual models outperform monolingual ones for certain languages, the opposite trend is seen for others.

1 Introduction

Emotions are both familiar and enigmatic. We experience and manage them daily, yet they remain complex, nuanced, and often difficult to articulate (Muhammad et al., 2025a). Additionally, language is used in intricate ways to convey emotions (Wiebe et al., 2005; Mohammad and Kiritchenko, 2018). Moreover, there is significant variability in how individuals perceive and express emotions, even within the same culture or social group.

Automatic emotion recognition encompasses several tasks, such as detecting a speaker's emotions, identifying the emotions conveyed in written text, and recognizing emotions evoked in a reader (Mohammad, 2021; Teodorescu and Mohammad, 2023). SemEval-2025 Task 11 focuses on identifying the emotion that most people believe the speaker is experiencing based on a given sentence or short text snippet (Muhammad et al., 2025b). The first track addresses multi-label emotion detection across 30 languages (Belay et al., 2025; Muhammad et al., 2025a). The goal is to predict one or more emotions that most people believe the speaker is experiencing, based on a sentence or short text snippet uttered by the speaker. The emotions include : joy, sadness, fear, anger, surprise, and disgust.

In this paper, we explore two distinct approaches: fine-tuning-based approach (FBA) and prompt-based approach (PBA). For FBA, We frame the task as a multi-label sentence classification task and we experiment with monolingual and multilingual encoder models. We provide further details in Section 3.2. For PBA, we leverage a pre-trained large language model (LLM) to classify emotions without fine-tuning. We experiment with few-shot prompting, where we provide positive and negative examples for each emotion label in the same language as the input text. The model then analyzes the input text based on these examples and produces emotion scores. For all PBA experiments, we use GGPT-4o mini as our primary LLM (OpenAI et al., 2024) and further details of this approach are provided in Section 3.1.

In summary,

- We compare the effectiveness of monolingual and multilingual models in the task of multi-label emotion classification. We observe mixed results, with some multilingual models performing better for certain languages while performing worse for others.

- We explore the effectiveness of a prompt-based approach for multi-label emotion classification. The prompt-based approach shows lower performance compared to fine-tuned models across most languages. For some languages, such as German and Chinese, it results in a two-point F1-score drop.

[†]These authors contributed equally to this work.

Proceedings of the The 19th International Workshop on Semantic Evaluation (SemEval-2025), pages 1549–1556

- We analyze the cross-lingual differences in both of our approaches and observe a correlation between the resource availability of a language and its performance in the classification task, with high-resource languages showing higher performance.

2 Background

Emotion detection has become crucial due to its psychological, social, and commercial significance (Maruf et al., 2024; Muhammad et al., 2025a). Individuals express their emotions through language, body gestures, and facial expressions. Automatic emotion recognition encompasses various sub-tasks, such as detecting a speaker's emotions, identifying the emotions conveyed in written text, or recognizing emotions evoked in a reader of a target text (Mohammad, 2021; Teodorescu and Mohammad, 2023).

2.1 Emotion Detection

In the context of text analysis, emotion detection models analyze text to identify the underlying sentiment of the author. By relying on different aspects of the text, such as word choice and tonality, a model can learn to associate a text with various emotional categories, such as happiness, sadness, anger, or fear (Acheampong et al., 2020; Nandwani and Verma, 2021). These models have applications in various domains, including social media monitoring, customer feedback analysis, and sentiment analysis of product reviews (Acheampong et al., 2020).

2.2 Task Description

SemEval-2025 Task 11 focus on identifying the emotion that most people would associate with a speaker based on a text snippet (Muhammad et al., 2025b). The first track of the task involves multi-label emotion detection across various languages. Given a specific text snippet, the goal is to predict one or more emotions that the speaker is perceived to be experiencing. The target emotions include joy, sadness, fear, anger, surprise, and disgust. The distribution of the data across languages is shown in Figure 1.

2.3 Pre-trained Models

XLM-R is a RoBERTa based model pre-trained on 100 languages using CommonCrawl-100 data (Lample and Conneau, 2019). XLM-R obtain state-of-the-art results on many down stream task. We use XLM-R as one of our base model for fine-tuning.

AfroXLMR is a multilingual model created by MLM adaptation of XLM-R-large model on 17 under-resourced languages (Alabi et al., 2022). We use AfroXLMR for language that do not have monolingual language model such as Amharic, Tigrinya, and Oromo.

AraBERT is an Arabic pre-trained language model based on the BERT architecture (Antoun et al.). We use AraBERT to compare the effectiveness of Arabic monolingual and multilingual model variants (XLM-R) by fine-tuning both on Arabic data.

MacBERT is a monolingual Chinese model pre-trained with a novel MLM as correction pre-training task Cui et al. (2020). MacBERT enhances performance in Chinese NLP tasks by incorporating a novel masking strategy and whole word masking which has successfully combined the advantages from models like RoBERTa (Liu et al., 2019) and ALBERTa (Lan et al., 2019). We use MacBERT similar to AraBERT to compare the effectiveness of Chinese monolingual and multilingual model variants (XLM-R) by fine-tuning both on Chinese data.

ModernBERT is a recently released encoder model which shows improvement many down-stream tasks across many benchmarks (Warner et al., 2024). We use ModernBERT to compare the effectiveness of English monolingual and multilingual model variants (XLM-R) by fine-tuning both on English data.

RubertBaseCased is a monolingual Russian model pre-trained based on BERT architecture Kuratov and Arkhipov (2019). We use RubertBaseCased to compare the effectiveness of Russian monolingual and multilingual model variants (XLM-R) by fine-tuning both on Russian data.

3 System Description

We explore two distinct approaches: the Fine-Tuning-Based Approach (FBA) and the Prompt-Based Approach (PBA). For the FBA, we experiment with monolingual and multilingual encoder models by framing the task as a multi-label sentence classification problem. For the PBA, we uti-

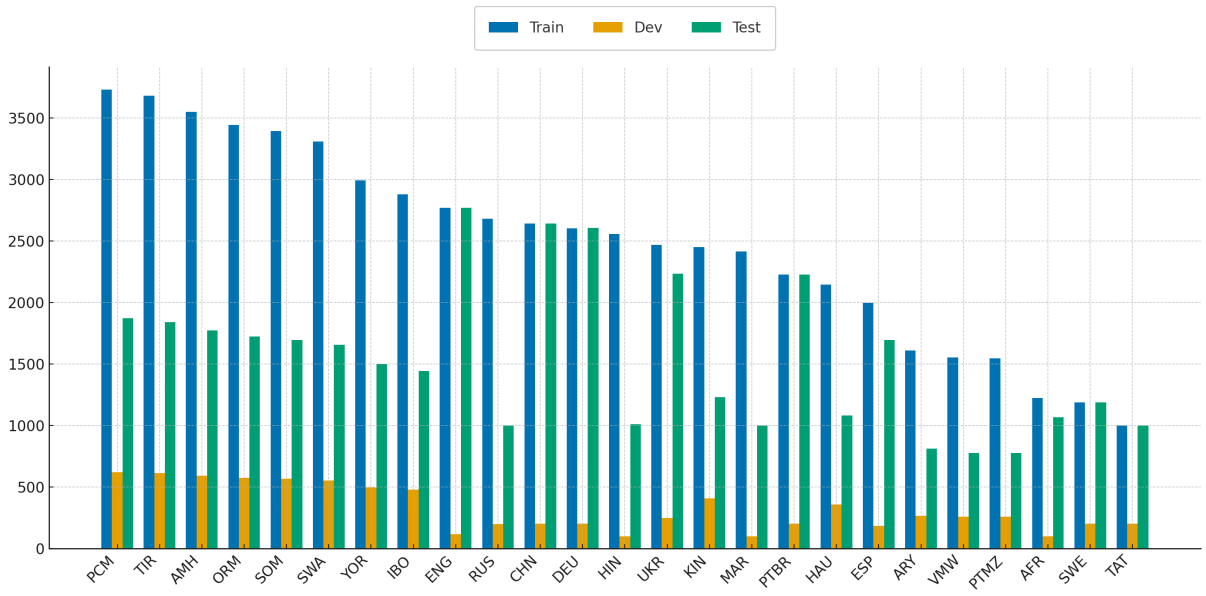

Figure 1: Distribution of train, development, and test samples across the 25 target languages. The figure highlights significant data imbalance, with across, which may impact model performance and cross-lingual generalization.

	HIN	RUS	ESP	MAR	ENG	AMH	TAT	DEU	UKR	HAU	CHN	ORM	AFR	SWE	PCM
XLM-R	87.28	**86.91**	78.74	72.7	70.98	69.44	69.23	66.78	66.43	66.34	62.93	59.31	**58.74**	58.18	57.96
XLM-5	**87.79**	86.64	**79.7**	**86.67**	**72.94**	**71.74**	**71.72**	**68.36**	**67.28**	**69.38**	**64.65**	**61.2**	58.03	**60.87**	**61.08**
GPT4	78.66	76.6	75.05	67.84	69.23	-	-	57.91	47.2	47.22	44.08	34.35	49.83	44.74	45.73

	PTBR	ARY	TIR	IBO	SOM	PTMZ	KIN	SWA	YOR	VMW
XLM-R	56.98	56.16	53.76	53.69	52.86	47.26	37.25	31.51	26.01	20.72
XLM-5	**60.07**	**59.21**	**56.87**	**54.08**	**55.85**	**51.08**	**41.5**	**32.33**	**30.8**	24.97
GPT4	52.5	41.12	39.37	36.85	49.9	42.39	24.34	13.92	-	**47.2**

Table 1: F1-Macro Scores of XLM-R, ensemble of five XLM-R models (XLM-5), and GPT4 models on the test set. The table is sorted by the XLM-R score. Results on the development set are available in Table 3.

lizes a pre-trained large language model (LLM) to classify emotions without fine-tuning. We use an English prompt for all of the samples. We use few-shot prompting approach by providing positive and negative examples per emotion label in the same language as the input text. The model processes the input text using these examples and generates emotion scores (0 to 1). We use GPT-4o mini as our main LLM for all of the PBA experiments (OpenAI et al., 2024).

3.1 Prompt Based Approach

Given the general nature of the task, we hypothesized that a well-trained LLM could effectively detect emotions without requiring fine-tuning. To align with the competition guidelines, we adapted the model by providing it with 20 randomly selected positive and negative examples for each emotion label, ensuring that the examples matched the language of the input text. We choose these numbers to balance providing sufficient context for the LLM while maintaining reasonable response times.

3.1.1 In Context Learning

For each prompt, we included the input text to be classified, 20 positive and 20 negative examples in the same language as the input text for each emotion label, and a request for the LLM to output a JSON object with emotion scores ranging from 0 to 1 for each label. We used the GPT-4o mini model with a temperature of 0 to ensure deterministic outputs. The prompt template is available in Appendix A.4.

3.1.2 Threshold Selection

To determine the optimal classification thresholds, we compute the F1-score for each emotion label across the development dataset. We evaluate thresholds ranging from 0 to 1 in increments of 0.01 and select the values that maximized the F1 metric. Finally, we apply these thresholds to generate predictions on the test dataset.

3.2 Fine-tuning Approach

In the fine-tuning approach, we experiment with both monolingual and multilingual models. For English, Chinese, Arabic, Russian, and Ukrainian, we test both monolingual and multilingual models. For under-resourced languages such as Amharic and Tigrinya, we experiment with two competitive multilingual models: XLM-R and AfroXLMR. For the remaining languages, we use XLM-R.

3.3 Model Ensemble

We use an ensemble of five models, each initialized with a different seed of the XLM-R model. We apply both hard voting and soft voting strategies. In hard voting, we aggregate the class label votes from all models and predict the class with the highest vote count. In soft voting, we sum the predicted probabilities for each class label and select the label with the highest cumulative probability.

4 Analysis and Conclusion

4.1 Monolingual and Multilingual Model

We compare monolingual and multilingual models to determine whether using a specialized model improves the performance of a multilingual model for a target language. We select six languages that have both a monolingual model and are included in the XLM-R multilingual model: English, Chinese, Arabic (Algerian and Moroccan), Russian, and Ukrainian. We follow a similar fine-tuning procedure for the multilingual and the monolingual models in their respective target languages. Specifically, we use ModernBERT (Warner et al., 2024) for English, AraBERT (Antoun et al.) for Arabic, and MacBERT (Cui et al., 2020) for Chinese, and RubertBaseCased for Russian. Table 2 presents the results of the test data. Overall, the multilingual model performs better than the monolingual model in four out of six languages. We observe a significant difference in Russian, where the multilingual model demonstrates strong performance, whereas English is an outlier, with the monolingual model outperforming its multilingual counterpart.

In our study on multi-label emotion classification in Chinese, we selected MacBERT (Masked Language Model as Correction BERT) (Cui et al., 2020) as the pre-trained model.

As shown in Table 5, the Moroccan Arabic (ARY) F1-Macro score for five models of the monolingual ensemble is slightly higher than the multilingual model XLM-R and the XLM-5 ensemble in the development set. However, its performance is lower in the test set. There are some possible reasons for this phenomenon. First, the limited dataset specific to one language might restrict the generalization of the model whereas the multilingual models benefit from exposure to a vast corpus from multiple languages. Second, multilingual models can take advantage of similarities between languages to transfer knowledge from high-resource languages to low-resource ones, an advantage that monolingual models lack.

Language	Multi-5	Mono-5
ARQ	54.71	**55.53**
CHN	**64.65**	59.69
ENG	72.94	**73.06**
RUS	**86.64**	54.85
UKR	**67.28**	57.48
ARY	**59.21**	57.42

Table 2: Comparison of Multilingual and Monolingual ensemble model for Arabic, Chinese and English. Multi-5 represents an ensemble of a multilingual model in this case XLM-R and Mono-5 represent an ensemble of a monolingual model - AraBERT for Arabic, MacBERT for Chinese, ModernBERT for English, RuBERT for Russian and for Ukranian.

4.2 Model Prompting vs Fine-tuning

Table 1 shows the F1-scores across 25 languages for three methods. XLM-R exhibits wide variation across languages, with strong performance for languages like Hindi and Russian and lower scores for low-resource languages such as Yoruba and Makhuwa. The ensemble method, XLM-5, provides a noticeable performance improvement, particularly for Marathi (86.67 vs. 72.7), although it still struggles with low-resource languages such as Yoruba and Makhuwa. The GPT-4-based approach shows lower performance compared to fine-tuned models across most languages, with some languages, such as German and Chinese, showing a two-point F1-score difference. Overall, the ensemble approach outperforms the single-model approach (XLM-R) across most languages, especially for high-resource languages like Hindi, Russian, and Spanish. The prompt-based approach performs significantly worse than both XLM-R and XLM-5, particularly for low-resource languages. Its advantage may lie in flexibility for few-shot learning, but it appears less effective for structured tasks like multi-label classification. Both XLM-R and

XLM-5 show a decline in performance for under-resourced languages, while GPT-4 struggles even more with these languages.

4.3 Cross-lingual Analysis

We compare performance variations across languages using the language classification schema proposed by (Joshi et al., 2020). This taxonomy categorizes languages into five classes, ranging from class 0, representing low-resource languages, to class 5, representing high-resource languages. We observe a correlation between a language's resource level and its performance in the classification task. High-resource languages in classes 4 and 5, such as Hindi, Russian, and Spanish, tend to achieve higher scores. Conversely, low-resource languages in class 1, including Igbo, Somali, and Kinyarwanda, exhibit lower F1-scores. The lowest F1-scores are typically observed for languages from taxonomy classes 0 and 1, such as Emakhuwa, Yoruba, and Swahili. The full list of languages and their taxonomy classifications is provided in Table 4 in Appendix A.2.

4.4 Conclusion

In this work, we compare the effectiveness of prompt-based and fine-tuning-based approaches. We further analyze the performance of monolingual and multilingual models for multi-label emotion classification. Our analysis indicates mixed results, with multilingual models outperforming monolingual models for some languages while underperforming for others. Additionally, we explore the effectiveness of the prompt-based approach and find that it generally performs worse than fine-tuned models across most languages. Furthermore, our analysis of cross-lingual differences reveals a correlation between a language's resource availability and its classification performance, with high-resource languages achieving higher F1-scores. These findings highlight the limitations of prompt-based approaches and emphasize the importance of language-specific adaptations in multilingual emotion classification.

Limitations

Despite achieving strong results across multiple languages, our work has several limitations,Our work has several limitations. First, the prompt-based approach exhibited slight output variance despite a low temperature setting, which we did not quantify. Second, threshold selection was based solely on the development set, introducing potential overfitting risks. Third, we did not conduct detailed error analysis to explain model failures across languages. Our cross-lingual analysis relies on an older language resource taxonomy, which may not fully capture current multilingual capabilities.

Acknowledgments

This work was conducted as part of an internship by Wondimagegnhue Tufa at Huawei Technologies Finland Research Center. We thank the Huawei Research team for providing computational resources and technical support throughout the project. We also express our gratitude to the SemEval-2025 Task 11 organizers for creating the multilingual emotion detection benchmark and offering valuable guidelines. Finally, we appreciate the feedback from anonymous reviewers, which helped improve the quality of this paper.

References

Francisca Adoma Acheampong, Chen Wenyu, and Henry Nunoo-Mensah. 2020. Text-based emotion detection: Advances, challenges, and opportunities. *Engineering Reports*, 2.

Jesujoba O. Alabi, David Ifeoluwa Adelani, Marius Mosbach, and Dietrich Klakow. 2022. Adapting pre-trained language models to African languages via multilingual adaptive fine-tuning. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 4336–4349, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Wissam Antoun, Fady Baly, and Hazem Hajj. Arabert: Transformer-based model for arabic language understanding. In *LREC 2020 Workshop Language Resources and Evaluation Conference 11–16 May 2020*, page 9.

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.

Yiming Cui, Wanxiang Che, Ting Liu, Bing Qin, Shijin Wang, and Guoping Hu. 2020. Revisiting pre-trained models for chinese natural language processing. *arXiv preprint arXiv:2004.13922*.

Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. 2020. The state and

fate of linguistic diversity and inclusion in the NLP world. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6282–6293, Online. Association for Computational Linguistics.

Yuri Kuratov and Mikhail Arkhipov. 2019. Adaptation of deep bidirectional multilingual transformers for russian language. *Preprint*, arXiv:1905.07213.

Guillaume Lample and Alexis Conneau. 2019. Cross-lingual language model pretraining. *Preprint*, arXiv:1901.07291.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2019. Albert: A lite bert for self-supervised learning of language representations. *arXiv preprint arXiv:1909.11942*.

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*.

Abdullah Al Maruf, Fahima Khanam, Md. Mahmudul Haque, Zakaria Masud Jiyad, M. F. Mridha, and Zeyar Aung. 2024. Challenges and opportunities of text-based emotion detection: A survey. *IEEE Access*, 12:18416–18450.

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).

Saif M. Mohammad. 2021. Sentiment analysis: Automatically detecting valence, emotions, and other affectual states from text. *Preprint*, arXiv:2005.11882.

Shamsuddeen Hassan Muhammad, Nedjma Ousidhoum, Idris Abdulmumin, and Others. 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.

Pansy Nandwani and Rupali Verma. 2021. A review on sentiment analysis and emotion detection from text. *Social Network Analysis and Mining*, 11.

OpenAI, Josh Achiam, et al. 2024. Gpt-4 technical report. *Preprint*, arXiv:2303.08774.

Daniela Teodorescu and Saif M. Mohammad. 2023. Evaluating emotion arcs across languages: Bridging the global divide in sentiment analysis. *Preprint*, arXiv:2306.02213.

Benjamin Warner, Antoine Chaffin, Benjamin Clavié, Orion Weller, Oskar Hallström, Said Taghadouini, Alexis Gallagher, Raja Biswas, Faisal Ladhak, Tom Aarsen, Nathan Cooper, Griffin Adams, Jeremy Howard, and Iacopo Poli. 2024. Smarter, better, faster, longer: A modern bidirectional encoder for fast, memory efficient, and long context finetuning and inference. *Preprint*, arXiv:2412.13663.

Janyce Wiebe, Theresa Wilson, and Claire Cardie. 2005. Annotating expressions of opinions and emotions in language. *Language Resources and Evaluation*, 39:165–210.

A Appendix

A.1 Evaluation on the development set

	mar	rus	hin	esp	eng	hau	tat	amh	deu	ukr	chn	afr	arq	
XLM-R	**96.84**	86.96	83.83	77.41	73.88	68.77	66.9	66.8	66.78	65.62	61.14	**57.95**	57.63	
XLM-5	95.83	**87.39**	**88.58**	**80.18**	**77.38**	**71.55**	**74.37**	**71.29**	**69.54**	**67.03**	**61.37**	57.97	**60.88**	
GPT4	86.35	77.84	67.41	76.09	72.57	50.79	-	-	56.67	42.32	52.48	43.59	47.96	
	pcm	**ary**	**sun**	**tir**	**orm**	**ptmz**	**som**	**swe**	**ibo**	**ptbr**	**kin**	**swa**	**yor**	**vmw**
XLM-R	56.57	**54.73**	54.46	54.12	53.75	52.81	52.26	**51.77**	50.74	48.54	37.7	31.62	30.6	22.67
XLM-5	**60.59**	52.99	**56.44**	**55.76**	**56.58**	**56.71**	**54.99**	49.94	**53.12**	**55.66**	**43.29**	**35.08**	**33.43**	28.05
GPT4	47.25	-	54.25	-	39.65	49.68	32.83	50.34	50.06	28.57	24.23	11.06	-	**47.2**

Table 3: F1-Macro of the Models on the development set

A.2 Language classification

Language Code	Class
pcm	0
ibo, jav, kin, som, sund, orm, tat	1
amh, hau, swa, tir, yor	2
afr, arq, ind, ary, ron, ukr	3
ptbr, hin, ptmz, rus, swe	4
zh, eng, deu, spa	5
emk, xho, zul	UNK

Table 4: Classification of language based on resource status. The higher class represents high-resourced languages and the lower class represents under-resourced languages. Source (Joshi et al., 2020).

A.3 Monolingual models

Language	Mono	Mono-5	Mono-10	XLM-5
amh	-	**70.12**	-	71.29
arq	53.76	**55.53**	55.25	60.88
ary	56.08	57.42	**58.50**	52.99
chn	-	**59.69**	-	61.37
eng	72.74	73.06	**74.23**	77.38
orm	-	**60.76**	-	56.58
rus	-	**54.85**	-	87.39
ukr	-	**57.48**	-	67.03

Table 5: Mono, Mono-5, and Mono-10 values for various languages

A.4 Prompt template

```
Given the following positive and
    negative examples for each emotion
    label, classify the emotion of the
    input text.
The response should be a JSON object
    with scores for each emotion label,
    where each score is between 0 and 1.
The labels are: anger_label,
    disgust_label, fear_label, joy_label
    , sadness_label, surprise_label.

Positive samples for anger_label:
- [20 randomly selected samples]

Negative samples for anger_label:
- [20 randomly selected samples]

...

Now, classify the following input text
    and return a JSON object with
    emotion scores for each of the
    labels.

### Input Text:
[input text]

### JSON Output (example format):
{
  "anger_label": 0.2,
  "disgust_label": 0.3,
  "fear_label": 0.8,
  "joy_label": 0.1,
  "sadness_label": 0.6,
  "surprise_label": 0.7
}
```