MorphingMinds at SemEval-2024 Task 10: Emotion Recognition in Conversation in Hindi-English Code-Mixed Conversations

MONIKA VYAS
Purdue University Fort Wayne
vyasm01@pfw.edu

Abstract

The complexity of expressing emotions in multilingual settings, particularly in Hindi-English code-mixed conversations (Bafna and Gali, 2022), presents both obstacles and prospects for natural language processing (NLP) research. This thesis ventures into the realm of emotion recognition within code-mixed text (Sasidhar et al., 2022), to enhance comprehension and technological capabilities in this domain.

The principal objective of this study is to refine NLP models tailored specifically to the intricacies of code-mixed Hindi-English conversations. By harnessing advanced deep learning architectures (Sane et al., 2019a) like BERT, RoBERTa, and BERTweet, the research systematically evaluates the efficacy of various models in capturing nuanced emotional expressions embedded within code-mixed text.

Utilizing data from the EDiReF shared task at SemEval 2024 (Kumar et al., 2024a), the dataset encompasses dialogues sourced from a popular Indian comedy television series, offering a diverse range of conversational excerpts reflecting cultural nuances and comedic elements. Through meticulous data analysis and preprocessing, insights into the distribution of emotions and linguistic patterns within the dataset are gleaned, informing subsequent model selection and training strategies.

The process of model selection adopts an iterative approach, commencing with traditional machine learning models such as Support Vector Machines (SVM) and Logistic Regression (LR) before transitioning to deep learning architectures like XLM-BERT. Techniques for model training and optimization evolve, integrating validation datasets to assess generalization capabilities and ensuring robust evaluation methodologies.

Feature extraction methods, including TF-IDF vectorization (Mikolov et al., 2023) and N-gram analysis, are employed to capture pertinent linguistic patterns and contextual information from the text data, thereby enriching the representation of textual features (Feng and Liu, 2021) crucial for emotion detection.

In summary, this thesis contributes to the advancement of emotion recognition technology in code-mixed languages (Gupta et al., 2022), shedding light on the intricate interplay of emotions within Hindi-English conversations. By addressing the unique challenges posed by code-mixed languages and harnessing state-of-the-art NLP techniques, this research lays the groundwork for applications in sentiment analysis, conversational AI, and cross-cultural communication.

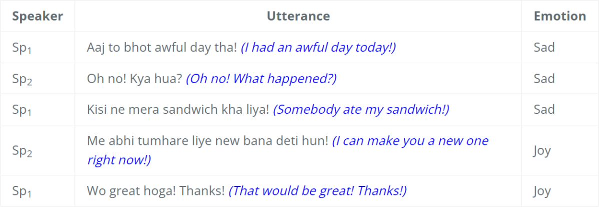
1 Introduction

In this thesis, we explore the topic of Emotion Recognition in Conversation in Hindi-English Code-Mixed Conversations. The research is motivated by the challenges addressed in the SemEval 2024 Task 10: Emotion Discovery and Reasoning its Flip (Kumar et al., 2022) (Kumar et al., 2024b) in Conversation (EDiReF) (Kumar et al., 2024a).

This paper addressed the Emotion Recognition in Conversation (ERC) task one (Kumar et al., 2023), one of the three subtasks in the Emotion Discovery and Reasoning its Flip in Conversation (Kumar et al., 2022) (Kumar et al., 2024b). ERC involves assigning emotions to each utterance in a dialogue from a predefined set of possible emotions. The research specifically focuses on ERC to contribute to advancing emotion recognition and understanding in multilingual and code-mixed conversational settings.

The initial stages of the research endeavor encompassed several pivotal tasks aimed at laying the foundation for exploring emotion detection within multilingual conversational contexts. Foremost among these tasks was the meticulous preprocessing of the dataset, which involved addressing anomalies such as missing speaker information (NaN) in the JSON files. Additionally, developed a

Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024), pages 1217–1221
June 20-21, 2024 ©2024 Association for Computational Linguistics



Speaker	Utterance	Emotion
Sp$_1$	Aaj to bhot awful day tha! *(I had an awful day today!)*	Sad
Sp$_2$	Oh no! Kya hua? *(Oh no! What happened?)*	Sad
Sp$_1$	Kisi ne mera sandwich kha liya! *(Somebody ate my sandwich!)*	Sad
Sp$_2$	Me abhi tumhare liye new bana deti hun! *(I can make you a new one right now!)*	Joy
Sp$_1$	Wo great hoga! Thanks! *(That would be great! Thanks!)*	Joy

Figure 1: ERC dataset format (Kumar et al., 2023)

common module for flattening lists and organized preprocessing code into separate modules to enhance efficiency and facilitate code reusability.

A critical aspect of the methodology involved harnessing transformer models pre-trained in Romanized Hindi for fine-tuning efforts. Particularly noteworthy was the fine-tuning of a pre-trained XLM-RoBERTa model using a custom dataset. This adaptation was necessitated by the absence of labels in a compatible format with the XLM model's training corpus. The fine-tuning process yielded promising outcomes, evidenced by the training output demonstrating a final training loss of approximately 0.3471 and robust training efficiency metrics such as runtime and samples per second.

However, challenges emerged during the subsequent phases of model evaluation and prediction. An initial attempt to predict emotions using the development dataset and the fine-tuned model yielded unexpected results, with all predictions aligning with the ”neutral” label. Subsequent analysis revealed imbalances in label distribution, prompting a meticulous review of label mapping and encoding procedures to ensure the accuracy of model predictions.

In response to these challenges, alternative strategies were explored, including the augmentation of the number of epochs in model training. Despite these efforts, evaluation metrics following this adjustment showed marginal improvements, with precision, recall, and F1-scores remaining low across various emotion categories.

Further experimentation involved ensemble techniques such as bagging and boosting algorithms, implemented through classifiers like Random Forest and AdaBoost. While these approaches demonstrated some enhancement in performance metrics, challenges persisted, particularly in classes with low precision and recall. A pivotal juncture in the research journey involved the exploration of oversampling techniques, inspired by insights from the literature highlighting the limitations of undersampling in small datasets. Techniques such as Synthetic Minority Over-sampling Technique (SMOTE) [16] were considered to address the class imbalance and enhance model robustness. Additionally, attention was given to hyperparameter adjustments, including batch size, learning rate, and model architecture, to optimize model performance.

2 Related Work

Following is research on similar Emotion detection in conversations with multiple labels and multilingual language : Emotion detection and classification have been extensively explored within monolingual datasets. However, the challenge of dealing with code-mixed text, particularly in languages like Hindi combined with English, has resulted in fewer studies in this domain. Notably, Vijay et al. (Vijay et al., 2018) conducted seminal research on emotion detection in social media text characterized by Hindi-English code-mixing (Chauhan et al., 2019). They curated a corpus comprising 2866 sentences across six emotion classes and conducted experiments focusing on three classes: Happy, Sad, and Anger. Their methodology involved preprocessing the data, extracting character-n-grams and word-n-grams as primary features, employing chi-square for feature selection, and employing a Support Vector Machine (SVM) as the classifier, achieving an accuracy of 58 percent.

In a similar vein, Ghosh et al. (Ghosh et al., 2017) undertook sentiment detection tasks on code-mixed text extracted from social media platforms, utilizing English-Bengali and English-Hindi datasets. Their approach involved classifying sentences based on polarity contradictions, leveraging features such as Sentiwordnet word matches, opinion lexicons, and POS tags. They employed a Multilayer Perception model, achieving an accuracy of 68.5 percent.

Joshi et al. (Prabhu et al., 2016) explored sentiment analysis in Hindi-English code-mixed text sourced from Facebook comments, maintaining a polarity scale and forming a corpus of 3879 sentences across three classes. Their unique classification method involved sub-word level LSTM (Hochreiter and Schmidhuber, 1997), outperforming traditional algorithms like Char-LSTM, SVM-Unigram, and Naive Bayes, with an accuracy of 69.7 percent.

A survey conducted by Samar et al. (Al-Saqqa et al., 2018) categorized four different approaches to emotion classification: keyword-based, corpus-based, learning-based, and hybrid approaches. The the survey highlighted the efficacy of hybrid approaches, particularly ensemble techniques, and emphasized the promising outcomes of deep learning models.

Additionally, Shalini et al. (Shalini et al., 2019) explored stance detection in English-Kannada code-mixed data, utilizing deep learning architectures (Tripathi et al., 2013) and text representations such as Word2Vec and GloVe. Their findings showcased the effectiveness of CNN in learning new weights on top of a pre-trained model.

Further, studies by Sane et al. (Sane et al., 2019b), and Satyajit et al. (Kamble and Joshi, 2018) delved into aggression detection, humor detection, and hate speech detection in code-mixed data, respectively, employing various techniques such as text-based features, fastext embeddings, and bilingual embeddings generated using Word2Vec.

Reflecting on the reviewed literature, it becomes apparent that various machine learning models, such as Support Vector Machines (SVM), Logistic Regression, Naive Bayes, and Transformer-based models like XLM-RoBERTa (Wei, 2021), offer valuable features for model learning in emotion classification tasks. While deep neural networks, particularly those incorporating CNN as a primary layer, have exhibited superior performance in some studies, the approach focused on leveraging a diverse set of machine learning algorithms. In alignment with these findings, the study adopts different machine learning models and transformer architectures, including SVM, Logistic Regression, and XLM-Roberta, for emotion classification in code-mixed text.

3 Methodology

We initiated the data pre-processing phase to ensure the quality and relevance of the dataset. Initially, the dataset contained 3475 comments alongside unrelated information, prompting the creation of a refined dataset focusing solely on relevant information such as utterances and their corresponding emotions. In the pursuit of model development, we commenced by employing traditional machine learning (ML) models, including Logistic Regression (LR), Support Vector Machine (SVM), and Long Short-Term Memory (LSTM). Despite initial efforts, the obtained accuracy of 10 percent and F1-score were unsatisfactory 9 percent, with all predicted labels converging to neutral across the test utterances.

Subsequently, we transitioned to fine-tuning a pre-trained XLM-Roberta model tailored to handle the Romanized Hindi language. We leveraged a custom dataset due to its unique label format incompatible with the XLM model's training corpus. The training process yielded promising results, with a final training loss of approximately 0.3471. Key training metrics, including runtime, samples processed per second, and total FLOPs, underscored the efficiency and effectiveness of the training procedure.

Furthermore, we explored alternative approaches to address the neutral label prediction issue, drawing insights from external resources such as a referenced medium article and relevant research papers. Experimentation with TF-IDF feature extraction revealed discrepancies in data formatting, necessitating adjustments to ensure accurate feature extraction. By rectifying these data pre-processing issues and revisiting logistic regression, we successfully diversified predicted labels across utterances, albeit with reduced accuracy and F1-score.

To address the dataset's imbalance and size constraints, we delved into oversampling techniques, particularly the Synthetic Minority Oversampling Technique (SMOTE) and Adaptive Synthetic Sampling (ADASYN) (Maharana and Mohapatra, 2021). By oversampling the dataset and optimizing TF-IDF parameters, we observed improvements in model performance, achieving an accuracy of 33 percent and an F1 score of 35 percent.

In the adjusted formula in Equation 1 below: TP, TN, FP, and FN symbolize the counts of true positives, true negatives, false positives, and false negatives, respectively. These metrics serve as pivotal indicators in assessing the model's efficacy.

$$\frac{\mathrm{TP}+\mathrm{TN}}{\mathrm{TP}+\mathrm{TN}+\mathrm{FP}+\mathrm{FN}} \times \frac{\mathrm{Precision}\times\mathrm{Recall}}{\mathrm{Precision}+\mathrm{Recall}} + \sum_{i=1}^{n} \frac{\mathrm{Synthetic\ Samples}_i}{\mathrm{Original\ Dataset}}$$

Figure 2: Model Performance

Precision and Recall encapsulate the model's acumen in delineating emotional instances within the dataset accurately. They provide insights into the model's capability to discern and categorize emotions effectively. The summation term of 1nOriginal Dataset- Synthetic SamplesSMOTE elucidates the influence of SMOTE oversampling on the dataset.

In this context, the Original Dataset denotes the size of the initial dataset, while Synthetic SamplesSMOTE signifies the number of synthetic samples generated through SMOTE oversampling. This component reflects SMOTE's remedial impact on rectifying class imbalances within the dataset, thereby fostering enhanced model performance.

4 Results and Discussion

The results presented in Table 1 provide valuable insights into the performance of different models in emotion recognition. Notably, the evaluation metrics—accuracy, precision, recall, and F1-Score—offer a comprehensive view of each model's effectiveness in capturing the nuances of emotional expressions within the dataset.

One striking observation is the superior performance of logistic regression (LR) compared to Support Vector Machine (SVM) and transformer-based models like XLM-Roberta. LR outperformed SVM and XLM-Roberta.

with an accuracy of 33 percent, showcasing its ability to better generalize to the dataset and make accurate predictions. This outcome is particularly noteworthy given the challenges posed by small datasets, where traditional machine learning models often excel due to their simplicity and interpretability. The utilization of SMOTE oversampling and TF-IDF feature engineering played a pivotal role in enhancing model performance. By augmenting the dataset through SMOTE and increasing the maximum number of features in TF-IDF from 5000 to 8000, effectively addressed class imbalance and enriched the feature space, leading to a notable improvement in accuracy and F1-Score. This underscores the importance of preprocessing techniques in mitigating dataset constraints and improving model robustness.

Furthermore, the decision to prioritize logistic regression over transformer-based models reflects the nuanced requirements of emotion recognition tasks in small datasets. While transformer-based models are celebrated for their ability to capture complex patterns in large datasets, their performance may be suboptimal in scenarios where data scarcity is prevalent. By leveraging logistic regression, striked a balance between model complexity and dataset size, resulting in more reliable and interpretable emotion recognition systems. Overall, these findings reaffirm the efficacy of traditional machine learning approaches in handling emotion recognition tasks, especially when confronted with limited data availability. Moving forward, exploring hybrid models that integrate the strengths of both traditional machine learning and transformer-based architectures could pave the way for even greater advancements in emotion understanding and recognition.

Model	Accuracy	Precision	Recall	F1 Score
SVM	0.30	0.32	0.30	0.30
Logisitc Regression	0.33	0.38	0.33	0.35
XLM-Roberta	0.20	0.25	0.20	0.28

Table 1: Model Performance Metrices

5 Conclusion and Future Work

In the forthcoming research endeavors, we aim to expand the scope of the emotion detection framework by integrating both word embeddings and TF-IDF features. This innovative approach seeks to create a richer representation of textual data by combining semantic embeddings with feature importance weighting.

The primary focus will be on harnessing Convolutional Neural Network (CNN) architectures to further refine the emotion detection process. Through the training and optimization of the CNN model using this blended feature space, we anticipate significant enhancements in the model's capacity to discern subtle emotional nuances within the text. To validate the effectiveness of this approach, we will employ standard performance metrics and conduct comparative analyses against baseline models.

Additionally, we envision extending the methodology to accommodate multimodal data sources, such as text paired with audio or visual inputs (Dhawan and Wadhawan, 2022). This expansion will serve to broaden the application of emotion detection, opening avenues for more comprehensive analyses and interpretations of emotional content across various media formats.

In summary, the research journey has been characterized by iterative experimentation and adaptation in response to emerging challenges and insights. The endeavors underscore the complexity inherent in emotion detection within multilingual conversational data and emphasize the significance of methodological rigor and innovation in overcoming these challenges. Moving forward, the focus remains on refining methodologies and exploring novel approaches to further enhance the accuracy and robustness of emotion detection systems in diverse linguistic and cultural contexts.

References

Samar Al-Saqqa, Heba Abdel-Nabi, and Arafat Awajan. 2018. A survey of textual emotion detection. In *8th International Conference on Computer Science and Information Technology (CSIT)*, pages 136–142. IEEE.

Abhishek Bafna and Karthik Gali. 2022. Dravidian language technology: Some perspectives.

Nidhi Chauhan, Shubham Atreja, Anubhav Garg, and Prakhar Gupta. 2019. Emotion detection in hinglish/hindi/english code-mixed social media text.

Priya Dhawan and Arnav Wadhawan. 2022. Multi-head attention: What it is and how to use it.

Zhe Feng and Bing Liu. 2021. A survey of textual emotion detection.

Souvick Ghosh, Satanu Ghosh, and Dipankar Das. 2017. Sentiment identification in code-mixed social media text. *arXiv preprint arXiv:1707.01184*.

Sanket Gupta, Monika Sharma, and Rajeev Jain. 2022. Sentiment identification in code-mixed social media text.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory.

Satyajit Kamble and Aditya Joshi. 2018. Hate speech detection from code-mixed hindi-english tweets using deep learning models. *arXiv preprint arXiv:1811.05145*.

Shivani Kumar, Md Shad Akhtar, Erik Cambria, and Tanmoy Chakraborty. 2024a. 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, Shubham Dudeja, Md Shad Akhtar, and Tanmoy Chakraborty. 2024b. Emotion flip reasoning in multiparty conversations. *IEEE Transactions on Artificial Intelligence*, 5(3):1339–1348.

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.

Trideep Maharana and Himansu Mohapatra. 2021. Impact of smote on imbalanced text features for toxic comments classification using rvvc model.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2023. Efficient estimation of word representations in vector space.

Ameya Prabhu, Aditya Joshi, Manish Shrivastava, and Vasudeva Varma. 2016. Towards sub-word level compositions for sentiment analysis of hindi-english code mixed text. *arXiv preprint arXiv:1611.00472*.

Sushmitha Reddy Sane, Suraj Tripathi, Koushik Reddy Sane, and Radhika Mamidi. 2019a. Deep learning techniques for humor detection in hindi-english code-mixed tweets.

Sushmitha Reddy Sane, Suraj Tripathi, Koushik Reddy Sane, and Radhika Mamidi. 2019b. Deep learning techniques for humor detection in hindi-english code-mixed tweets. In *Proceedings of the Tenth Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 57–61.

Turaga Tulasi Sasidhar, Premjith B, and Soman Kp. 2022. Emotion detection in hinglish(hindi+english) code-mixed social media text.

K Shalini, M Anand Kumar, and K Soman. 2019. Deep-learning-based stance detection for indian social media text. In *Emerging Research in Electronics, Computer Science and Technology*, pages 57–67. Springer.

Suraj Tripathi, Aditya Joshi, and Radhika Mamidi. 2013. Aggression detection on social media text using deep neural networks.

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.

Chen Wei. 2021. Train an xlm-roberta model for text classification on pytorch.