Knowledge-Infused Hierarchy-Aware Emotion Recognition in Code-mixed Mental Health Counseling Conversations

Aseem Srivastava, Kushagra Mittal, Anusha Tiwari, Md. Shad Akhtar
IIIT-Delhi, New Delhi, India
{aseems, kushagra20075, anusha20362, shad.akhtar}@iiitd.ac.in

Abstract

Effective counseling is often best achieved in a client's preferred language, allowing better emotional resonance. Despite this, most existing research in emotion recognition in counseling focuses predominantly on English, overlooking the rich emotional and linguistic complexities of other widely spoken languages. Hinglish, a code-mixed blend of Hindi and English, is one such underexplored linguistic medium that millions use to express their emotions authentically. To address this gap, our research lays a foundational step in developing a mental-health conversation dataset in code-mixed Hinglish language, aka. `IndieMH`. We manually translate counseling conversations from publicly available sources into Hinglish. Moreover, we employ the dataset for emotion classification task for counseling patients. We prepare an exhaustive annotation guideline to annotate `IndieMH` with 13 emotional states under 3 broad emotion categories. Our rigorous sanity check ensures that the quality of `IndieMH` adheres to research standards. Furthermore, we propose a novel knowledge-cum-hierarchy aware method named `Healer` for counseling emotion classification in the Hinglish language. To evaluate the model's performance, we benchmark `Healer` against 11 potential baseline methods and report standard classification metrics, including accuracy, weighted-F1, and weighted-precision.

Keywords: Mental Health Counseling Conversation, Code-mixed, Emotion Recognition in Conversation

1. Introduction

Mental health counseling requires an effective communication, where language plays a pivotal role. However, the accessibility of mental health support remains uneven across linguistic and cultural boundaries, particularly for speakers of low-resource and code-mixed languages. Hinglish, a blend of Hindi and English, is one such linguistic medium that millions rely on to express themselves. Despite its ubiquity, mental health resources, including virtual mental health assistants (VMHAs), predominantly cater to English-speaking populations, leaving Hinglish speakers underserved. VMHAs, such as Woebot and Elomia[1], have gained prominence as scalable solutions to address the growing demand for mental health support. These AI-driven bots offer asynchronous, round-the-clock support for individuals facing mental health challenges. However, the effectiveness of such systems depends heavily on their ability to engage clients in a linguistically and emotionally resonant manner (Pathak et al., 2024). VMHAs primarily first understand patients on multiple verticals like emotion, dialogue-act, etc., followed by providing a suitable support delivery. However, for Hinglish speakers, this presents a critical gap.

Previous research in the domain of multilingual mental health has made notable advancements. Contributions for standalone Hindi language and

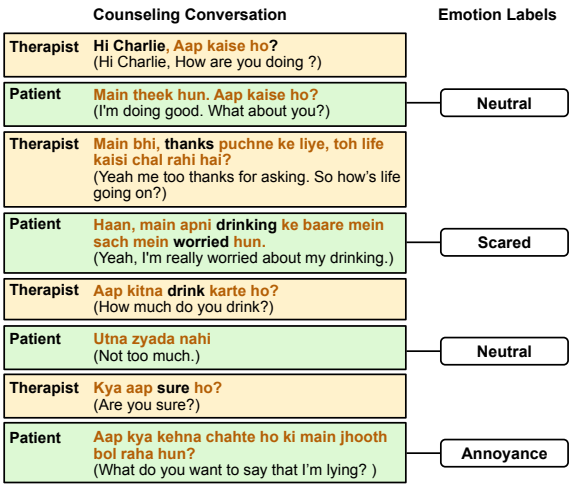

Figure 1: A sample code-mixed counseling conversation between a therapist and a patient tagged with the patient's emotion labels. We present conversation in both the expert-translated Hinglish and the original English language. Orange tokens are Hinglish tokens.

diverse applications on top have been enormous for a series of tasks such as sentiment analysis (Akhtar et al., 2016b, 2018, 2016a), event identification (Sahoo et al., 2020), humor recognition (Chauhan et al., 2022) and in certain cases considered to be a low-resource language (Mamta et al., 2022). On the contrary, code-mixing is also studied for limited tasks such as sentimental classification on code-mixed Hinglish tweets (Ghosh et al., 2023)

[1]www.woebot.com; www.elomia.com

Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026), pages 2887–2898
11–16 May 2026.

and open-domain code-mixed question-answering (Gupta et al., 2020). A number of other studies discuss code-mixing for many other application areas, but to date, there is a noticeable lack of studies in the space of emotion classification for code-mixed mental health counseling. The intricate interplay of codemixed nuances and emotional well-being forms a complex narrative that requires our attention and understanding.

To this end, we propose IndieMH, a novel emotion-annotated code-mixed mental health counseling dialogue dataset. IndieMH comprises 190 counseling sessions that underwent meticulous translations, hierarchical emotion annotations, and extensive expert validation and sanity checks. This is done following a three-phase pipeline to ensure the adequate quality, safety, and robustness of the dataset construction. Each iterative phase concluded with a meticulous review, cross-validation, and feedback loop involving experts. This ensures that the code-mixing from English to Hinglish captures the essence of natural communication, reflecting the colloquial linguistic nuances of a real therapeutic experience. To the best of our knowledge, this is the first-of-its-kind Hinglish (Hindi-English code-mixed) counseling dialogue dataset. We rely on a publicly available counseling dataset – HOPE (Malhotra et al., 2022), which experts manually clean and translate in order to compile the proposed dataset. IndieMH contains $\sim 11.5K$ utterances, out of which $5.7K$ are patient's utterances and $5.8K$ are therapist's utterances.

In Figure 1, the therapist begins by greeting the client, and the conversation develops from there. Each patient utterance is labeled with an emotion, while the therapist generally maintains a neutral tone. Patients, on the other hand, express a range of emotions throughout the dialogue. It's essential for VMHA (just like experts) to first gauge the patient's emotional state and then respond. Hence, we propose Healer, a novel knowledge-enhanced method for emotion classification in counseling patients. Healer leverages two essential components: (a) contextual knowledge and (b) domain-centric knowledge. Unlike primitive methods that rely solely on dialogue, which can leave important nuances unaddressed in sensitive areas such as mental health, Healer incorporates additional domain knowledge to understand and interpret these nuances effectively. Additionally, patients' emotions also tend to show a bias toward certain emotional states, causing deep learning models to catch class bias. As a result, we trained our model in a hierarchical manner following the hierarchy order set by experts in the annotation guidelines. To compare the performance, we benchmark Healer against 11 competitive baselines, all relevant to our scope of research, and report standard classification metric scores, including accuracy, weighted-F1, and weighted-precision. Our evaluation shows substantial improvement in performance compared to baselines. Further, we present detailed analysis to attest Healer's performance. We summarize the main contributions below:

- We propose a first-of-its-kind, code-mixed counseling dataset, IndieMH, for patient emotion recognition to cater to Hinglish's rich linguistic identity.
- We perform an extensive analysis on the dataset and baselines to present intricate findings. We scrutinize the dataset's diversity to understand the multifaceted inherent emotional dynamics.
- We propose a novel hierarchy-cum-knowledge aware method, called Healer for counseling emotion recognition task.
- We perform extensive benchmarking of Healer on the proposed dataset, IndieMH against 11 potential baselines. The exhaustive evaluation and analysis further

Reproducibility. Dataset and code is open-sourced at github.com/flamenlp/IndieMH.

2. Related Work

We provide an overview of the existing literature in two relevant spaces: (i) mental health counseling and (ii) code-mixing for emotion classification.

Mental Health Counseling. Counseling has transitioned from traditional face-to-face interactions to text-based formats for several compelling reasons. The shift is driven by increased accessibility, allowing individuals to seek support from the comfort of their own environments. Such VMHA-based solutions to facilitate counseling solutions have been much explored (Saha et al., 2022a,b). For instance, client understanding in the counseling, where they propose a model achieving state-of-the-art results on identifying dialogue-acts (Malhotra et al., 2022). Another work proposed a new English counseling summarization dataset for counseling dialogue summarization using domain-specific knowledge (Srivastava et al., 2022). Virtual mental health assistants are also a popular source for various health applications, such as delivering therapy for depression and anxiety (Fitzpatrick et al., 2017), improving physical activity and diet (Maher et al., 2020), and conducting virtual client monitoring (Cleres et al., 2021).

Code-mix and Emotion Classification. Understanding emotions in text-counseling is a multi-faceted task. Earlier works have explored mental health counseling and emotion classification within

Split	#Dial	Speaker	Emotion Labels													Total
			AGR	ANY	ANT	COF	CON	DIS	JOY	NTR	SAD	SER	SCR	SUR	TRU	
Train	130	**Patient**	32	245	124	170	58	25	62	2246	689	42	363	14	48	4118
Test	39	**Patient**	4	23	28	41	8	3	12	618	178	8	63	3	15	1004
Val	21	**Patient**	1	24	18	31	7	3	6	314	101	11	67	3	8	594
Total	190	**Patient**	37	292	170	242	73	31	80	3178	968	61	493	20	71	5716

Table 1: Emotion labels distribution in IndieMH. The train, test, and validation splits are 70:20:10.

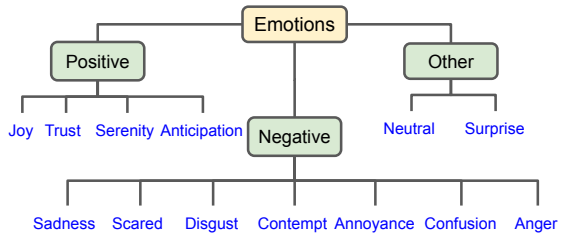

Figure 2: Distribution of the 13 emotion labels among 3 emotion categories

text-based settings. Earlier, (Madhu Midhan et al., 2023) worked on recognizing emotions for text-based conversations, focusing on generic machine learning algorithms. However, code-mixing is a less-touched space by the sight of research. This further extends to the Hinglish dataset, which is much less researched when compared to common languages, viz. English. There exist works in detecting entailment in code-mixed Hinglish dialogue focusing on monolingual aspect (Chakravarthy et al., 2020). (Gupta et al., 2018) presents a benchmark for multilingual question answering in English and Hindi. Another work proposes Hinglish code-mixed memes for sentiment, emotion, and emotion intensity (Mishra et al., 2023). Other works include resources and emotion classification tasks on the Hinglish dataset (Ghosh et al., 2023; Vijay et al., 2018), which mostly focus on social media and open domain text corpus. The research further extends to other languages, which includes Chinese-English (Lovenia et al., 2022), Polish-English (Zygadło et al., 2021), Italian-English (Bianchi et al., 2021) and many more. However, we barely find any research dedicated to emotion classification in code-mixed Hinglish language for counseling, marking a significant gap in improving research on mental health support for Hinglish speakers.

3. Dataset

In this section, we discuss our proposed code-mixed counseling conversation dataset for emotion recognition, named IndieMH. In total, we annotate $\sim 5.7K$ utterances with 13 emotion labels carefully designed to cater to the emotions observed in counseling. The remaining section furnishes the details of data collection, annotation schemes, emotion labels, and necessary statistics.

3.1. Dataset Source

One of the existing hurdles in the process of data selection is the unavailability of public counseling sessions, mainly due to the fact that they usually contain sensitive personal information. Hence, to develop a dataset for code-mixed Hinglish counseling, we utilize HOPE, a publicly available counseling dataset available in English (Malhotra et al., 2022). A dedicated team manually translates the conversations from English to Hinglish, adhering to translation and annotation guidelines. The resulting dataset consists of $11.5K$ utterances from 190 dyadic conversations.

3.2. Code-Mix Translation

The code-mix translation task involves the translation from English to code-mixed (Hinglish), aiming to capture the linguistic nuances inherent to Hinglish-speaking communities in the context of counseling. To ensure the adequacy of the translations, we go through three phases to arrive at the final high-quality code-mix corpus. Next, we discuss these phases and quality assessment.

In the first two phases, we explored existing language resources and tools, aiming to rely on LLMs to achieve natural translations. However, our rigorous evaluation reveals that the available tools fall short of delivering translation quality. Hence, we progress to the third phase, where we introduce an expert-driven, manual annotation process.

3.3. Manual Translation

Our approach to translating the original dataset into code-mixed Hinglish involves a structured process that ensures linguistic authenticity and cultural relevance. We begin by enlisting expert annotators

Dataset	Avg. Length			CMI
	Train	Test	Val	
HOPE	21.50	23.42	22.63	-
IndieMH	21.88	23.82	24.39	74.28

Table 2: Code-mixing Quality. The average utterance length in the original HOPE dataset and the code-mixed Hinglish dataset, IndieMH. CMI represents the Code Mixing Index.

Utterance	Emotion
Main toh awesome hoon. (*I'm doing awesome.*)	Joy
Haan, ye bahut helpful rahega. (*Yeah, that'd be really helpful.*)	Trust
Mujhe lagta hai critic samajhta hai...Mujhe ye dikh raha hai. (*I think the critic feels like it understands...I can see that.*)	Serenity
Mai 20 pounds ghatana chahoonga...bahut achcha hoga. (*I would like 20 pounds...that would be great.*)	Anticipation
Wow! Tum music seekh rahi ho? (*Wow! You're learning music?*)	Surprise
Ye bahut painful hai, bahut hi zyada..., jisse mai bahar nikalna chahti hoon. (*It's painful. Definitely. Definitely... I want out.*)	Sadness
Haan, isse abhi bhi panic hota hai...jab mai inhe leta hoon. (*Yeah, it's still. It's still panic when I take them...*)	Scared
Yeah, hume koi tarika nikalna hoga...Wo bas mere piche pada hai. (*Yeah, we got to figure out some way. He's just on me.*)	Annoyance
Woh hai, jo jerk hai. (*He's the one being a jerk.*)	Anger
Mujhe ek loser ki tarah feel ho raha tha. (*I felt like a loser.*)	Disgust
You know, mere ideas unki samajh se bahar the. (*You know, my ideas were just kind of higher than what they could understand*)	Contempt
Mujhe pakka hai ki hua hai, but mujhe theek se yaad nahi aa raha... (*I'm sure there has been but I can't really remember one.*)	Confusion
To mai tumse kuch questions poochoonga... (*So I'll be asking you a series of questions...*)	Neutral

Table 3: Annotated examples in IndieMH.

with a deep understanding of counseling and its linguistic nuances. The annotators receive comprehensive training to acquaint themselves with the unique linguistic nuances and code-mixed aspects of counseling sessions for annotation purposes. Throughout this phase, annotators recognize that Hinglish communication often consists of a base grammar rooted in conversational Hindi, augmented by seamless switches to commonly used English terms like *symptoms*, *thought*, *problem* as opposed to their Hindi counterpart *lakshan*, *vichar* and *samasya*. This linguistic adaptation mirrors the natural flow of dialogue in this context.

Code-Mixing Quality. The average number of tokens per utterance post-manual translation does not show a significant difference in average utterance length. As evident in Table 2, where we present average utterance length pre- and post-manual annotation and translation. We also calculate the corpus level code-mixing index (CMI) (Gambäck and Das, 2016) to validate the quality of code-mixing. The resultant CMI score of 74.28% indicates that most of the utterances in IndieMH are adequately code-mixed.

3.4. Emotion Annotation Guidelines

IndieMH contains dyadic conversations involving two speakers: *therapists* and *patients*. To design a robust annotation guideline, we employed in-house experts to ensure both safety and robustness in our emotion labeling pipeline. The complete guideline preparation was carried out in three major phases, during which we conducted sample annotation rounds. Each random sample annotation round involved annotating 20 dialogues and revising the guidelines based on the observations collected. The final set of emotion labels follows a hierarchy as shown in Figure 2, that includes three broad categories: (i) positive, (ii) negative, and (iii) neutral. A total of 13 emotion labels are categorized under this hierarchy. Precisely, the *positive* emotion category contains the following 4 emotion labels: **joy** (JOY), **trust** (TRU), **serenity** (SER), and **anticipation** (ANT); likewise, *neutral* category of emotions contain following 2 labels: **neutral** (NTR), and **surprise** (SUR); whereas *negative* category of emotions contains 7 labels: **sadness** (SAD), **scared** (SCR), **annoyance** (ANY), **anger** (AGR), **disgust** (DIS), **contempt** (CON), and **confusion** (COF). We present an example of a code-mixed counseling session, consisting of 13 rounds of exchanges between the therapist and patient, along

Language	Train			Test			Val		
	#Vocab	Avg /Utt	Avg /Dial	#Vocab	Avg /Utt	Avg /Dial	#Vocab	Avg /Utt	Avg /Dial
English	34380	4.36	272.85	10507	5.23	269.41	5705	4.49	271.67
Hindi (Romanized)	142590	18.08	1131.67	38695	19.27	992.18	23986	20.15	1142.19
Total	176970			49202			29691		

Table 4: Vocabulary distribution of Hinglish code-mixing in IndieMH. The table showcases vocabulary size (#Vocab), the average number of tokens on utterance (Avg /Utt), and dialogue (Avg /Dial) level for each data split.

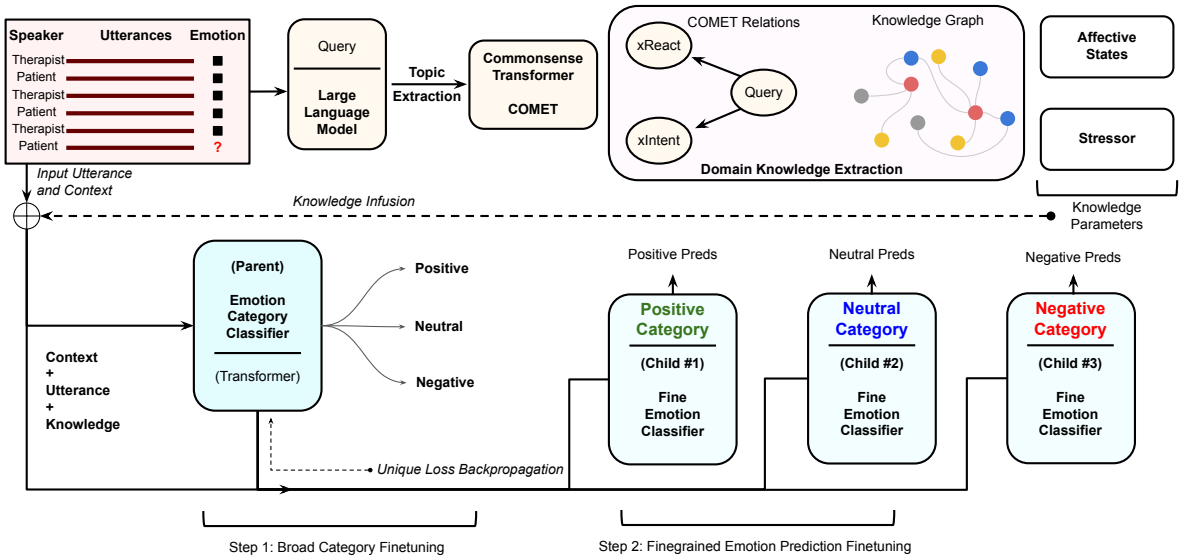

Figure 3: A schematic diagram of Healerthat leverages two major components: (a) knowledge-infusion and (b) context-infusion. For knowledge-infusion, Healer relies on LLM (MentaLlama) and COMET to allow the domain-centered knowledge graph to fetch domain-relevant information. As a result, this extracted knowledge, along with standard context, is used for hierarchical training for emotion recognition.

with the corresponding emotion labels, in Table 3

Annotator Details. We employ ten expert annotators, taking care of a balanced gender distribution with six females and four males. All annotators are fluent in both English and Hindi falling within the age group of 20-40 years. They possess fluency in both English and Hindi, ensuring a seamless transition in the code-mix translation process. Additionally, the annotators inherit diverse cultural and socio-economic backgrounds, further enriching the annotation process while effectively mitigating the potential for unintentional biases. After translating, two of these experts annotate emotion labels for the entire dataset. The iterative process of annotation and discussion continues until we achieve an acceptable level of agreement. We evaluate the annotation quality using Cohen's Kappa score (Cohen, 1960). We obtain an average inter-rater agreement score of 0.82 – which falls under the *substantial* category.

3.5. Dataset Statistics

The dataset statistics of the code-mixed dialogue in IndieMH is summarised in Table 1. We observe that the most common emotion label in patients' utterances, apart from *NTR*, is a negative emotion *sadness* (SAD). It is followed by *serenity* (SER), *annoyance* (ANY), and *confusion* (COF). It is worth noting that one may expect SAD and *scared* (SCR) to be the most common emotions among clients, considering prevalent mental health issues. However, SCR ranks ninth most common emotion in IndieMH, and the reason is the client's behavior, which is often miserable about past episodes of severe issues; hence, talking about them in a counseling session emerges complex emotions such as *sadness* as opposed to *scared*. Nonetheless, IndieMH is significantly skewed toward *neutral* emotion label, with 2246/4118, 618/1004, and 314/594 instances appearing across train, test, and val split.

3.6. Linguistic statistics

The dataset statistics of the code-mixed dialogue is summarised in Table 4. The average number of English tokens per utterance is 4.35, 5.23, and 4.49 in the train, test, and validation sets, respectively. This is an accurate reflection of the conversational Hinglish speakers, as people often use fewer English words and majorly use Hindi vocabulary.

4. Methodology

In a typical setting, a counseling conversation dialogue, D consisting of a sequence of utterances $U[1:m] = \langle U_1, U_2, ..., U_m \rangle$, where m is the number of utterances in a dialogue. The objective of Healer is to assign a correct emotion label y_t to every utterance U_t. As shown in Figure 3, the complete architecture is divided into two major threads: (a) knowledge extraction and (b) contextualization. Collectively, these contribute to the hierarchical training of Healer.

4.1. Knowledge Extraction

This process is bifurcated into two major phases. Phase-I involves extracting commonsense knowl-

edge, followed by Phase-II, where we extract knowledge using a domain-relevant knowledge graph to achieve structured and domain-specific insights.

Phase-I: Commonsense Knowledge. To initiate the knowledge extraction process, we first form a *query* by concatenating the immediate context and the current utterance. This query is input to an LLM specifically meant to understand mental health-related context, which acts as our topic extraction module using MentaLLaMA (Yang et al., 2024). Using a dedicated prompt, MentaLLaMA extracts the top five relevant topics from the input query. These relevant topics are then used as input to the commonsense transformer (COMET) model (Bosselut et al., 2019). It generates responses based on the extractions from *xReact* and *xIntent* relations, which are essential for understanding the emotional reactions and intentions behind the utterances.

Phase-II: Domain-centric Knowledge. We utilize the HEAL knowledge graph, which includes five types of nodes-stressors, expectations, response types, feedback types, and affective states for counseling management conversations (Welivita and Pu, 2022). In our study, we employ *affective* states and *stressor* nodes relevant to our downstream task for topical and contextual understanding. We discuss this more about the selection of knowledge graph in Section 6. HEAL yields keywords associated with each node's properties. We fuse the usage of HEAL with commonsense reasoning via COMET. The *xReact* COMET beams are employed to calculate similarity with all *affective* state nodes in the knowledge graph. Similarly, the *xIntent* beams are used to calculate similarity with all *stressor* nodes. Once the similarity scores are obtained, we identify the *stressor* node with the highest similarity score. We then traverse all the edges connected to this *stressor* node to find the *affective* state node that is connected to the *stressor* and has the highest similarity. This ensures that the most relevant *affective* states and *stressors* are identified based on the context and content.

The outputs from the *knowledge extraction* module provide enhanced information that will be integrated into the hierarchical learning method. This fusion of knowledge extraction and contextualization enables `Healer` to effectively classify emotions in the `IndieMH` dataset.

4.2. Contextualization

In `Healer`, the context considered is the local context. For an utterance U_i, the *local context* is defined as immediate utterances till the last therapist's utterance that appears before u_i. This is done to avoid the bias introduced by global context in the knowledge-extraction module. This bias arises from topics getting diluted across multiple utterances, which is characteristic of typical counseling conversations. These dialogues often contain many filler utterances, such as *"Really?"* or *"Yes, I could do that."*, which do not significantly contribute to defining the dialogue topics.

4.3. Hierarchical Learning

The `IndieMH` dataset has a significant class imbalance, which causes classical deep learning frameworks to fall into the trap of biased predictions. Hierarchical learning allows the proposed model, `Healer`, to first learn the broad-level categorization of emotion labels in `IndieMH`through the emotion category classifier, followed by the fine-grained learning of specific emotion labels. This hierarchical learning approach is achieved by employing *HingRoberta* and fine-tuning it for three broad emotion class predictions, namely, *positive*, *negative*, and *other*, for each counseling utterance by the patient. Further, we employ an additional layer of the previously pretrained emotion category classifier (broad-level) to perform the task of fine-grained emotion classification by replacing the last linear layer with a new layer for the classification of 13 fine-grained emotion labels. This two-step hierarchical learning process enables `Healer` to effectively handle the class imbalance and achieve better emotion classification.

5. Experiments and Results

Here, we present the baseline selection, performance comparison, ablation, and analyses.

Baselines. We choose the following systems as our baselines: ►**MURIL** (Khanuja et al., 2021) is a BERT-based model pretrained on a corpora across 17 languages, including Hindi[2]. ►**IndicBERT** (Kakwani et al., 2020) is a multilingual ALBERT model pretrained on 12 languages (including Hindi) with 9B tokens[3]. ►**Multilingual MiniLM** (Wang et al., 2020) is a DistilBERT-based model fine-tuned on the multilingual corpus[4]. ►**mBERT** (Devlin et al., 2018) is a BERT-Base architecture pretrained on 104 languages, including Hindi[5]. ►**XLM-RoBERTa** (Conneau et al., 2019) is a RoBERTa-based architecture pretrained on 2.5TB of corpus containing 100 languages[6]. ►**Knowledge Enriched Transformer (KET)** (Zhong et al., 2019) uses hierar-

[2]https://huggingface.co/google/muril-base-cased
[3]https://github.com/AI4Bharat/Indic-BERT-v1
[4]https://github.com/microsoft/unilm/tree/master/minilm
[5]Link: https://github.com/google-research/bert
[6]https://t.ly/5ANph

Model	Pretraining	Acc.	$F1_w$	P_w
BERT (Devlin et al., 2018)	English	57.47	55.69	54.21
RoBERTa (Devlin et al., 2018)	English	57.77	58.55	59.82
XLM-RoBERTa (Conneau et al., 2019)	English	57.67	58.66	60.22
MURIL (Khanuja et al., 2021)	Hinglish	51.29	53.39	57.70
IndicBERT (Kakwani et al., 2020)	Hinglish	57.07	52.31	48.89
Multiling. MiniLM (Wang et al., 2020)	Hinglish	57.47	54.55	55.69
mBERT (Devlin et al., 2018)	Hinglish	48.90	51.94	59.75
HingBERT (Nayak and Joshi, 2022)	Hinglish	60.26	60.53	61.36
HingBERT LID (Nayak and Joshi, 2022)	Hinglish	59.36	59.90	60.97
HingRoBERTa (Nayak and Joshi, 2022)	Hinglish	61.75	62.59	64.13
HingMBERT (Nayak and Joshi, 2022)	Hinglish	56.27	58.38	62.26
Healer	Hinglish	**63.45**	**63.58**	64.26
− Context − Knowledge	Hinglish	57.37	59.72	**64.94**
+ Context − Knowledge	Hinglish	56.08	58.87	63.36
− Context + Knowledge	Hinglish	58.47	60.48	64.05
$\Delta_{\text{Healer}-BEST}(\%)$	–	$\uparrow 2.75$	$\uparrow 1.58$	$\uparrow 0.20$

Table 5: Result and Ablation: We benchmark Healer on IndieMH, across 11 potentially relevant baseline from both pertained language: English and code-mixed (Hinglish), evaluating their performance in comparison to 3 standard classification metrics: accuracy (Acc.), weighted-F1 ($F1_w$) and weighted-precision (P_w) for the emotion recognition task.

chical self-attention and leverages commonsense using context-aware affective graph attention mechanism for emotion detection in dialogues[7]. ►**Hing Model Family** (Nayak and Joshi, 2022) is a series of transformer models pretrained on *L3Cube-Hing code-mixed corpus*, consisting 52.93M sentences and 1.04B tokens [8]. ►**BERT**[9](Devlin et al., 2018) and **RoBERTa**[10] (Devlin et al., 2018) are encoder only transformers models .

Performance Comparison. For comparative analysis, we divided our comparison into two segments based on their pretraining language: English and Code-mixed (Hinglish) language. The experimental results of Healer are presented in Table 5. Among the English language models, *XLM-Roberta* achieved the highest $F1_w$ score of 58.66. However, within the code-mixed category, the *HingRoberta* model outperformed all other baselines, obtaining a $F1_w$ score of 62.59, a P_w score of 64.13, and accuracy of 61.75%. Our proposed method, Healer, achieves the best performance, obtaining a $F1_w$ score of $63.58(+1.58\%)$, a P_w score of $64.26(+0.20\%)$, and an accuracy of $63.45\%(+2.75\%)$.

Ablation Study. To assess the contribution of each component, we perform a detailed ablation study. In our study, the role of both components, *knowledge* and *context* in the pipeline plays a significant role. As presented in Table 5, the performance deteriorates if either of these components are unplugged. For instance, unplugging context reduces the Healer's performance by -7.84%, -4.87%, and -0.32% across accuracy, $F1_w$, and P_w, respectively, whereas, unplugging knowledge changes the model's performance by -11.61%, -7.40%, and -1.40% across accuracy, $F1_w$, and P_w, respectively. On the other hand, if we detach both context and knowledge components together, the Healer shows a reduction in accuracy by -9.58%, $F1_w$, by -6.07% and P_w by $+1.05\%$.

Qualitative and Error Analysis. To assess the quality of the predictions by Healer, we present a few samples in Table 6. we present a 6 dedicated examples from IndieMH, highlighting correct and incorrect predictions. The correctly classified utterances (#1, #2, and #3) showcase Healer's ability to understand the overall emotion expressed in the utterances instead of being biased towards 'emotionally charged keywords'. For instance, in the first instance (#1), the model correctly predicts the emotion label as "Sadness," which aligns with the gold label. This correct prediction shows that the model successfully interprets the speaker's emotional tone in the context of financial stress and seeking help. The identified *affective state* is *caring*, which matches the emotional content of the speaker, who is expressing vulnerability and reliance on others. Additionally, the *stressor* (support, advice, friend, helping, professional) underscores the speaker's need for emotional and financial assistance, which the model accurately associates with sadness. Similarly, the second example (#2) could be considered by humans as an expression of fear, given the phrases *"log mujhe follow kar rahe hain"*. However, Healer is able to take into account the context of the conversation, wherein this utterance was the patient's reply to the therapist's question of *"...Kuch aur cheezein hain jinko lekar tumhe problems hain?"*, which indicates that the patient is in a complaining mood and hence Healer predicts 'annoyance'.

Despite being proficient at capturing figurative emotions in utterances, Healer, in some cases, falls short in identifying a few complex cases. For instance, in the fifth example (#5), the patient's utterance uses phrases like *"Maaf karo"* and *"..aise aur nahi reh sakta."* which could either be uttered with a sad or annoyed temperament; and possibly, additional modalities like audio or video could shed light to correctly identify the emotion of such cases.

6. Discussion

Discussion on Knowledge Graph. In our study, we employ the HEAL knowledge graph to enhance the performance of emotion classification within the domain of code-mixed counseling di-

[7]https://github.com/zhongpeixiang/KET
[8]https://huggingface.co/l3cube-pune/hing-bert
[9]https://github.com/google-research/bert
[10]https://huggingface.co/roberta-base

#	Utterances	Gold	Predicted
1	Financial outlook ki baat karte hain, well, main state se help leta hoon. Aur I mean, I usually have enough to pay for everything. I mean, kuch din aise bhi hote hain jab mujhe bolna padta hai, you know, mere roommate se, "Kya tum mujhe thoda paisa udhaar de sakte ho?" Ya meri sister is always there agar mujhe usse chahiye. (*Let's talk about finances. I get help from the state. Usually I can pay for everything, but some days I have to ask my roommate for a small loan, or my sister helps if I need it.*) [**HEAL Affective States:** Caring] [**HEAL Stressor:** support, advice, friend, helping, professional]	Sadness	Sadness
2	Street pe follow kar rahe hain, jab main shop ja raha hoon toh log mujhe follow kar rahe hain. You know, yeh cheezein roz ho rahi hain, aur aur zyada ho rahi hain. After, maine apne friends ko iske baare mein samjhane ki koshish ki hai, lekin woh bahut pareshan ho jaate hain, woh sochte hain, you know, ki mujhe mental problems hain, lekin unhein nahi dikh raha hai. You know, woh har din mere saath nahi rehte hain. You know. (*People follow me on the street, even when I go to the shop, it's happening daily and increasing. I tried explaining to my friends, but they get upset and think I'm mentally unwell; they don't see what I face every day.*) [**HEAL Affective States:** Lonely] [**HEAL Stressor:** listen, listeners, vent, blows, listens]	Annoyance	Annoyance
3	Haan. Main toh bilkul nahi chahta ki hamari upar ki management, tum jaante ho, ismein involve ho jaye aur yeh bada issue ban jaye. (*Yes. I really don't want upper management to get involved and for this to become a big issue.*) [**HEAL Affective States:** Questioning] [**HEAL Stressor:** mask, wear, identity, fantasying, purposes]	Scared	Scared
4	Jab mujhe kaam par jaana hota hai, phir ghar aake, you know, I just sleep a lot, aur lagta hai ki yeh abnormal hai. Aur ab kuch dino se maine kam khana shuru kar diya hai, I've lost a bit of weight, and it's affecting my life, isiliye main bus soch raha hoon ki main aapke paas aata hoon. (*When I go to work and then come home, I sleep a lot, it feels abnormal. For a few days I've been eating less, lost some weight, and it's affecting my life, so I thought I should come to you.*) [**HEAL Affective States:** Ashamed] [**HEAL Stressor:** eating, weight, eat, lose, fat]	Scared	Sadness
5	Maaf karo, aur woh nahi chahte ki main position badlu. Lekin main nahi kar sakta, main aise aur nahi reh sakta. (*Sorry, and they don't want me to change my position. But I can't do this anymore; I can't live like this.*) [**HEAL Affective States:** Questioning] [**HEAL Stressor:** success, stories, motivate, outcome, overcame]	Sadness	Annoyance
6	Um, woh to main bilkul sure nahi hoon. Lagta hai mahine ke end par hoga. (*Um, I'm not completely sure. I think it will be at the end of the month.*) [**HEAL Affective States:** Sad] [**HEAL Stressor:** happiness, remember, forgot, genuinely, happy]	Neutral	Confusion

Table 6: Qualitative and Error Analysis: The table highlights instances from the `IndieMH` dataset and their `Healer`'s emotion predictions. We compare those predictions with gold labels and put findings pertaining to the reasoning behind the model's working. Incorrect and correct predictions are color coded.

alogues. HEAL, developed from over 1 million distress-oriented conversations on Reddit, consists of five core node types: stressors, expectations, response types, feedback types, and affective states. Each node type plays a vital role in understanding the dynamics of distress management conversations, making HEAL an ideal resource for our downstream emotion classification task.

The *stressors* node identifies the causes of emotional distress, providing a link to the origin of the patient's emotions. By aligning patient utterances with relevant stressors, we can better contextualize the emotional states that arise. The *expectations* node captures the specific inquiries or needs expressed by the individual in distress. Understanding these expectations helps the model align responses to the patient's needs, improving the overall accuracy. The *response types* node offers insights into how listeners typically respond to distress, which helps our model predict and validate emotion transitions. Similarly, *feedback types* reflect the speaker's reaction after receiving responses, offering a mechanism to assess whether the emotional state has shifted positively or negatively. This feedback loop is essential for accurately modeling emotional trajectories throughout the conversation. Finally, the *affective states* node directly associates emotions with other nodes, allowing for emotion-based reasoning. With HEAL's structured representation of distress dialogues, we identify relevant stressors and emotional states that improve the emotion classification task.

Challenges in Code-mixing. Code-mixing is a complex problem, and when intersected with mental health counseling, it becomes sensitive as well. After extensive Phase-I experimentation, we observe that there barely exist any well-maintained APIs or language models for adequate code-mix translation of English language sentences to Hinglish code-mixed language. Even if we settle for the weaker or less popular approaches, they lack the required nuances. Hence, we explore various other methods of translation. We discuss attempts that did not work for us in the Supplementary. Nonetheless, after recurring attempts, we finalized the manual annotation method.

7. Conclusion

With thousands of languages and dialects spoken globally, the lack of resources tailored to this diversity obscures mental health needs. In this study, we proposed IndieMH, a novel and first-of-its-kind counseling dataset that addresses the unique emotional dynamics of Hinglish, a widely used code-mixed language. The dataset underwent rigorous annotation and validation processes to ensure the quality of both translation and emotion annotations. To ensure we understand the emotional dynamics in Hinglish counseling, we propose a novel hierarchy-cum-knowledge aware counseling emotion recognition method called Healer. We benchmark Healerwith 11 competitive baselines. The evaluation shows that our Healer performs better on all standard metrics, clearly surpassing the predefined standard for Hinglish. We conclude with a discussion on the core challenges of dealing with code-mixing.

8. Ethics Statement

This work involves sensitive mental-health counseling dialogue and therefore requires careful handling of privacy, consent, and responsible sharing. Our dataset is derived from the HOPE corpus, a previously published resource. We did not collect or curate any source videos; we used only the English dialogue transcripts provided by HOPE to create a Hinglish version and corresponding annotations. Any selection criteria applied to the original online content were performed as part of HOPE's dataset construction, and we inherit the data as released in that work. We preserve HOPE's de-identification approach and apply additional screening to reduce the risk of accidental disclosure of identifiers introduced during translation. Because the upstream resource traces back to online media, we treat redistribution and potential re-identification as primary residual risks. To mitigate these risks, we distribute the derived resource under a controlled-access, research-only, non-commercial data-use agreement that prohibits re-identification attempts and inappropriate linkage to external sources (which is fairly complex given the scale of manual translation this work has gone through). We also provide documentation of intended use, known limitations, and a point of contact for concerns or takedown requests. Finally, models trained on counseling-style conversations can encode biases and may be misused in high-stakes settings. This resource is intended to set up a benchmark for research on emotion understanding in counseling dialogue in Hinglish and should not be deployed as a standalone clinical decision-making component without further validation and safeguards.

9. Limitations

Our dataset focuses on Hinglish (Hindi-English) code-mixed counseling conversations and may not generalize to other code-mixed language pairs or cultural contexts. Further, emotion labels are derived from text-only utterances, which can underspecify affect compared to settings with prosody and visual cues. Expanding coverage to additional code-mixed languages and incorporating richer context signals are important directions for future work.

10. Acknowledgment

We acknowledge the support of Neeraj for her support in shaping the early sample annotation cycles for guideline buidling phase. We also acknowledge the support of Infosys Foundation through the Center for AI (CAI) at IIIT Delhi.

11. Acknowledgment

We thank Neeraj, PhD Scholar at IIIT-Delhi, for her support during the early annotation rounds to refine the guidelines through iterative feedback. We also acknowledge the support of the Infosys Foundation through the Center for Artificial Intelligence (CAI) at IIIT Delhi.

12. Bibliographical References

Md Shad Akhtar, Asif Ekbal, and Pushpak Bhattacharyya. 2016a. Aspect based sentiment analysis in Hindi: Resource creation and evaluation. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 2703–2709, Portorož, Slovenia. European Language Resources Association (ELRA).

Md Shad Akhtar, Asif Ekbal, and Pushpak Bhattacharyya. 2018. Aspect based sentiment analysis: Category detection and sentiment classification for hindi. In *Computational Linguistics and Intelligent Text Processing*, pages 246–257, Cham. Springer International Publishing.

Md Shad Akhtar, Ayush Kumar, Asif Ekbal, and Pushpak Bhattacharyya. 2016b. A hybrid deep learning architecture for sentiment analysis. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 482–493, Osaka,

Japan. The COLING 2016 Organizing Committee.

Federico Bianchi, Debora Nozza, and Dirk Hovy. 2021. FEEL-IT: Emotion and sentiment classification for the Italian language. In *Proceedings of the Eleventh Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 76–83, Online. Association for Computational Linguistics.

Antoine Bosselut, Hannah Rashkin, Maarten Sap, Chaitanya Malaviya, Asli Celikyilmaz, and Yejin Choi. 2019. COMET: Commonsense transformers for automatic knowledge graph construction. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4762–4779, Florence, Italy. Association for Computational Linguistics.

Sharanya Chakravarthy, Anjana Umapathy, and Alan W Black. 2020. Detecting entailment in code-mixed Hindi-English conversations. In *Proceedings of the Sixth Workshop on Noisy User-generated Text (W-NUT 2020)*, pages 165–170, Online. Association for Computational Linguistics.

Dushyant Singh Chauhan, Gopendra Vikram Singh, Aseem Arora, Asif Ekbal, and Pushpak Bhattacharyya. 2022. A sentiment and emotion aware multimodal multiparty humor recognition in multilingual conversational setting. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 6752–6761, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

David Cleres, Frank Rassouli, Martin Brutsche, Tobias Kowatsch, and Filipe Barata. 2021. Lena: a voice-based conversational agent for remote patient monitoring in chronic obstructive pulmonary disease.

Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1):37–46.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Unsupervised cross-lingual representation learning at scale. *CoRR*, abs/1911.02116.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805.

Kathleen Kara Fitzpatrick, Alison Darcy, and Molly Vierhile. 2017. Delivering cognitive behavior therapy to young adults with symptoms of depression and anxiety using a fully automated conversational agent (woebot): A randomized controlled trial. *JMIR Ment Health*, 4(2):e19.

Björn Gambäck and Amitava Das. 2016. Comparing the level of code-switching in corpora. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 1850–1855, Portorož, Slovenia. European Language Resources Association (ELRA).

Soumitra Ghosh, Amit Priyankar, Asif Ekbal, and Pushpak Bhattacharyya. 2023. Multitasking of sentiment detection and emotion recognition in code-mixed hinglish data. *Knowledge-Based Systems*, 260:110182.

Deepak Gupta, Surabhi Kumari, Asif Ekbal, and Pushpak Bhattacharyya. 2018. MMQA: A multi-domain multi-lingual question-answering framework for English and Hindi. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Deepak Gupta, Pabitra Lenka, Asif Ekbal, and Pushpak Bhattacharyya. 2020. A unified framework for multilingual and code-mixed visual question answering. In *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing*, pages 900–913, Suzhou, China. Association for Computational Linguistics.

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 EMNLP*.

Simran Khanuja, Diksha Bansal, Sarvesh Mehtani, Savya Khosla, Atreyee Dey, Balaji Gopalan, Dilip Kumar Margam, Pooja Aggarwal, Rajiv Teja Nagipogu, Shachi Dave, Shruti Gupta, Subhash Chandra Bose Gali, Vish Subramanian, and Partha Talukdar. 2021. Muril: Multilingual representations for indian languages.

Holy Lovenia, Samuel Cahyawijaya, Genta Winata, Peng Xu, Yan Xu, Zihan Liu, Rita Frieske, Tiezheng Yu, Wenliang Dai, Elham J. Barezi, Qifeng Chen, Xiaojuan Ma, Bertram Shi, and Pascale Fung. 2022. ASCEND: A spontaneous Chinese-English dataset for code-switching in

multi-turn conversation. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 7259–7268, Marseille, France. European Language Resources Association.

T Madhu Midhan, Padma Selvaraj, M Harshavardan Kumar Raju., M Bhanu Prakash Reddy., and T Bhaskar. 2023. Classification of mental health and emotion of human from text using machine learning approaches. In *2023 6th International Conference on Information Systems and Computer Networks (ISCON)*, pages 1–7.

Carol Ann Maher, Courtney Rose Davis, Rachel Grace Curtis, Camille Elizabeth Short, and Karen Joy Murphy. 2020. A physical activity and diet program delivered by artificially intelligent virtual health coach: Proof-of-concept study. *JMIR Mhealth Uhealth*, 8(7):e17558.

Ganeshan Malhotra, Abdul Waheed, Aseem Srivastava, Md Shad Akhtar, and Tanmoy Chakraborty. 2022. Speaker and time-aware joint contextual learning for dialogue-act classification in counselling conversations. In *Proceedings of the Fifteenth ACM International Conference on Web Search and Data Mining*, WSDM '22, page 735–745, New York, NY, USA. Association for Computing Machinery.

. Mamta, Asif Ekbal, Pushpak Bhattacharyya, Tista Saha, Alka Kumar, and Shikha Srivastava. 2022. HindiMD: A multi-domain corpora for low-resource sentiment analysis. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 7061–7070, Marseille, France. European Language Resources Association.

Shreyash Mishra, S Suryavardan, Megha Chakraborty, Parth Patwa, Anku Rani, Aman Chadha, Aishwarya Reganti, Amitava Das, Amit Sheth, Manoj Chinnakotla, Asif Ekbal, and Srijan Kumar. 2023. Overview of memotion 3: Sentiment and emotion analysis of codemixed hinglish memes.

Ravindra Nayak and Raviraj Joshi. 2022. L3Cube-HingCorpus and HingBERT: A code mixed Hindi-English dataset and BERT language models. In *Proceedings of the WILDRE-6 Workshop within the 13th Language Resources and Evaluation Conference*, pages 7–12, Marseille, France. European Language Resources Association.

Agnibh Pathak, Soham Bhattacharjee, Tulika Saha, and Sriparna Saha. 2024. Does sentiment and emotion affect mental health? a multi-task classification framework for comprehensive understanding of mental health, emotion, and sentiment from motivational conversations. *ACM Trans. Comput. Healthcare*. Just Accepted.

Tulika Saha, Vaibhav Gakhreja, Anindya Sundar Das, Souhitya Chakraborty, and Sriparna Saha. 2022a. Towards motivational and empathetic response generation in online mental health support. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '22, page 2650–2656, New York, NY, USA. Association for Computing Machinery.

Tulika Saha, Saichethan Reddy, Anindya Das, Sriparna Saha, and Pushpak Bhattacharyya. 2022b. A shoulder to cry on: Towards a motivational virtual assistant for assuaging mental agony. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2436–2449, Seattle, United States. Association for Computational Linguistics.

Sovan Kumar Sahoo, Saumajit Saha, Asif Ekbal, and Pushpak Bhattacharyya. 2020. A platform for event extraction in Hindi. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 2241–2250, Marseille, France. European Language Resources Association.

Aseem Srivastava, Tharun Suresh, Sarah P. Lord, Md Shad Akhtar, and Tanmoy Chakraborty. 2022. Counseling summarization using mental health knowledge guided utterance filtering. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, KDD '22, page 3920–3930, New York, NY, USA. Association for Computing Machinery.

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.

Wenhui Wang, Furu Wei, Li Dong, Hangbo Bao, Nan Yang, and Ming Zhou. 2020. Minilm: Deep self-attention distillation for task-agnostic compression of pre-trained transformers.

Anuradha Welivita and Pearl Pu. 2022. Heal: A knowledge graph for distress management conversations. *Proceedings of the AAAI Conference on Artificial Intelligence*, 36(10):11459–11467.

Kailai Yang, Tianlin Zhang, Ziyan Kuang, Qianqian Xie, Jimin Huang, and Sophia Ananiadou. 2024. Mentallama: Interpretable mental health analysis on social media with large language models. In *Proceedings of the ACM Web Conference 2024*, WWW '24, page 4489–4500, New York, NY, USA. Association for Computing Machinery.

Peixiang Zhong, Di Wang, and Chunyan Miao. 2019. Knowledge-enriched transformer for emotion detection in textual conversations. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 165–176. Association for Computational Linguistics.

Artur Zygadło, Marek Kozłowski, and Artur Janicki. 2021. Text-based emotion recognition in english and polish for therapeutic chatbot. *Applied Sciences*, 11(21).

Supplementary

A. Additional Discussion

A.1. Additional Attempts with LLMs

Basic English-to-Hindi translation does not work in our case as we needed romanized utterances for code-mixed data construction to employ baseline models based on BERT and other similar encoder-only models. We initially theorized an approach of English-to-Hindi translation followed by Hindi-to-English transliteration (writing Hindi words using the English alphabet) to solve our code-mixing problem. We employed the Google Translate library for the first step, i.e., to translate the English sentences to Hindi, and obtained Devanagari translations for each utterance. However, there barely exist any well-documented Hindi-to-English transliteration APIs for the second step that could adequately suffice in our case.

We implemented a variety of prompt engineering techniques to enhance ChatGPT's translation accuracy. Notably, it outperformed other methods by introducing a blend of Hindi and English words into the translations. Despite this progress, several issues were encountered. The translations frequently displayed misgendering of speakers, exhibiting a bias toward male speakers. Additionally, inconsistencies arose in the utilization of second-person pronouns in Hindi, such as *aap* or *tum*. Some translations introduced complex Hindi words like *mahatvapurn* or *mansik swasthya*, while others exhibited a lack of naturalness. This means that the translations did not align with the language typically used in daily conversations of Hinglish speakers. They occasionally converted commonly used English words into more challenging Hindi terms and vice versa, resulting in a disconnect from the linguistic patterns employed by Hinglish speakers in their everyday interactions.

A.2. Why Hierarchical Learning?

To address the significant class imbalance present in the original 13-class dataset, we employed a hierarchical training method. This involved consolidating the 13 classes into 3 more balanced, meaningful categories. This approach mitigates the bias toward the dominant class and ensures the model learns fairly. Subsequently, with the 3 broad emotions categories, we trained separate models for each. This allowed each model to focus on the specific class imbalance within its assigned category rather than the overall imbalance of the initial 13 classes. This two-step process facilitated more effective learning and classification for all emotions.