Participation at SemEval 2025 Task 11: Bridging the Gap in text-based emotion detection

1st Owais Waheed
Computer Science
Habib University
Karachi, Pakistan
ow07611@st.habib.edu.pk

2nd Hammad Sajid
Computer Science
Habib University
Karachi, Pakistan
hs07606@st.habib.edu.pk

3rd Muhammad Areeb Kazmi
Computer Science
Habib University
Karachi, Pakistan
mk07202@st.habib.edu.pk

4th Kushal Chandani
Computer Science
Habib University
Karachi, Pakistan
kc07535@st.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.**

***Index Terms*—Emotion Detection, Emotion Intensity Prediction, Cross-lingual Emotion Detection, BERT, Multilingual NLP, Sentiment Analysis**

I. 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, which involve multi-label emotion classification and intensity prediction in multilingual text. 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.

Given the increasing reliance on AI-driven applications for interpreting human emotions in various contexts, the outcomes of this research will contribute to improving emotion-aware systems. By exploring different model architectures and learning techniques, we seek to establish best practices for addressing emotion detection challenges in both monolingual and cross-lingual settings.

II. Research Question

In this study, we apply deep learning methods to the challenging task of emotion detection and intensity prediction from short text snippets across multiple languages, focusing particularly on multi-label classification and cross-lingual emotion recognition. The primary objective of this study is to determine the most effective models and methods for the above mentioned purpose as part of the SemEval Task 11.

Emotion detection is a complex and nuanced task, as emotions can be expressed and interpreted differently based on language, cultural context, and the speaker's intent. In this study, we aim to develop a model that identifies not just the primary emotion, but the intensity of multiple emotions present in a single text snippet. Our focus is on perceived emotions—those most people believe the speaker might be feeling—rather than the actual or evoked emotions. This distinction ensures the study remains focused, practical, and achievable in scope.

To ensure robust cross-lingual emotion detection, we incorporate datasets spanning multiple languages, highlighting the model's ability to generalize beyond monolingual text. This research will contribute to advancing the field of sentiment analysis and emotion recognition in text, with applications in natural language processing, human-computer interaction, and multilingual AI systems.

III. Literature Review

Given the niche of our task and problem at hand, we have kept the literature review to the most recent papers. While they may not directly work in our case, their usage has been similar to our task and we will discuss them ahead.

A. Emotion Recognition in Conversations (ERC)

ERC has emerged as an important domain within NLP, where the goal is to detect emotional states across sequential utterances in conversations rather than isolated sentences. The complexity of ERC lies in capturing emotion transitions across multiple turns of conversation and understanding the context surrounding the dialogue.

SemEval 2024 Task 10 introduced the Emotion Discovery and Reasoning its Flip in Conversation (EDiReF) challenge [1], which explored ERC in both English and code-mixed Hindi-English dialogues. The dataset provided for the challenge contained manually annotated conversations with emotion labels and triggers for emotion shifts. EDiReF's goal was to enhance emotion recognition by identifying triggers behind emotional transitions. Models developed for this task achieved F1-scores ranging from 0.70 to 0.79 across various subtasks, demonstrating the value of leveraging conversation history in emotion detection. Another team of the same task [4] UMUTeam's approach for ERC used pre-trained transformer models like BERT, showing their effectiveness in recognizing emotions in Hindi-English code-mixed dialogues. Their methodology involved fine-tuning transformers on annotated datasets, which helped capture contextual information crucial for emotion recognition. Despite challenges such as recognizing nuanced shifts in emotions, the model's performance on Hindi-English data (F1 score of 43%) emphasized the importance of code-mixed conversation datasets.

B. Cross-lingual Emotion Detection

A significant contribution in this domain was made by Wang et al. [2], and others have explored how transformers perform in such tasks, highlighting the potential of models like mBERT to effectively manage multilingual datasets. SemEval 2024 Task 10 utilized similar models, demonstrating that cross-lingual capabilities are crucial for tasks involving less-resourced languages, such as those included in Task 11. The ability to predict emotions across languages based on training in one language is vital for such tasks. More recently, transformer-based models, such as BERT and its multilingual variations like mBERT and indicBERT, have been effectively utilized in emotion recognition tasks. For example, Wadhawan and Aggarwal [5] demonstrated state-of-the-art performance using BERT-based models for emotion recognition in Hindi-English code-mixed texts. This aligns with SemEval 2024 Task 10, which focused on emotion recognition in multilingual and code-mixed dialogues, showcasing the relevance of fine-tuned BERT models in understanding nuanced emotional expressions across different linguistic and sociocultural contexts

C. Emotional Flip Reasoning

Emotion flip reasoning (EFR) focuses on understanding why emotional shifts occur within conversations. Unlike ERC, which merely detects emotions, EFR seeks to identify the trigger utterances responsible for emotional flips. This process is crucial for enhancing user experiences in conversational agents by generating more empathetic and contextually aware responses.

Kumar et al.'s [1] task at SemEval 2024 was the first to introduce EFR explicitly. Their work centered on identifying the utterances responsible for triggering emotional shifts within conversations. The EFR dataset annotated conversations with emotions and the corresponding triggers of emotional flips, providing a valuable resource for understanding the causes of emotional transitions. Their approach to EFR involved creating models capable of identifying the specific triggers leading to emotion changes, such as a speaker shifting from joy to sadness in response to a particular utterance.

Another study by the UMUTeam [4] introduced models for EFR in both Hindi-English and English dialogues. They used a combination of transformers and context-based processing to identify the triggers of emotional flips in multi-party conversations. Despite achieving moderate success with an F1 score of 26% for code-mixed dialogues, the results highlighted the difficulty of pinpointing exact triggers due to the complexity of conversational dynamics

D. Ensemble Approach for Semantic textual relatedness

MasonTigers' entry for SemEval 2024 Task 1 [3] stood out for semantic textual relatedness which can be handy for our Task 2. They adhered to the approaches utilizing an ensemble of statistical machine learning combined with language-specific BERT based models and sentence transformers. They were able to bring 0.84 on test Spearman correlation for English.

IV. Dataset Clarity

The dataset used for the our project supports three primary sub-tasks: **Multi-label Emotion Detection**, **Emotion Intensity Prediction**, and **Cross-lingual Emotion Detection**. The structure and clarity of the dataset are essential for building accurate models that can capture the subtleties of emotional expression in text. Below is an overview of the input/output formats, metadata, and categories for each task.

A. Track A: Multi-label Emotion Detection

In this task, the dataset allows for detecting multiple emotions within a single text snippet. The input format includes the text snippet and corresponding binary labels for the emotions: joy, sadness, fear, anger, and surprise. Each emotion is represented as 0 (absent) or 1 (present). The output format consists of the *text_id* and the predicted emotion labels.

- **Dimensions**: The English dataset consists of **2800 rows x 7 columns** (one column for the text snippet, and six columns for the five emotion labels and the text ID).
- **Size**: 2800 samples.

This dataset is commonly used in recent emotion detection research due to its clear multi-label format, making it suitable for training models to detect overlapping emotional expressions. In addition, other similar datasets include SemEval 2018 Task 1: Affect in Tweets, which also uses multi-label emotion detection. However, that dataset focused on both text and image modalities for emotion recognition. [7]

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
sample_03	How stupid of him.	1	0	0	0	0
sample_04	None of us did.	0	0	0	0	0
sample_05	I can't believe it!	0	0	0	0	1

TABLE I
EMOTION LABEL TABLE

B. Track B: Emotion Intensity Prediction

This dataset extends beyond detecting the presence of emotions to predicting their intensity. Each emotion is labeled with an ordinal value ranging from 0 (no emotion) to 3 (high intensity). The input format is similar to Track A, but instead of binary labels, each emotion is associated with its intensity value.

- **Dimensions**: The English dataset contains **2800 rows x 7 columns**, with the same structure as Track A but with intensity values replacing binary labels.
- **Size**: 2800 samples.

This dataset has been instrumental in research focusing on emotional intensity, providing insights into not only which emotions are present but also their degree of expression. Furthermore a similar dataset that we found was the GoEmotions dataset by Google which also provides fine-grained emotion intensity labels for emotion classification tasks. [6]

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
sample_03	How stupid of him.	2	0	0	0	0
sample_04	None of us did.	0	0	0	0	0
sample_05	I can't believe it!	0	0	0	0	3

TABLE II
EMOTION INTENSITY LABEL TABLE

C. Track C: Cross-lingual Emotion Detection

In Track C, the dataset is similar to Track A, emotion detection across multiple languages. This track introduces an additional emotion, *disgust*, in some languages, expanding the emotional categories beyond those in the English dataset. The input format includes text snippets with corresponding emotion labels, while the output format consists of the *text_id* and the predicted labels.

- **Dimensions**: The dataset includes **2603 rows x 8 columns**, with one additional column for the disgust emotion compared to the English datasets.
- **Size**: 2603 samples.

This dataset is valuable in cross-lingual emotion detection research, allowing models to generalize across different languages and cultures while maintaining a consistent set of emotions. Another similar dataset was appeard in SemEval 2024: Emotion Discovery and Reasoning Task (EDiReF). It involved emotion recognition and emotion flip reasoning in Hindi-English code-mixed dialogues. [1]

The test dataset will be in another language, in our case, Russian, to predict the emotion labels we used.

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
Die Kriegsbegeisterung kennt anscheinend	0	0	0	1	0	0
Selenskiy = reichster Bettler	0	0	0	0	0	1
Ich liebe diese Doppelmoral, Russland greift	0	1	0	0	0	0

TABLE III
GERMAN TEXT EMOTION LABELS

D. Input/Output Formats

- **Input**: Text snippets with corresponding labels for each emotion (binary labels for Track A and C and intensity labels for Track B).
- **Output**: Text IDs with predicted emotion labels for each snippet.

V. METHODOLOGY

A. Track A: Multi-label Emotion Detection

Initially, we trained the **BERT-base-uncased** model comprising of 110 M parameters for the baseline of our task. BERT based models are the most suitable for classification tasks due to them being encoder based and having bi-directional context understanding capabilities. This baseline model was trained on commonly used hyperparameters mentioned in the table IV.

To improve the efficiency of our tasks, ensembling approach was used that combined multiple individual models to improve overall predictive performance and robustness. The models chosen were:

- **DistilBERT-base:** Faster and smaller than standard BERT model so it is efficient in NLP tasks.
- **DeBERTa-base:** Good for High-accuracy NLP tasks.
- **XLM RoBERTa-base:** Contains large multilingual corpora to improve classification accuracy.

Ensembling of models combines the predictions of multiple models to produce a more robust and accurate outcome than any individual model, leveraging the strengths and compensating for the weaknesses of different models.

Initially, we used the **Gemini API** for dataset augmentation, leveraging its capabilities to generate enriched data. However, its 60 requests per minute limit and session constraints on Google Colab made the process slow and less scalable, hindering efficient dataset preparation.

To address these issues, we switched to the **GPT-3.5 Turbo API**, which provided greater flexibility and scalability. Additionally, we applied class upsampling to mitigate class imbalances by increasing data for underrepresented classes. These improvements resulted in a more balanced, diverse, and high-quality dataset.

The size of the dataset increased from merely 2800 rows to over 5000 rows with some classes balanced (despite upsampling not all classes could be balanced).

TABLE IV
HYPERPARAMETERS FOR TRAINING OF BERT AND ENSEMBLED MODELS

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

B. *Track B: Emotion Intensity*

Initially, 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. To map these continuous predictions to the discrete labels (0, 1, 2, 3), the values were rounded to the nearest integer. This approach served as a baseline for the task.

To enhance performance, we implemented an ensembling approach by training five different models. Each model was evaluated based on its Mean Squared Error (MSE), and three top-performing models were selected for the final ensemble. The ensembling process combined the predictions of these models to produce more robust and accurate intensity values.

TABLE V
TRAINING LOSS WITH MSE ACROSS MODELS

Model	Training Loss	MSE
microsoft/deberta-v3-base	0.1202	0.2259
microsoft/deberta-v3-large	0.0466	0.2331
roberta-base	0.0861	0.2572
bert-base-uncased	0.0742	0.2716
distilbert-base-uncased	0.1356	0.2927
FacebookAI/xlm-roberta-base	0.4045	0.3630

The hyperparameters used for training all models were consistent:

- **Epochs**: 5
- **Batch size**: 16
- **Learning rate**: 5×10^{-5}

This ensemble approach improved the predictive accuracy of emotion intensity by leveraging the unique strengths of each model while mitigating their individual weaknesses.

Note: We used Augmentation techniques in this task as well, in order to extend the data synthetically. However, the results came out to be below the baselines of the task. The reason being, in this task, we have multiple labels, depending on emotional intensity (0, 1, 2, and 3), for each class, hence when augmenting the data, the data came out to be of different emotional intensity than the original (despite the prompt clearly indicating to keep the intensity of the emotion same), hence leading to creating confusion for the model. Hence, although working upon it, we have not included the results of training the model on Augmented data.

C. *Track C: Cross-lingual Emotion Detection*

Our approach to Track C was similar to that of Track A, with minor changes. We trained the **BERT-based** model comprising of 110 M parameters for the baseline for the task. Here too, BERT based models are the most suitable for our task due to them being encoder based and having bi-directional context understanding capabilities. We trained our dataset with the chosen models below:

- **DistilBERT-base-multilingual:** Faster and smaller than standard BERT model so it is efficient in NLP tasks and performs well on different languages.
- **multilingual-BERT-base-cased:** Good for High-accuracy NLP tasks involving multiple languages.
- **XLM RoBERTa-base:** Contains large multilingual corpora to improve classification accuracy.

We note performance in the table below where we split the data into train and test with the same consistency that we have done throughout the project:

TABLE VI
TRAINING LOSS AND VALIDATION LOSS ACROSS MODELS

Model	Training Loss	Validation Loss
FacebookAI/xlm-roberta-base	0.3402	0.2975
distilbert-base-multilingual	0.3156	0.2789
multilingual-bert-base-uncased	0.3400	0.3006

The hyperparameters used for training all models were consistent:

- **Epochs**: 4
- **Batch size**: 16
- **Learning rate**: 5×10^{-5}

However, as an experiment to improve our results, we worked on another technique

D. *Cross-lingual Emotion Detection using Translation Method*

We carry out the the same work we did for Track C except for a middle layer of translation.

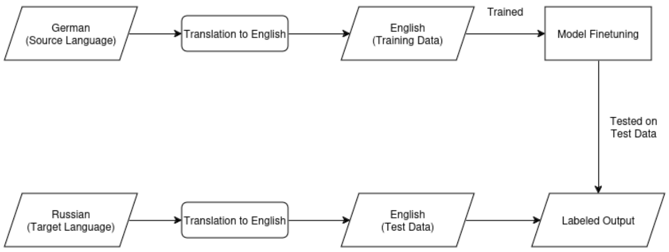

Fig. 1. Cross-lingual Emotion Detection Using Translation

Here, in Figure 1, we translate our Training Language, German, into English using University of Helskinki's translation model available at HuggingFace.co and finetune our model by training the data on it. Next, we translate our Target Language, Russian, into English, and test it on our fine-tuned model to produce the labeled output.

TABLE VII
TRAINING LOSS AND VALIDATION LOSS ACROSS MODELS

Model	Training Loss	Validation Loss
FacebookAI/xlm-roberta-base	0.3252	0.3801
distilbert-base-multilingual	0.3538	0.3574
multilingual-bert-base-uncased	0.3232	0.3711

We note the training loss and validation loss in the table below:

The hyperparameters used for training all models were consistent:

- **Epochs**: 4
- **Batch size**: 16
- **Learning rate**: 5×10^{-5}

VI. RESULTS

A. Track A: Multi-label Emotion Detection

Initially, we fine-tuned the **BERT-base-uncased** model to establish a baseline for performance. 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 XI, 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.

TABLE VIII
PERFORMANCE OF VARIOUS FINE-TUNED MODELS EMPLOYED FOR PREDICTING EMOTION LABELS.

Model(s)	Accuracy	Micro F1 Score	Macro F1 Score
BERT-base-uncased (baseline)	0.48	0.65	0.61
Ensembled (DistilBERT, DeBERTa, Distilbert)	0.52	0.69	0.67

B. Track B: Emotion Intensity

For Track B, we evaluated the performance of our models on emotion intensity prediction using Pearson correlation (r) as the evaluation metric. The baseline model, **BERT-base-uncased**, achieved promising results, with the average Pearson correlation across emotions being **0.674**. To improve upon these results, we applied an ensemble approach combining best 3 models which were **DeBERTa-v3-base**, **DeBERTa-v3-large**, and **roberta-base**. By aggregating the predictions of these models, the ensemble demonstrated a significant improvement in performance, achieving an average Pearson correlation of **0.7279**.

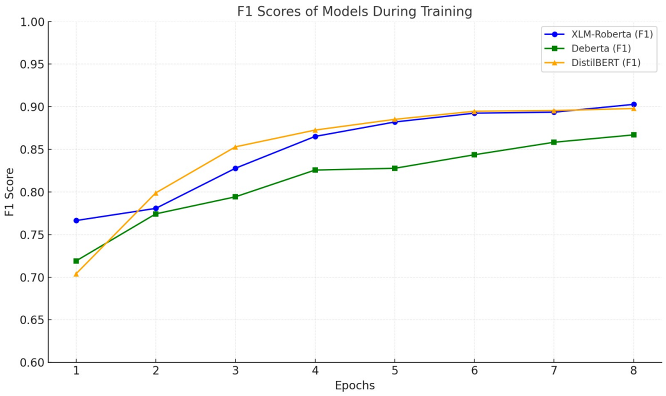

Fig. 2. Visual representation of F1 scores across ensembles models. XLM Roberta shows the best performance.

Overall, the ensemble approach consistently outperformed the baseline across all emotion categories, demonstrating its efficacy in capturing nuanced intensity levels. These results highlight the potential of ensemble learning to enhance the prediction of emotion intensity, especially for subtle emotional states.

TABLE IX
EVALUATION SCORES FOR EMOTION INTENSITY (PEARSON R).

Emotion	Baseline Pearson r	Ensemble Pearson r
Anger	0.6883	0.7201
Fear	0.5566	0.6403
Joy	0.6804	0.7154
Sadness	0.7892	0.8350
Surprise	0.6556	0.7287
Average	**0.6740**	**0.7279**

C. Track C: Cross-lingual Emotion Detection

1) Track C simple approach: For Track C, we initially fine-tuned on the models mentioned earlier. The macro and micro F1 scores reported in Table X represent the result on the test dataset as obtained by submitting the model predictions to **CodeBench**.

Here, the baseline model achieved a macro F1 score of **0.09566** maximum which was in the case of **xlm-roBERTa-base**.

TABLE X
PERFORMANCE OF VARIOUS FINE-TUNED MODELS EMPLOYED FOR PREDICTING EMOTION LABELS.

Model(s)	Accuracy	Micro F1 Score	Macro F1 Score
mBERT-base-uncased	0.05943	0.12208	0.09566
DistilBERT-multilingual	0.06691	0.13382	0.09508
xlm-roBERTa-base	0.06691	0.13382	0.09508

2) Track C using Translation method: Using the translation method, as described in the methodology, we recorded the macro and micro F1 scores in Table X1 which represents the result on the test dataset obtained.

TABLE XI
PERFORMANCE OF VARIOUS FINE-TUNED MODELS EMPLOYED FOR PREDICTING EMOTION LABELS.

Model(s)	Accuracy	Micro F1 Score	Macro F1 Score
mBERT-base-uncased	0.08603	0.15249	0.12367
DistilBERT-multilingual	0.09226	0.13770	0.10687
xlm-roBERTa-base	0.09226	0.17176	0.13268

Here, the best model which achieved a macro F1 score of **0.13268** maximum which was in the case of **xlm-roBERTa-base**. The **mBERT base** was second with a score of **0.12367** while **distilBERT-multilingual** performed least with a macro F1 score of **0.10687**.

In both cases, **xlm-roBERTa-base** turned out to be on top in terms of macro F1 score and helped us to be on the top of Track C's leaderboard as of now.

VII. CONCLUSION

Our paper highlights the efficiency of transformer-based models in addressing the multifaceted challenges of emotion detection. For Multi-label Emotion Detection and Emotion Intensity Prediction, ensemble learning significantly enhanced performance, while the inclusion of innovative techniques like data augmentation ensured balanced training datasets. In the Cross-lingual Emotion Detection task, translation-based methods coupled with fine-tuning on multilingual transformers like XLM-RoBERTa yielded promising results, reinforcing the potential of leveraging intermediate language translations to bridge linguistic gaps.

The findings highlight the importance of ensemble learning and data augmentation for tackling class imbalance and improving generalization across linguistic contexts. Our methods demonstrated adaptability to low-resource languages, providing practical solutions for real-world applications such as sentiment analysis and emotion-aware AI systems.

VIII. FUTURE WORK

In the future, we plan to extend this work by incorporating advanced techniques like contrastive learning to enhance model performance through better feature representation. Additionally, we aim to refine our data preprocessing pipeline by implementing strategic upsampling and downsampling methods to balance the datasets as required for specific tasks. These improvements will enable us to optimize the dataset for various scenarios, enhancing the effectiveness of our augmentation techniques. By integrating these strategies, we anticipate achieving more generalizable results. Furthermore, we plan to use large advanced models like LLaMA 3.2 and T5 for improved performance. For Task 3, we will explore and utilize different languages to ensure a multilingual approach, allowing for broader applicability across diverse datasets.

REFERENCES

[1] S. Kumar, M. S. Akhtar, E. Cambria, and T. Chakraborty, "SemEval 2024 – Task 10: Emotion Discovery and Reasoning its Flip in Conversation (EDiReF)," *arXiv preprint arXiv:2402.18944*, 2024. Available: https://arxiv.org/abs/2402.18944.

[2] Y. Wang, Y. Li, P. P. Liang, L.-P. Morency, P. Bell, and C. Lai, "Cross-Attention is Not Enough: Incongruity-Aware Dynamic Hierarchical Fusion for Multimodal Affect Recognition," in *Proc. of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*, pp. 703–709, June 2024. Available: https://arxiv.org/abs/2305.13583.

[3] D. Goswami, S. S. C. Puspo, M. N. Raihan, A. N. B. Emran, A. Ganguly, and M. Zampieri, "MasonTigers at SemEval-2024 Task 1: An Ensemble Approach for Semantic Textual Relatedness," in *Proc. of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*, pp. 1380–1390, June 2024.

[4] R. Pan, J. A. García-Díaz, D. Roldán, and R. Valencia-García, "UMUTeam at SemEval-2024 Task 10: Discovering and Reasoning about Emotions in Conversation using Transformers," in *Proc. of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*, pp. 703–709, June 2024.

[5] A. Wadhawan and A. Aggarwal, "Towards emotion recognition in Hindi-English code-mixed data: A transformer based approach," in *Proc. Eleventh Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, Online, 2021, pp. 195-202.

[6] D. Demszky, et al., "GoEmotions: A dataset of fine-grained emotions," in *Proc. 58th Annual Meeting of the Association for Computational Linguistics (ACL)*, Online, 2021, pp. 4040-4054.

[7] S. Mohammad, F. Bravo-Marquez, M. Salameh, and S. Kiritchenko, "SemEval-2018 Task 1: Affect in Tweets," in *Proc. 12th International Workshop on Semantic Evaluation (SemEval-2018)*, New Orleans, LA, USA, 2018, pp. 1-17.