IASBS at SemEval-2024 Task 10: Delving into Emotion Discovery and Reasoning in Code-Mixed Conversations

Mehrzad Tareh, Aydin Mohandesi, Ebrahim Ansari
Institute for Advanced Studies in Basic Sciences (IASBS)
{m.tareh, mohandesi, ansari}@iasbs.ac.ir

Abstract

In this study, we introduce the SemEval 2024 Task 10, entitled "Emotion Discovery and Reasoning its Flip in Conversation (EDiReF)". Our research presents a comprehensive framework to analyze emotional dynamics within Hindi-English mixed-language and English conversations. We extend beyond traditional emotion identification to uncover the triggers behind shifts in emotional states using advanced Natural Language Processing (NLP) techniques. Employing a systematic methodology that encompasses data preprocessing, feature engineering, and the deployment of language models such as GPT-4 and DistilBERT, we unravel the complex interplay of emotions in communication. Our approach yields significant insights, enhancing applications from social media analytics to mental health, thus marking a notable advancement in the integration of emotional intelligence into AI. Noteworthy is our system's achievement of third place on the leaderboard, demonstrating robust performance with a weighted F1-Score of 0.70. This study not only contributes to the field of emotional AI but also paves the way for future research on the nuanced understanding of emotion in mixed-language communications.

1 Introduction

In the age of global digital communication, the English language, with its pervasive influence, has led to a notable increase in bilingual or code-mixed conversations, especially on various social media and messaging platforms. Among these, Hindi-English code-mixing, or "Hinglish", has emerged as a prominent linguistic phenomenon in the Indian subcontinent, embodying an increasingly globalized society's cultural and linguistic interplay. Despite the widespread occurrence of code-mixed communication, a gap remains in the research landscape, particularly in understanding the emotional dynamics of such interactions (Ramalingam et al., 2023; Attri et al., 2020). Studies in Natural Language Processing (NLP) are progressively delving into the significance of emotions in human dialogues, offering promising applications across various domains including human-computer interaction (Kulkarni et al., 2023), social media scrutiny (Sharma et al., 2020), and healthcare (Takale, 2024). The Emotional Dynamics in Realistic Environments and Friends (EDiReF) initiative plays a significant role in this area by examining emotional expression and shifts in both bilingual (Hindi-English) and monolingual (English) conversations.

Emotion recognition, a facet of affective computing, endeavors to understand human emotions utilizing diverse technological methods. Its potential implications reach far beyond traditional industries, spanning fields as diverse as transportation, finance, and entertainment, promising to revolutionize the way we interact with technology in our daily lives (Guo et al., 2024). Yet, conventional methodologies predominantly focus on analyzing monolingual, single-sentence data, while the EDiReF initiative broadens this horizon to encompass mixed-language dialogues and shifts in emotion within conversations. This acknowledgment of the intricate nature of emotional expression within diverse linguistic and cultural frameworks sets EDiReF apart.

The EDiReF (Kumar et al., 2024a) initiative consolidates three distinct subtasks: Emotion Recognition in Conversation (ERC) in Hindi-English code-mixed conversations (Kumar et al., 2023), Emotion Flip Reasoning (EFR) in Hindi-English code-mixed conversations, and EFR in English conversations (Kumar et al., 2022, 2024b). Each subtask presents its own set of challenges in comprehending and interpreting emotions within conversational contexts. In the ERC subtask, we were tasked with assigning emotion labels to each utterance within a dialogue, drawing from a predefined spectrum of emotions. This demands algorithms capable of discerning and distinguishing subtle emotional cues embedded within mixed-language exchanges, thereby facilitating a deeper comprehension of emotional dynamics in bilingual interactions.

The EDiReF shared task serves as a crucial avenue for researchers to delve into innovative methodologies, exchange insights, and benchmark their models using real-world conversational data. This initiative contributes significantly to the overarching objective of advancing affective computing and deepening our comprehension of human emotions within natural language interactions.

2 Related Work

The exploration of emotion in human dialogue, especially within the realm of code-mixed conversations, represents a burgeoning field of study that intersects with computational linguistics, affective computing, and cross-cultural communication. This section reviews seminal works and recent advancements that set the stage for our investigation into EDiReF.

Emotion recognition, an integral part of the expanding domain of affective computing, endeavors to decipher and interpret human emotions through technological means. This interdisciplinary field amalgamates aspects of computer science, psychology, and neuroscience to forge innovative devices capable of recognizing, understanding, and reacting to human emotional states (Montag et al., 2020). Such advancements hold the potential to significantly transform human-computer interactions, promising to enhance user experiences across various sectors such as retail, finance, and entertainment, thereby enabling personalized and intuitive interactions (Matin and Valles, 2020).

Historically, the focus has predominantly been on monolingual, single-sentence analyses; however, the EDiReF task expands this horizon by exploring mixed-language dialogues and the dynamics of emotion flips within conversations. This forward-looking approach acknowledges the intricacies of communication, emphasizing the significance of context, cultural differences, and linguistic diversity in the accurate interpretation of emotions. By incorporating mixed-language data, the task addresses the growing occurrence of bilingual conversations in global communications, aiming to develop more inclusive and precise emotion detection algorithms that reflect the true complexity of human interactions (Muhammadiyeva, 2022).

The availability and quality of annotated datasets for training and evaluation emerge as significant hurdles, especially for less represented languages. The quest for consistency in annotations across languages and emotions further adds to the complexity (Garg, 2020). Furthermore, the requirement for effective cross-lingual representation learning highlights the need for models to accurately capture language-specific features and emotions, necessitating sophisticated approaches in transfer learning (Ranaldi and Pucci, 2023). Additionally, identifying trigger utterances for emotion flips introduces another layer of complexity, requiring a nuanced understanding of dialogue dynamics and contextual cues (Kumar et al., 2024b). The scalability and generalization of models across different conversational contexts, languages, and domains remain formidable challenges.

In summary, while existing research provides valuable insights into emotion recognition and code-mixing, there remains a notable paucity of studies specifically addressing the dynamics of emotion in code-mixed conversations. Our work seeks to fill this gap, leveraging state-of-the-art NLP techniques to analyze emotional content and reasoning in Hindi-English mixed-language dialogues. By doing so, we contribute to the broader discourse on advancing affective computing in multilingual and multicultural contexts.

3 Task Description

This task comprises three subtasks, each addressing distinct aspects of emotional understanding and analysis in dialogues.

3.1 Emotion Recognition in Conversations (ERC)

Emotion Recognition in Conversations stands at the forefront of computational linguistics and artificial intelligence, employing advanced algorithms and methodologies to decipher emotional nuances within textual exchanges. By meticulously collecting and preprocessing data, ERC systems utilize machine learning models to classify emotional states expressed in conversations, ranging from joy to sadness and anger to fear (Dessai and Virani, 2023). As shown in Table 1, the distribution of emotions in the used dataset highlights the predominance of neutral emotions, followed by joy and anger. Such advancements hold transformative potential across diverse sectors, from revolutionizing customer service interactions to enhancing mental health support through an early intervention based on text analysis. The applications of ERC extend beyond sentiment analysis, playing pivotal roles in education, market research, and human-computer interaction (Loveland, 2011). Moreover, ERC serves as a potent tool for market research, enabling companies to gauge public sentiment towards products or brands by analyzing social media conversations and consumer reviews (Razouk et al., 2023).

Emotion	Proportion(%)
Neutral	45.4
Joy	19.0
Anger	9.4
Sadness	7.3
Fear	6.3
Contempt	6.1
Surprise	4.9
Disgust	1.4

Table 1: Distribution of emotions in dataset for ERC

Despite the immense potential of ERC, ethical considerations regarding privacy, biases in emotion detection algorithms, and responsible data handling are paramount. However, as ERC continues to evolve, particularly in navigating the complexities of Hindi-English code-mixed conversations, it offers promise in bridging cultural divides and enabling more nuanced sentiment analysis in multicultural settings. Nevertheless, addressing ethical concerns through responsible research and implementation is crucial to fostering a more inclusive and empathetic digital landscape (Bagora et al., 2022; Sitaram et al., 2015).

3.2 Emotion Flip Reasoning (EFR)

Emotion Flip Reasoning is a fascinating phenomenon observed in conversations, where emotional states have the potential to alter the trajectory of logical reasoning. This interplay between emotions and cognition highlights the dynamic nature of human interaction, showcasing how shifts in emotional states can influence our thought processes and ultimately impact the course of a conversation (Kumar et al., 2024b). EFR underscores the intricate relationship between emotions and reasoning, shedding light on how our affective states shape the way we interpret information, make decisions, and engage in dialogue.

In the context of Hindi-English code-mixed conversations, the complexity of EFR is heightened due to the fusion of two languages and their respective cultural nuances. This linguistic code switching adds layers of meaning and emotion to the conversation, further intertwining the emotional and rational components of communication (Gautam et al., 2021).

EFR manifests in Hindi-English code-mixed conversations through the interplay of emotional cues and linguistic expressions from both languages (Bagora et al., 2022). For example, we may use Hindi to convey deeply rooted emotions or cultural concepts, triggering a shift in the logical progression of the dialogue. Similarly, an emotional response articulated in English may prompt a reassessment of previously held beliefs or arguments. Recognizing and understanding EFR in code-mixed conversations is crucial for effective communication in multicultural settings, requiring sensitivity to both linguistic and emotional cues. As shown in Table 2, the distribution of triggers for EFR highlights the prevalence of such shifts in Subtasks 2 and 3. Embracing the complexities of EFR can lead to more meaningful and productive interactions, fostering mutual understanding and collaboration across cultural and linguistic boundaries (Kumar et al., 2024b).

Dataset	0.0(%)	1.0(%)
Subtask 2	93.5	6.5
Subtask 3	84.7	15.3

Table 2: Distribution of triggers in dataset for EFR

3.3 Code-mixing in Conversations

Code-mixing in conversations occurs when speakers blend elements from two or more languages within the same discourse. This phenomenon is prevalent in multilingual communities where individuals are fluent in multiple languages and switch between them based on social context, familiarity, or communicative needs. In such conversations, speakers may switch between languages mid-sentence or incorporate phrases, expressions, or even entire sentences from one language into another. Code-mixing adds richness and depth to communication, allowing speakers to draw from a wider linguistic repertoire to express their thoughts and convey nuances that may not be readily available in a single language. Hindi-English code-mixing, commonly known as Hinglish, is a prominent example of code-mixing in conversations, especially in regions with significant bilingual populations like India (Kodali et al., 2022). In Hinglish conversations, speakers seamlessly integrate Hindi and English elements, creating a unique linguistic fusion that reflects the cultural and linguistic diversity of the Indian subcontinent. Hinglish code-mixing serves as a linguistic bridge, allowing speakers to navigate between their cultural identities and accommodate the diverse linguistic backgrounds of their interlocutors (Jawahar et al., 2021).

The use of code-mixing, whether in general conversations or specifically in Hinglish, serves several communicative functions. Firstly, it facilitates smoother communication by allowing speakers to express themselves using the most appropriate linguistic resources available to them. Additionally, code-mixing can convey social and cultural affiliations, signaling aspects of the speaker's identity such as ethnicity, education level, or social status. Moreover, code-mixing can serve pragmatic functions, such as clarifying meanings, emphasizing certain points, or creating humorous effects (Vogh, 2022).

3.4 Hindi-English Code-mixed Conversations

Code-mixed conversations, blending Hindi and English, are a common phenomenon in bilingual societies like India. This linguistic fusion reflects the cultural and social dynamics of the populace. In everyday interactions, individuals effortlessly switch between the two languages, often using Hindi for informal contexts and English for formal or technical discussions. These fluid exchanges showcase the flexibility and richness of language usage in diverse settings (Yadav et al., 2020; Mukherjee, 2019).

Code-mixing isn't just about linguistic versatility; it's also deeply ingrained in identity expression. By integrating Hindi and English, speakers navigate their cultural affiliations and social environments. This blending of languages is not only a means of communication but also a reflection of one's hybrid cultural identity (Attri et al., 2020). Furthermore, code-mixed conversations play a crucial role in digital communication, especially on social media platforms and messaging apps. In the virtual realm, users often employ a mix of Hindi and English to cater to a wider audience while maintaining a sense of familiarity and belonging. As shown in Table 3, the proportions of Hindi and English usage vary significantly across different tasks, illustrating the dynamic nature of code-mixed communication. This phenomenon has led to the emergence of unique online subcultures and linguistic trends. In essence, code-mixed conversations not only bridge linguistic divides but also serve as a vibrant expression of cultural fusion in a globalized world (Dabrowska, 2019).

Task	Hindi(%)	English(%)
Subtask 1	59.8	40.2
Subtask 2	39.8	60.2
Subtask 3	17.7	82.3

Table 3: Hindi vs English proportions in dataset

4 Dataset Description

The organizers have supplied and divided the datasets for participants into train, development, and test sets. The Table 4 provided contains details regarding the number of instances used for training, development, and testing.

Dataset	Train	Dev	Test
Subtask 1	8,506	1,354	1,580
Subtask 2	98,777	7,462	7,690
Subtask 3	35,000	3,522	8,642

Table 4: Summary of dataset split

4.1 Introduction to MaSac Dataset

The MaSaC dataset is a carefully curated collection designed to investigate code-mixed dialogue interactions in the Indian context, drawing from the popular television series "Sarabhai v/s Sarabhai." This dataset captures the authentic dynamics of conversations in a multi-party, multimodal setting, predominantly featuring a blend of Hindi and English languages. With over 11,000 utterances dedicated to task 1 and 114,000 to task 2, it provides researchers with a rich corpus to explore various aspects of language usage and communication dynamics. Each utterance in a dialogue has been labeled by any of these eight emotions: Anger, Disgust, Sadness, Joy, Neutral, Surprise, Fear, and Contempt. Its multimodal nature, incorporating textual, auditory, and visual elements from the television show, offers a comprehensive view of dialogue interactions, while its labeling of emotions for each utterance enriches understanding of the emotional nuances within the conversations (Kumar et al., 2023).

Furthermore, the MaSaC dataset holds significant promise for diverse research endeavors. In the realm of natural language understanding, it facilitates the development and evaluation of models capable of comprehending and generating code-mixed utterances, thereby enhancing language processing capabilities in multilingual environments. Additionally, computational linguistics, enables investigations into code-switching phenomena and sociolinguistic variations, shedding light on language usage patterns and communicative strategies. Beyond academia, the dataset's exploration of socio-cultural aspects embedded within language interactions offers valuable insights for sociolinguists and cultural researchers, fostering a deeper understanding of identity expression, social dynamics, and cultural nuances depicted in televised narratives (Bedi et al., 2023).

4.2 Introduction to MELD Dataset

The Multimodal Emotion Lines Dataset (MELD) stands out as a pivotal resource for researchers delving into the intricate realm of emotion recognition, particularly within the context of multiparty conversations. Comprising over 47,000 utterances extracted from the beloved television series Friends, MELD is an extension and enhancement of EmotionLines (Hsu et al., 2018), and presents an extensive collection of genuine exchanges, showcasing diverse interactions among numerous speakers engaged in dynamic dialogues. Unlike its predecessor, MELD adopts a multimodal approach by incorporating textual transcripts.

However, it is crucial to note that the dataset provided by the organizers did not include audio-visual cues, focusing instead on the textual aspect to understand emotional communication. Each conversation in the dataset is carefully annotated, presenting detailed labels for emotions expressed, making it a valuable resource for supervised learning techniques. The dataset labels each utterance with one of seven emotions: Anger, Disgust, Sadness, Joy, Neutral, Surprise, and Fear, aiming to capture the full spectrum of emotional dynamics in conversation (Kumar et al., 2024b).

5 Experimental Setup

This section outlines the various aspects of the experimental setup, including dataset preparation, evaluation metrics, baseline systems, and training methodologies.

5.1 JSON Parsing

In our exploration of this task, we encounter datasets structured in JSON format, encapsulating dialogues among multiple speakers annotated with emotions and triggers. To effectively manage this data, we employ a detailed parsing process. Initially, we load the JSON data into memory, followed by iterative processing of each dialogue entry. We extract pertinent details such as episode ID, speaker ID, emotion label, trigger label, and utterance text, organizing them into a structured DataFrame format. Additionally, to ensure each entry's unique identification, we generate non-negative integer IDs for each dialogue instance, facilitating seamless referencing during subsequent data analysis and model development stages. By effectively extracting essential information and generating unique IDs for dataset entries, we can navigate through the data with ease, enabling profound understanding and fostering advancements in emotion recognition and reasoning within code-mixed and multi-party conversation dialogues.

5.2 Dataset Preprocessing

We meticulously executed preprocessing steps on the dataset to ensure uniformity of data and streamline subsequent analysis and model development. Initially organized by task organizers (table 4), the dataset underwent thorough text normalization techniques including lowercase conversion and removal of redundant characters and excessive punctuation to enhance readability and consistency. Subsequent tokenization segmented the text for deeper analysis, followed by language-specific procedures such as stopword removal, lemmatization, and stemming to refine textual content. Language identification techniques were also employed to differentiate between Hindi and English segments, allowing for targeted preprocessing steps. Our collaborative efforts standardized and formatted the dataset in alignment with the prescribed framework of the shared task, laying a robust foundation for effective participation and further analysis.

5.2.1 Dataset Cleaning and Standardization

- Addressed data integrity by handling less than 5 samples with invalid values, assigning the most common value within the dataset to maintain consistency.
- Standardized speakers' names to ensure uniformity by resolving discrepancies like varying capitalization.
- Enhanced dataset structure by assigning non-negative IDs to episodes and speakers, enabling more efficient data processing.
- In the EFR task, focused solely on triggers with a value of 0.0, resulting in an emotional inversion where the trigger value shifted to 1.0.

5.3 Dataset Translation

Initially, while relying on Google Translate, our comprehensive manual verification process unveiled discrepancies and inaccuracies, compelling us to explore alternatives of greater reliability and precision. In this context, GPT-4 emerged as a pivotal tool, distinguished for its contextual comprehension and translation proficiency (et. al, 2024). Our translation methodology was meticulously crafted to prioritize fidelity to the original dialogues, thereby ensuring the preservation of nuanced semantic and syntactic elements across linguistic boundaries.

By harnessing the capabilities of GPT-4 in conjunction with supplementary support from Google Translate, we embarked on a systematic translation endeavor aimed at capturing the inherent complexity and subtlety of the conversations. The outcomes were noteworthy, as GPT-4 consistently exceeded expectations, exhibiting an exceptional aptitude for encapsulating the intricate nuances of emotional expression and linguistic subtleties. Armed with these carefully refined translations, we deftly integrated them into the final English datasets, confident in their accuracy and faithfulness to the original discourse. This meticulous approach not only ensured linguistic coherence but also paid homage to the rich cultural nuances embedded within the conversations (Nakayama et al., 2019).

5.3.1 Dataset Normalization

Hindi-English code-mixed conversations posed a significant challenge, especially with restrictions on available tools due to sanctions. Our initial attempts with a robot browser yielded unsatisfactory results, prompting us to leverage the GPT-4 API in conjunction with the spaCy [1]. This combination significantly outperformed traditional translation services, like Google Translate, in accuracy. Post-translation, we employed a pre-trained model (Kunchukuttan et al., 2020) to normalize utterances to standard Hindi/Romani, followed by tokenization and analysis with Morph to adapt the analyzed form.

5.3.2 Feature Engineering

For sentiment analysis, we utilized the INT8 DistilBERT model (Xin He, 2022) through the Cloudflare API [2], streamlining the process to calculate positive and negative scores for each utterance. A novel feature introduced was the calculation of polarity differences between consecutive utterances within an episode, aiding in the detection of emotion flips.

5.4 Dataset Polarity

In the Dataset Polarity section, we present the essential analysis conducted on the provided dataset. Here, we outline the methodology we employed to evaluate the polarity of utterances within the dataset. Utilizing cutting-edge natural language processing technology, specifically DistilBERT from Intel, we undertook the task of calculating polarity scores for each utterance.

Polarity scores serve as a quantitative measure of the sentiment conveyed within individual utterances. Our approach involved leveraging DistilBERT's pre-trained language understanding capabilities to discern the underlying sentiment expressed in the text. By employing this state-of-the-art model, we aimed to capture subtle emotional undertones present in the conversations, particularly in the context of Hinglish dialogues. Furthermore, to enhance our understanding of the dataset, we computed the difference in polarity scores between utterances. This differential analysis provides insights into the shifts in sentiment within conversations, a crucial aspect for tasks such as ERC and EFR. By discerning fluctuations in sentiment, we gain valuable information for identifying trigger utterances for emotion flips in multi-party dialogues.

The integration of DistilBERT from Intel in our polarity analysis underscores our commitment to

[1] https://spacy.io/

[2] https://developers.cloudflare.com/workers-ai/models/

leveraging state-of-the-art techniques for robust sentiment analysis. Through this punctilious approach, we aim to provide a comprehensive understanding of the emotional dynamics inherent in the dataset, thereby facilitating advancements in emotion recognition and reasoning tasks within code-mixed conversations.

6 Methodology

To explore the intricacies of ERC and EFR within Hindi-English code-mixed conversations, as well as EFR in English conversations, our methodology integrates a blend of traditional machine learning models with the cutting-edge capabilities of transformer-based architectures.

At the outset, we harness the advanced linguistic comprehension of GPT-4, a state-of-the-art language model, to ascertain the emotions embedded in each utterance of the dialogue. The prowess of GPT-4 lies in its nuanced grasp of context and the subtleties of natural language, rendering it highly effective for the preliminary prediction of emotions within conversations. Subsequently, we employ a spectrum of classical machine learning techniques, including Random Forest, Support Vector Machines (SVM), Logistic Regression, and Naive Bayes classifiers. These algorithms are foundational to the field of natural language processing and serve as a benchmark for evaluating the advanced methodologies employed later. This hybrid modeling approach aims to capitalize on the depth and context awareness provided by transformer-based models, like GPT-4, while also valuing the interpretability and established nature of classical machine learning techniques. By leveraging this diverse array of models, our objective is to harness the strengths of both modern and traditional approaches to optimize performance across the ERC and EFR tasks in conversations conducted in both Hinglish and English.

For the task of Emotion Recognition in Conversation, we adopt the weighted F1-score as our primary evaluation metric (see Table 7 and 8). This metric is chosen for its ability to provide a balanced measure of the model's precision and recall, while also accounting for class imbalances that are common in real-world datasets. This nuanced evaluation allows us to assess the model's ability to accurately recognize emotions across a diverse set of conversations. In the case of Emotion Flip Reasoning, our focus shifts toward precision as the primary evaluation metric (see Table 6). Precision is particularly relevant for EFR tasks as it measures the model's accuracy in identifying the specific instances where an emotional flip occurs within the conversation. This metric enables us to refine the model's performance in pinpointing these critical junctures, thereby ensuring high reliability in the model's reasoning capabilities.

By employing this diverse range of models, we aimed to leverage both the sophistication of transformer-based architectures and the interpretability of classical machine learning algorithms. This hybrid approach allowed us to explore different facets of the data and optimize performance across the ERC and EFR tasks in both Hinglish and English conversations.

7 Results and Discussions

In our study, we present a comprehensive analysis of our findings, which is predicated on our noteworthy accomplishment of securing the third position on the Codalab leaderboard among forty participants.

Team	Subtask 1	Subtask 2	Subtask 3
MasonTigers	0.78	0.79	0.79
Knowdee	0.73	0.66	0.61
IASBS	**0.70**	**0.12**	**0.25**
Alden_Jenish	0.66	0.07	0.04

Table 5: Top four participants' scores on CodaLab

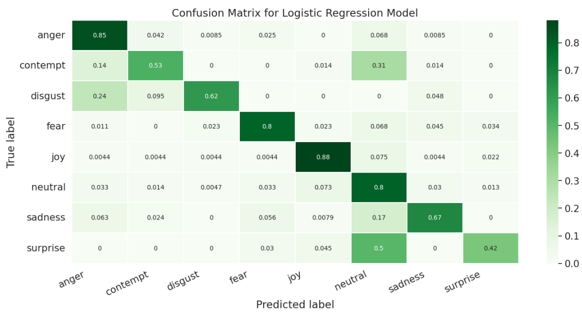

Figure 1: Confusion matrix of task 1

This achievement underscores the efficacy of our comprehensive framework, which integrates advanced language models like GPT-4 and DistilBERT, alongside sophisticated data preprocessing and feature engineering techniques, to explore the nuanced interplay of emotions in conversations.

In our investigation, the evaluation of various classification models across the three subtasks of emotion recognition and reasoning exhibited dis-

Model	Precision	Recall	Accuracy	F1-Score
Logistic Regression	**0.77**	**0.77**	**0.77**	**0.77**
SVM	0.77	0.77	0.77	0.77
GaussianNB	0.78	0.76	0.76	0.76
MLP	0.76	0.76	0.76	0.76
LDA	0.77	0.76	0.76	0.76
KNN	0.75	0.75	0.75	0.75
Random Forest	0.76	0.76	0.76	0.74
AdaBoost	0.75	0.75	0.75	0.74
QDA	0.62	0.71	0.71	0.66

Table 6: Comparative Evaluation Results of Various Classification Models for ERC (Hinglish)

Model	Precision	Recall	Accuracy	F1-Score
BernouliNB	**0.62**	**0.64**	**0.64**	**0.63**
SVM	0.63	0.71	0.71	0.62
QDA	0.64	0.71	0.71	0.62
KNN	0.59	0.68	0.68	0.61
Logistic Regression	0.74	0.72	0.72	0.6
Random Forest	0.64	0.71	0.71	0.6
LDA	0.59	0.7	0.7	0.6
MLP	0.51	0.71	0.71	0.59
AdaBoost	0.51	0.71	0.71	0.59

Table 7: Comparative Evaluation Results of Various Classification Models for EFR (Hinglish)

Model	Precision	Recall	Accuracy	F1-Score
SVM	**0.79**	**0.89**	**0.89**	**0.83**
LDA	0.79	0.89	0.89	0.83
AdaBoost	0.79	0.89	0.89	0.83
Logistic Regression	0.8	0.77	0.77	0.78
KNN	0.8	0.73	0.73	0.76
Random Forest	0.81	0.72	0.72	0.76
BernoulliNB	0.82	0.7	0.7	0.75
QDA	0.81	0.71	0.71	0.75
MLP	0.81	0.69	0.69	0.74

Table 8: Comparative Evaluation Results of Various Classification Models for EFR (English)

tinct performance characteristics, as detailed in Table 6, 7 and 8. For ERC, the models demonstrated a competitive range of F1-Scores, with Logistic Regression slightly outperforming others in terms of precision and recall, reflecting a balanced capacity for emotion classification within dialogues. Figure 1 shows the confusion matrix of this task. The nuanced demands of EFR in both Hinglish and English conversations required precision as a critical metric due to the importance of accurately identifying emotional shifts. In this context, the SVM model showed notable efficacy, particularly in English conversations, as it achieved a precision of 0.79, indicating a strong ability to discern the nuanced triggers of emotion flips. Further analysis in Table 7 revealed that for EFR within Hinglish code-mixed conversations, the Logistic Regression model surfaced as a frontrunner, achieving a precision of 0.74, underscoring its capability to handle the intricacies of code-switching and the emotional dynamics inherent in bilingual discourse.

Meanwhile, Table 8 underscores the adaptability of machine learning models to the linguistic complexity and emotional subtleties in English conversations, with the Random Forest model marking a precision of 0.81, reflecting its robustness in parsing and understanding the nuanced indicators of emotional transitions.

These findings not only elucidate the strengths and limitations inherent in various computational models' ability to comprehend the complexities of emotional nuances within conversational contexts but also underscore the imperative for further refinement. It is crucial to enhance the sensitivity and precision of these models, especially within the intricate landscape of multilingual and multicultural communication.

8 Conclusion

Our study delves into how language and culture intertwine with emotional dynamics in bilingual conversations, revealing new insights through an examination of Emotion Recognition in Conversations and Emotion Flip Reasoning. It highlights the need for computational models to accurately interpret emotional nuances across different cultures and languages, advocating for interdisciplinary efforts to enhance AI's empathy and cultural awareness. The research aims to improve global understanding and connectivity, contributing to better human-computer interaction and societal unity. Future directions include expanding the research to more languages and cultures, integrating sociolinguistic and anthropological insights into computational models, and exploring the role of multimodal communication in emotion recognition to develop more sophisticated AI systems.

9 Future Work

Moving forward, our research trajectory entails several promising avenues for exploration. Firstly, we aim to broaden our linguistic and cultural scope by expanding our investigations to encompass a more extensive array of languages and cultural backgrounds. Additionally, we seek to enrich our computational models by integrating insights from fields such as sociolinguistics and anthropology, thereby fostering a more nuanced understanding of emotional expression within diverse societal frameworks. Moreover, we endeavor to delve into the realm of multimodal communication to unravel the intricate interplay between verbal and nonverbal cues in emotion recognition. By embracing these future directions, we aspire to cultivate more sophisticated AI systems capable of seamlessly navigating the intricate tapestry of human emotion across diverse cultural landscapes.

References

Shree Harsh Attri, TV Prasad, and Gavirneni RamaKrishna. 2020. Hiphet: A hybrid approach to translate code mixed language (hinglish) to pure languages (hindi and english). *Computer Science*, 21(3).

Aditi Bagora, Kamal Shrestha, Kaushal Maurya, and Maunendra Sankar Desarkar. 2022. Hostility detection in online hindi-english code-mixed conversations. In *14th ACM Web Science Conference 2022*, WebSci '22, page 390–400, New York, NY, USA. Association for Computing Machinery.

Manjot Bedi, Shivani Kumar, Md Shad Akhtar, and Tanmoy Chakraborty. 2023. Multi-modal sarcasm detection and humor classification in code-mixed conversations. *IEEE Transactions on Affective Computing*, 14(2):1363–1375.

Marta Dabrowska. 2019. What is indian in indian english? markers of indianness in hindi-speaking users' social media communication. *Journal on Asian Linguistic Anthropology*.

Amita Umesh Dessai and Hassanali G. Virani. 2023. *Emotion Detection and Classification Using Machine Learning Techniques*, pages 11–31. Multidisciplinary Applications of Deep Learning-Based Artificial Emotional Intelligence. IGI Global, Hershey, PA, USA.

OpenAI et. al. 2024. Gpt-4 technical report. *Preprint*, arXiv:2303.08774.

Neha Garg. 2020. Annotated corpus creation for sentiment analysis in code-mixed hindi-english (hinglish) social network data. *Indian Journal of Science and Technology*, 13:4216–4224.

Devansh Gautam, Kshitij Gupta, and Manish Shrivastava. 2021. Translate and classify: Improving sequence level classification for English-Hindi code-mixed data. In *Proceedings of the Fifth Workshop on Computational Approaches to Linguistic Code-Switching*, pages 15–25, Online. Association for Computational Linguistics.

Runfang Guo, Hongfei Guo, Liwen Wang, Mengmeng Chen, Dong Yang, and Bin Li. 2024. Development and application of emotion recognition technology — a systematic literature review. *BMC Psychology*, 12(1):95.

Chao-Chun Hsu, Sheng-Yeh Chen, Chuan-Chun Kuo, Ting-Hao Huang, and Lun-Wei Ku. 2018. EmotionLines: An emotion corpus of multi-party conversations. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Ganesh Jawahar, El Moatez Billah Nagoudi, Muhammad Abdul-Mageed, and Laks Lakshmanan, V.S. 2021. Exploring text-to-text transformers for English to Hinglish machine translation with synthetic code-mixing. In *Proceedings of the Fifth Workshop on Computational Approaches to Linguistic Code-Switching*, pages 36–46, Online. Association for Computational Linguistics.

Prashant Kodali, Anmol Goel, Monojit Choudhury, Manish Shrivastava, and Ponnurangam Kumaraguru. 2022. SyMCoM - syntactic measure of code mixing a study of English-Hindi code-mixing. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 472–480, Dublin, Ireland. Association for Computational Linguistics.

Rahul Kulkarni, Dr Varade, and Rajendra Yelalwar. 2023. Nlp and human-computer interaction: Enhancing user experience through language technology. *International Journal for Research in Applied Science and Engineering Technology*, 11:1393–1396.

Shivani Kumar, Md Shad Akhtar, Erik Cambria, and Tanmoy Chakraborty. 2024a. Semeval 2024 – task 10: Emotion discovery and reasoning its flip in conversation (ediref). *Preprint*, arXiv:2402.18944.

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.

Anoop Kunchukuttan, Divyanshu Kakwani, Satish Golla, Gokul N. C., Avik Bhattacharyya, Mitesh M. Khapra, and Pratyush Kumar. 2020. Ai4bharat-indicnlp corpus: Monolingual corpora and word embeddings for indic languages. *Preprint*, arXiv:2005.00085.

Susan Loveland. 2011. Human computer interaction that reaches beyond desktop applications. In *Proceedings of the 42nd ACM Technical Symposium on Computer Science Education*, SIGCSE '11, page 595–600, New York, NY, USA. Association for Computing Machinery.

Rezwan Matin and Damian Valles. 2020. A speech emotion recognition solution-based on support vector machine for children with autism spectrum disorder to help identify human emotions. In *2020 Intermountain Engineering, Technology and Computing (IETC)*, pages 1–6.

Christian Montag, Kenneth L. Davis, and Jaak Panksepp. 2020. *Affective Neuroscience Theory and Other Approaches to Understanding Emotions*, pages 121–130. Punctum Books.

O Muhammadiyeva. 2022. Use of language elements in the process of social communication. *Zamonaviy lingvistik tadqiqotlar: xorijiy tajribalar, istiqbolli izlanishlar va tillarni o'qitishning innovatsion usullari*, 1:321–323.

Siddhartha Mukherjee. 2019. Deep learning technique for sentiment analysis of hindi-english code-mixed text using late fusion of character and word features. In *2019 IEEE 16th India Council International Conference (INDICON)*, pages 1–4.

Sahoko Nakayama, Takatomo Kano, Andros Tjandra, Sakriani Sakti, and Satoshi Nakamura. 2019. Recognition and translation of code-switching speech utterances. In *2019 22nd Conference of the Oriental COCOSDA International Committee for the Coordination and Standardisation of Speech Databases and Assessment Techniques (O-COCOSDA)*, pages 1–6.

Anita Ramalingam, Nirav Agarwal, and Harshvardhan Arvind Singh. 2023. Detection of virulent messages written in code-mixed hindi-english language. In *Recent Trends in Data Science and its Applications*. River Publishers.

Leonardo Ranaldi and Giulia Pucci. 2023. Does the English matter? elicit cross-lingual abilities of large language models. In *Proceedings of the 3rd Workshop on Multi-lingual Representation Learning (MRL)*, pages 173–183, Singapore. Association for Computational Linguistics.

Ayoub Razouk, Hamza Melliani, Jallal Mohamed El Adnani, and Moulay El Mehdi Falloul. 2023. Analyzing public sentiment towards islamic finance through social media: Using sentiment analysis on twitter data. *Revue d'Intelligence Artificielle*, 37.

Dipti Sharma, Munish Sabharwal, Vinay Goyal, and Mohit Vij. 2020. Sentiment analysis techniques for social media data: A review. In *First International Conference on Sustainable Technologies for Computational Intelligence*, pages 75–90, Singapore. Springer Singapore.

Dinkar Sitaram, Savitha Murthy, Debraj Ray, Devansh Sharma, and Kashyap Dhar. 2015. Sentiment analysis of mixed language employing hindi-english code switching. In *2015 International Conference on Machine Learning and Cybernetics (ICMLC)*, volume 1, pages 271–276.

Dattatray G. Takale. 2024. A study of natural language processing in healthcare industries. *Journal of Web Applications and Cyber Security (e-ISSN: 2584-0908)*, 2(2):1–6.

Kendall Vogh. 2022. *Chapter 14. Code-mixing and semantico-pragmatic resources in francophone Maine: Meanings-in-use of yeah/yes and ouais/oui*, pages 244–256. John Benjamins Publishing Company.

Yu Wenz Xin He. 2022. distilbert-base-uncased-finetuned-sst-2-english-int8-static.

Konark Yadav, Aashish Lamba, Dhruv Gupta, Ansh Gupta, Purnendu Karmakar, and Sandeep Saini. 2020. Bi-lstm and ensemble based bilingual sentiment analysis for a code-mixed hindi-english social media text. In *2020 IEEE 17th India Council International Conference (INDICON)*, pages 1–6.