SSN_Semeval10 at SemEval-2024 Task 10: Emotion Discovery and Reasoning its Flip in Conversations

Antony Rajesh A **Supriya Abirami A** **Chandrabose Aravindan** **Senthil Kumar B**

Department of Information Technology
Sri Sivasubramaniya Nadar College of Engineering
Chennai, Tamilnadu, INDIA
{*antony*2010532, *supriyaabirami*2010354, *AravindanC*, *Senthil*} @*ssn.edu.in*

Abstract

This paper presents a transformer-based classifier for recognizing emotions in Hindi-English code-mixed conversations, adhering to the SemEval task constraints. Leveraging BERT-based transformers, we fine-tune pre-trained models (mBERT and indicBERT) on the dataset, incorporating tokenization and attention mechanisms. Our approach achieved competitive performance (weighted F1-score of 0.4), showcasing the effectiveness of BERT in nuanced emotion analysis tasks within code-mixed conversational contexts. This F1-score was ranked 16^{th} among the 39 submissions.

1 Introduction

Recognition of emotions from conversation enables advancements in sentiment analysis, mental health monitoring, chatbot development and ultimately enhances user experiences and well-being. The EDiReF shared task (Task 10) at SemEval 2024(Kumar et al., 2024) comprises three subtasks: Emotion Recognition in Conversation (ERC)(Kumar et al., 2023) and Emotion Flip Reasoning (EFR)(Kumar et al., 2022) in both Hindi-English code-mixed conversations and English conversations. ERC involves assigning emotions to each utterance from a predefined set, while EFR aims to identify trigger utterances for emotion flips in multi-party conversations. This task is vital for understanding emotional dynamics in conversational contexts, particularly in multilingual settings like Hindi-English code-mixed conversations.

This paper proposes a classifier for ERC that adopts a BERT-based transformer architecture (Lee, 2022) for emotion recognition task. By fine-tuning pre-trained BERT models, like mBERT (DevlinJ et al., 2018) and indicBERT (Kakwani et al., 2020), on the given dataset, we leverage transfer learning to understand and reason about emotions effectively in multilingual conversational contexts like Hindi-English code-mixed conversations.

We participated in sub-task 1 (ERC) of Task 10 (EDiReF) and competed with 38 other teams within the provided time frame. Our system achieved rank 16 for this sub-task with a range of weighted F1-scores between 0.3 and 0.4 using BERT-based models. While we successfully utilized BERT-based models for emotion recognition in Hindi-English code-mixed conversations, our system encountered challenges in accurately capturing emotional contexts, which affected our overall performance.

2 Background

Sub-task 1 challenges participants to provide emotions as output for particular utterances in conversations. Both training and validation datasets are provided, with both datasets in textual format. The training set includes 343 conversations with 8505 utterances, while the validation set contains 46 conversations with 1354 utterances. Each conversation in both datasets comprises episodes, speakers, utterances, and emotions. Utterances are in Hindi-English code-mixed (e.g., *"Namaste, how are you?"*). The emotion distribution and utterance length distribution for both datasets are summarized in Figure 1 and Figure 2, respectively. Notably, the emotion distribution in both datasets is prominently skewed towards 'neutral', as indicated by the larger area in the distribution. Upon analysis, the emotions involved in both datasets are identified as *['neutral', 'contempt', 'sadness', 'fear', 'joy', 'surprise', 'anger', 'disgust']*.

3 Related Work

In recent years, emotion recognition in conversational contexts has seen significant contributions. (Maheshwari and Varma, 2022) focused on emo-

Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024), pages 553–557

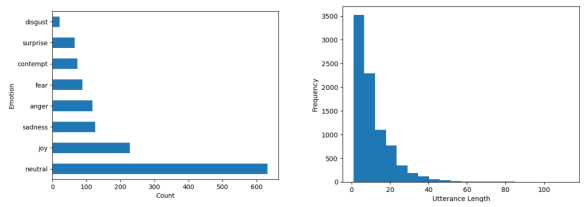

Figure 1: Emotion distribution and Utterance length distribution in training dataset

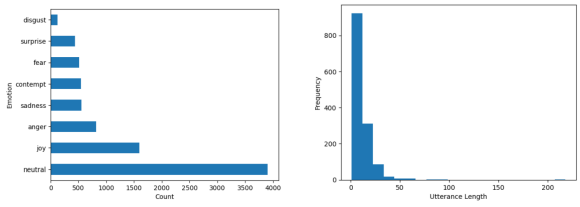

Figure 2: Emotion distribution and Utterance length distribution in validation dataset

tion recognition in tweets, emphasizing the importance of context. (Poria et al., 2019) survey offers a comprehensive overview of emotion recognition systems in dialogues, covering deep learning approaches and challenges. (Wang et al., 2023) study explores using transformers for emotion recognition in conversations, highlighting their effectiveness.

While deep learning has revolutionized the field, earlier works laid the foundation. (Thelwall et al., 2012) and (Pang and Lee, 2008) explored traditional approaches to Emotion Recognition (ER) using hand-crafted features and rule-based systems. (Mohammad and Turney, 2013) and (Tang et al., 2016) marked a shift towards deep learning for ER, focusing on Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs) for learning emotion representations from text data.

(Wadhawan and Aggarwal, 2021) introduces a new dataset for analyzing emotions in Hindi-English tweets and proposes a transformer-based approach using BERT to achieve state-of-the-art accuracy in emotion detection, outperforming other deep learning models like CNNs, LSTMs, Bi-LSTMs.

(D. et al., 2019) also contribute to the field by applying traditional and deep machine learning approaches to identify offensive language in social media, demonstrating the versatility of these techniques in analyzing online sentiment. This aligns with our work on emotion recognition in code-mixed social media data, as both studies explore methods for sentiment analysis in similar contexts.

And the recent case study, (Tatariya et al., 2024) mentions the challenges in code-mixed data for emotion classification. The study investigates the effectiveness of pre-trained language models in understanding sociolinguistic contexts. The findings underscore the importance of considering linguistic diversity and sociolinguistic factors in developing and interpreting emotion recognition models.

(Vijay et al., 2018) pioneered the work on emotion recognition in Hindi-English code-mixed social media text. Their work established a benchmark by creating a corpus of annotated data and proposing a classification system for emotion detection.

Building on Wadhawan and Aggarwal's success with BERT with the help of works done by (**?**) in SemEval 2021 and (Lee, 2022) in emotion recognition in conversations, our mBERT model aims to further improve emotion detection by addressing the cultural nuances and fine-tuning on a larger code-mixed hindi-english dataset while addressing the limitations highlighted by Tatariya et al.

4 System Overview

This section provides an overview of our BERT-based transformer system and justifies our selection of pre-trained BERT models. After data-preprocessing, our system takes conversations as input in form of sequence of tokens and produces emotion class as output for emotion classification. This processs is illustrated in Figure 3.

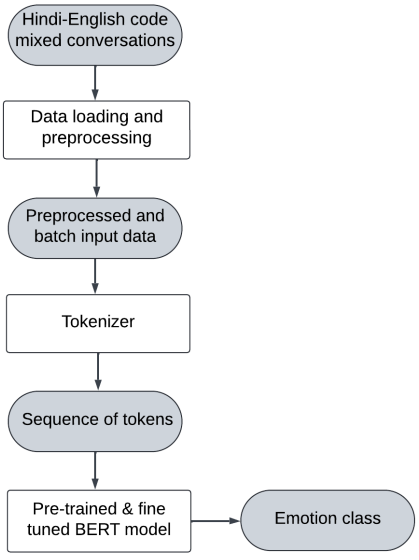

Figure 3: BERT-based Transformer System Overview

For Hindi-English Code-mixed language, we

used mBERT (DevlinJ et al., 2018) and indicBERT (Kakwani et al., 2020). Due to multilingual understanding of mBERT and IndicBERT which is designed for Indic languages, enabling them to process both Hindi and English as well as their mixtures effectively. Their cross-lingual transfer learning capabilities ensure robust performance with minimal fine-tuning, while their rich representations of language capture essential contextual information across language boundaries.

5 Experimental Setup

In this section, we present the implementation details of our system. During the training phase, we used the provided training dataset to fine-tune the BERT models and the validation set for evaluation. Later, both the training and the validation sets were used to fine-tune the BERT models which were submitted for testing using the test set.The stages involved in experiments are detailed below.

5.1 Data Pre-processing steps

5.1.1 JSON Parsing

To facilitate quick access to data samples, the given JSON dataset, containing information on episodes, speakers, utterances, and emotions, will be transformed into a text file with three columns: speakers, utterances, and emotions. A blank line in the text file will serve as the separator between different conversations.

5.1.2 Emotion loading for Specific Utterances and Retaining Previous Dialogue Context

Following JSON parsing, the data loader organizes input by loading the emotion associated with each dialogue alongside its utterances. Utterances undergo text cleaning, removing punctuations and stopwords in Hindi-English code-mixed languages. To grasp the current dialogue's emotion context, the loader loads the sequence of previous dialogues, including their utterances and speaker names. Speaker names are indexed starting from zero *(e.g., 0 for Ram, 1 for Divya)*, facilitating the mapping of utterances to their corresponding speakers. Additionally, the data loader maintains a set of emotions involved in the previous dialogue context.

5.2 Neural Architecture for emotion recognition in conversations

We downloaded the pre-trained mBERT[1] and indicBERT[2] models from the huggingface transformer library. We adopted the code of the transformer model for the Emotion Recognition Challenge (Lee, 2022) to implement our system. We processed the batch tokens through multiple layers of Transformer blocks, including self-attention mechanisms and feed-forward neural networks. For evaluating the performance metrics of every epoch while training, we used precision, recall, and weighted f1 score of the validation set. We used cross entropy loss for loss function and with the help of AdamW optimizer, we have updated the weights involved. The final layer of the BERT model determines the number of emotion classes using a data loader that tracks emotions in the input data, outputting the class label for classification. Class labels are then converted into emotions for analysis in the ERC task.

We conducted our experiments on Google Colab using the T4 GPU runtime mode. We trained the chosen BERT models with a batch size of 1 and a learning rate of 1e-6.

6 Results

We trained both the mBERT and indicBERT models according to the experimental setup. During training, we recorded and stored the weighted F1 scores of both the models on the validation set, which are detailed in Table 1. Notably, mBERT's performance improves until around 7-8 epochs, while indicBERT's score remains stable. Additionally, in Table 2, the final scores for precision, recall, and weighted F1 are detailed. Using the trained mBERT model, we achieved the 16th rank with a weighted F1 score of 0.4 in subtask 1 of task 10, whereas the top ranked system achieved a score of 0.78.

We examined the emotions predicted by two models, mBERT and indicBERT, and compared them to the actual test labels. We visualized the results using a confusion matrix of mBERT in Figure 4.

After normalizing the emotion distribution of the training dataset and mBERT correct predictions, we observed that their patterns appear similar in

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

[2]https://huggingface.co/ai4bharat/indic-bert

Epochs	mBERT	indicBERT
1	29.44	28.34
2	35.35	29.1
3	35.49	29.1
4	38.68	29.1
5	40.72	29.1
6	41.4	29.12
7	41.56	29.18

Table 1: weighted f1 scores of mBERT and indicBERT for 7 epochs

Model	Precision (%)	Recall (%)	Weighted F1 Score (%)
M-BERT	41.46	47.56	42.47
IndicBERT	21.85	46.75	29.78

Table 2: mBERT and indicBERT - Final Performance metrics

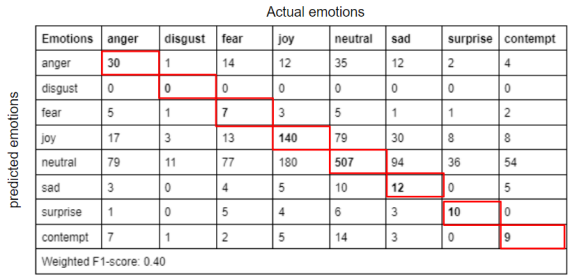

Figure 4: Confusion matrix with Highlighted correctly predicted emotions by mBERT

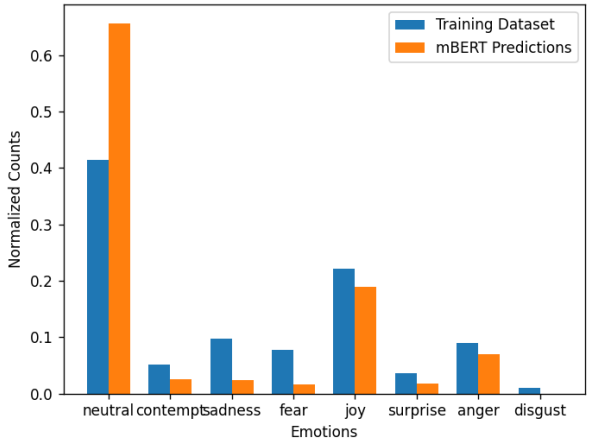

Figure 5: Normalized Emotion distribution of training dataset and mBERT correct predictions

Figure 5. However, *disgust*, *contempt*, *fear*, *sadness*, and *surprise* exhibit the lowest areas in the distribution, indicating that it is challenging for the model to identify utterances with these emotions. Therefore training datasets with a more balanced emotion distribution may possibly enhance the performance of mBERT.

The confusion matrix of indicBERT showed 0 for all the entries except for neutral where 656 test cases were predicted correctly. This clearly indicates that indicBERT has not learnt the contextual representations of utterances in the training dataset. This is primarily due to the fact that indicBERT was trained using Hindi unicode, whereas our dataset uses transliterated Hindi. We will try to resolve the issue in the future.

7 Conclusion

In our study on understanding emotions in Hindi-English conversations for SemEval 2024 Task 10, we used BERT-based models. Our system ranked 16th in subtask 1. However, accurately capturing nuanced emotions posed challenges, suggesting areas for improvement.

For future work, we plan to enhance our system in several ways. First, we aim to expand our dataset with more Hindi-English code-mixed tweets to expose the model to a wider range of expressions. Second, we'll refine our data preprocessing by translating Hindi-English utterances into plain English to reduce ambiguity. Additionally, we'll explore models beyond BERT, like LLAMA and GPT-2, known for text generation and question answering tasks. We'll also investigate specialized models like HingBERT and its family models for improved accuracy in Hindi-English code-mixed text analysis.

In essence, our future research focuses on dataset expansion, preprocessing improvements, and exploring diverse models to better understand emotions in multilingual conversations.

References

Thenmozhi D., Senthil Kumar B., Srinethe Sharavanan, and Aravindan Chandrabose. 2019. SSN_NLP at SemEval-2019 task 6: Offensive language identification in social media using traditional and deep machine learning approaches. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 739–744, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

DevlinJ, Chang M, Lee K, and Toutanova K. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *Computing Research Repository*, arXiv:1810.04805.

Divyanshu Kakwani, Anoop Kunchukuttan, Satish Golla, Gokul N.C., Avik Bhattacharyya, Mitesh M. Khapra, and Pratyush Kumar. 2020. IndicNLPSuite: Monolingual corpora, evaluation benchmarks and pre-trained multilingual language models for Indian languages. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4948–4961, Online. Association for Computational Linguistics.

Shivani Kumar, Md Shad Akhtar, Erik Cambria, and Tanmoy Chakraborty. 2024. Semeval 2024 – task 10: Emotion discovery and reasoning its flip in conversation (ediref). In *Proceedings of the 2024 Annual Conference of the North American Chapter of the Association for Computational Linguistics*. Association for Computational Linguistics.

Shivani Kumar, Ramaneswaran S, Md Akhtar, and Tanmoy Chakraborty. 2023. From multilingual complexity to emotional clarity: Leveraging commonsense to unveil emotions in code-mixed dialogues. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 9638–9652, Singapore. Association for Computational Linguistics.

Shivani Kumar, Anubhav Shrimal, Md Shad Akhtar, and Tanmoy Chakraborty. 2022. Discovering emotion and reasoning its flip in multi-party conversations using masked memory network and transformer. *Knowledge-Based Systems*, 240:108112.

Joosung Lee. 2022. The Emotion is Not One-hot Encoding: Learning with Grayscale Label for Emotion Recognition in Conversation. In *Proc. Interspeech 2022*, pages 141–145.

Himanshu Maheshwari and Vasudeva Varma. 2022. An ensemble approach to detect emotions at an essay level. In *Proceedings of the 12th Workshop on Computational Approaches to Subjectivity, Sentiment & Social Media Analysis*, pages 276–279, Dublin, Ireland. Association for Computational Linguistics.

Saif M. Mohammad and Peter D. Turney. 2013. Crowdsourcing a word-emotion association lexicon.

Bo Pang and Lillian Lee. 2008. Opinion mining and sentiment analysis. *Found. Trends Inf. Retr.*, 2:1–135.

Soujanya Poria, Navonil Majumder, Rada Mihalcea, and Eduard Hovy. 2019. Emotion recognition in conversation: Research challenges, datasets, and recent advances.

Duyu Tang, Furu Wei, Bing Qin, Nan Yang, Ting Liu, and Ming Zhou. 2016. Sentiment embeddings with applications to sentiment analysis. *IEEE Transactions on Knowledge and Data Engineering*, 28(2):496–509.

Kushal Tatariya, Heather Lent, Johannes Bjerva, and Miryam de Lhoneux. 2024. Sociolinguistically informed interpretability: A case study on hinglish emotion classification.

Mike Thelwall, Kevan Buckley, and Georgios Paltoglou. 2012. Sentiment strength detection for the social web. *J. Am. Soc. Inf. Sci. Technol.*, 63(1):163–173.

Deepanshu Vijay, Aditya Bohra, Vinay Singh, Syed Sarfaraz Akhtar, and Manish Shrivastava. 2018. Corpus creation and emotion prediction for Hindi-English code-mixed social media text. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Student Research Workshop*, pages 128–135, New Orleans, Louisiana, USA. Association for Computational Linguistics.

Anshul Wadhawan and Akshita 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, Online. Association for Computational Linguistics.

Yaoting Wang, Yuanchao Li, Paul Pu Liang, Louis-Philippe Morency, Peter Bell, and Catherine Lai. 2023. Cross-attention is not enough: Incongruity-aware dynamic hierarchical fusion for multimodal affect recognition.