Zero_Shot at SemEval-2025 Task 11: Fine-Tuning Deep Learning and Transformer-based Models for Emotion Detection in Multi-label Classification, Intensity Estimation, and Cross-lingual Adaptation

Ashraful Islam Paran*, **Sabik Aftahee***, **Md. Refaj Hossan***
Jawad Hossain and Mohammed Moshiul Hoque
Department of Computer Science and Engineering
Chittagong University of Engineering and Technology
{u1904029, u1904024, u1904007, u1704039}@student.cuet.ac.bd
moshiul_240@cuet.ac.bd

Abstract

Language is a rich medium employed to convey emotions subtly and intricately, as abundant as human emotional experiences themselves. Emotion recognition in natural language processing (NLP) is now a core element in facilitating human-computer interaction and interpreting intricate human behavior via text. It has potential applications in every sector i.e., sentiment analysis, mental health surveillance. However, prior research on emotion recognition is primarily from high-resource languages while low-resource languages (LRLs) are not well represented. This disparity has been a limitation to the development of universally applicable emotion detection models. To address this, the SemEval-2025 Shared Task 11 focused on perceived emotions, aiming to identify the emotions conveyed by a text snippet. It includes three tracks: Multi-label Emotion Detection (Track A), Emotion Intensity (Track B), and Cross-lingual Emotion Detection (Track C). This paper explores various models, including machine learning (LR, SVM, RF, NB), deep learning (BiLSTM+CNN, BiLSTM+BiGRU), and transformer-based models (XLM-R, mBERT, ModernBERT). The results showed that XLM-R outperformed other models in Tracks A and B, while BiLSTM+CNN performed better for Track C across most languages.

1 Introduction

Language, as a means of communication, plays a central role in the conveyance and perception of emotions (Mohammad et al., 2018). Emotions are an intrinsic part of human communication, affecting how we perceive and respond to others' messages. Although we all feel and deal with emotions daily, the detection of emotions in text remains a challenging task in NLP (Muhammad et al., 2025a). Emotions are difficult to convey explicitly, and the ways people perceive and express emotions vary greatly, with differences in culture, context, and personality (Wiebe et al., 2005; Mohammad and Kiritchenko, 2018; Acheampong et al., 2020). Therefore, emotion detection from text is one of the most difficult tasks in NLP. The ability to detect emotions from text is increasingly crucial for applications such as virtual assistants, mental health monitoring systems, and social media analytics (Acheampong et al., 2020). The majority of existing studies on emotion detection have focused on high-resource languages, which are supported by large datasets and extensive research. This focus has created an enormous research gap in terms of low-resource languages, which lack high-quality annotated data (Tafreshi et al., 2024; Muhammad et al., 2025a). To address these challenges, the SemEval-2025 Shared Task 11 titled *Bridging the Gap in Text-Based Emotion Detection*[1] focused on approximately 32 low-resource languages, including Afrikaans (afr), Amharic (amh), Oromo (Orm), Hausa (Hau), among others (Muhammad et al., 2025b). The task includes three tracks: multi-label and cross-lingual emotion detection, and detection of emotion intensity. These challenges involve processing textual features to grasp the hidden meaning in cultural and linguistic contexts and classify the text into 6 classes i.e., *joy, sadness, fear, anger, surprise, or disgust* simultaneously in track A, detecting emotion intensity as *no emotion, low, moderate, or high* in track B, and adapting the model to multiple languages in track C. Therefore, the contributions of this work are as follows:

- Developed deep learning and transformer-based approaches that effectively process textual features to detect emotion in low-resource languages.
- Investigated various ML, DL, and transformer-

*Authors contributed equally to this work.

[1] https://github.com/emotion-analysis-project/SemEval2025-task11

Proceedings of the The 19th International Workshop on Semantic Evaluation (SemEval-2025), pages 1890–1904

based models to identify the emotion and its intensity while evaluating performance metrics and conducting error analysis to determine the best strategy.

The implementation details of the tasks will be found in the GitHub repository[2]. The rest of the paper is organized as follows: Section 2 discusses related work, Section 3 describes the dataset and task, Section 4 outlines the system overview, Section 5 presents the results analysis, Section 6 summarizes insights and future research directions, and Section 7 outlines the limitations of our research.

2 Related Work

Multi-label emotion detection is essential for discerning complex emotional states from text, where instances may display multiple emotions concurrently.

2.1 Multi-label Emotion Detection

The emotion detection task is challenging due to the nuanced and context-sensitive nature of the text. Jabreel and Moreno (2019) enhanced emotion classification in tweets by using deep learning in the SemEval-2018 Task 1 dataset, achieving a 59% accuracy. Another work focused on improving recognition accuracy in contextual settings using multi-task learning and the EMOTIC dataset (Bendjoudi et al., 2020). A framework was developed for multimodal emotion detection, showing superior results on the CMU-MOSEI dataset (Zhang et al., 2020). Le et al. (2023) employed transformer-based techniques for video content, using IEMOCAP and CMU-MOSEI datasets, and showing significant advancements. Abdul-Mageed and Ungar (2017) created EmoNet for fine-grained emotion detection on Twitter, achieving high accuracies. Mansy et al. (2022) introduced an ensemble model for Arabic tweets, outperforming previous methods.

2.2 Multi-label Emotion Intensity Detection

Ganesh and Kamarason (2020) developed a CNN-based model for multi-labeled emotion intensity analysis on Twitter, emphasizing the need for refined emotion analysis in social media. Mashal and Asnani (2017) and Rodríguez and Garza (2019) further explored algorithms to accurately measure and predict emotion intensities in informal texts and social networks, respectively. Firdaus et al. (2020) introduced the MEISD dataset for multimodal emotion and sentiment analysis, catering to the demand for sophisticated emotion detection systems. Additionally, Singh et al. (2022) created EmoInHindi, an annotated Hindi dataset, to facilitate multi-label emotion recognition in resource-scarce languages. Another study discussed the use of machine learning for classifying emotions in tweets, highlighting the challenges in less-researched linguistic contexts like Urdu (Ashraf et al., 2022; Mashal and Asnani, 2020).

2.3 Cross-lingual Emotion Detection

Neumann and Vu (2018) explored multilingual speech emotion recognition using CNNs for English and French, evaluating the cross-language adaptability of emotional indicators. Another work proposed a novel approach for estimating sentiment prevalence across languages without target language data (Esuli et al., 2020). Transfer learning evaluated for speech emotion recognition across languages and corpora, emphasizing multi-task learning to boost model versatility (Goel and Beigi, 2020). Their models show improved accuracy in cross-lingual recognition, particularly when incorporating auxiliary tasks like language identification. Navas Alejo et al. (2020) investigated emotion intensity prediction across languages using translation and embedding techniques. Kanclerz et al. (2020) showed deep transfer learning's efficacy in sentiment analysis, using language-agnostic representations to effectively predict sentiments in low-resource languages.

3 Dataset and Task Description

The shared task on emotion detection consists of three tracks, namely Track A (multi-label emotion classification), Track B (emotion intensity prediction), and Track C (cross-lingual emotion detection). Track A predicts perceived emotions (*joy, sadness, fear, anger, surprise, disgust*) in a text, with *disgust* excluded for some languages. Track B assigns an intensity level (*no, low, moderate, high*) to each emotion. The dataset (Muhammad et al., 2025a; Belay et al., 2025) for this track provides labeled instances indicating the degree of emotion expressed in the text. Finally, Track C focuses on cross-lingual emotion detection, requiring models to classify emotions in a target language using a labeled dataset from another language. It lacks a labeled training corpus as compared to other tracks.

[2] https://github.com/RJ-Hossan/SemEval_2025_T11

The dataset structure remains consistent across tracks, allowing exploration of multi-label classification, intensity prediction, and cross-lingual generalization. Tables A.1, A.2, and A.3 in Appendix A show the class-wise distribution of train, validation, and test set for these tasks.

4 System Overview

Figure 1 illustrates the exploration of various ML, DL, and transformer-based models to develop a framework for detecting multi-label and cross-lingual emotion, along with emotion intensity estimation.

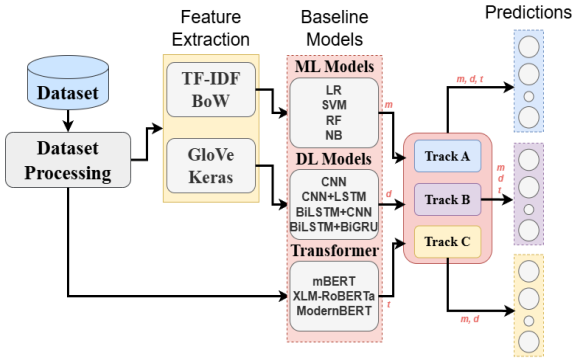

Figure 1: Schematic process of multi-label emotion classification, emotion intensity prediction, and cross-lingual emotion detection.

4.1 Data Preprocessing

The text preprocessing pipeline entailed several crucial steps toward cleaning and normalizing the data. Emojis were removed using *emoji.replace_emoji*, which replaced them with an empty string, while special characters were eliminated with `re.sub(r'[\w\s]', '', text)` without retaining anything but alphanumeric content. Subsequently, the text was lowercase for consistency and then tokenized using *.split()*. The removal of stopwords was done using the *nltk* library. Finally, cleaned tokens were reconstructed into sentences so that there would be a uniform text format for the model to perform better in cross-lingual sentiment analysis.

4.2 Feature Extraction

Feature extraction is necessary for ML and DL models to learn from text. We utilized TF-IDF (Takenobu, 1994) to extract features for various ML algorithms. For DL models, we employed GloVe (Pennington et al., 2014) and Keras-based embeddings to obtain features.

4.3 ML Models

Several machine learning (ML) models were explored to identify multi-label and cross-lingual emotions. Specifically, we used LR, SVM, NB, and RF classifiers for emotion detection in different languages. Furthermore, we applied hyperparameter tuning to enhance model performance, such as experimenting with *linear* and *RBF* kernels for SVM, varying *max_iter* for LR, and optimizing the value *alpha* for NB. GridSearchCV[3] was used to systematically explore the optimal hyperparameters, ensuring improved classification accuracy in detecting multi-label and cross-lingual emotion. Table 1 provides the tuned hyperparameters used in the experiments for ML models for Track C.

Model	Hyperparameters
Logistic Regression	max_iter = 256
Random Forest Classifier	n_estimators = 120, max_depth = 12
Support Vector Machine	kernel = rbf, C = 2
Naive Bayes	alpha = 1.0

Table 1: Tuned hyperparameters used for ML models (Track C).

4.4 DL Models

Several deep learning models were explored for these emotion detection tasks, including CNN, BiLSTM+CNN, and BiLSTM+BiGRU, among others, to effectively capture the sequential dependencies in textual data. To detect emotion in a cross-lingual setting, the BiLSTM+CNN model begins by transforming tokenized input text into dense vectors using an embedding layer with a vocabulary size of 10,000, an embedding dimension of 128. The text is processed through a Bidirectional LSTM layer of 128 units to capture sequential dependencies, followed by a dropout layer of 0.3. The model includes two Conv1D layers with 128 and 64 filters, kernel sizes 5 and 3, MaxPooling1D layers, and BatchNormalization for stable learning. The output is flattened and passed through a Dense layer (128 units, ReLU), with a final output layer of 6 units using sigmoid activation for multi-label classification. The tuned hyperparameters for this task are presented in Table 2.

[3] https://scikit-learn.org/dev/modules/generated/sklearn.model_selection.GridSearchCV.html

Parameters	GRU	CNN	BiLSTM+CNN	BiLSTM+BiGRU
Learning rate	0.0001	0.0001	0.0001	0.0001
Batch size	32	32	32	32
Optimizer	Adam	Adam	Adam	Adam
Epochs	40	45	45	45
Embedding_dim	128	-	-	-
max_length	100	-	-	-
SpatialDropout1D	0.2	-	-	-
GRU units	128, 64, 32	-	-	-
Dropout rate	0.3	0.3	0.3	0.3
BatchNormalization	Yes	No	No	Yes
Dense units	128	128	128	128
Conv1D filters	-	128, 64	128, 64	-
kernel size	-	5, 3	5, 3	-
MaxPooling1D	-	size (2)	size (2)	-

Table 2: Tuned hyperparameters used for DL models (Track C).

4.5 Transformer-based Models

Various transformer-based models such as XLM-R, mBERT, and ModernBERT were employed to leverage their powerful attention mechanisms for multi-label emotion detection and intensity prediction tasks. By fine-tuning these models on our specific datasets, we aimed to achieve better performances in these tasks. We set up the multi-label emotion classification pipeline with the *FacebookAI/xlm-roberta-base*[4] and *google-bert/bert-base-multilingual-uncased*[5] models. The *EmotionsDataset* class was created to tokenize text inputs using *AutoTokenizer* and get corresponding emotion labels. Details about the tuned hyperparameters for these tasks are presented in Tables 3 and 4.

Hyperparameter	Value
Learning Rate	2e-5
Per Device Batch Size	8
Number of Epochs	5
Max Sequence Length	128
Loss Function	Binary Cross-Entropy
Optimizer	AdamW
Weight Decay	0.01

Table 3: Tuned hyperparameters used for multi-label emotion classification task (Track A) using XLM-R.

The hyperparameters were selected based on standard fine-tuning practices, i.e., a learning rate of 2e-5 for observed stable convergence, a batch size of 8 to avoid overfitting on small datasets, 5 training epochs for balanced performance, Binary Cross-Entropy for multi-label classification (Track A), and Cross-Entropy Loss for intensity estimation (Track B). We chose transformer-based models like XLM-R for Tracks A and B due to their proven multilingual contextual understanding, as XLM-R excels in handling nuanced emotional expressions across diverse linguistic structures.

Hyperparameter	Value
Learning Rate	2e-5
Per Device Batch Size	8
Number of Epochs	5
Max Sequence Length	128
Loss Function	Cross-Entropy Loss
Optimizer	AdamW
Weight Decay	0.01

Table 4: Hyperparameters used for emotion intensity detection task (Track B) using XLM-R.

4.6 System Requirements

The BiLSTM+CNN model for cross-lingual emotion detection and the XLM-R model for multi-label emotion and emotion intensity recognition were trained on a dual-GPU setup (NVIDIA Tesla T4x2), utilizing parallel processing for enhanced performance. The BiLSTM+CNN model utilized approximately 5-7 GB of GPU memory, whereas the XLM-R model utilized approximately 8-10 GB of GPU memory. Overall, the BiLSTM+CNN, along with other DL models, were trained for 45 epochs for 90-100 minutes, depending on training time by the number of data sets and computation of class weights. It enabled the efficient execution of challenging tasks to achieve flawless output for any language and the objectives of emotion detection.

5 Result Analysis

Table 5 presents the evaluation results of ML, DL, and transformer-based models for multi-label emotion detection across five languages: Amharic, Hindi, Igbo, Marathi, and Russian. Among ML models, SVM demonstrates the highest F1 scores in most cases, particularly excelling in Russian (60.85), Marathi (50.33), and Hindi (48.78), while Random Forest (RF) performs slightly better for Igbo (41.81). In DL models, CNN consistently outperforms other architectures, achieving the best results across all languages, with the highest F1 score in Marathi (51.70). However, transformer models significantly surpass both ML and DL models, with XLM-R achieving the highest F1 scores in four out of five languages, including Hindi (84.60), Marathi (78.74), and Russian (84.38), while m-BERT performs best for Igbo (49.35).

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

[5] https://huggingface.co/google-bert/bert-base-multilingual-uncased

Language	Classifier	P (%)	R (%)	F1 (%)	A (%)
Amharic	SVM	60.56	38.45	46.44	39.06
	LR	66.87	19.76	29.10	32.64
	RF	60.61	27.34	36.97	35.23
	NB	60.90	5.45	9.65	23.28
	CNN	56.46	28.64	36.45	33.77
	CNN+LSTM	23.50	23.43	17.54	18.83
	CNN+BiLSTM	35.48	24.07	28.58	30.55
	XLM-R	54.93	50.67	**52.61**	56.82
	m-BERT	29.34	7.60	8.48	37.54
	ModernBERT	32.98	15.68	20.23	38.05
Hindi	SVM	77.61	36.02	48.78	41.78
	LR	85.59	12.75	20.79	25.05
	RF	74.76	22.39	33.01	31.78
	NB	16.67	0.57	1.10	14.85
	CNN	75.22	35.65	47.77	40.89
	CNN+LSTM	11.81	1.76	3.06	15.94
	CNN+BiLSTM	50.80	13.83	20.95	24.65
	XLM-R	85.42	83.86	**84.60**	81.29
	m-BERT	77.82	75.23	76.44	73.37
	ModernBERT	65.14	49.57	55.95	52.28
Igbo	SVM	61.41	38.39	47.10	51.59
	LR	69.12	22.07	33.01	40.86
	RF	77.65	31.43	41.81	45.98
	NB	57.23	8.68	14.48	30.40
	CNN	76.16	31.10	41.08	44.25
	CNN+LSTM	12.30	5.41	7.52	25.69
	CNN+BiLSTM	52.26	29.32	34.88	42.31
	XLM-R	37.53	37.23	37.35	55.89
	m-BERT	51.96	47.12	**49.35**	62.81
	ModernBERT	63.65	39.98	46.65	58.31
Marathi	SVM	85.03	37.47	50.33	46.10
	LR	87.03	15.22	24.01	29.90
	RF	82.93	28.13	40.98	39.30
	NB	50.00	1.42	2.71	19.80
	CNN	84.07	38.82	51.70	46.00
	CNN+BiLSTM	66.70	23.32	34.07	32.30
	XLM-R	84.41	74.35	**78.74**	75.80
	m-BERT	77.16	68.84	72.38	70.00
	ModernBERT	68.01	50.98	57.92	57.10
Russian	SVM	88.66	47.32	60.85	54.40
	LR	90.08	17.73	27.32	34.80
	RF	85.29	41.59	54.08	49.10
	NB	83.33	4.54	8.51	25.40
	CNN	66.59	27.69	36.90	38.60
	CNN+BiLSTM	45.74	18.55	24.16	32.80
	XLM-R	89.28	80.07	**84.38**	81.10
	m-BERT	86.85	81.22	83.90	80.90
	ModernBERT	79.78	63.67	70.72	64.90

Table 5: Performance of the employed models for detecting multi-label emotion in several languages where P, R, F1, and A denote precision, recall, F1 score (macro), and accuracy, respectively.

Table 6 presents the evaluation results of ML, DL, and transformer-based models for multi-label emotion intensity detection in four languages: Algerian Arabic, Chinese, Hausa, and Russian. Among ML models, LR and SVM show competitive performance, with LR achieving the highest F1 scores in Chinese (27.83) and Russian (29.19), while SVM performs best in Hausa (28.94), but all ML models, including RF and NB, struggle in Algerian Arabic (F1 around 23.10). In DL models, CNN+BiLSTM consistently outperforms CNN+GRU, with the highest F1 scores across all languages, peaking at 35.04 in Hausa. However, the transformer-based model XLM-RoBERTa significantly surpasses both ML and DL models, achieving the highest F1 scores in all four languages: Algerian Arabic (29.17), Chinese (46.71), Hausa (57.34), and an outstanding 83.74 in Russian. These results underscore the superior effectiveness of transformer-based models, particularly XLM-RoBERTa, for both multi-label emotion detection and intensity detection, delivering markedly better performance across diverse linguistic contexts.

Language	Classifier	P (%)	R (%)	F1 (%)	A (%)
Algerian Arabic	LR	21.96	27.83	23.43	12.00
	SVM	19.86	27.78	23.10	12.00
	RF	19.57	27.78	23.10	12.00
	NB	19.86	27.78	23.10	12.00
	CNN+BiLSTM	23.42	28.14	23.90	12.00
	CNN+GRU	19.89	27.78	23.12	12.00
	XLM-RoBERTa	28.00	30.40	29.17	15.00
Chinese	LR	29.83	30.62	27.83	23.00
	SVM	29.83	30.62	27.83	23.00
	RF	25.65	30.56	27.70	22.50
	NB	29.83	30.62	27.83	23.00
	CNN+BiLSTM	25.65	30.56	27.70	22.50
	CNN+GRU	25.65	30.56	27.70	22.50
	XLM-RoBERTa	45.00	48.50	46.71	30.00
Hausa	LR	35.13	29.36	29.12	19.66
	SVM	37.36	29.12	28.94	18.82
	RF	24.56	25.13	22.68	13.20
	NB	20.39	25.00	22.43	12.92
	CNN+BiLSTM	40.81	35.00	35.04	26.40
	CNN+GRU	35.90	32.04	31.53	23.31
	XLM-RoBERTa	54.90	60.10	57.34	35.00
Russian	LR	38.43	28.97	29.19	30.32
	SVM	35.79	27.83	27.47	27.99
	RF	21.67	25.00	23.21	22.74
	NB	21.68	25.00	23.21	22.74
	CNN+BiLSTM	36.42	32.42	32.53	33.24
	CNN+GRU	34.00	32.83	32.61	37.32
	XLM-RoBERTa	82.05	85.47	83.74	70.00

Table 6: Performance of the employed models for detecting multi-label emotion intensity in several languages where P, R, F1, and A denote precision, recall, F1 score (macro), and exact match accuracy, respectively.

Table 7 demonstrates the evaluation results of ML and DL models to detect cross-lingual emotion across five languages, i.e., Amharic, Algerian Arabic, Hausa, Oromo, and Somali. Among machine learning (ML) models, SVM consistently achieves the highest F1 scores, performing best for Amharic (41.37), Hausa (47.80), Oromo (32.04), and Somali (20.39). In contrast, Random Forest (RF) and Naïve Bayes (NB) show poor performance. For deep learning (DL) models, BiLSTM+BiGRU outperforms other architectures in most cases, achieving the highest F1 scores for Hausa (53.48) and Oromo (42.62). CNN-based models also perform

competitively, particularly in Amharic (40.25). Overall, transformer models were not implemented for this task, but BiLSTM+BiGRU emerges as the strongest deep learning model, while SVM remains the best-performing ML model across most languages. Appendix B presents an in-depth error analysis of the employed models, whereas Appendix C outlines a comparative performance ranking between our proposed system and the baseline model (RemBERT), evaluated using F1 scores.

Language	Classifier	P (%)	R (%)	F1 (%)	A (%)
Amharic	LR	66.54	20.64	29.77	34.27
	RF	50.00	0.10	0.20	17.98
	NB	60.02	6.69	11.34	24.18
	SVM	65.76	31.98	41.37	38.44
	CNN	48.24	37.74	40.25	35.51
	GRU	44.10	42.98	43.40	34.16
	BiLSTM+BiGRU	51.36	46.70	48.57	37.20
	BiLSTM+CNN	46.35	45.90	**45.94**	33.71
Algerian Arabic	LR	48.90	10.52	15.00	13.53
	RF	11.31	2.35	3.89	11.64
	NB	55.45	9.40	12.90	13.64
	SVM	58.22	20.64	27.22	14.52
	CNN	43.62	27.87	31.21	11.97
	GRU	46.61	29.83	34.25	12.08
	BiLSTM+BiGRU	47.04	33.14	35.05	13.30
	BiLSTM+CNN	51.19	39.36	**43.62**	15.52
Hausa	LR	82.79	19.41	29.68	26.85
	RF	65.62	2.32	4.32	15.28
	NB	80.99	5.81	10.27	18.06
	SVM	79.25	35.13	47.80	38.06
	CNN	57.07	47.26	49.51	35.19
	GRU	57.85	45.64	50.49	35.65
	BiLSTM+BiGRU	57.45	52.16	53.48	39.17
	BiLSTM+CNN	53.33	50.39	**49.93**	34.91
Oromo	LR	66.62	14.26	18.97	43.64
	RF	40.97	1.63	3.02	26.15
	NB	39.46	11.19	13.39	42.01
	SVM	80.22	23.84	32.04	49.27
	CNN	56.54	30.43	33.92	48.00
	GRU	46.35	33.70	37.78	48.69
	BiLSTM+BiGRU	46.14	40.72	42.62	48.05
	BiLSTM+CNN	37.20	42.30	**39.13**	42.42
Somali	LR	45.03	5.26	9.20	41.51
	NB	33.33	0.14	0.28	38.27
	SVM	68.21	12.99	20.39	46.17
	CNN	40.34	21.79	27.31	44.04
	GRU	41.36	27.96	32.90	41.63
	BiLSTM+BiGRU	37.77	33.73	35.41	42.33
	BiLSTM+CNN	36.14	37.41	**36.52**	40.62

Table 7: Performance of the employed models for detecting cross-lingual emotion in several languages where P, R, F1, and A denote precision, recall, F1 score (macro), and accuracy, respectively.

6 Conclusion

This paper demonstrated a multi-label emotion classification and intensity prediction model with the best performance for Task A (multi-label emotion classification) and Task B (emotion intensity prediction) through XLM-RoBERTa, while a BiLSTM+CNN model was found superior in Task C (cross-lingual emotion detection) across different LRLs. The outcome of our work demonstrates the capabilities of transformer models for structured emotion prediction and hybrid deep learning models for cross-lingual transfer learning. Future work will explore enhancing model generalizability across languages with scarce labeled data by merging self-supervised learning and contrastive learning techniques. We also plan to research domain adaptation methods and data augmentation strategies to improve emotion recognition in low-resource languages and multi-lingual social media settings.

7 Limitations

Although the study presents valuable information on emotion detection in LRLs, certain limitations inevitably affect the generalizability and robustness of its findings.

- The emotion intensity prediction task faces challenges due to subjective labeling, leading to inconsistencies in the dataset.
- The model struggles with capturing fine-grained emotion variations and overlapping emotions, particularly in multilingual and code-mixed scenarios, where expressions of emotions vary across languages and cultures.
- The cross-lingual emotion detection task is constrained by the absence of a labeled training dataset, making it heavily dependent on transfer learning techniques, which may not generalize well across distant language pairs.
- The study is affected by class imbalance, where certain emotions are underrepresented, limiting the model's ability to learn and predict rare emotions effectively. Advanced data augmentation strategies could help mitigate this issue.

Acknowledgments

We thank the SemEval-2025 shared task organizers for running this task. This work was supported by the Directorate of Research & Extension (DRE), Chittagong University of Engineering & Technology (CUET).

References

Muhammad Abdul-Mageed and Lyle Ungar. 2017. Emonet: Fine-grained emotion detection with gated recurrent neural networks. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*, pages 718–728. Association for Computational Linguistics.

Francisca Adoma Acheampong, Chen Wenyu, and Henry Nunoo-Mensah. 2020. Text-based emotion detection: Advances, challenges, and opportunities. *Engineering Reports*, 2(7):e12189.

Noman Ashraf, Lal Khan, Sabur Butt, Hsien-Tsung Chang, Grigori Sidorov, and Alexander Gelbukh. 2022. Multi-label emotion classification of urdu tweets. In *PeerJ Computer Science*, volume 8, page e896.

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.

Ilyes Bendjoudi, Frederic Vanderhaegen, Denis Hamad, and Fadi Dornaika. 2020. Multi-label, multi-task cnn approach for context-based emotion recognition. *Information Fusion*, xx(xx):xx.

Andrea Esuli, Alejandro Moreo, and Fabrizio Sebastiani. 2020. Cross-lingual sentiment quantification. *Journal of Artificial Intelligence Research*, 67:569–607.

Mauajama Firdaus, Hardik Chauhan, Asif Ekbal, and Pushpak Bhattacharyya. 2020. Meisd: A multimodal multi-label emotion, intensity and sentiment dialogue dataset for emotion recognition and sentiment analysis in conversations. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 4441–4453.

V. Ganesh and M. Kamarason. 2020. Multi-labelled emotion with intensity based sentiment classification model in tweets using convolution neural networks. *International Journal of Advanced Trends in Computer Science and Engineering*, 9(2):1650–1656.

Shivali Goel and Homayoon Beigi. 2020. Cross-lingual cross-corpus speech emotion recognition. *IEEE Transactions on Affective Computing*, 11(2):287–299.

Mohammed Jabreel and Antonio Moreno. 2019. A deep learning-based approach for multi-label emotion classification in tweets. *Applied Sciences*, 9(6):1123.

Kamil Kanclerz, Piotr Miłkowski, and Jan Kocon. 2020. Cross-lingual deep neural transfer learning in sentiment analysis. *Knowledge-Based Systems*, 195:105746.

Hoai-Duy Le, Guee-Sang Lee, Soo-Hyung Kim, Seungwon Kim, and Hyung-Jeong Yang. 2023. Multi-label multimodal emotion recognition with transformer-based fusion and emotion-level representation learning. *IEEE Access*, 11:14742–14752.

Alaa Mansy, Sherine Rady, and Tarek Gharib. 2022. An ensemble deep learning approach for emotion detection in arabic tweets. *International Journal of Advanced Computer Science and Applications*, 13(4):980–989.

Sonia Xylina Mashal and Kavita Asnani. 2017. Emotion intensity detection for social media data. In *IEEE International Conference on Computing Methodologies and Communication*.

Sonia Xylina Mashal and Kavita Asnani. 2020. Emotion analysis of social media data using machine learning techniques. *IOSR Journal of Computer Engineering*, 22:17–20.

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.

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

Shamsuddeen Hassan Muhammad, Nedjma Ousidhoum, Idris Abdulmumin, Jan Philip Wahle, Terry Ruas, Meriem Beloucif, Christine de Kock, Nirmal Surange, Daniela Teodorescu, Ibrahim Said Ahmad, David Ifeoluwa Adelani, Alham Fikri Aji, Felermino D. M. A. Ali, Ilseyar Alimova, Vladimir Araujo, Nikolay Babakov, Naomi Baes, Ana-Maria Bucur, Andiswa Bukula, Guanqun Cao, Rodrigo Tufino Cardenas, Rendi Chevi, Chiamaka Ijeoma Chukwuneke, Alexandra Ciobotaru, Daryna Dementieva, Murja Sani Gadanya, Robert Geislinger, Bela Gipp, Oumaima Hourrane, Oana Ignat, Falalu Ibrahim Lawan, Rooweither Mabuya, Rahmad Mahendra, Vukosi Marivate, Andrew Piper, Alexander Panchenko, Charles Henrique Porto Ferreira, Vitaly Protasov, Samuel Rutunda, Manish Shrivastava, Aura Cristina Udrea, Lilian Diana Awuor Wanzare, Sophie Wu, Florian Valentin Wunderlich, Hanif Muhammad Zhafran, Tianhui Zhang, Yi Zhou, and Saif M. Mohammad. 2025a. Brighter: Bridging the gap in human-annotated textual emotion recognition datasets for 28 languages. *Preprint*, arXiv:2502.11926.

Shamsuddeen Hassan Muhammad, Nedjma Ousidhoum, Idris Abdulmumin, Seid Muhie Yimam, Jan Philip Wahle, Terry Ruas, Meriem Beloucif, Christine

De Kock, Tadesse Destaw Belay, Ibrahim Said Ahmad, Nirmal Surange, Daniela Teodorescu, David Ifeoluwa Adelani, Alham Fikri Aji, Felermino Ali, Vladimir Araujo, Abinew Ali Ayele, Oana Ignat, Alexander Panchenko, Yi Zhou, and Saif M. Mohammad. 2025b. SemEval task 11: Bridging the gap in text-based emotion detection. In *Proceedings of the 19th International Workshop on Semantic Evaluation (SemEval-2025)*, Vienna, Austria. Association for Computational Linguistics.

Irean Navas Alejo, Toni Badia, and Jeremy Barnes. 2020. Cross-lingual emotion intensity prediction. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2143–2152.

Michael Neumann and Ngoc Thang Vu. 2018. Cross-lingual and multilingual speech emotion recognition on english and french. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC)*.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar. Association for Computational Linguistics.

Fernando M. Rodríguez and Sara E. Garza. 2019. Predicting emotional intensity in social networks. *Journal of Intelligent & Fuzzy Systems*, 36:4709–4719.

Gopendra Vikram Singh, Priyanshu Priya, Mauajama Firdaus, Asif Ekbal, and Pushpak Bhattacharyya. 2022. Emoinhindi: A multi-label emotion and intensity annotated dataset in hindi for emotion recognition in dialogues. *Preprint accepted at LREC*.

Shabnam Tafreshi, Shubham Vatsal, and Mona Diab. 2024. Emotion classification in low and moderate resource languages. *arXiv preprint arXiv:2402.18424*.

Tokunaga Takenobu. 1994. Text categorization based on weighted inverse document frequency. *Information Processing Society of Japan, SIGNL*, 94(100):33–40.

Janyce Wiebe, Theresa Wilson, and Claire Cardie. 2005. Annotating expressions of opinions and emotions in language. *Language Resources and Evaluation*, 39(2):165–210.

Dong Zhang, Xincheng Ju, Junhui Li, Shoushan Li, Qiaoming Zhu, and Guodong Zhou. 2020. Multi-modal multi-label emotion detection with modality and label dependence. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, pages 3584–3593. Association for Computational Linguistics.

A Class-wise Distribution of Dataset

Tables A.1, A.2, and A.3 show the class-wise distribution of emotion datasets across different languages for train, validation, and test sets.

Language	Classes	Train	Valid	Test	W_T
Marathi	Anger	350	14	164	6846
	Disgust	299	11	98	5243
	Fear	382	15	147	6820
	Joy	461	19	175	7958
	Sadness	431	17	207	8137
	Total	1923	76	791	35004
Hindi	Anger	422	16	161	12425
	Disgust	265	10	111	6819
	Fear	380	14	146	9550
	Joy	442	11	191	11031
	Sadness	449	17	169	11127
	Total	1958	68	778	50952
Russian	Anger	543	47	226	8351
	Disgust	273	26	122	4759
	Fear	328	21	108	4464
	Joy	555	34	193	6498
	Sadness	421	39	141	6235
	Total	2120	167	790	30307
Amharic	Anger	1188	207	582	48616
	Disgust	1268	209	628	45846
	Fear	109	22	54	3339
	Joy	549	93	276	17436
	Sadness	771	127	355	25488
	Total	3885	658	1895	140725
Igbo	Anger	578	97	290	9090
	Disgust	538	89	271	9015
	Fear	219	36	111	3494
	Joy	467	77	234	11345
	Sadness	493	82	247	7835
	Total	2295	381	1153	40779

Table A.1: Class-wise distribution of train, validation, and test set for Track A, where W_T denotes class-wise total words in the train set.

Language	Classes	Train	Valid	Test	W_T
Algerian Arabic	Anger	296	31	293	4060
	Disgust	206	28	202	3140
	Fear	223	26	216	3442
	Joy	153	13	160	2403
	Sadness	404	40	405	6552
	Surprise	313	32	305	4663
	Total	1595	170	1581	24260
Chinese	Anger	1178	92	1162	1300
	Disgust	403	32	417	440
	Fear	71	5	74	80
	Joy	529	37	537	785
	Sadness	354	22	386	394
	Surprise	178	17	193	188
	Total	2713	205	2769	3187
Hausa	Anger	408	67	209	7207
	Disgust	329	55	168	4290
	Fear	327	53	169	4577
	Joy	320	53	162	4320
	Sadness	647	109	328	10390
	Surprise	349	57	177	4078
	Total	2380	394	1213	34862
Russian	Anger	349	68	116	3415
	Disgust	154	32	56	1701
	Fear	284	44	68	2702
	Joy	429	52	119	3521
	Sadness	290	44	73	2860
	Surprise	231	34	56	1879
	Total	1737	274	488	16078

Table A.2: Class-wise distribution of train, validation, and test set for Track B, where W_T denotes class-wise total words in the train set.

Language	Classes	Train	Valid	Test	W_T
Amharic	Anger	1188	207	582	28987
	Disgust	1268	209	628	27307
	Fear	109	22	54	1975
	Joy	549	93	276	10329
	Sadness	771	127	355	15657
	Surprise	151	27	82	2828
	Total	3549	592	1774	69926
Algerian Arabic	Anger	296	31	293	4060
	Disgust	206	28	202	3140
	Fear	223	26	216	3442
	Joy	153	13	160	2403
	Sadness	404	40	405	6552
	Surprise	313	32	305	4663
	Total	901	100	902	12914
Hausa	Anger	408	67	209	7207
	Disgust	329	55	168	4290
	Fear	327	53	169	4577
	Joy	320	53	162	4320
	Sadness	647	109	328	10390
	Surprise	349	57	177	4078
	Total	2145	356	1080	29279
Oromo	Anger	646	108	323	18533
	Disgust	557	94	275	11338
	Fear	123	21	65	2911
	Joy	1091	183	547	18403
	Sadness	298	52	159	6631
	Surprise	129	27	69	2357
	Total	3442	574	1721	67780
Somali	Anger	328	55	163	9611
	Disgust	477	83	241	12816
	Fear	305	50	149	7167
	Joy	595	99	297	12073
	Sadness	391	67	194	10151
	Surprise	179	28	88	3462
	Total	3392	566	1696	78451

Table A.3: Class-wise distribution of train, validation, and the test set for track C, where W_T denotes class-wise total words in the train set.

Table A.1 and A.2 show data for languages like Marathi, Hindi, Russian, Amharic, Igbo, and Algerian Arabic, and large imbalances in class distribution. For example, in Amharic, there are 1268 samples for *Disgust* but only 109 for the *Fear* class. Furthermore, Table A.3 is extended to include *Surprise* as an additional emotion class and covers Amharic, Algerian Arabic, Hausa, Oromo, and Somali. Class imbalance is particularly evident in languages like Oromo, where *Joy* (547 samples) dwarfs *Fear* (65 samples). The tables also contain W_T values that are the total number of words in the training set, higher for more-resourced languages (e.g., 69926 for Amharic in Table A.3) than lower-resourced languages (e.g., 2911 for *Fear* in Oromo).

B Error Analysis

Both quantitative and qualitative error analyses were conducted to gain a deeper understanding of the performance of the best-performing model.

B.1 Quantitative Analysis

This section offers a detailed quantitative error analysis of the results from the best-implemented model across different languages for the subtasks.

Multi-label Emotion Detection

Figures B.1, B.2, B.3, B.4, and B.5 present the label-wise confusion matrices for five languages (Amharic, Igbo, Marathi, Russian, and Hindi, respectively). In comparison, the true positive detection for the joy class is significantly higher in Marathi (138) compared to Amharic (169), indicating better performance in detecting joy-related expressions in Marathi.

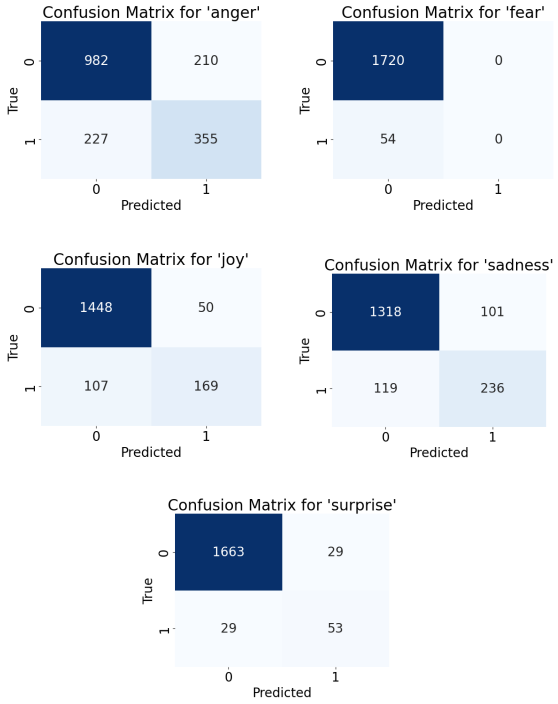

Figure B.1: Confusion matrix of the proposed model (XLM-R) for Amharic language in Track A.

For instance, in Amharic, the model predicts 1720 non-fear instances but fails to detect any actual fear cases, which indicates a severe class imbalance issue. The relatively improved joy detection in Marathi may suggest better linguistic features for identifying joy in the Marathi dataset or better representation in training data. The poor performance in detecting fear class can be attributed to fear being expressed implicitly or contextually, making it difficult for models to detect. Moreover, cultural differences in how fear is expressed might have played a role. The fear class had fewer labeled instances,

limiting the model's ability to learn features associated with it, and model bias in multi-label settings favored more frequent emotions, further suppressing fear detection.

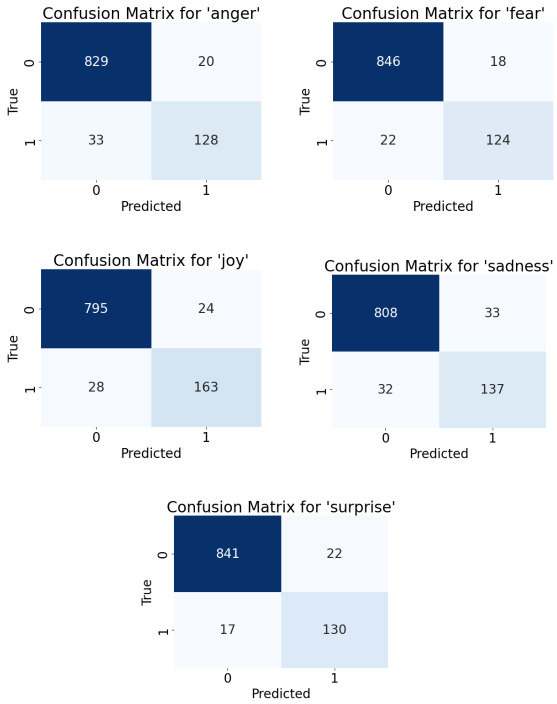

Figure B.2: Confusion matrix of the proposed model (XLM-R) for Hindi language in Track A.

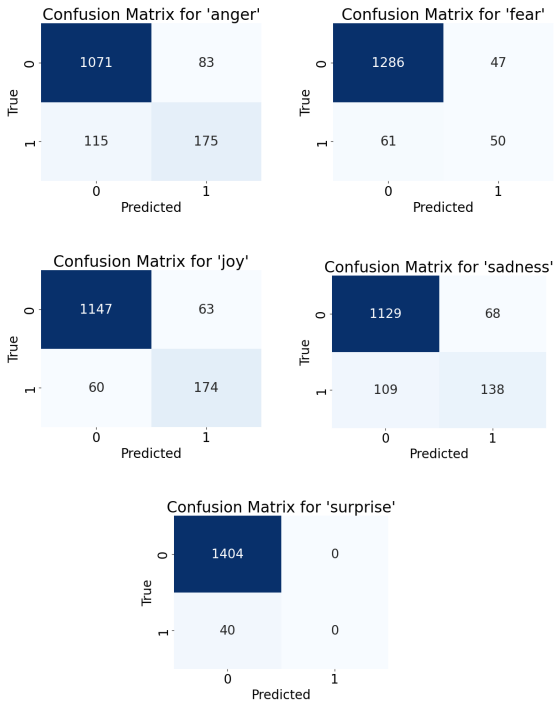

Figure B.3: Confusion matrix of the proposed model (m-BERT) for Igbo language in Track A.

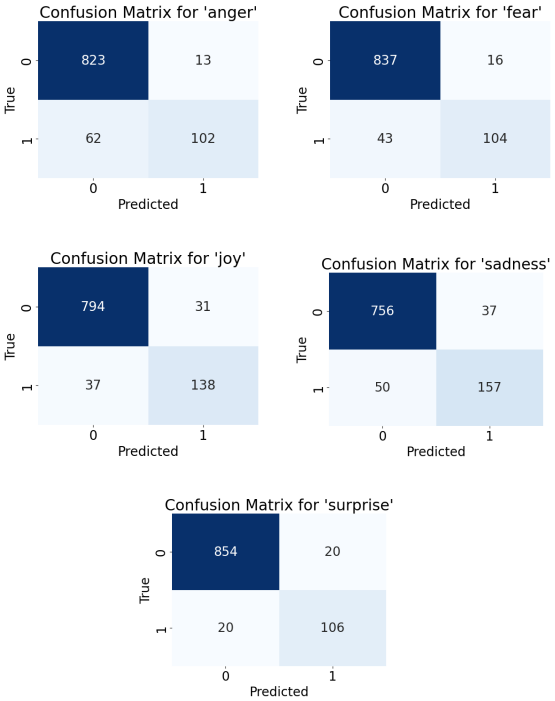

Figure B.4: Confusion matrix of the proposed model (XLM-R) for Marathi language in Track A.

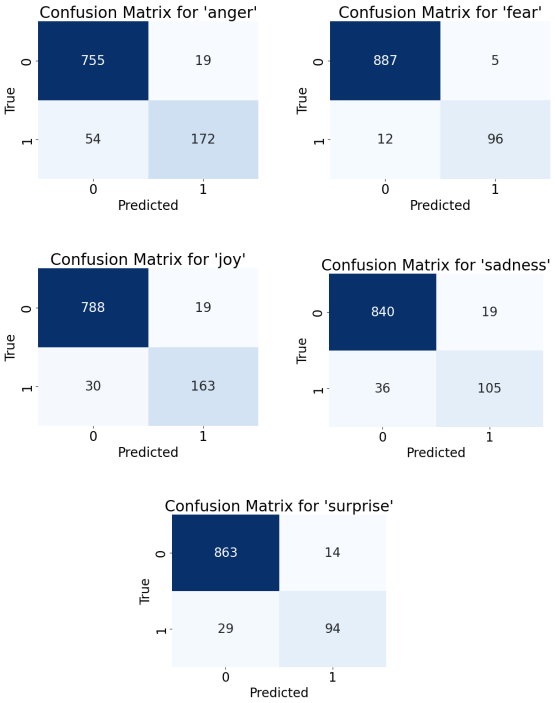

Figure B.5: Confusion matrix of the proposed model (XLM-R) for Russian language in Track A.

On the other hand, anger detection performs better in Amharic (355) than in Marathi (102). Fear detection remains notably poor across all languages, with very low true positive values (0 for Amharic, 50 for Igbo, 104 for Marathi, 96 for

Russian, and 124 for Hindi). The model appears to be biased towards predicting negative instances (class 0), as evidenced by the consistently high true negative values across all languages.

Multi-label Emotion Intensity Detection

Figures B.6, B.7, B.8, and B.9 illustrate the label-wise confusion matrices for four languages: Russian, Algerian Arabic, Chinese, and Hausa. Each demonstrated varied performance in detecting six emotions: *anger, disgust, fear, joy, sadness, and surprise*. In Russian, the model exhibits high accuracy in recognizing *anger* and *disgust*, with 275 and 311 true positives and no false negatives, but it shows limitations in detecting *joy* and *fear*, with 52 and 44 false negatives, respectively. Algerian Arabic performs well in accurately detecting *anger* and *disgust*, with 69 and 72 true positives and no false negatives for both, yet struggles with the identification of *joy* and *fear*, as evidenced by 13 and 26 false negatives. The Chinese model is proficient in distinguishing *sadness* and *disgust* but has difficulties with *fear* and *surprise*, where false negatives are noticeable at 5 and 17.

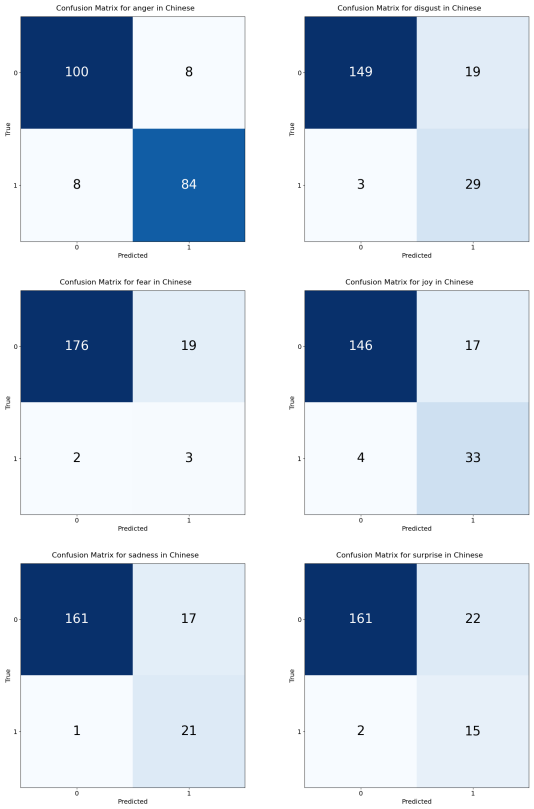

Figure B.7: Confusion matrix of the proposed model (XLM-RoBERTa) for Chinese language in Track B.

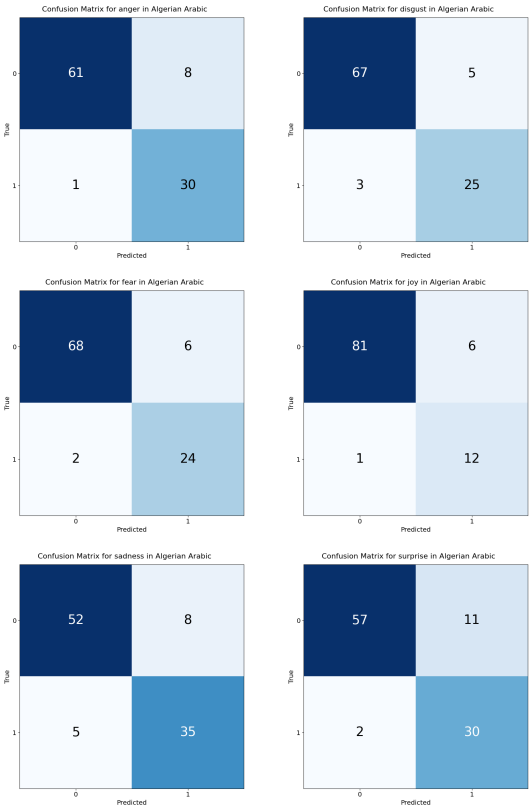

Figure B.6: Confusion matrix of the proposed model (XLM-R) for Algerian Arabic language in Track B.

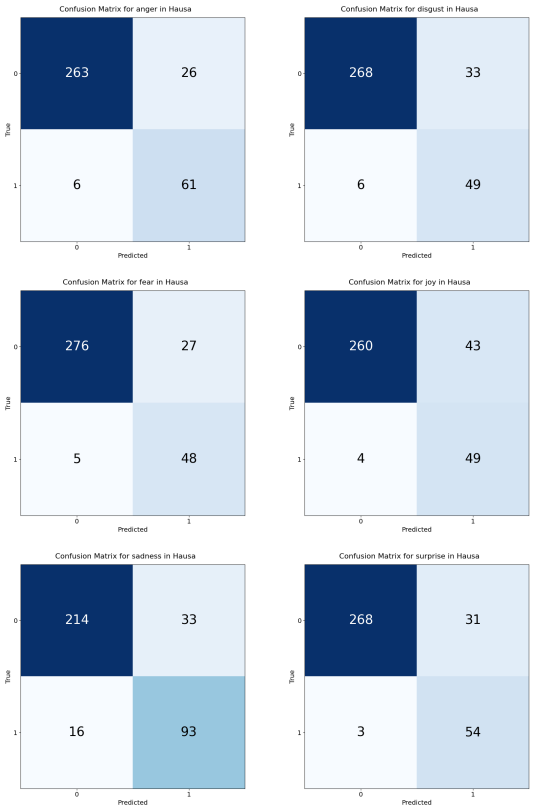

Figure B.8: Confusion matrix of the proposed model (XLM-RoBERTa) for Hausa language in Track B.

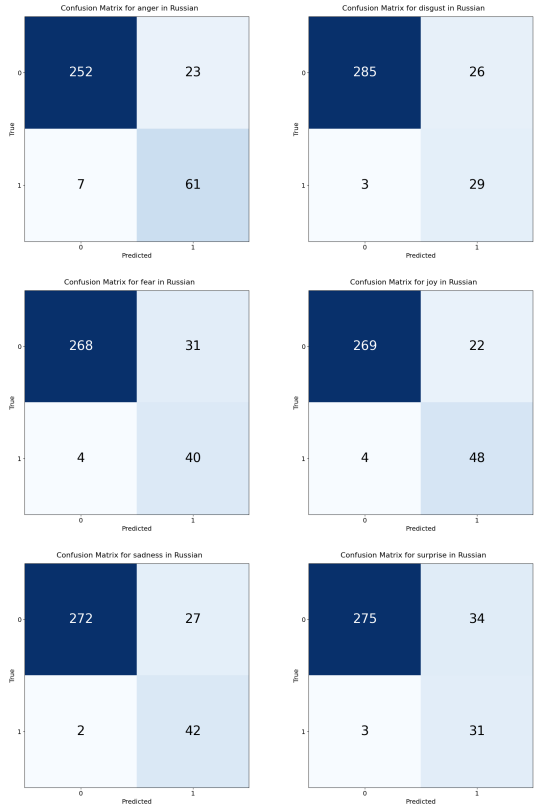

Figure B.9: Confusion matrix of the proposed model (XLM-RoBERTa) for the Russian language in Track B.

However, Hausa shows robust classification in *disgust* and *sadness* with 301 and 247 true positives, respectively, but faces challenges in *joy* and *surprise*. These matrices show the effectiveness of the model in recognizing certain emotions while pointing out specific areas of improvement, which may be affected by cultural or linguistic factors in the training data.

Cross-lingual Emotion Detection

Figures B.10, B.11, B.12, B.13, and B.14 illustrate the confusion matrix on the label for five languages (Amharic, Algerian Arabic, Hausa, Oromo, and Somali, respectively). However, between Oromo and Amharic languages, true positive detection (class 1 predicted correct as 1), *joy* performs significantly better in Oromo (409) compared to Amharic (146), showing that the model is better used to detect Oromo expressions of *joy*. *Anger* detection, however, is better in Amharic (333) than in Oromo (150). Detection of *fear* is extremely poor for both languages, with extremely low true positive values (5 for Amharic, 17 for Oromo). The model is highly biased towards predicting negative instances (class 0) for both languages, as evident from the consistently high true negative scores (class 0 predicted as 0). This suggests the likelihood of training data skewness or model calibration issues.

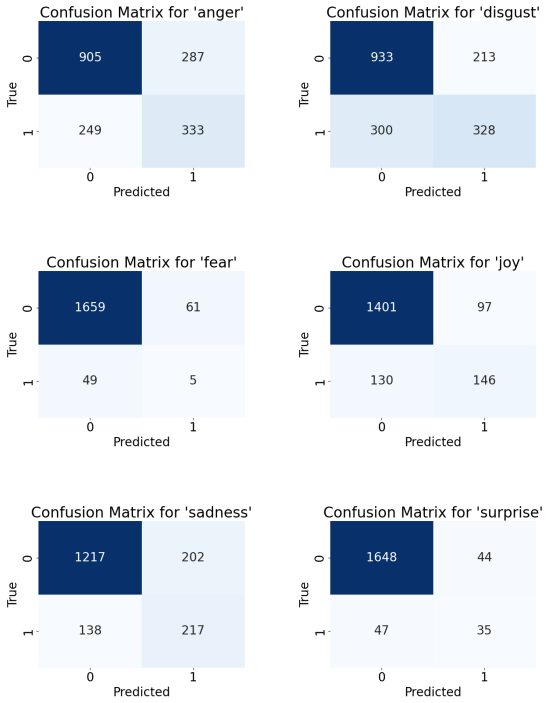

Figure B.10: Confusion matrix of the proposed model (BiLSTM+CNN) for Amharic language in Track C.

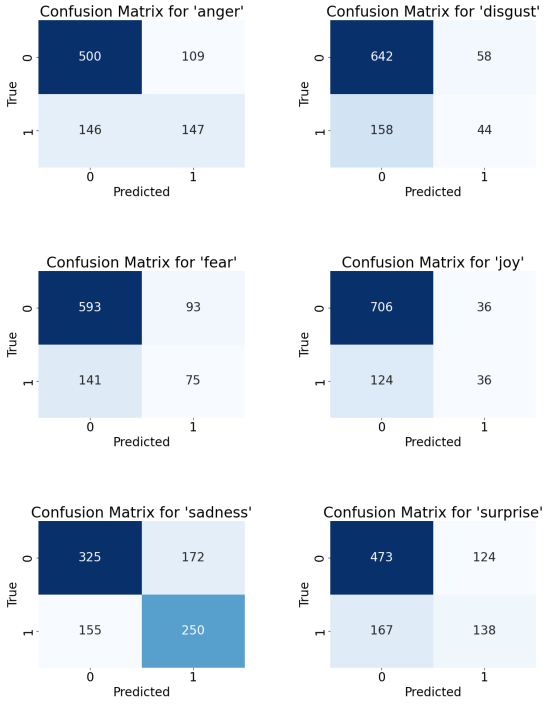

Figure B.11: Confusion matrix of the proposed model (BiLSTM+CNN) for Algerian Arabic language in Track C.

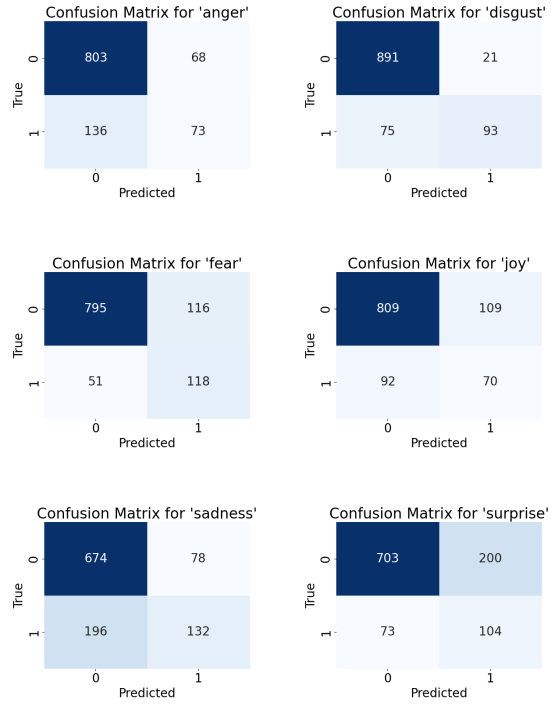

Figure B.12: Confusion matrix of the proposed model (BiLSTM+CNN) for Hausa language in Track C.

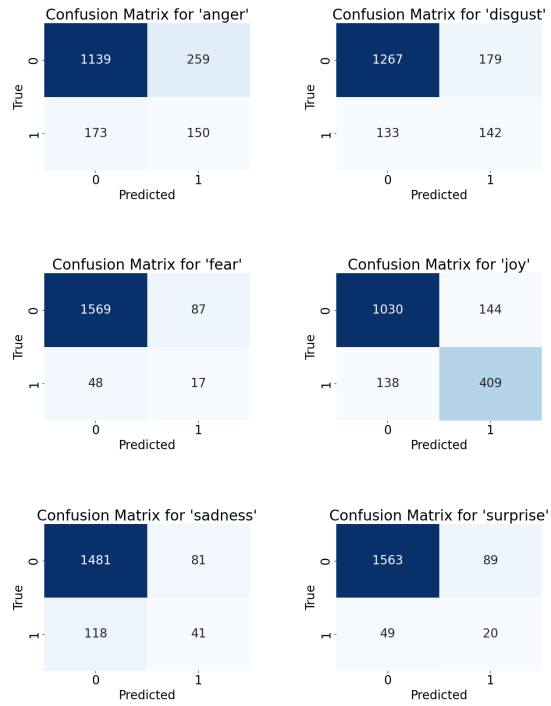

Figure B.13: Confusion matrix of the proposed model (BiLSTM+CNN) for Oromo language in Track C.

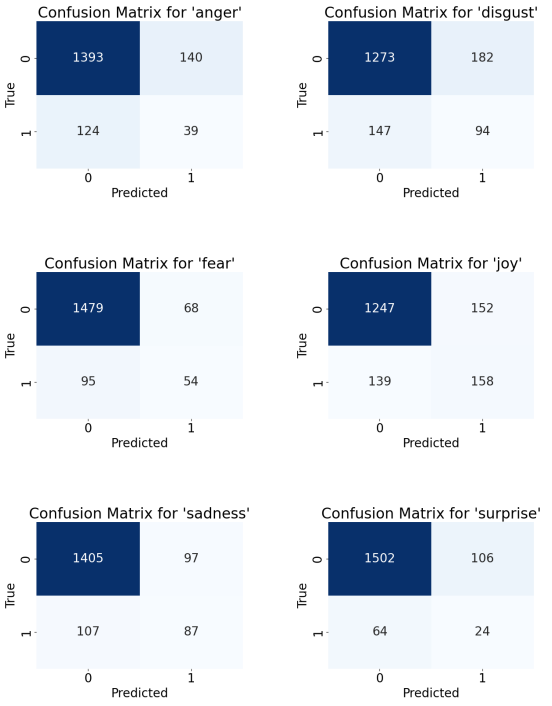

Figure B.14: Confusion matrix of the proposed model (BiLSTM+CNN) for Somali language in Track C.

For instance, for Amharic, the model predicts 1659 *non-fear* instances but only 5 actual *fear* instances, which is a case of extreme class skewness. The relatively improved performance for *joy* detection in Oromo may be explained by more discriminated linguistic features for *joy* in this language or even greater coverage in the training data. The consistently poor performance in detecting *fear* in both languages shows that expressions of *fear* may be more culturally coded or context-bound and hence more challenging to identify in a cross-lingual setting.

B.2 Qualitative Analysis

This section offers a detailed qualitative error analysis of the results of the best-implemented model in different languages for the subtasks.

Multi-label Emotion Detection

Figures B.15 and B.16 illustrate the multi-label emotion detection results, highlighting both strengths and weaknesses in the model's ability to identify emotions in Russian and Amharic text. The model demonstrates strong accuracy in cases where emotions are clearly stated, as seen in Russian Sample 1 and Amharic Sample 2, where the

predicted emotions align perfectly with the actual labels.

Sample Text	Actual Label	Predicted Label
Sample 1: ጦርነት ሲጀመር ቀላል ይመስላ። እግዚያብሔር ለሁለቱም ወገን ሠላሙን ያምጣ። ኢትዮጲያውያን ግን ጉዳዩን ከሀይማኖት ጋር አያይዛችሁ ሀገራችሁን ለመበጥበጥ አጀንዳ አትፍጠሩ። ይህ ጦርነት የድንበር የይዞታ ጉዳይ እንጂ ሀይማኖታዊ አይደለም (When a war starts, it seems easy. May God bring peace to both parties. But Ethiopians, do not link the issue with religion and create an agenda to destroy your country. This war is a border issue, not a religious one)	Anger, Sadness	Anger
Sample 2: አንድን ህዝብ በጅምላ እየሰደብክ ነገ ጉራጌን እንደማትሰድብ ምን ማስተማመኛ አለ?ጉራጌ እጅግ ከሚጠላቸው ነገሮች አንዱ ዘረኝነት ነው።የኦሮሞንም ሆነ የአማራን ህዝብ እኩል ህዝቤ ብሎ (What is the assurance that you will not insult Gurage while mass insulting a nation? One of the things Gurage hates the most is racism.)	Anger	Anger
Sample 3: በእውነት መሳይ መኮንን እዚህ መካከል ሳየው ህፃን መንሱርን መርሳታቸው ገርሞኛል የሚገርም መንግስት! (When I see a really handsome officer here, I am surprised that they forgot baby Mansoor. Amazing government!)	Surprise	Surprise

Figure B.15: Few sample predictions by the XLM-R for the Amharic language in Track A.

Sample Text	Actual Label	Predicted Label
Sample 1: у меня не было этой игры, но чет немного страшно .-. #анжелапедофил (I didn't have this game, but it's a little scary.-. #angelapedophile)	Fear	Fear
Sample 2: наконец то у нас всё хорошо (finally everything is fine with us)	Joy	Joy
Sample 3: Выходные они такие...КОРОТКИЕ!Ужасно короткие! (Weekends are so...SHORT! Terribly short!)	Sadness	Sadness

Figure B.16: Few sample predictions by the XLM-R for the Russian language in Track A.

However, it struggles with complex or nuanced expressions, particularly in cases of mixed emotions. For instance, in Amharic Sample 1, the actual emotions include both *Anger* and *Sadness*, but the model predicts only *Anger*, suggesting difficulty in capturing layered emotional content. Similarly, the model's ability to differentiate emotional intensity varies, as seen in Russian (Sample 3), where the sentiment is correctly classified as *Sadness*, but subtleties in emotional expression may require further refinement. These findings highlight the challenges of multilingual emotion detection, particularly for languages with unique linguistic structures and cultural expressions of emotions. Improving contextual sensitivity could enhance the robustness of such models for diverse languages.

Multi-label Emotion Intensity Detection

Figure B.17 illustrates the results of an emotion intensity detection task in Hausa language samples with rich qualitative data. For Sample 1, which describes the definition of terrorists and bandits (with crying emojis), the model correctly detects *Anger* but at a lower intensity (1) compared to the ground truth (2), correctly detecting *Sadness* as well. Notably, the model also predicts *Fear (1)*, which, in the case of the threatening content, is to be expected but was not annotated by humans.

Sample Text	Actual Label	Predicted Label
Sample 1: Ga manyan Yan ta'ada can da Yan bindiga suna holewa abanza Sai wadanda suke kokarin neman yancin kansu😭😭😭 (There are big terrorists and bandits who are trying to find their freedom 😭😭😭)	Anger (2), Sadness (2), The rest are (0)	Anger (1), Fear (1), Sadness (1), The rest are (0)
Sample 2: Yana shiga cikin gonar ya taka kashin mutun seda ya koma bakin gari ya wanke. (He went into the garden and stepped on the dead man's bone and went back to the city to wash.)	Disgust (2), The rest are (0)	Disgust (3), The rest are (0)
Sample 3: wasu yan takarar pdp na kyamar hoto da jonathan (some pdp candidates hate jonathan's image)	All emotions (0)	All emotions (0)

Figure B.17: Few sample predictions by the Bi-LSTM+CNN for the Hausa language in Track B.

Sample 2, a disturbing account of treading upon human bodies, shows the model correctly detecting *Disgust* as the most prominent emotion but overestimating its strength (3 versus the ground truth of 2). This suggests that the model might be more sensitive to *Disgust*-fostering content than humans are. Sample 3, about political candidates hating someone's picture, was correctly labeled as having no emotions (0) in all categories, which indicated the model's ability to differentiate between emotive content and factual information. These results indicate the model's strengths in emotion type detection but not intensity estimation, which highlights the subtle challenge in multilin-

gual emotion intensity detection, particularly for resource-scarce languages.

Cross-lingual Emotion Detection

Figures B.18 and B.19 illustrate the cross-lingual emotion detection results with strengths and weaknesses to detect emotions between the Hausa and Oromo languages.

Sample Text	Actual Label	Predicted Label
Sample 1: Allah yarabamu da corona ka h bamu Corola🙏🚗 (God bless us with corona and give us Corola 🙏🚗)	Fear	Fear, Sadness
Sample 2: Ministar Tinubu Ta Bayyana Malamin Addinin da Ya Mata Addu'a Ta Samu Mukami, Ta Yi Godiya (Minister Tinubu Explained The Religious Teacher Who Prayed For Her To Get The Position, She Was Thankful)	Joy	Joy, Surprise
Sample 3: Akatambayi buzu ko ya Iya s allah Sai yace ah ah Sai akace to ko kokas an rasulullah Sai yace 😁😁😁😁😁😁😁😁m Kun Kara badda ni. (He asked Buzu if he can pray and he said ah ah. Then he was asked if he was prayin g. Then he said 😁😁😁😁😁😁😁😁m Kun Kara badda me.)	Joy	Anger
Sample 4: Ibrahim ya bata kayan sa da kashin alade (Abraham lost his clothes with a pig's bone)	Disgust	Disgust

Figure B.18: Few sample predictions by the Bi-LSTM+CNN for the Hausa language in Track C.

Sample Text	Actual Label	Predicted Label
Sample 1: Bakar Waaree sodaa nafxenyaa f xiqqollee hin suukannooyne. Arra garuu Warra Sodaa Nafxenyaaf iyyuutu Goota. N afxenyaan awwalamte akka lammada hin d eebinetti. Isii irraanfadhaa. Qalbii takkaan r afaa. Warri biyya harkaa qabu oromoodha. *(Bakar Waaree was not the least bit terrified of the fear of the Nazis. Today, however, even those who are afraid of the Nazis are Heroes. Nafxanya was buried so that she would neve r come back. Forget about her. Sleep with on e heart. Those in charge of the country are O romo.)*	Anger	Anger, Disgust
Sample 2: This team work make the great re sult which broud us.Galatoomaa ijoollee biyy aa nu boonsitan nutis hojii gaarii hojjettaniin baay'ee gammaneerra. *(This team work make t he great result which broud us.Thank you guy s for making us proud and we are very happy with the good work you have done.)*	Joy	Joy
Sample 3: Halkan edaaKonya Horro Guduru u wallaggaa Ona Amuuruu Araddaa oborraatt iIrree fi gaachanni oromoo WBO waan ajaa'ib aa hojjetee jira injifannoon goolabeeraas.Um manni horro Guduruu wallaggaa baga Gamm addan kanaaf tumsa barbaachisu gochuun bira dhaabbadhaa WBO bira. *(The WBO has done a wonderful job in the village of Amuuruu in t he district of Horro Guduruu Wallagga last ni ght.)*	Joy	Disgust, Joy

Figure B.19: Few sample predictions by the Bi-LSTM+CNN for the Oromo language in Track C.

While the model performs flawlessly for certain emotions (perfect accuracy for *Disgust* in Hausa Sample 4), it falls short at more subtly expressed emotions, particularly where emojis are used (misclassifying *Joy* as *Anger* in Hausa Sample 3 despite numerous smiling emojis). The model tends to predict more than a single emotion, showing uncertainty in its prediction, as seen in predictions like *Fear, Sadness* and *Joy, Surprise.* There are language-specific tendencies where Oromo samples have mixed emotion predictions (like Sample 3 with *Disgust, Joy*), echoing potential challenges with the contextual understanding of this language. These findings highlight the need for improved cultural and contextual sensitivities in multilingual emotion detection systems, particularly for low-resource African languages, whose emotional expressions may be very different from those of high-resource languages on which the models are typically trained.

C Performance Ranking

Table C.1 outlines the F1 scores and rankings of the proposed framework compared to the baseline model (RemBERT) across corresponding languages and tracks.

Language	Models	F1 Score	Rank
Track A			
Marathi	Proposed Framework	0.807	22
	Baseline (RemBERT)	0.822	18
Hindi	Proposed Framework	0.838	23
	Baseline (RemBERT)	0.855	17
Russian	Proposed Framework	0.843	22
	Baseline (RemBERT)	0.838	25
Amharic	Proposed Framework	0.542	25
	Baseline (RemBERT)	0.638	15
Igbo	Proposed Framework	0.470	20
	Baseline (RemBERT)	0.479	15
Track B			
Algerian Arabic	Proposed Framework	0.292	18
	Baseline (RemBERT)	0.016	23
Chinese	Proposed Framework	0.4671	20
	Baseline (RemBERT)	0.4053	21
Hausa	Proposed Framework	0.573	14
	Baseline (RemBERT)	0.270	23
Russian	Proposed Framework	0.837	14
	Baseline (RemBERT)	0.876	9
Track C			
Amharic	Proposed Framework	0.459	8
	Baseline (RemBERT)	0.486	6
Algerian Arabic	Proposed Framework	0.436	8
	Baseline (RemBERT)	0.338	12
Hausa	Proposed Framework	0.499	8
	Baseline (RemBERT)	0.319	11
Oromo	Proposed Framework	0.391	4
	Baseline (RemBERT)	0.262	8
Somali	Proposed Framework	0.365	5
	Baseline (RemBERT)	0.273	9

Table C.1: Comparison of results between the proposed framework and baseline models across all tasks.