IASBS at SemEval-2025 Task 11: Ensembling Transformers for Bridging the Gap in Text-Based Emotion Detection

Erfan Mohammadzadeh[†], **Aydin Mohandesi**[†], **Mehrzad Tareh**[†], **Ebrahim Ansari**
Institute for Advanced Studies in Basic Sciences (IASBS)
{e.mohammadzadeh, mohandesi, m.tareh, ansari}@iasbs.ac.ir
[†] denotes equal contribution

Abstract

This paper addresses the SemEval 2025 - Task 11 entitled "Bridging the Gap in Text-Based Emotion Detection", with a focus on perceived emotions from short text snippets, as spoken by individuals across various languages. This study involves three tracks: (1) multi-label emotion detection (ED), (2) emotion intensity prediction, and (3) cross-lingual ED. A comprehensive analysis of multiple languages, including Arabic, Amharic, Chinese, English, and other languages, is conducted utilizing a variety of machine and deep learning techniques. For Track A, a hybrid approach using an ensemble of advanced pre-trained transformer models, coupled with majority voting on predictions, yields significant insights. Track B leverages a multilingual transformer model alongside prompt engineering, using Average Ensemble Voting (AEV) for emotion intensity prediction. In Track C, a cross-lingual ED task, a classification of languages based on linguistic families is employed to enhance the performance of multilingual models. The methodologies incorporated, such as model selection, prompt engineering, and voting mechanisms, are evaluated using F1-score and Pearson correlation metrics. This research contributes to the broader field of ED, highlighting the importance of cross-lingual approaches and model optimization for accurate emotion prediction across diverse linguistic landscapes.

1 Introduction

Text-based emotion detection has become a critical task in the field of natural language processing (NLP), enabling systems to understand and respond to human emotions based on written language. The ability to accurately infer the emotions of speakers from their text offers vast potential in applications ranging from customer service automation to mental health monitoring and Sentiment analysis (SA) in social media (Saadati et al., 2024). However, despite significant advancements, this task remains challenging due to the complexity of emotional expressions across cultures, contexts, and languages. The present task (Muhammad et al., 2025b), "Bridging the Gap in Text-Based Emotion Detection", focuses on understanding the perceived emotions of a speaker through short text snippets. Unlike traditional SA, which often centers on the emotional impact on the reader or the identification of the speaker's true emotions—both of which are difficult to ascertain from a limited text—the focus here is on detecting the emotion that is most likely perceived by an average reader or listener. Specifically, the task aims to identify emotions such as joy, sadness, fear, anger, surprise, and disgust from a variety of languages, including widely spoken ones like English, Spanish, and Chinese, as well as lesser-studied languages such as Emakhuwa, Igbo, and Hausa.

The task is organized into three distinct tracks: Track A (Multi-label Emotion Detection), Track B (Emotion Intensity), and Track C (Cross-lingual Emotion Detection). Track A requires systems to predict the presence or absence of specific emotions in text, encompassing the prediction of one or more perceived emotions in a given text snippet. Track B involves determining the intensity of the identified emotions, ranging from no emotion to high degrees of emotion. Track C challenges systems to generalize emotion detection across multiple languages, testing the models' ability to predict emotion labels for unseen languages based on training data.

The challenge further emphasizes the importance of handling multilingual data and the complexities of translating emotional expressions across languages and traditions. With this in mind, various methodologies have been explored, including the use of state-of-the-art models such as Microsoft's DeBERTa (Chehreh et al., 2024), Google's Multilingual DistilBERT (M-DistilBERT) (Sanh et al., 2019), and Google's Gemini (Reid et al., 2024). These models fine-

Proceedings of the The 19th International Workshop on Semantic Evaluation (SemEval-2025), pages 695–702
July 31 - August 1, 2025 ©2025 Association for Computational Linguistics

tuned for different tracks, and employ techniques such as translation, prompt engineering, and model ensemble strategies like Majority Voting and AEV, to enhance performance and address the unique challenges presented by the diverse language sets involved.

The paper is organized into seven sections: Section 2 covers the dataset, Section 3 reviews related work, and Section 4 presents the system architecture. Section 5 describes the experimental setup, followed by Section 6, which presents and analyzes the results. Section 7 concludes with key findings and future research directions. This paper aims to present the methodology, evaluation strategies, and results of our approach to tackling the multifaceted problem of text-based emotion detection, while also exploring how advancements in multilingual NLP can bridge the gap in understanding human emotions across different languages.

2 Dataset

The shared task utilizes a comprehensive and multilingual dataset designed to explore how emotions are perceived across different languages. The main objective is to determine which emotion people are most likely to associate with a speaker based on a brief text snippet. Provided by the organizers of SemEval 2025 - Task 11 on Text-Based Emotion Detection, this dataset covers multiple languages and captures a broad range of linguistic and cultural nuances. For all languages except Amharic, Oromo, Somali, and Tigrinya, we have used the dataset referenced in (Muhammad et al., 2025a). However, for any analysis involving one of these four languages, we have used the dataset from (Belay et al., 2025) instead. This distinction ensures that the dataset remains culturally and contextually appropriate for each language. The emotions considered in this study include anger, disgust, fear, joy, sadness, and surprise. In track A, 28 languages are covered. For track B, 11 languages are included. In track C, 32 languages are available. Our output for tracks A and C must be in a binary format (0 or 1), as 1 indicating the presence and 0 indicating the absence of a given emotion in the text. In track B, however, the output must be one of the integers from 0 to 3 for each emotion, representing its intensity level (0: no emotion, 1: low degree of emotion, 2: moderate degree of emotion, and 3: high degree of emotion). The dataset is been pre-processed by the task organizers and is divided into three key subsets: training, validation, and testing.

3 Related Work

Sentiment analysis identifies the overall emotional tone of a text (positive, negative, or neutral), while ED focuses on identifying specific emotions. SA is broader, while ED provides more detailed and intense emotional insights. Emotion detection from text, particularly perceived emotions, is a rich field of research in NLP, playing a crucial role in analyzing human expressions and understanding people's attitudes toward specific subjects. Several studies have focused on detecting emotions based on textual cues, with a variety of approaches ranging from lexicon-based methods to advanced machine and deep learning techniques. However, the challenge of bridging the gap between text and the perception of emotions across different languages, cultures, and contexts remains an open issue. In this section, we review relevant work related to text-based emotion detection, emphasizing approaches related to multi-label emotion detection, emotion intensity prediction, and cross-lingual emotion detection.

3.1 Text-Based Emotion Detection

Early emotion detection methods primarily relied on lexical resources like the Emotion Lexicon (EmoLex), which linked words to specific emotion categories, but these approaches struggled to capture the complexity of mixed emotions or irony (Doan and Luu, 2022). With the rise of deep learning, neural network-based models such as RNNs, CNNs, and Transformer architectures like BERT and its multilingual variants have significantly improved ED by better understanding contextual nuances in the text (Rezapour, 2024). A key challenge in this area is multi-label classification, where texts may convey multiple emotions simultaneously; recent studies (Ameer et al., 2023) have explored frameworks for predicting multiple emotions from a single sentence, showing notable advancements over single-label methods.

3.2 Emotion Intensity Detection

Recent advancements in emotion intensity prediction have gained significant attention in NLP, with datasets like Emotion Intensity (Kajiwara et al., 2021) underscoring its growing importance. This task has become a critical component of emotion recognition, especially in NLP challenges like SemEval. Notably, models like LE-PC-DNN combine convolutional and fully connected layers with lexicon-based features and transfer learning to predict emotion intensity in tweets, aiming for state-of-the-art performance through deep multi-task learning (Kulshreshtha et al., 2018). Similarly, the CrystalFeel system leverages parts-of-speech, n-grams, word embeddings, and affective lexicons to predict intensity levels for emotions, achieving strong results while revealing insightful word- and message-level associations (Gupta and Yang, 2018). These innovations reflect the increasing sophistication of emotion intensity prediction, combining deep learning and linguistic features for more accurate and efficient emotion analysis in text.

3.3 Cross-Lingual Emotion Detection

Cross-lingual emotion detection is an emerging field within emotion detection, where models are designed to generalize across languages with diverse structures, cultural contexts, and expressions. This contrasts with traditional systems that typically rely on a single language or closely related language sets. Research in this area has investigated the transfer of models across different languages. For example, the work of (Kadiyala, 2024) highlights performance drops when training and testing data originate from different languages. A key approach to improving cross-lingual emotion detection is the use of multilingual models like Google's Multilingual BERT (M-BERT), XLM-RoBERTa (XLM-R), and multilingual T5, which have demonstrated stronger performance in multilingual tasks. For example, (Hassan et al., 2022) adapted M-BERT (Devlin et al., 2018) model for cross-lingual emotion detection, leveraging shared embeddings across languages to achieve competitive results. However, challenges remain, including cultural differences in emotional expression and the need for large, multilingual labeled datasets.

4 System Architecture

The system architecture for multilingual emotion detection uses back translation, multilingual models, transfer learning, LLMs to improve accuracy across languages. Prompt engineering optimizes task processing, ensuring effective and adaptable ED across diverse datasets.

4.1 Back Translation

In our paper for SemEval 2024 - Task 10 (Tareh et al., 2024), we utilized back translation, a widely adopted technique in multilingual NLP, which contributed significantly to achieving the best performance. This method ensures that the meaning of the original text is preserved across languages, particularly in ED tasks. By addressing challenges in handling multilingual data, especially when training models on text from various languages, we were able to improve the model's robustness. The back translation process involves translating the text from the target language to English and then back to the original language, creating a parallel corpus that helps identify any inconsistencies. This technique was crucial in enhancing the accuracy of our ED models, as it helped pinpoint misinterpretations or discrepancies during translation, as detailed in our paper.

For this particular task, back translation was incorporated into the data preprocessing pipeline. Text from various languages was first translated into English using the Llama3.3-70b[1] translation model, which, with its 70 billion parameters, was selected for its effectiveness with a broad range of languages, including those from underrepresented language families. Once the data was in English, it was back-translated to the original language, allowing for a comparison between the original and retranslated texts. This comparison helped detect errors in the translation, ensuring that emotional expressions were accurately preserved and improving the overall quality of the data used for training the ED models (Wendler et al., 2024).

Despite its benefits, back translation has limitations. The quality of the back-translation depends on the initial translation model and the characteristics of the language pairs involved, especially when languages have different sentence structures or cultural contexts. Additionally, the process is computationally expensive and time-consuming, particularly when dealing with large datasets across many languages. However, the advantages of preserving emotional content and reducing translation biases outweighed these challenges, making back translation an essential technique for enhancing multilingual ED models and ensuring more accurate predictions in cross-lingual tasks (Yoon, 2022).

4.2 Multilingual Models and Transfer Learning

In recent years, multilingual transformer models have become the standard for addressing cross-lingual emotion detection. For example, M-BERT and M-DistillBERT have been used for several ED

[1] https://developers.cloudflare.com

tasks, including cross-lingual tasks, due to their ability to handle multiple languages simultaneously and share knowledge across languages. In the context of the present work, models like DeBERTa (Aziz et al., 2023) and Gemini-1.5-flash have been used to tackle the challenges of multi-label classification, emotion intensity, and cross-lingual detection, demonstrating the effectiveness of leveraging pre-trained language models fine-tuned with task-specific datasets. Transfer learning, which involves fine-tuning pre-trained models on task-specific datasets, has also been widely used to bridge the gap between languages. This technique has been especially useful in tackling the challenge of cross-lingual emotion detection, where training data may not be available for all languages. (Mozhdehi and Moghadam, 2023) investigated the impact of transfer learning on cross-lingual performance, demonstrating that fine-tuning multilingual models with domain-specific data enhances their effectiveness.

4.3 Prompt Engineering

Prompt engineering is a key approach for optimizing large language models (LLMs) to perform various tasks effectively. It involves creating well-structured prompts that guide LLMs to generate accurate and relevant responses. This includes clear task descriptions, well-organized input data, and defined output formats. The goal is to enhance the LLM's ability to process and complete tasks, especially through in-context learning, where instructions and examples are provided in natural language (Brown et al., 2020). Effective prompt engineering also requires careful structuring of input data, particularly for specialized formats like knowledge graphs, and ensuring the output format aligns with expectations (Zhang et al., 2023).

As an emerging field, prompt engineering plays a significant role in improving LLM performance across various applications, including vision-language tasks, SA, and academic research. Key prompting techniques, such as few-shot learning, and chain-of-thought, enhance reasoning and task-specific accuracy (Chen et al., 2024). While prompt engineering offers substantial benefits, it also presents challenges, such as ambiguity, bias, and issues of generalizability. As LLMs continue to be integrated into multiple domains, including healthcare and scientific research, mastering prompt engineering has become increasingly important for professionals aiming to leverage AI for enhanced problem-solving and workflow efficiency (Lamba, 2024).

5 Experimental Setup

The experimental setup includes multiple stages aimed at optimizing performance across multilingual tasks, focusing on model transformation, fine-tuning, integration, and strategies like prompt engineering and ensemble learning. Here's an overview of the steps and methods used.

5.1 Text Translation for Multilingual Data

Due to the dataset's diverse linguistic nature, we utilize Llama3.3, a powerful multilingual language model, to translate all text into English. Standardizing the training language enables us to leverage a unified model instead of training separate models for each language, significantly improving efficiency and consistency.

5.2 Model Selection and Fine-Tuning

We carefully select and fine-tune pre-trained transformer models for each track, ensuring they are optimized to meet the specific requirements of the task. Figure 1 shows the architecture.

- **Track A:** Fine-tuned DeBERTa—enhances BERT and RoBERTa with disentangled attention and an improved mask decoder for better efficiency and performance—recognized for its strong emotion classification capabilities.

- **Track B:** Fine-tuned M-DistilBERT—a lightweight yet effective model optimized for multilingual tasks. It is trained on a concatenation of Wikipedia data in 104 languages and features 6 layers, 768 dimensions, and 12 attention heads, totaling 134M parameters.

- **Track C:** Fine-tuned M-BERT and DistilBERT, capable of learning cross-lingual representations.

In tracks A and B, each model is trained on the translated English dataset, with the validation set used for hyperparameter tuning to maximize performance. In track C, We categorize the languages based on linguistic families and their relevance in NLP research. The languages are grouped into seven categories based on linguistic families and their relevance to NLP, as shown in Table 1.

5.3 Prompt Engineering with Gemini

The Gemini family represents a significant advancement in multimodal AI models, offering impressive capabilities across image, audio, video, and text understanding. For our ED task, we deployed Gemini-1.5-flash-002, a more efficient alternative to the computationally intensive Pro version, while still maintaining the 2M+ context length and multimodal capabilities. This transformer decoder model is specifically optimized for tensor processing units (TPUs), featuring parallel computation of attention and feedforward components, online distillation from Gemini-1.5-pro, and training with higher-order preconditioned methods (Reid et al., 2024). We interfaced with the model through its API, which facilitated efficient request handling and response storage. To maximize performance on our emotion tasks, we configured all safety parameters to "*None*", enabling unrestricted access to the model's full potential during inference while balancing performance and cost-effectiveness.

5.4 Ensemble Strategy with Voting Mechanism

To improve robustness and reduce model biases, we leverage ensemble learning techniques:

- **Majority Voting Mechanism for Track A:** DeBERTa, Gemini-1.5, and SVM generate independent predictions, and a majority vote determines the final emotion labels, ensuring balanced and unbiased results. Refer to Equation 1 for the majority voting mechanism formula.

- **AEV Mechanism for Tracks B and C:** Since emotion intensity prediction and cross-lingual detection require nuanced outputs, we apply performance-based weighting, prioritizing predictions from the model with the highest validation accuracy. Refer to Equation 2 for the AEV mechanism formula.

By leveraging transformer models, prompt engineering, and ensemble techniques, our approach ensures accurate, multilingual emotion detection, enhancing both efficiency and generalization across diverse texts.

6 Results

In this section, we present the evaluation results for the three subtasks across multiple languages. The performance of each model is reported using F1-scores and Pearson correlation scores, as appropriate for each subtask. Additionally, we discuss the rankings and highlight key insights drawn from the results.

Track A evaluates the performance of various models, such as DeBERTa, SVM, and Gemini-1.5, and investigates the effectiveness of the majority voting mechanism, which combines multiple strategies, across different languages. The F1-scores for each approach are reported in Table 3.

Track B evaluates the performance of M-DistilBERT, Gemini-1.5, and an AEV strategy using Pearson correlation scores. The results are summarized in Table 4. The AEV strategy yielded the best performance in 7 out of 11 languages, indicating that a hybrid model can enhance generalization.

Track C investigates the ability of models to generalize across languages using DistilBERT, M-BERT, and an AEV strategy. Although the F1-scores of M-BERT and DistilBERT in Table 5 were generally comparable, their predictions showed significant variability in certain cases. M-BERT consistently outperformed DistilBERT in most languages, highlighting its robust performance in cross-lingual tasks. However, there were instances where DistilBERT provided superior results. To leverage the strengths of both models and improve overall performance, we implemented a voting mechanism to combine their predictions.

7 Conclusion

This study addresses the challenges of text-based emotion detection across multiple languages, focusing on multi-label classification, emotion intensity prediction, and cross-lingual detection. By fine-tuning advanced NLP models (Gemini, DeBERTa, M-BERT, M-DistillBERT) and combining them with traditional methods like SVM, strong results were achieved. A voting-based ensemble method enhanced model reliability. The approach demonstrated the effectiveness of multilingual models, especially for low-resource languages, ranking in the top 10 teams of the competition. However, improvements are needed in emotion intensity prediction and for low-resource languages. Future work will refine model architectures, fine-tune LLMs, and explore new techniques like RAG and CAG for further enhancement.

References

Iqra Ameer, Necva Bölücü, Muhammad Hammad Fahim Siddiqui, Burcu Can, Grigori Sidorov, and Alexander Gelbukh. 2023. Multi-label emotion classification in texts using transfer learning. *Expert Systems with Applications*, 213:118534.

Abdul Aziz, Md. Akram Hossain, and Abu Nowshed Chy. 2023. CSECU-DSG at SemEval-2023 task 4: Fine-tuning DeBERTa transformer model with cross-fold training and multi-sample dropout for human values identification. In *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*, pages 1988–1994, 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. 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 Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Isun Chehreh, Farzaneh Saadati, Ebrahim Ansari, and Bahram Sadeghi Bigham. 2024. Enhanced multi-label question tagging on stack overflow: A two-stage clustering and deberta-based approach. In *Proceedings of the 36th Conference of Open Innovations Association FRUCT*, pages 858–863, Helsinki, Finland. FRUCT Oy.

Banghao Chen, Zhaofeng Zhang, Nicolas Langrené, and Shengxin Zhu. 2024. Unleashing the potential of prompt engineering in large language models: a comprehensive review. *Preprint*, arXiv:2310.14735.

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.

An Long Doan and Son T. Luu. 2022. Improving sentiment analysis by emotion lexicon approach on vietnamese texts. In *2022 International Conference on Asian Language Processing (IALP)*, pages 39–44.

Raj Kumar Gupta and Yinping Yang. 2018. CrystalFeel at SemEval-2018 task 1: Understanding and detecting emotion intensity using affective lexicons. In *Proceedings of the 12th International Workshop on Semantic Evaluation*, pages 256–263, New Orleans, Louisiana. Association for Computational Linguistics.

Sabit Hassan, Shaden Shaar, and Kareem Darwish. 2022. Cross-lingual emotion detection. *Preprint*, arXiv:2106.06017.

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.

Tomoyuki Kajiwara, Chenhui Chu, Noriko Takemura, Yuta Nakashima, and Hajime Nagahara. 2021. WRIME: A new dataset for emotional intensity estimation with subjective and objective annotations. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2095–2104, Online. Association for Computational Linguistics.

Devang Kulshreshtha, Pranav Goel, and Anil Kumar Singh. 2018. How emotional are you? neural architectures for emotion intensity prediction in microblogs. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 2914–2926, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Divya Lamba. 2024. The role of prompt engineering in improving language understanding and generation. *International Journal For Multidisciplinary Research*.

Mahsa Mozhdehi and AmirMasoud Moghadam. 2023. Textual emotion detection utilizing a transfer learning approach. *The Journal of Supercomputing*, 79:1–15.

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.

Machel Reid, Nikolay Savinov, Denis Teplyashin, Dmitry Lepikhin, Timothy Lillicrap, Jean-baptiste Alayrac, Radu Soricut, Angeliki Lazaridou, Orhan Firat, Julian Schrittwieser, et al. 2024. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*.

Mahdi Rezapour. 2024. Emotion detection with transformers: A comparative study. *Preprint*, arXiv:2403.15454.

Farzaneh Saadati, Isun Chehreh, and Ebrahim Ansari. 2024. The role of social media platforms in spreading misinformation targeting specific racial and ethnic groups: A brief review. In *Proceedings of the 36th Conference of Open Innovations Association FRUCT*, pages 892–902, Helsinki, Finland. FRUCT Oy.

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

Mehrzad Tareh, Aydin Mohandesi, and Ebrahim Ansari. 2024. IASBS at SemEval-2024 task 10: Delving into emotion discovery and reasoning in code-mixed conversations. In *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*, pages 1229–1238, Mexico City, Mexico. Association for Computational Linguistics.

Chris Wendler, Veniamin Veselovsky, Giovanni Monea, and Robert West. 2024. Do llamas work in english? on the latent language of multilingual transformers. *Preprint*, arXiv:2402.10588.

Yeo Yoon. 2022. Can we exploit all datasets? multimodal emotion recognition using cross-modal translation. *IEEE Access*, 10:1–1.

Wenxuan Zhang, Yue Deng, Bing Liu, Sinno Jialin Pan, and Lidong Bing. 2023. Sentiment analysis in the era of large language models: A reality check. *Preprint*, arXiv:2305.15005.

A Language Groups for Track C

The languages are grouped as follows:

Language Groups
Group 1: English, German, Swedish, Afrikaans
Group 2: Spanish, Portuguese, Romanian
Group 3: Russian, Ukrainian
Group 4: Hindi, Marathi
Group 5: Chinese
Group 6: Arabic
Group 7: Hausa

Table 1: Categorization of languages based on linguistic families and NLP relevance

B Majority Voting Formula

The formula calculates the predicted emotion detection $\hat{y}_i^{(e)}$ based on the majority agreement from three models. If at least two models predict 1, the final prediction is 1; otherwise, the prediction is 0.

$$\hat{y}_i^{(e)} = \begin{cases} 1 & \text{if } y_{i,\text{model1}}^{(e)} + y_{i,\text{model2}}^{(e)} + y_{i,\text{model3}}^{(e)} \geq 2 \\ 0 & \text{otherwise} \end{cases} \tag{1}$$

C Average Ensemble Voting Formula

The AEV formula calculates the predicted emotion intensity $\hat{y}_i^{(e)}$ by averaging the outputs from two models and rounding to the nearest integer.

$$\hat{y}_i^{(e)} = \left\lfloor \frac{y_{i,\text{model1}}^{(e)} + y_{i,\text{model2}}^{(e)}}{2} + 0.5 \right\rfloor \tag{2}$$

D Hyperparameters

The hyperparameters for the model were set to optimize performance and training stability, as shown in Table 2.

Hyperparameter	**Value**
Seed	42
Batch size	16
Weight decay	0.01
Learning rate	2e-5
Warmup ratio	0.06
Warmup steps	500
Max sequence length	128
Number of training epochs	11
temperature	0.1
candidate count	1
max output tokens	500

Table 2: System hyperparameter settings

E Findings

These results highlight the impact of different architectures and voting techniques on multilingual NLP tasks.

Language	DeBERTa	SVM	Gemini	Majority Voting
Hindi	0.7743	0.7332	0.8108	0.7712
Russian	0.7683	0.7793	0.8224	0.7945
English	0.7280	0.5210	0.6930	0.7351
Romanian	0.7116	0.6159	0.6596	0.7255
Spanish	0.6871	0.6569	0.7279	0.7198
Marathi	0.6745	0.7384	0.8006	0.7995
German	0.5847	0.4378	0.5673	0.6137
Ukrainian	0.5389	0.4103	0.5884	0.5666
Swedish	0.5118	0.4347	0.5131	0.5296
Chinese	0.5025	0.4097	0.5421	0.5537
Hausa	0.4694	0.6282	0.5743	0.6582
Afrikaans	0.4395	0.3110	0.5617	0.5906
Igbo	0.3950	0.5442	0.3600	0.5155
Amharic	0.3128	0.4716	0.5232	0.4666

Table 3: F1-scores for Track A in different languages and models

Language	M-DistilBERT	Gemini	AEV
Russian	0.765	0.7795	0.8599
English	0.5492	0.6926	0.6670
Romanian	0.5323	0.611	0.6006
Spanish	0.589	0.6429	0.6922
German	0.4504	0.5973	0.5634
Ukrainian	0.4556	0.5267	0.5648
Chinese	0.4135	0.5055	0.5188
Hausa	0.4378	0.5225	0.6575
Amharic	0.3199	0.4612	0.4612
Portuguese	0.3979	0.5144	0.5046
Arabic	0.2702	0.4612	0.4514

Table 4: Average Pearson correlation for different models and languages

F System Configuration

This section details the hardware and software specifications of the system used for the experiments. The following tables summarize the CPU, RAM, GPU, and operating system configurations:

G Model Architecture

The model combines Transformer-based models (DeBERTa, DistilBERT, Gemini, M-Bert) and SVM with ensemble methods (voting, averaging) for multilingual NLP tasks.

Language	DistilBERT	M-BERT	AEV
English	0.5799	0.6062	0.6571
German	0.4379	0.4569	0.4551
Swedish	0.4569	0.4521	0.4707
Afrikaans	0.3179	0.3200	0.3187
Spanish	0.6559	0.6780	0.6960
Portuguese	0.3745	0.3920	0.3540
Romanian	0.6162	0.6276	0.6384
Russian	0.7830	0.8095	0.8167
Ukrainian	0.4852	0.5075	0.4945
Hindi	0.7217	0.7584	0.7643
Marathi	0.7382	0.7736	0.7673
Chinese	0.5291	0.5344	0.5434
Arabic	0.4274	0.4780	0.4370
Hausa	0.5784	0.5677	0.5982

Table 5: F1-scores for Track C in different languages and models

Type	KEY	VALUE
CPU	Model	Intel(R) Xeon(R) E5-2620 v4
	Frequency	2.10 GHz
	On-line CPU(s) list	16
	Sockets	1
	Core(s) per socket	8
	Thread(s) per core	2
	Op-mode	64-bit
RAM	Block Size	128 MB
	Total Capacity	16 GB
GPU	Brand	NVIDIA
	Model	RTX 2080 Rev. A
	Memory	8 GB

Table 6: System Specifications

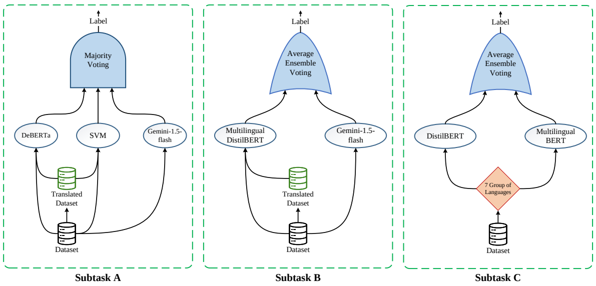

Figure 1: Model architecture with layers, transformers, and ensemble methods.