Habib University at SemEval-2025 Task 11: Bridging the Gap in Text-Based Emotion Detection

Owais Waheed[†*], **Hammad Sajid**[†], **Muhammad Areeb Kazmi**[†], **Kushal Chandani**[†],
Sandesh Kumar[†], Abdul Samad[†]
[†]Habib University, Dhanani School of Science & Engineering, Pakistan
{ow07611, hs07606, mk07202, kc07535}@st.habib.edu.pk,
{sandesh.kumar, abdul.samad}@sse.habib.edu.pk

Abstract

Emotion detection in text has emerged as a pivotal challenge in Natural Language Processing (NLP), particularly in multilingual and cross-lingual contexts. This paper presents our participation in SemEval 2025 Task 11, focusing on three subtasks: Multi-label Emotion Detection, Emotion Intensity Prediction, and Cross-lingual Emotion Detection. Leveraging state-of-the-art transformer models such as BERT and XLM-RoBERTa, we implemented baseline models and ensemble techniques to enhance predictive accuracy. Additionally, innovative approaches like data augmentation and translation-based cross-lingual emotion detection were used to address linguistic and class imbalances. Our results demonstrated significant improvements in F1 scores and Pearson correlations, showcasing the effectiveness of ensemble learning and transformer-based architectures in emotion recognition. This work advances the field by providing robust methods for emotion detection, particularly in low-resource and multilingual settings.

1 Introduction

Emotion detection in text has become an essential task in Natural Language Processing (NLP), particularly with the rapid growth of digital communication and social media. Identifying emotions accurately in textual data can enhance applications such as mental health monitoring, customer service, and user sentiment analysis. However, this task is complicated by the subtlety and complexity of emotional expression across languages and cultures. While traditional machine learning methods have provided baseline solutions, recent advances in deep learning and transformer models, like BERT, have shown promise in improving emotion recognition capabilities.

This research builds on this foundation by focusing on the challenges presented in SemEval Task 11. Our work addresses three primary subtasks as defined in SemEval 2025 Task 11: (1) **Multi-label Emotion Detection** (Track A), which involves detecting multiple emotions such as joy, sadness, fear, anger, and surprise from text snippets; (2) **Emotion Intensity Prediction** (Track B), where the goal is to predict the intensity of each emotion on an ordinal scale from 0 (none) to 3 (high); and (3) **Cross-lingual Emotion Detection** (Track C), which extends emotion detection to multilingual settings, introducing an additional emotion category (disgust) in certain languages (Muhammad et al., 2025b). By participating in this competitive NLP task, we aim to identify the most effective models and techniques for emotion detection, including cross-lingual approaches that address the diversity of emotional expression across languages. The complexity of emotion recognition lies not only in identifying emotions but also in accurately determining their intensity, making this task both practical and academically valuable.

2 Related Works

Recent work in emotion recognition has tackled several interrelated challenges. In the domain of Emotion Recognition in Conversations (ERC), tasks such as SemEval 2024 Task 10 introduced the Emotion Discovery and Reasoning its Flip in Conversation (EDiReF) challenge (Kumar et al., 2024), which explored recognizing emotion transitions in both English and code-mixed Hindi-English dialogues. Approaches by teams like UMUTeam (Pan et al., 2024) demonstrated the effectiveness of fine-tuning pre-trained transformers (e.g., BERT) on annotated conversational datasets despite challenges such as subtle emotional shifts.

Parallel efforts have focused on cross-lingual emotion detection, where researchers have leveraged multilingual models like mBERT

[*]corresponding author

Proceedings of the The 19th International Workshop on Semantic Evaluation (SemEval-2025), pages 1656–1661
July 31 - August 1, 2025 ©2025 Association for Computational Linguistics

and XLM-RoBERTa to predict emotions across languages (Wang et al., 2024). Models such as those proposed by Wadhawan and Aggarwal (Wadhawan and Aggarwal, 2021) have shown state-of-the-art performance in code-mixed texts, reinforcing the importance of multilingual training for tasks like Task 11.

Further, emotional flip reasoning (EFR) – the process of identifying trigger utterances responsible for emotional shifts – has been investigated to enhance conversational agents' empathy and contextual understanding (Kumar et al., 2024; Pan et al., 2024). Ensemble approaches, as seen in MasonTigers' work on semantic textual relatedness (Goswami et al., 2024), also indicate that combining multiple models can lead to more robust performance. Additional datasets such as GoEmotions (Demszky et al., 2021) and SemEval 2018 Task 1 (Mohammad et al., 2018) have contributed fine-grained emotion intensity labels, providing further context to our investigations.

3 System Overview

3.1 Dataset overview:

The dataset used for these tasks is **BRIGHTER** dataset which is a collection of multi-labeled emotion-annotated datasets in 28 different languages. (Muhammad et al., 2025a) There are 5 emotion labels for some languages that is joy, sadness, fear, anger, and surprise. Whereas disgust is labeled as a sixth emotion in other languages

Tables 1, 2, and 3 give a glimpse of input/output formats and emotion labels for each track.

Id	Text	Anger	Fear	Joy	Sadness	Surprise
sample_01	Never saw him again.	0	0	0	1	0
sample_02	I love telling this story.	0	0	1	0	0

Table 1: Track A: Multi-label Emotion Detection Format (binary labels).

Id	Text	Anger	Fear	Joy	Sadness	Surprise
sample_01	Never saw him again.	0	0	0	2	0
sample_02	I love telling this story.	0	0	2	0	0

Table 2: Track B: Emotion Intensity Prediction Format (ordinal labels 0–3).

Text	Anger	Disgust	Fear	Joy	Sadness	Surprise
Auf die Frage an Präsident Biden.	1	0	0	0	0	0
Sind Organe von adipösen Menschen	0	0	0	0	0	0

Table 3: Track C: Cross-lingual Emotion Detection Format (additional *disgust* label).

3.2 Track A: Multi-label Emotion Detection

For Track A, we initially trained the BERT-base-uncased model (110M parameters) as a baseline due to its strong bidirectional context understanding. The training utilized standard hyperparameters. To improve performance, we adopted an ensemble learning approach, combining multiple models:

- **DistilBERT-base**: A faster and more efficient version of BERT, optimized for NLP tasks.
- **DeBERTa-base**: Enhances NLP task accuracy through improved attention mechanisms.
- **XLM RoBERTa-base**: A multilingual model designed to improve classification performance across diverse languages.

Ensembling these models improved robustness by leveraging their individual strengths. For only English language, We did data augmentation. Initially, we used the Gemini API for dataset augmentation, but due to its request limits, we transitioned to GPT-3.5 Turbo API, enabling scalable data augmentation. We also applied class upsampling to mitigate class imbalance, increasing dataset size from 2800 to 5000 rows, though perfect balance was not achieved. We also applied NLP data augmentation techniques to enhance our dataset and improve model generalization. Specifically, we used synonym replacement and back translation as augmentation methods. Synonym replacement involved substituting words with their synonyms while preserving the overall meaning of the text, helping the model learn different lexical variations. Back translation was used to translate text into another language and then back into the original language, introducing natural variations while maintaining the original context. These techniques increased the diversity of our training data, and improving the model's ability to generalize to unseen examples. For implementation, we used the nlpaug library, which provided efficient and flexible augmentation methods for both synonym replacement and back translation, ensuring high-quality transformations of textual data.

3.3 Track B: Emotion Intensity Regression

For Track B, we used the **BERT-base-uncased** model to encode text snippets. A regression layer was added on top of the encoder to predict the intensity values directly in the range of 0 to 3. For this purpose, we employed standard Mean Squared Error (MSE) loss:

$$\mathcal{L} = \frac{1}{N} \sum_{i=1}^{N} (\hat{y}_i - y_i)^2 \tag{1}$$

where $\hat{y}_i$ is the predicted intensity score and y_i is the ground truth. To map these continuous predictions to the discrete labels (0, 1, 2, 3), the values were rounded to the nearest integer. We then applied an ensemble strategy, selecting the top three models based on Mean Squared Error (MSE):

- **DeBERTa-v3-base**
- **DeBERTa-v3-large**
- **RoBERTa-base**

For multilingual extension, we followed Track C's translation pipeline, training each language-specific model separately. To deal with class imbalance, We also experimented with assigning class weights based on inverse frequency for two languages (German and Portuguese), leading to improved performance.

The approach for Track C mirrored Track A with slight modifications. The models trained included **DistilBERT-base-multilingual**, **Multilingual-BERT-base-cased**, and **XLM RoBERTa-base**

To enhance cross-lingual performance, we introduced a translation-based method as shown in Figure 1.

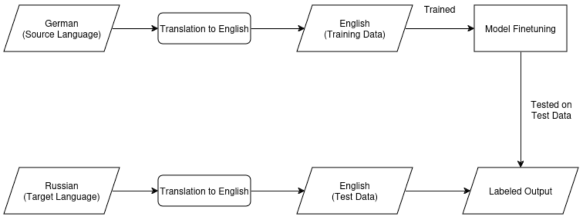

Figure 1: Cross-lingual Emotion Detection Using Translation

This involved translating training data into English using Facebook's Seamless M4T-v2-large model before training, followed by translating target languages into English before classification. We strategically paired languages based on semantic similarities as shown in Table 4. The pairings of target and training languages for the cross-lingual emotion detection task are justified based on linguistic similarity, cultural proximity, and available training data. For example, Afrikaans and German, both Germanic languages, share vocabulary and grammar, which aids cross-lingual transfer. Similarly, Amharic and Somali, despite being from different language families, share a regional and cultural context, allowing for similar emotional expressions. Igbo and Hausa, spoken in Nigeria, also have cultural overlap, making emotional expression transfer feasible. Somali and Amharic, geographically and culturally close, exhibit shared emotional expression patterns. Chinese and Latin American Spanish, spoken by large and diverse populations, offer rich data for training cross-lingual models. Swahili and Somali, both from East Africa, share cultural context and emotional expression similarities. Hindi and Marathi, as Indo-Aryan languages with shared vocabulary and grammar, also have similar cultural expressions, making them ideal pairings. The reverse pairing of Marathi and Hindi is equally valid due to their linguistic and cultural overlap. German and Swedish, both Germanic languages, share syntax and cultural context in Northern Europe, which facilitates cross-lingual emotion detection. Latin American Spanish and Romanian have similar communicative styles due to their Romance language roots, making them suitable for cross-lingual models. The reverse pairing of Romanian and Latin American Spanish works for the same reasons. Russian and Ukrainian, both Slavic languages, have high lexical similarity and cultural overlap, which supports their use in emotion detection. Similarly, Swedish and German, both Germanic languages, share similar structures and emotional expression. Finally, Ukrainian and Russian, with significant linguistic and cultural similarities, provide a strong basis for cross-lingual emotion detection. These pairings are based on shared linguistic features, cultural contexts, and available resources, facilitating effective emotion detection across languages.

3.4 Resources Beyond Training Data

- **Lexicons and Augmentation Tools**: Gemini API and GPT-3.5 Turbo API for dataset enrichment.
- **Translation Pipeline**: Facebook's Seamless M4T-v2-large for cross-lingual adaptation.
- **Language Pairing Strategy**: Matching target and training languages to optimize cross-lingual performance (Table 4).

Target Language	Best Training Language
Afrikaans	German
Amharic	Somali
Igbo	Hausa
Somali	Amharic
Chinese	Latin American Spanish
Swahili	Somali
Hindi	Marathi
Marathi	Hindi
German	Swedish
Latin American Spanish	Romanian
Romanian	Latin American Spanish
Russian	Ukrainian
Swedish	German
Ukrainian	Russian

Table 4: Best Training Language for Each Language Pair

4 Experimental Setup

Our experiments were conducted separately for each track, leveraging transformer-based models and ensemble learning strategies. We utilized the provided training and validation sets, with the validation set incorporated into training for the final submission. Preprocessing, hyperparameter tuning, and ensemble strategies varied across tasks.

4.1 Track A: Multi-label Emotion Detection

For the baseline, we fine-tuned a **BERT-base-uncased** model (110M parameters) using common hyperparameters (4 epochs, batch size 16, learning rate 5×10^{-5}). To improve performance, we employed an ensemble of:

- **DistilBERT-base-uncased**
- **DeBERTa-base**
- **XLM-RoBERTa-base**

Ensembling was performed by averaging the output logits from each model. Table 5 summarizes the training hyperparameters.

Model(s)	Epochs	Batch Size	Learning Rate
BERT-base-uncased	4	16	5×10^{-5}
Ensembled (DistilBERT, DeBERTa, XLM-RoBERTa)	8	16	5×10^{-5}

Table 5: Training hyperparameters for Track A.

4.2 Track B: Emotion Intensity Prediction

For Track B, the baseline used the **BERT-base-uncased** encoder with an added regression layer to predict continuous intensity values (0–3), which were then rounded. An ensemble of five models was trained, and the top three (including **DeBERTa-v3-base**, **DeBERTa-v3-large**, and **RoBERTa-base**) was selected based on Mean Squared Error (MSE). Table 6 lists the training losses and MSE for the evaluated models.

Model	Training Loss	MSE
microsoft/deberta-v3-base	0.12	0.22
microsoft/deberta-v3-large	0.04	0.23
roberta-base	0.08	0.25
bert-base-uncased	0.07	0.27
distilbert-base-uncased	0.13	0.29
FacebookAI/xlm-roberta-base	0.40	0.36

Table 6: Training loss and MSE for various models in Track B.

4.3 Track C: Cross-lingual Emotion Detection

For Track C, we initially fine-tuned multilingual versions of the same model used in Track A. Using similar hyperparameters (4 epochs, batch size 16, learning rate 5×10^{-5}). We obtained baseline results (e.g., a maximum macro F1 score of 0.09566 with XLM-RoBERTa-base).

An alternative approach incorporated an intermediate translation step. Training data in German was translated to English using a model from the University of Helsinki (via HuggingFace), and the target Russian texts were similarly translated before inference. Table 7 shows the training and validation losses for this method.

Model	Training Loss	Validation Loss
FacebookAI/xlm-roberta-base	0.32	0.38
distilbert-base-multilingual	0.35	0.35
multilingual-bert-base-uncased	0.32	0.37

Table 7: Training and validation losses for Track C (simple approach).

5 Results

5.1 Track A: Multi-label Emotion Detection

Initially, we fine-tuned the **BERT-base-uncased** model to establish a baseline for performance on English Dataset. During training, we recorded key metrics such as training loss, validation loss, accuracy, and F1 score for each epoch, using an 80-20 training-validation split. The macro and micro F1 scores, reported in Table 8, represent the result on the test dataset, as obtained by submitting the model predictions to **CodaBench**. This baseline model achieved a macro F1 score of **0.65** and a micro F1 score of **0.61**.

Subsequently, we adopted an ensemble learning approach, combining multiple models to improve performance. The models included in the ensemble were **DistilBERT-base-uncased**, **DeBERTa-base-uncased**, and **XLM Roberta-base**. By averaging the logits from these models to generate final predictions, we significantly improved the macro and micro F1 scores, achieving **0.67** and **0.69** respectively. The Table 8 also shows the scores of multiple languages on the **ensemble approach**.

Language	Micro F1 Score	Macro F1 Score
English (baseline)	0.65	0.61
English (ensemble)	0.74	0.71
Afrikaans (ensemble)	0.62	0.45
German (ensemble)	0.61	0.44
Amharic (ensemble)	0.63	0.45
Spanish (ensemble)	0.70	0.71
Hindi (ensemble)	0.81	0.80
Marathi (ensemble)	0.76	0.76
Russian (ensemble)	0.71	0.72
Arabic (ensemble)	0.20	0.14
Swedish (ensemble)	0.51	0.21

Table 8: Performance of various fine-tuned models employed for predicting emotion labels.

5.2 Track B: Emotion Intensity Prediction

For emotion intensity prediction, we used a fine-tuned ensemble model on multilingual datasets. Our best average Pearson correlation score was **0.60** for Amharic and **0.57** for German. Joy and sadness were the easiest emotions to predict, while surprise and fear showed lower correlation scores. The overall results are shown in Table 9

Language	Overall
AMH	0.60
DEU	0.53
ENG	0.74
ESP	0.64
PTBR	0.38
RUS	0.76
ARQ	0.30
CHN	0.47
HAU	0.39
UKR	0.42
RON	0.49

Table 9: Average Pearson correlation score for different languages in Track B

5.3 Track C: Cross-lingual Emotion Detection

Using the translation method, subsequently, we adopted an ensemble learning approach the same in Track A, combining multiple models to improve performance. The models included in the ensemble were **DistilBERT-base-uncased**, **DeBERTa-base-uncased**, and **XLM Roberta-base**. By averaging the logits from these models to generate final predictions. The Table 10 also shows the scores of multiple languages on the **ensemble approach**.

Language	Macro F1	Micro F1
Afrikaans	0.30	0.35
Amharic	0.26	0.30
German	0.08	0.16
Spanish	0.24	0.33
Hindi	0.67	0.69
Marathi	0.76	0.76
Russian	0.20	0.22
Somali	0.27	0.38
Chinese (Mandarin)	0.39	0.45
Swahili	0.11	0.12
Swedish	0.22	0.51
Ukrainian	0.17	0.21
Igbo	0.09	0.17
Romanian	0.47	0.53

Table 10: Macro F1 and Micro F1 Scores for Different Languages

The best results were observed in Hindi and Marathi, likely due to their regional relevance, as well as the high quality of the translations in the translation method employed.

6 Conclusion

Our work demonstrates that transformer-based models, particularly when combined in ensemble frameworks, can effectively address the challenges of emotion detection in text. For both multi-label classification and intensity prediction, ensemble learning significantly improved performance over single-model baselines. In the cross-lingual setting, the translation-based approach yielded notable improvements, underlining the potential of intermediate language translation to bridge linguistic gaps. Future work will explore refined data balancing strategies, and the integration of larger advanced models (e.g., LLaMA 3.2, T5) to further enhance performance.

7 Acknowledgments

We would like to acknowledge the support provided by our supervisor Dr. Abdul Samad at Habib University, Karachi, Pakistan for guiding us throughout this project through his experience and valuable insights. We would also like to thank SemEval chairs for their guidance and organization of this task.

References

D. Demszky et al. 2021. Goemotions: A dataset of fine-grained emotions. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 4040–4054.

D. Goswami, S. S. C. Puspo, M. N. Raihan, A. N. B. Emran, A. Ganguly, and M. Zampieri. 2024. Masontigers at semeval-2024 task 1: An ensemble approach for semantic textual relatedness. In *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*, pages 1380–1390.

S. Kumar, M. S. Akhtar, E. Cambria, and T. Chakraborty. 2024. Semeval 2024 – task 10: Emotion discovery and reasoning its flip in conversation (ediref). *arXiv preprint arXiv:2402.18944*.

S. Mohammad, F. Bravo-Marquez, M. Salameh, and S. Kiritchenko. 2018. Semeval-2018 task 1: Affect in tweets. In *Proceedings of the 12th International Workshop on Semantic Evaluation (SemEval-2018)*, pages 1–17.

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.

R. Pan, J. A. García-Díaz, D. Roldán, and R. Valencia-García. 2024. Umuteam at semeval-2024 task 10: Discovering and reasoning about emotions in conversation using transformers. In *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*, pages 703–709.

A. Wadhawan and A. Aggarwal. 2021. Towards emotion recognition in hindi-english code-mixed data: A transformer based approach. In *Proceedings of the Eleventh Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 195–202.

Y. Wang, Y. Li, P. P. Liang, L.-P. Morency, P. Bell, and C. Lai. 2024. Cross-attention is not enough: Incongruity-aware dynamic hierarchical fusion for multimodal affect recognition. In *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*, pages 703–709.