UB_Tel-U at SemEval-2025 Task 11: Emotions Without Borders - A Unified Framework for Multilingual Classification Using Augmentation and Ensemble

Tirana Noor Fatyanosa[•] **Putra Pandu Adikara**[•*] **Rochmanu Purnomohadi Erfitra**[•]
Muhammad Rajendra Alkautsar Dikna[•] **Sari Dewi Budiwati**[◦] **Cahyana**[◦]
[•] Brawijaya University, [◦] Telkom University,
{fatyanosa, adikara.putra}@ub.ac.id
{prnamatra4, mr.adikna}@gmail.com
{saridewi, cahyanayana}@telkomuniversity.ac.id

Abstract

This paper presents UB_Tel-U's submission to SemEval 2025 Task 11, which addresses three tracks: Multi-label Emotion Detection, Emotion Intensity, and Cross-lingual Emotion Detection. Our approach leverages a unified multilingual training strategy enriched by diverse external corpora and data augmentation techniques, enhancing both the diversity and robustness of the dataset. Rather than building separate models for each language, we consolidate all data into a single multilingual dataset, allowing the model to learn cross-lingual emotional patterns effectively. Our ensemble framework combines the multilingual capacity of BERT, DistilBERT, XLM-RoBERTa, the zero-shot generalization capabilities of LLaMA 3.3, and an English-specific model fine-tuned for emotion classification. The proposed system achieved competitive results, ranking 11th for Afrikaans (afr) in Track A, 9th for Ukrainian (ukr) in Track B, and 3rd for Amharic (amh), Chinese (chn), Hindi (hin), Marathi (mar), Brazilian Portuguese (ptbr), Russian (rus), and Ukrainian (ukr) in Track C.

1 Introduction

This paper presents UB_Tel-U's submission to SemEval 2025 Task 11, addressing three tracks: Multi-label Emotion Detection, Emotion Intensity, and Cross-lingual Emotion Detection (Muhammad et al., 2025b). While emotion recognition plays a critical role in various NLP applications, existing research has predominantly focused on high-resource languages, leaving a significant performance gap for low-resource languages that often lack sufficient annotated data (Muhammad et al., 2025a,b). The shared task provides a large multilingual dataset (Muhammad et al., 2025a; Belay et al., 2025a), offering an opportunity to explore scalable and cross-lingual emotion detection systems.

To address the multilingual challenge, we adopt a unified modeling strategy by merging data across all languages into a single multilingual training set, enabling the model to learn language-agnostic emotional representations. The UB_Tel-U system incorporates a combination of data preprocessing, data augmentation, and ensemble learning to enhance robustness and generalization. We leverage multiple transformer-based models, including multilingual, zero-shot, and English-specific architectures, and enrich the training data with external corpora such as SemEval 2018 Task 1. Final predictions are generated using ensemble methods, which help integrate complementary model outputs and mitigate individual weaknesses.

Our system demonstrated strong performance in Track C (Cross-lingual Emotion Detection), ranking third for several languages, including Amharic (amh), Chinese (chn), Hindi (hin), Marathi (mar), Brazilian Portuguese (ptbr), Russian (rus), and Ukrainian (ukr). In Track A, our best performance was in Spanish (esp) with a score of 0.7083, while in Track B, Russian (rus) achieved the highest Pearson correlation (0.7817). Despite these successes, the system struggled with certain low-resource languages, such as Emakhuwa (vmw) and Yoruba (yor), highlighting the limitations of current techniques in the absence of sufficient training data. Moreover, emotion intensity prediction (Track B) remained particularly challenging, with considerable variance across languages.

In summary, this paper presents a comprehensive multilingual approach to emotion detection. Our system aims to deliver scalable and adaptable performance across both high- and low-resource languages by unifying multilingual data, employing augmentation, and leveraging model ensembling. The results of our approach highlight both the strengths and limitations of current approaches while pointing toward future directions for improving the understanding of cross-lingual emotions.

[*]Corresponding author: adikara.putra@ub.ac.id

Proceedings of the The 19th International Workshop on Semantic Evaluation (SemEval-2025), pages 1948–1956

2 Background

2.1 Task Description

The SemEval 2025 Task 11 competition comprises three subtasks, each with distinct input and output formats (Muhammad et al., 2025a; Belay et al., 2025a). Track A: Multi-label Emotion Detection requires detecting the presence of six emotions—joy, sadness, fear, anger, surprise, and disgust—from a given text snippet. The output consists of binary labels indicating whether each emotion is present (1) or absent (0). For example, the input "I can't believe he forgot my birthday." might produce the output {'anger': 1, 'sadness': 1, 'joy': 0, 'fear': 0, 'surprise': 0, 'disgust': 0}.

Track B: Emotion Intensity Prediction focuses on quantifying the strength of an expressed emotion. Given a text snippet, the system outputs a numerical value between 0 and 3, where 0 indicates no emotion and 3 represents strong emotion. For instance, the input "I am thrilled about my new job!" could result in the output {'joy': 3}.

Track C: Cross-lingual Emotion Detection extends emotion classification to unseen languages. The input is a text snippet in a target language without available training data, and the output follows the same format as Track A, predicting emotion labels in a multilingual context.

2.2 Related Work

Recent advancements in multilingual NLP have significantly improved multi-label emotion detection, emotion intensity prediction, and cross-lingual emotion classification. Transformer-based architectures, such as BERT and RoBERTa, combined with domain-specific preprocessing, have enhanced the accuracy of emotion classification in social media text (Ying et al., 2019). The development of multilingual transformers like mBERT has also led to improved sentiment analysis for code-mixed and low-resource languages (Nazir et al., 2025).

Additionally, dynamic weighting frameworks have been introduced to address label imbalance in large-scale multilingual datasets (Yilmaz et al., 2023), while comprehensive multilingual datasets provide valuable benchmarks for evaluating emotion detection models (Augustyniak et al., 2023). However, emotion intensity detection remains particularly challenging in low-resource languages, where labeled data is scarce, and models struggle to generalize (Plisiecki et al., 2024; Zhang et al., 2024; Belay et al., 2025b).

Supervised models often outperform general-purpose LLMs in accuracy, but LLMs provide a viable alternative when labeled data is limited (Plisiecki et al., 2024). Fine-tuned models demonstrate superior performance in multi-label emotion classification, underscoring the importance of language-specific adaptation (Belay et al., 2025b). Similarly, in multilingual machine translation, fine-tuning has been shown to enhance model performance across diverse languages (Budiwati et al., 2021). In parallel, advancements in Affective Computing (AC) emphasize the roles of instruction tuning, prompt engineering, and hybrid AI frameworks as promising strategies for improving emotion intensity detection (Zhang et al., 2024).

Cross-lingual emotion detection aims to transfer emotion classification models across languages while addressing challenges such as limited labeled data, linguistic variation, and cultural influences on emotion expression (Zhao et al., 2024; Cheng et al., 2024; Barnes, 2023; Navas Alejo et al., 2020). Existing approaches include machine translation-based methods, embedding-based models, and transfer learning strategies to enhance multilingual sentiment adaptation (Zhao et al., 2024).

Ensemble methods combining LLMs with traditional classifiers have shown promising results, outperforming baselines in multilingual emotion detection tasks (Cheng et al., 2024). While rule-based methods remain effective in low-resource settings (Barnes, 2023), hybrid approaches that integrate linguistic features with deep learning present the most robust solutions for diverse language contexts (Navas Alejo et al., 2020).

3 System Overview

3.1 Transformer-Based Model Comparison

Transformer-based models leverage self-attention mechanisms to discern intricate contextual relationships within text. In this study, we examine four prominent transformer-based models for multilingual and emotion-specific tasks: *bert-base-multilingual-cased* (Devlin et al., 2019), *distilbert-base-multilingual-cased* (Sanh et al., 2019), *xlm-roberta-base* (Conneau et al., 2020), and *j-hartmann/emotion-english-distilroberta-base* (Hartmann, 2021).

We fine-tune each model using a multi-label classification setup. Each model's output layer was configured with a sigmoid activation function and the number of output neurons equal to the number

of emotion labels. This setup allows the model to independently predict the presence of each emotion per input instance. We used binary cross-entropy loss as the objective function, appropriate for multi-label tasks. A threshold of 0.5 was applied to the sigmoid outputs during evaluation to determine label assignment. Evaluation metrics included macro-averaged F1-score and overall accuracy. Models were trained for five epochs using the AdamW optimizer with a batch size of 32 and model checkpoints saved at each epoch.

3.2 Data Preprocessing

Our data preprocessing involves three key steps: lowercasing the text, merging all languages into a single dataset, and converting emoticons and emojis into standardized tokens. First, we convert all text to lowercase to reduce variability. Then, to simplify training, we merge data from multiple languages into one unified dataset. Next, we create a dictionary mapping common emoticons (e.g., ":)", ":-D") to specific emotion labels (e.g., "happy", "very happy") and replace each occurrence in the text with its corresponding label. Finally, we convert emojis into standardized textual descriptions.

3.3 Data Augmentation

To boost our model's multilingual emotion classification capabilities, we applied corpus-based data augmentation by incorporating external datasets that closely align with our original data. Specifically, we enrich our training set with data from SemEval 2018 Task 1, which includes content in English, Spanish, and Arabic (Mohammad et al., 2018). Previous research (Wei and Zou, 2019; Ma, 2019) has demonstrated that adding both real and synthetic training data can substantially improve model performance. Therefore, we applied corpus-based data augmentation to enrich the training data and enhance the model's generalization.

3.4 Zero-shot Classification

We utilize a zero-shot classification to automatically detect emotion from the given text. This zero-shot classification is based on LlaMa 3.3. We use a specific prompt and try multiple prompts as part of prompt engineering (Appendix A). Since the model is mostly trained using English dataset, we ask the model in English to make a prediction. The response of the prompt must concise and should only output the labels: 'anger', 'disgust', 'fear', 'joy', 'sadness', 'surprise'. However, sometimes the response can be a sentence or even a paragraph, so we add a post-processing step. The post-processing takes only the last sentence and filter out other words and leave only the expected labels. Furthermore, since this is a multilingual task emotion detection and not only in English, we ask the model to predict the language of the given text first, if the text is not in English, translate the text to English first, if the the model cannot directly translate to English, the model may translate it to another language transitively to English. As an illustration, the text may be translated first from African, to French, and from French to English. However, due to a difference in cultures and languages, some idioms or other cultural expressions, especially that have emotions, maybe lost in translation when translated to English.

3.5 Model Ensemble

We explore two ensemble techniques: majority voting and the OR rule ensemble. Majority voting is a widely used ensemble method where each classifier casts a vote for a class label, and the label with the most votes is selected as the final prediction. Majority voting is particularly effective when classifiers are diverse and independent (Zhu, 2013), and can be implemented in two forms: hard voting (based on predicted labels) or soft voting (based on predicted probabilities). In this work, we adopt the hard voting approach.

In contrast, the OR rule ensemble (also known as disjunctive ensemble) follows a more inclusive strategy, where the final decision is made if at least one classifier predicts a positive outcome. Rather than relying on the majority, it applies a logical "OR" operation across classifiers' outputs to maximize label coverage.

Specifically for Track B, where predictions involve emotion intensity values, we adapt the ensembling by selecting the maximum predicted intensity across models for each emotion. For example, if Model 1 predicts an intensity of 1 and Model 2 predicts an intensity of 3 for the same emotion, the final ensemble prediction will choose the higher value, 3.

4 Experimental Setup

For model comparison, we use the development (dev) set as the test set and split the training set into the training and validation sets using *Multilabel-*

StratifiedShuffleSplit from *iterative-stratification* library (Sechidis et al., 2011) with 'n_splits=1', we allocate 20% of the data as the validation set (test_size=0.2) and set 'random_state=42' for reproducibility. The label columns include {'anger', 'disgust', 'fear', 'joy', 'sadness', 'surprise', 'lang'}, with the addition of the 'lang' label to ensure stratification preserves the language distribution in the training and validation sets.

The evaluation metric for Track A and Track C is the average F1-score (macro), computed by first calculating the macro-averaged F1-score for each language and then averaging these scores across all languages. For Track B, where emotion intensities are continuous values ranging from 0 to 3, the evaluation metric is the Pearson correlation coefficient, also averaged across all languages.

Our research utilizes a comprehensive set of Python libraries to support data preprocessing, model training, and evaluation within our multilingual emotion classification system. For data manipulation and processing, we employ emoji (Carreira, 2017) and pandas (Reback et al., 2020) to handle text data. Dataset management and preparation are supported by the Hugging Face Datasets library (Lhoest et al., 2021) for efficient data loading and augmentation, and by the *iterative-stratification* library (Sechidis et al., 2011) for performing stratified splitting on multi-label datasets.

For model development, we leverage the Hugging Face transformers library (Wolf et al., 2020) alongside PyTorch (Paszke et al., 2019) as the deep learning framework, enabling seamless integration of pre-trained models and efficient training processes. Finally, for model evaluation, we use scikit-learn (Pedregosa et al., 2011) to compute various performance metrics. Our code is publicly available on GitHub.[1]

5 Results

5.1 Transformer-based Model Comparison

Table 1 presents a comparison of transformer-based models across three evaluation tracks, where Tracks A and C are assessed using macro-averaged F1 scores, and Track B is evaluated using Pearson correlation. Among the models, *xlm-roberta-base* consistently achieves the best performance across all tracks, highlighting its effectiveness in multilingual emotion classification tasks.

LLaMA 3.3 also demonstrates strong performance, suggesting its ability to generalize effectively across tasks. In contrast, *bert-base-multilingual-cased* and *distilbert-base-multilingual-cased* achieve moderate results. Meanwhile, *emotion-english-distilroberta-base*, which is trained specifically on English data, falls behind when evaluated in the multilingual setting. These findings highlight the importance of cross-lingual pretraining and multilingual model design in achieving robust performance in emotion classification tasks.

5.2 Effectiveness of Ensemble Methods

We compare the performance of two ensemble strategies, majority voting and the OR rule, across all tracks, as shown in Table 2. The OR Rule achieves the best results in Track A and Track C, with average F1 macro scores of 0.4997 and 0.4619, respectively, indicating improved sensitivity in detecting multiple emotion labels. Meanwhile, majority voting yields the highest Pearson correlation in Track B (0.5827), suggesting stronger performance in predicting emotion intensity. These findings indicate that while the OR rule is more effective for multi-label classification tasks, majority voting may be better suited for capturing continuous emotional dimensions.

Moreover, Table 3 shows that the ensemble consistently performs best on the 'joy' label across all tracks, achieving the highest average F1 Macro scores in Track A (0.7554) and Track C (0.7263), and the highest average Pearson correlation in Track B (0.6889). Emotions such as 'sadness', 'anger', and 'fear' also yield strong results, while 'disgust' and especially 'surprise' show relatively lower scores, particularly in Track B. These results indicate the ensemble's strength in detecting prominent emotions like 'joy' and 'sadness', but highlight challenges in handling emotions with more subtle or ambiguous expressions.

5.3 Overall System Performance

Table 4 demonstrates the progressive impact of each method on the overall system performance across all three tracks. We observe that incorporat-

[1]https://github.com/UB-Tel-U/semeval-2025-task-11

[2]https://huggingface.co/google-bert/bert-base-multilingual-cased

[3]https://huggingface.co/distilbert/distilbert-base-multilingual-cased

[4]https://huggingface.co/FacebookAI/xlm-roberta-base

[5]https://huggingface.co/j-hartmann/emotion-english-distilroberta-base

[6]https://huggingface.co/meta-llama/Llama-3.3-70B-Instruct

Table 1: Transformer-based model comparison results.

Model	Track A	Track B	Track C
1 - bert-base-multilingual-cased[2]	0.3846	0.4156	0.3540
2 - distilbert-base-multilingual-cased[3]	0.3804	0.3915	0.3511
3 - FacebookAI/xlm-roberta-base[4]	**0.4775**	**0.5485**	**0.4388**
4 - j-hartmann/emotion-english-distilroberta-base[5]	0.3346	0.3547	0.3065
5 - llama3.3[6]	0.4328	0.5447	0.4201

Table 2: Model ensemble results.

Ensemble Type	Track A		Track B		Track C	
	Best Combinations	Average F1 Macro	Best Combinations	Average Pearson	Best Combinations	Average F1 Macro
Majority Voting	1, 3, 5	0.4767	3, 5	**0.5827**	1, 3, 5	0.4397
OR Rule	1, 3, 4	**0.4997**	3, 5	0.5782	1, 3, 4	**0.4619**

Table 3: Ensemble performance per emotion label.

Label	Track A	Track B	Track C
anger	0.6849	0.6302	0.6644
disgust	0.6784	0.5477	0.6569
fear	0.6811	0.5464	0.6594
joy	**0.7554**	**0.6889**	**0.7263**
sadness	0.7232	0.6083	0.7073
surprise	0.6782	0.4696	0.6538

ing data preprocessing alone yields only marginal improvements. In contrast, data augmentation leads to a more notable and consistent performance boost across all tracks. The combination of preprocessing and data augmentation results in further performance gains, demonstrating the effectiveness of enriched and diversified training inputs. This combination proves particularly useful in multilingual settings, where the variation in text quality and structure across languages can be significant.

The highest scores are achieved through the ensemble approach, reaching 0.4997 in Track A, 0.5827 in Track B, and 0.4619 in Track C. Ensemble learning effectively leverages the complementary strengths of each individual model, compensating for their weaknesses and reducing the risk of overfitting to specific languages or emotion labels.

5.4 Submission

Due to time constraints and submission limitations, we were unable to submit the best-performing model identified in this study. It is important to note that all submitted models were trained using the combined train + dev set, whereas the models reported in our analysis were trained on the train set and evaluated on the dev set solely for comparison purposes.

All submitted models utilized both preprocessing and augmentation techniques. For Track A, we submitted an OR Rule ensemble consisting of *xlm-roberta-base* (Conneau et al., 2020), *j-hartmann/emotion-english-distilroberta-base* (Hartmann, 2021), and *llama3.3* (Meta AI, 2024). For Track B and Track C, we submitted *xlm-roberta-base* individually as our final model.

The results in Table 5 reveal varied performance across all three tasks and multiple languages, highlighting both the strengths and limitations of our multilingual emotion classification system. In Track A, the system achieved scores ranging from 0.1399 to 0.7083. Notably, Afrikaans (afr) attained a score of 0.5512, securing a relatively high rank of 11, which places it among the top-performing languages. In contrast, some languages like Emakhuwa (vmw) and Yoruba (yor) exhibited lower performance, with scores of 0.1399 and 0.225, respectively. These results highlight the system's strong capability in high-resource or moderately supported languages, while also emphasizing ongoing challenges in achieving robust performance for low-resource languages.

In this track, the system achieved Pearson correlation scores ranging from 0.3321 to 0.7817. Notably, Ukrainian (ukr) obtained a score of 0.5365, achieving a high rank of 9 among the participating languages. Russian (rus) achieved the highest score of 0.7817, showcasing the system's strong capability in handling emotion intensity tasks for certain languages. However, performance varied significantly across languages, suggesting that accurately modeling continuous emotion intensities remains more challenging compared to multi-label classification.

Table 4: Overall system performance on dev set.

Method	Track A	Track B	Track C
Baseline	0.4524	0.5485	0.4173
+ Data Preprocessing	0.4430	0.5500	0.4111
+ Data Augmentation	0.4761	0.5558	0.4380
+ Data Preprocessing + Data Augmentation	0.4775	0.5591	0.4388
+ Ensemble	**0.4997**	**0.5827**	**0.4619**

Table 5: Ranking results.

Lang	Track A		Track B		Track C	
	Score	Rank	Score	Rank	Score	Rank
afr	0.5512	11	-	-	0.3714	6
amh	0.4844	30	0.5689	10	0.6227	3
arq	0.529	13	0.3321	16	0.4227	9
ary	0.4326	24	-	-	0.4445	5
chn	0.5059	30	0.5657	11	0.5883	3
deu	0.5829	25	0.5512	13	0.6031	5
eng	0.7032	49	0.5311	31	0.6564	6
esp	0.7083	36	0.683	17	0.7484	4
hau	0.5157	28	0.5304	17	0.5862	6
hin	0.6673	34	-	-	0.8578	3
ibo	0.3865	21	-	-	0.4298	5
ind	-	-	-	-	0.5119	6
jav	-	-	-	-	0.3526	7
kin	0.3294	20	-	-	0.2911	6
mar	0.6838	32	-	-	0.833	3
orm	0.3589	24	-	-	0.3758	5
pcm	0.5384	16	-	-	0.528	4
ptbr	0.4707	25	0.4775	16	0.499	3
ptmz	0.3671	24	-	-	0.3776	5
ron	0.6924	28	0.556	15	0.7027	4
rus	0.6554	39	0.7817	18	0.8314	3
som	0.2989	25	-	-	0.3506	6
sun	0.419	20	-	-	0.3755	5
swa	0.3018	12	-	-	0.2018	7
swe	0.5058	22	-	-	0.5447	5
tat	0.5456	20	-	-	0.6386	4
tir	0.4047	16	-	-	0.3643	5
ukr	0.4214	29	0.5365	9	0.5789	3
vmw	0.1399	15	-	-	0.0423	5
xho	-	-	-	-	0.163	5
yor	0.225	20	-	-	0.1394	7
zul	-	-	-	-	0.1075	8

In Track C, our cross-lingual approach is tested on languages that lack direct training data. The performance in this track shows a promising range, with scores from 0.0423 to 0.8578. Notably, Amharic (amh), Chinese (chn), Hindi (hin), Marathi (mar), Brazilian Portuguese (ptbr), Russian (rus), and Ukrainian (ukr) ranked third in Track C, demonstrating the strength of our model in transferring emotion labels across languages.

6 Conclusion

In this study, we presented a unified multilingual framework for emotion classification, evaluated across three diverse tracks: multi-label classification, emotion intensity regression, and cross-lingual generalization. Our experimental results show that combining multiple transformer-based models with strategic data preprocessing, augmentation, and ensemble learning substantially enhances system performance. Among these components, the ensemble approach proved particularly effective, consistently outperforming individual models by leveraging their complementary strengths. This integration improved robustness across languages and highlighted the importance of model diversity and data enrichment for multilingual emotion recognition.

Despite these advancements, several challenges remain. Emotion intensity prediction continues to exhibit variability across languages, and performance in low-resource settings is still limited by data scarcity and linguistic diversity. To address these issues, future research could explore more sophisticated data augmentation strategies such as back-translation across related languages, generative paraphrasing, or adversarial training. Finally, incorporating domain-adaptive pretraining or language-specific adapters could further refine model sensitivity to cultural and linguistic nuances.

References

Łukasz Augustyniak, Szymon Woźniak, Marcin Gruza, Piotr Gramacki, Krzysztof Rajda, Mikołaj Morzy, and Tomasz Kajdanowicz. 2023. Massively multilingual corpus of sentiment datasets and multi-faceted sentiment classification benchmark. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, Red Hook, NY, USA. Curran Associates Inc.

Jeremy Barnes. 2023. Sentiment and emotion classification in low-resource settings. In *Proceedings of the 13th Workshop on Computational Approaches to Subjectivity, Sentiment, & Social Media Analysis*,

pages 290–304, Toronto, Canada. Association for Computational Linguistics.

Tadesse Destaw Belay, Israel Abebe Azime, Abinew Ali Ayele, Grigori Sidorov, Dietrich Klakow, Philip Slusallek, Olga Kolesnikova, and Seid Muhie Yimam. 2025a. 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.

Tadesse Destaw Belay, Israel Abebe Azime, Abinew Ali Ayele, Grigori Sidorov, Dietrich Klakow, Philip Slusallek, Olga Kolesnikova, and Seid Muhie Yimam. 2025b. 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.

Sari Dewi Budiwati, Tirana Fatyanosa, Mahendra Data, Dedy Rahman Wijaya, Patrick Adolf Telnoni, Arie Ardiyanti Suryani, Agus Pratondo, and Masayoshi Aritsugi. 2021. To optimize, or not to optimize, that is the question: TelU-KU models for WMT21 large-scale multilingual machine translation. In *Proceedings of the Sixth Conference on Machine Translation*, pages 387–397, Online. Association for Computational Linguistics.

Vitor Carreira. 2017. emoji: Emoji for python. https://pypi.org/project/emoji/. Version 2.10.0.

Long Cheng, Qihao Shao, Christine Zhao, Sheng Bi, and Gina-Anne Levow. 2024. TEII: Think, explain, interact and iterate with large language models to solve cross-lingual emotion detection. In *Proceedings of the 14th Workshop on Computational Approaches to Subjectivity, Sentiment, & Social Media Analysis*, pages 495–504, Bangkok, Thailand. Association for Computational Linguistics.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. *arXiv preprint arXiv:1911.02116*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Jan Hartmann. 2021. Emotion english distilroberta base. https://huggingface.co/j-hartmann/emotion-english-distilroberta-base.

Quentin Lhoest, Albert Villanova del Moral, Yacine Jernite, Abhishek Thakur, Patrick von Platen, Suraj Patil, Julien Chaumond, Mariama Drame, Julien Plu, Victor Sanh, and Thomas Wolf. 2021. Datasets: A community library for natural language processing. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 175–184, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Edward Ma. 2019. NLP augmentation. https://github.com/makcedward/nlpaug.

Meta AI. 2024. Llama 3.3 - 70b instruct. https://huggingface.co/meta-llama/Llama-3.3-70B-Instruct.

Saif Mohammad, Felipe Bravo-Marquez, Mohammad Salameh, and Svetlana Kiritchenko. 2018. SemEval-2018 task 1: Affect in tweets. In *Proceedings of the 12th International Workshop on Semantic Evaluation*, pages 1–17, New Orleans, Louisiana. Association for Computational Linguistics.

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.

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.

Irean Navas Alejo, Toni Badia, and Jeremy Barnes. 2020. Cross-lingual emotion intensity prediction. In *Proceedings of the Third Workshop on Computational Modeling of People's Opinions, Personality, and Emotion's in Social Media*, pages 140–152, Barcelona, Spain (Online). Association for Computational Linguistics.

Muhammad Kashif Nazir, Cm Nadeem Faisal, Muhammad Asif Habib, and Haseeb Ahmad. 2025. Leveraging multilingual transformer for multiclass sentiment analysis in code-mixed data of low-resource languages. *IEEE Access*, 13:7538–7554.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Srinath Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. 2019. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems 32 (NeurIPS 2019)*, pages 8024–8035.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Édouard Duchesnay. 2011. Scikit-learn: Machine learning in python. *Journal of Machine Learning Research*, 12:2825–2830.

Hubert Plisiecki, Piotr Koc, Maria Flakus, and Artur Pokropek. 2024. Predicting emotion intensity in polish political texts: Comparing supervised models and large language models in a resource-poor language.

Jeff Reback, Wes McKinney, jbrockmendel, Joris Van den Bossche, Tom Augspurger, Phillip Cloud, Simon Hawkins, Gfyoung, Sinhrks, Maxwell Roeschke, et al. 2020. pandas-dev/pandas: Pandas. *Zenodo*.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*.

Konstantinos Sechidis, Grigorios Tsoumakas, and Ioannis Vlahavas. 2011. On the stratification of multi-label data. In *Machine Learning and Knowledge Discovery in Databases*, pages 145–158, Berlin, Heidelberg. Springer Berlin Heidelberg.

Jason Wei and Kai Zou. 2019. EDA: Easy data augmentation techniques for boosting performance on text classification tasks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 6382–6388, Hong Kong, China. Association for Computational Linguistics.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, and Jamie Brew. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45. Association for Computational Linguistics.

Selim F. Yilmaz, E. Batuhan Kaynak, Aykut Koç, Hamdi Dibeklioğlu, and Suleyman Serdar Kozat. 2023. Multi-label sentiment analysis on 100 languages with dynamic weighting for label imbalance. *IEEE Transactions on Neural Networks and Learning Systems*, 34(1):331–343.

Wenhao Ying, Rong Xiang, and Qin Lu. 2019. Improving multi-label emotion classification by integrating both general and domain-specific knowledge. In *Proceedings of the 5th Workshop on Noisy User-generated Text (W-NUT 2019)*, pages 316–321, Hong Kong, China. Association for Computational Linguistics.

Yiqun Zhang, Xiaocui Yang, Xingle Xu, Zeran Gao, Yijie Huang, Shiyi Mu, Shi Feng, Daling Wang, Yifei Zhang, Kaisong Song, and Ge Yu. 2024. Affective computing in the era of large language models: A survey from the nlp perspective.

Chuanjun Zhao, Meiling Wu, Xinyi Yang, Wenyue Zhang, Shaoxia Zhang, Suge Wang, and Deyu Li. 2024. A systematic review of cross-lingual sentiment analysis: Tasks, strategies, and prospects. *ACM Comput. Surv.*, 56(7).

Mu Zhu. 2013. When is the majority-vote classifier beneficial? *Cornell University*.

A Appendix

The zero-shot prompt used for all tracks is presented in Table 6. The two prompts differ in classification detail and output format. The first prompt (Track A and C) assigns only emotion labels, producing a simple comma-separated list, while the second (Track B) includes intensity scores (0-3) for each detected emotion. The first provides binary classification (emotion present or not), whereas the second captures emotion intensity, offering finer sentiment analysis. As a result, the first prompt is suited for general emotion detection, while the second focuses on detailed sentiment analysis, capturing both the type and intensity of emotions.

Table 6: Zero-prompt used in all tracks.

Track	Prompt
Track A and C	Classify emotion from the given text into one or more: Joy, Fear, Anger, Sadness, Disgust, Surprise. The output is only the multilabel classes (Joy, Fear, Anger, Sadness, Disgust, Surprise) and separated by a comma. Do not use other unspecified classes. Do not output the reasoning statement or unnecessary sentences. If you don't know or unsure, translate into English first. If you cannot translate directly to English, translate it first to another known language, and from that another known language to English. The translation can be transitive through more than one language.
Track B	Classify the given text into one or more emotion categories: Joy, Sadness, Fear, Anger, Surprise, or Disgust. Each emotion should be assigned an intensity score between 0 and 3, where 0 means no intensity and 3 means the highest intensity. The output format should be a comma-separated list of emotions with their intensity scores in parentheses, e.g., 'Joy(2), Fear(1), Anger(0)'. Do not include emotions that are not specified. Do not add any reasoning, explanation, or extra sentences. If the text is not in English, translate it first to English. If a direct English translation is not possible, use an intermediate language before translating to English. The translation can be transitive through multiple languages if necessary.