On the Way to Gentle AI Counselor: Politeness Cause Elicitation and Intensity Tagging in Code-mixed Hinglish Conversations for Social Good

Priyanshu Priya[1]*, **Gopendra Vikram Singh**[1]*, **Mauajama Firdaus**[2], **Jyotsna Agarwal**[3], **Asif Ekbal**[1]
[1]Department of Computer Science and Engineering, Indian Institute of Technology Patna, India
[2]University of Alberta, Canada
[3]National Institute of Mental Health and Neurosciences, Banglore, India
{[1]priyanshu_2021cs26,gopendra_1921cs15,asif}@iitp.ac.in
[2]mauzama.03@gmail.com,[3]jyo2049@nimhans.ac.in

Abstract

Politeness is a multifaceted concept influenced by individual perceptions of what is considered polite or impolite. With this objective, we introduce a novel task - Politeness Cause Elicitation and Intensity Tagging (**PCEIT**). This task focuses on conversations and aims to identify the underlying reasons behind the use of politeness and gauge the degree of politeness conveyed. To address this objective, we create **HING-POEM**, a new conversational dataset in *Hinglish* (a blend of Hindi and English) for mental health and legal counseling of crime victims. The rationale for the domain selection lies in the paramount importance of politeness in mental health and legal counseling of crime victims to ensure a compassionate and cordial atmosphere for them. We enrich the **HING-POEM** dataset by annotating it with politeness labels, politeness causal spans, and intensity values at the level of individual utterances. In the context of the introduced **PCEIT** task, we present **PAANTH** (**P**oliteness C**A**use Elicit**A**ion and I**N**tensity **T**agging in **H**inglish), a comprehensive framework based on Contextual Enhanced Attentive Convolution Transformer. We conduct extensive quantitative and qualitative evaluations to establish the effectiveness of our proposed approach using the newly constructed dataset. Our approach is compared against state-of-the-art baselines, and these analyses help demonstrate the superiority of our method[1].

1 Introduction

Crime is a severe social problem causing tremendous pain to victims. The World Health Organization (WHO) estimates that globally about one-third of the women and one billion children aged 2-17 have been subjected to some form of crime in their lifetime (WHO, 2023b,a) However, victims are often discouraged from seeking support due to self-disclosure (Quadara, 2008), acquainted perpetrators (Millar et al., 2002), fear of revenge (Planty et al., 2013), and social stigma about victimization and support-seeking (Kilpatrick and Acierno, 2003).

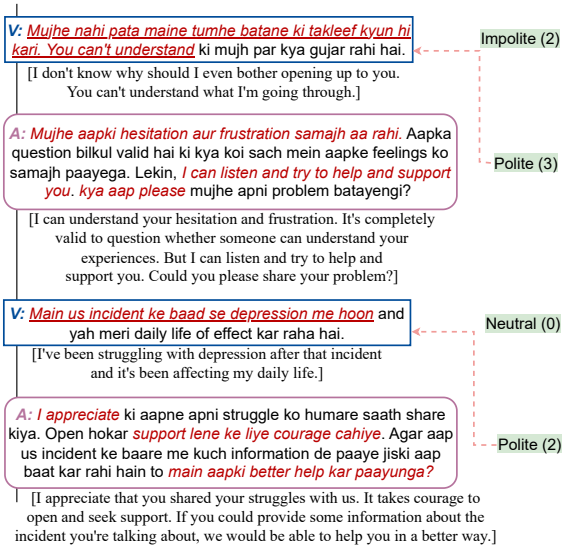

Figure 1: A dialogue snippet showcasing politeness and its corresponding intensity value (highlighted in green) and politeness causal span (underlined). ***V*** and ***A*** denote victim's and agent's utterances, respectively.

To tackle this overwhelming problem, recently, there has been an emerging interest in building conversational agents (or chatbots) that can extend support to the victims (Ahn et al., 2020b). In order to develop effective and interactive counseling systems that can be easily integrated with human experts, it is crucial to comprehend the victim's social and cognitive behavioral aspects. Politeness exhibits socially and cognitively-desirable behavior. The incorporation of politeness in interactions makes the victims feel respected, validated, and more willing to engage in the counseling process, ultimately promoting their healing and well-being (Kim et al., 2022; Mishra et al., 2023b). Also, in order to identify how to best avoid negative thought

*The authors are jointly first authors.
[1]Code and dataset are available at PAANTH.

Findings of the Association for Computational Linguistics: NAACL 2024, pages 4678–4696
June 16-21, 2024

patterns and maladaptive behavior, it is crucial to recognize not just the polite/impolite behavior but also the factor(s) or trigger(s) that contribute to that behavior and the intensity of that behavior during interaction. This allows for a more comprehensive understanding of the individual's emotional and mental state as well as communication dynamics, facilitating a tailored approach to counseling and promoting positive therapeutic outcomes.

For conversational systems to emulate intelligent behavior, they must not only be potent enough to identify politeness but also possess the ability to comprehend it in its entirety. With this objective in mind, we progress beyond the scope of politeness identification and introduce a novel task- ***Politeness Cause Elicitation and Intensity Tagging*** (**PCEIT**) in conversations. PCEIT aims to analyze the politeness, the underlying politeness cause(s)/factor(s) that lead individuals to employ polite or impolite behaviors and the extent to which polite or impolite language or behaviors are used during conversation. To illustrate, we depict a dialogue snippet between the victim and the agent in Figure 1. In the second turn of the dialogue snippet, the agent discerns the victim's hesitation and frustration. This discernment suggests that the victim harbors a concern that the agent may not fully comprehend the intricacies of their situation. Consequently, the agent behaves politely by acknowledging and validating the victim's feelings and extending support. The extraction of politeness causal spans enables the agent to respond with empathy, understanding, or reassurance as and when needed while ensuring an appropriate level of politeness. This eventually helps in establishing trust, validating the victim's emotional and mental state, and encouraging open communication, thereby enabling a more productive relationship.

Studies have shown that code-mixing enables more natural and engaging conversations among multilingual users (Bawa et al., 2020; Ahn et al., 2020a). Given the limited availability of code-mixed counseling conversational datasets, we present a novel and meticulously curated counseling conversational dataset in the code-mixed *Hinglish* language. We extend POEM (Priya et al., 2023a) - a counseling conversational dataset for mental health and legal counseling of crime victims by refurbishing its English text into code-mixed *Hinglish* embodiment. We name this dataset **HING-POEM**. POEM dataset lacks politeness cause and intensity information; hence, we annotate **HING-POEM** with politeness causal span and intensity information along with the politeness label. To address the task of PCEIT in conversations, we propose a **PAANTH**[2] (**P**oliteness C**A**use Elicit**A**ion and I**N**tensity **T**agging in **H**inglish) - a Contextual Enhanced Attentive Convolution Transformer (*CEACT*)-based framework. The system leverages the utterance-level politeness information for which the causes are to be extracted and intensity is to be predicted.

In summary, our contributions are *five-fold*: **(i)** We propose Politeness Cause Elicitation and Intensity Tagging (PCEIT) in conversations - a **novel task** that aims at analyzing the cause(s) that contribute to the use of polite/impolite behavior and the degree of polite/impolite behavior exhibited during a conversation; **(ii)** We extend an existing counseling conversational dataset, to curate **HING-POEM**, a **novel dataset** containing conversations between the victim and the counseling agent in code-mixed *Hinglish* language; **(iii)** We annotate **HING-POEM** with politeness label, politeness cause(s) and politeness intensity value at the utterance level; **(iv)** We develop a Contextual Enhanced Attentive Convolution Transformer (*CEACT*)-based framework for the PCEIT task in conversations; **(v)** We carry out extensive quantitative and qualitative analysis to prove the efficacy of the proposed approach.

Societal Implications and Reproducibility. The chatbots for mental health and legal support of the victims offer a potential solution by engaging effectively with victims and comprehending their needs for the overall development of society. Our present research focuses on the dialogue understanding module within mental health and legal conversational systems. The ongoing research in the mental health and legal domain for crime victims could leverage this work and enhance chatbots' ability to better comprehend counseling conversations and better emulate human-like behavior. The resources will be made available upon request to aid future research.

2 Related Work

In recent times, conversational system research for social good applications like healthcare (Pandey et al., 2022; Mishra et al., 2023c), education (Kasthuri and Balaji, 2021), charity donation

[2] **PAANTH** can be vaguely pronounced as Panth (Path) in Hindi.

(Samad et al., 2022), legal aid (Falduti and Tessaris, 2022), etc. have attracted significant attention from the natural language processing (NLP) community. Given the escalating demand to combat crimes against women and children to meet the Sustainable Development Goals (SDGs) 2030 (García-Moreno and Amin, 2016), a few research emphasize the need for initiating research on conversational systems for supporting crime victims (Ahn et al., 2020b; Socatiyanurak et al., 2021; Kim et al., 2022; Falduti and Tessaris, 2022). These systems are predominantly rule-based, which limits their scalability and generalizability.

In order to help victims feel ready to access professional support, a convincing approach is demanded (Maeng and Lee, 2022). The use of polite language displays a cordial and credible impression of the system, which helps in achieving positive outcomes during counseling (Lucas et al., 2014; Newbold et al., 2019; Mishra et al., 2023b; Priya et al., 2023b; Mishra et al., 2023c,a). In the past, a few studies developed computational approaches for identifying politeness in text (Danescu-Niculescu-Mizil et al., 2013; Aubakirova and Bansal, 2016; Chhaya et al., 2018; Madaan et al., 2020). Lately, computational methods for automatic detection of politeness in conversations have been proposed to enable the conversational system to effectively adapt to the ongoing conversation and generate responses according to users' situation (Kayaarma et al., 2019; Mishra et al., 2022; Khan et al., 2023). Likewise, (Priya et al., 2023a) introduced a politeness and emotion-annotated dialogue corpus and proposed a multi-task framework for detecting politeness and emotion simultaneously. All these existing politeness studies are in English.

The extensive utilization of social media has driven progress in studying code-mixed languages for a range of NLP tasks, including emotion detection (Vijay et al., 2018), sarcasm detection (Bedi et al., 2021), and sentiment analysis (Ghosh et al., 2023; Dowlagar and Mamidi, 2021), among others (Ramanarayanan and Suendermann-Oeft, 2017; Parekh et al., 2020; Singh et al., 2022a). There have been a handful of works that focus on politeness and its related cues, like emotion (Bothe and Wermter, 2022) in other languages (Kumar, 2014; Firdaus et al., 2020; Kumar, 2021; Li et al., 2020; Singh et al., 2022b). However, politeness research in code-mixed settings remains unexplored.

In a nutshell, our research pioneers the exploration of politeness cause elicitation and intensity tagging in code-mixed *Hinglish* conversations for mental health and legal counseling of crime victims.

3 Dataset

To promote the research and development of a code-mixed dialogue agent for a social good application containing code-mixed *Hinglish* (combining Hindi with English) conversations, we create **HING-POEM** dataset. This dataset is developed utilizing the existing dialogue dataset POEM (Priya et al., 2023a) by transforming its monolingual (English) utterances into code-mixed (Hinglish) manifestations. To the best of our knowledge, no large-scale code-mixed dataset is available to facilitate research in this direction.

3.1 HING-POEM Dataset Preparation

This section presents the process of creating code-mixed *Hinglish* dialogues. To reduce the human intervention, we prepare the entire dataset by prompting the large language model (LLM) in a few-shot manner, followed by manual intervention to ensure the quality of the generated dialogues. In particular, we first construct sample utterances manually, which are then utilized to prompt the LLM. These synthetic code-mixed dialogues are then manually verified by human experts to ensure good quality dialogues.

Sample Utterance Creation. We construct the sample utterances in *Hinglish* by translating the first six utterances (three Agent-Victim utterance pairs) of English dialogues in the POEM dataset to *Hinglish* following the *Matrix Language-Frame* model (Myers-Scotton, 1993). This theory allows for the insertion of grammatical constituents of an *embedded* language (here, English) into the utterances in *matrix* language (here, Hindi). The translation is done by three experienced human translators under the supervision of domain experts to ensure accuracy and appropriateness within the specified context. These translators are native Hindi speakers and equally fluent in English[3]. The translators possess Ph.D. degrees in Linguistics and relevant expertise in code-mixing. Before beginning the translation procedure, the guidelines for translation along with some sample *Hinglish* utterances translated from their English counter-

[3]The translators were paid according to institutional guidelines.

parts were explained to the translators. They are then instructed to recreate the same English utterances by switching between Hindi and English languages while adhering to the following guidelines: (i) Assume that the translators are interacting with a friend proficient in both Hindi and English; (ii) Use Roman script irrespective of whether the word being used belongs to English or Hindi; (iii) Do not attempt to convert the entire utterance into Hindi, instead switch to English whenever they feel it is appropriate, just as they would in their daily conversations; (iv) Translate adjectives and conjunctions into Hindi; (v) Avoid code-mixing named entities (names of persons, organizations, places, crimes, mental health issues, or legal terms) and noun phrases; (vi) Retain the placeholder words for the victim's personal information (<person_name>, <person_age>, <person_gender>, etc.) unchanged and not translate them.

Dialogue Creation via Prompting. The created sample *Hinglish* utterances for each dialogue are then utilized to prompt a multilingual LLM, BLOOMZ (Muennighoff et al., 2022) in a few-shot setting. In order to finalize the prompt, we experiment with six different prompts consisting of natural language instruction and the created sample utterances of a particular dialogue followed by the target utterance whose code-mixed version is to be generated. For each prompt, we generate 30 code-mixed dialogues by prompting BLOOMZ with Top-p sampling ($p = 0.75$) and temperature $\tau = 0.95$. The evaluation of the synthetic code-mixed dialogues and an example of the selected prompt are provided in the Dataset section in Appendix. We leverage the selected prompt along with sample utterances to prompt the BLOOMZ model to generate the code-mixed equivalent of all the dialogues in the POEM dataset. We obtain the Code-Mixing Index (CMI) (Gambäck and Das, 2016) of 0.82 for **HING-POEM**, which shows good quality code-mixing in the dataset.

3.2 Dataset Cleaning and Quality control

Once all the dialogues are converted to their *Hinglish* equivalent, manual verification is carried out for quality control. We then provide comprehensive guidelines to the evaluators and suitable examples for each possible case before beginning the manual verification. They are instructed to refer to the original English utterance and dialogue context while verifying the code-mixed counterpart to ensure a meaningful translation and preserve the context in the code-mixed equivalent. The entire evaluation process is done with two distinct groups of human evaluators (G1 and G2), each group consisting of three evaluators - one with a Ph.D. degree in Linguistics and two with a Master's degree in Computer Science. All the evaluators are native Hindi speakers with English as their education medium and are well-acquainted with the concept of code-mixing[4]. In the primary stage, each utterance in dialogues is rated for ***F***, ***A*** and ***C*** on the same scale of 1-5 by the evaluators in group G1. We obtain average ratings of 3.19, 3.04, and 3.27 for ***F***, ***A***, and ***C***, respectively, in this stage. Afterward, the utterances with ratings 1, 2, or 3 for either ***F***, ***A*** and ***C*** are filtered out for post-editing by referring to the source utterances by the same group of evaluators.

In the secondary stage, another group of evaluators (G2) are instructed to again rate the utterances for ***F***, ***A*** and ***C***. Besides, all evaluators are instructed to rate each dialogue on a scale of 1-5[5] for *Intelligibility* (***I***) to assess if the entire dialogue could be readily comprehended by a bilingual speaker proficient in both Hindi and English. Eventually, we achieve average ratings of 4.26, 4.73, 4.48, and 4.87 for ***F***, ***A***, ***C***, and ***I***, respectively, which indicates that the dataset is of standard quality. We also obtain the Code-Mixing Index (CMI) (Das and Gambäck, 2014) of 0.82, which further establishes the sufficiently good quality of the dataset in terms of the level of code-mixing. The CMI is calculated using the Equation 1.

$$CMI = \begin{cases} 1 - \frac{\max\{w_i\}}{n-u} & : n > u \\ 0 & : n = u \end{cases} \quad (1)$$

where, $\sum_{i=1}^{N}$ is the sum of all N languages present in the utterance of their respective number of words, $max\{w_i\}$ is the highest number of words present from any language (regardless of if more than one language has the same highest word count), n is the total number of tokens, and u is the number of tokens given other (language independent) tags.

Given the space constraints, comprehensive insights into the challenges during dataset preparation can be found in the Dataset section A.1 in the Appendix.

[4]The evaluators are different from those involved in dataset preparation and are paid according to institutional guidelines.

[5]***Intelligibility*** - 5: Very Good, 4: Good, 3: Average, 2: Poor, 1: Very Poor

3.3 Dataset Annotation

The annotations are performed by three annotators, two with a Ph.D. degree in Linguistics and one with a Master's degree[6]. All the annotators are proficient in both English and Hindi, sufficiently acquainted with labeling tasks, and well-versed with the concepts of code-mixing and politeness. They are briefed about the annotation guidelines and proper examples for each of annotation task. The utterances in **HINGLISH-POEM** are annotated with three distinct aspects, *viz.* politeness, politeness cause(s) and politeness intensity value. Due to space constraints, the complete dataset statistics and the politeness label distribution of the proposed **HINGLISH-POEM** dataset are provided in section A.1 of the appendix.

Politeness and Politeness Intensity Annotation. The utterances in **HINGLISH-POEM** dataset are annotated with politeness label and corresponding politeness intensity value in two steps. In the first step, we randomly sample 1,250 dialogues consisting of 30,450 utterances (avg. dialogue length 24.36) from the dataset and manually annotate each utterance of this subset with one of the three politeness labels, *viz. polite, neutral*, and *impolite*. Each politeness label is accompanied by one of the three ordinal intensity values (1,2 or 3), with 1 indicating the lowest intensity and 3 indicating the highest. The *neutral* label has an intensity value of 0. We then separately train two different pretrained XLM-Roberta (XLM-R) models (Conneau et al., 2020), one for politeness classification and the other for politeness intensity prediction, on this annotated data using the Masked Language Modelling (MLM) objective (we refer this model as **HINGPOEM**-XLM-R).

This model is further fine-tuned on the annotated dataset for politeness and politeness intensity prediction tasks. The results confirm the efficacy of the **HINGPOEM**-XLM-R model. We achieve accuracies of 73.28% and 68.19% using the fine-tuned XLM-R model and 78.34% and 71.02% using the fine-tuned **HINGPOEM**-XLM-R model (trained specifically on the code-mix corpus) for politeness classification and politeness intensity prediction, respectively. For MLM training, we train the models for 8 epochs with a learning rate of $2e^{-5}$, weight decay of 0.01, and a mask probability of 0.15. These models are further fine-tuned on the annotated dataset using the MLM training objective[7].

In the second step, we predict the politeness label and the corresponding intensity value of the utterances by passing the utterances in the remaining dialogues through their respective fine-tuned classifiers. The predicted labels are then cross-verified for their correctness by the same three annotators in order to create a gold-standard dataset. We observe a reliable multi-rater Kappa agreement ratio (McHugh, 2012) of 79.6% and 72.3% in the first step and 82.7% and 78.7% in the second step for politeness and politeness intensity annotations, respectively.

Politeness Cause Annotation. The utterances are marked manually for the causal span of a politeness label. We mark at most 4 causal spans for each utterance as we observe most of the utterances have a single cause and few of them have two or more causes. For an utterance u_t, the causal spans are marked from $c+1$ utterances, where c denotes the number of context utterances of u_t and $u_{c+1} = u_t$. We quantify the inter-rater agreement using the macro-F1 metric based on earlier work on span extraction (Poria et al., 2021), and we obtain an F1-score of 0.78, indicating that the annotations are of good quality.

3.4 Dataset Statistics

The dataset statistics of **HINGLISH-POEM** are shown in Table 1. The politeness label distribution in **HINGLISH-POEM** is depicted in Figure 3 in the Appendix.

4 Methodology

In this section, we outline the problem statement and subsequently delve into a comprehensive discussion of the proposed methodology.

Problem Formulation. Given a dialogue $D = \{u_1, u_2, ..., u_N\}$ consisting of a sequence of N utterances, where $u_t = \{w_{i1}, w_{i2}, ..., w_{iM}\}$ (M representing the word count in utterance u_t). Let $P = \{p_1, p_2, ..., p_N\}$, denote the utterance-level politeness in the dialogue D. For a target utterance u_t, the PCEIT task objective is to detect the politeness label, extract all possible causal spans, and identify the politeness intensity for the given politeness p_t.

Proposed Approach. In this section, we outline the various elements comprising our proposed approach for politeness cause elicitation and intensity

[6] Annotators are paid as per institute norms.

[7] We obtain an accuracy of 78.34% and 71.02% for politeness classification and politeness intensity prediction, respectively.

tagging (PCEIT) within code-mixed *Hinglish* conversations. The architecture of our approach is presented in Figure 2.

Contextual Character Embedding. Initially, we employ the SentencePiece tokenizer (Kudo and Richardson, 2018) to tokenize utterances. Given the complexities of code-mixed data like *Hinglish* where out-of-vocabulary (OOV) words are common, the standard pre-trained embeddings might not perform well (Pratapa et al., 2018). Hence, we adopt a hybrid strategy to generate embeddings, which combines character embeddings and context-dependent word embeddings. For character-level features in code-mixed utterances, we utilize a convolutional neural network (CNN) followed by max pooling to capture effective text representations and local dependencies at both word and sub-word levels (Chiu and Nichols, 2016). For context-dependent word embeddings, we employ a fusion of ELMo (Peters et al., 2018) and Tf-idf (ram, 2003) embeddings.

Contextual Enhanced Attentive Convolution Transformer (CEACT). We introduce *CEACT* to enhance the integration of context within input phrases, and *EnTrans* - an enhanced attention mechanism that replaces the conventional self-attention in the transformer encoder (Vaswani et al., 2017). The process involves two primary steps: Token Contribution Computation and Token Pruning. The forward propagation mechanism of *EnTrans* is shown in Figure 2. For clarity, we explain the operations using a single-head *EnTrans*. Importantly, we compute the attention map before token contribution to guide our approach of incorporating pre-pruning of Key and Value in the proposed *EnTrans*. To assess the combined impact of tokens arranged by columns or rows, we exploit the distributive nature of the vector inner product, thereby minimizing computational complexity effectively. Consider q_i and k_j as tokens in Query ($Q \in \mathbf{R}^{n \times x}$) and Key ($K \in \mathbf{R}^{m \times x}$), respectively, where *n* and *m* denote the dimensions of the query and key vectors. The recalibrated scores for row and column vectors denoted by Sco_r and Sco_c, respectively, are outlined as:

$$Sco_r = \sum_{i=1}^{n}\sum_{j=1}^{m} q_i k_{rj}^T (\sum_{i=1}^{n} q_i)(\sum_{j=1}^{n} k_{rj}^T), \quad r \in 1....n \quad (2)$$

$$Sco_c = \sum_{i=1}^{n}\sum_{j=1}^{m} q_i k_{jc}^T = (\sum_{i=1}^{n} q_i)(\sum_{j=1}^{n} k_{jc}^T), \quad c \in 1....m \quad (3)$$

where *r* and *c* symbolize the tokens in the query and key vectors, respectively, with *T* representing matrix transpose operations.

Metrics	Train	Validation	Test
# of Dialogues	2,859	1,080	1,061
# of Utterances	77,806	25,775	25,744
Avg. Utterances per Dialogue	27.21	23.87	24.26

Table 1: **HING-POEM** dataset statistics.

We further employ token-wise L2 normalization for both Query and Key, allowing us to assess the relevance of grouped tokens. The attention map's element values are confined to (-1, 1) due to normalization of token vectors in *Q* or *K*, mitigating the adverse impact of excessively dominant token vectors before the *Softmax* activation.

Token pruning involves computing contribution scores, denoted as $Sco_r \in \mathbf{R}^n$ and $Sco_c \in \mathbf{R}^m$, which ranks rows and columns based on their contribution levels. Subsequently, the rows and columns with the highest scores are selected, while the remaining ones are discarded. In our experimental setup, the number of selected rows or columns, represented as N_h (a hyper-parameter), is established as the square root of *n* as $Ind_r = argmax Sco_r[: N_h]$, $Ind_c = argmax Sco_c[: N_h]$. The reconfiguration of *K* and *V* is determined by $K = K^{[Ind_r, Ind_c]}$ and $V = V^{[Ind_r, Ind_c]}$. The process of selecting rows or columns is facilitated by employing the contribution scores along with *argmax* and $[: N_h]$ to identify the indices ranking at the top.

To effectively boost attention, we leverage a gated linear unit (GLU) in conjunction with a convolutional layer for input representation following (Wu et al., 2020). To optimize computational efficiency, we replace the standard convolution with a lighter version (Wu et al., 2019) that incorporates linear layers and depth-wise convolution. The convolutional output is then linearly combined with the output from *EnTrans*, and self-attention is subsequently applied to this combined representation. This integration of sources allows the model to capture intricate relationships and patterns in the data more effectively.

Auto-encoder. To better grasp emotional nuances within the input text, we use Context-Free-Grammar-Noun-Adjective-Pairs (Context Free ANP) to extract adjective-noun pairs from the utterances. This approach enables the model to identify textual concepts effectively. The ANP features extracted through Context-Free ANP are then input into an auto-encoder to generate a latent representa-

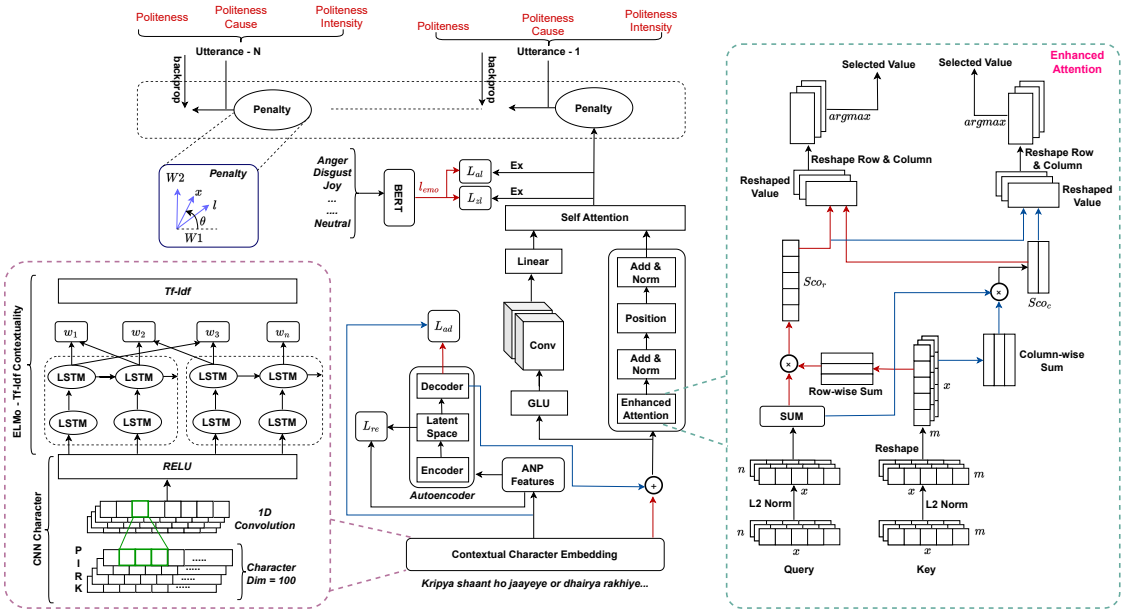

Figure 2: Architectural diagram of the proposed **PAANTH** framework.

tion. To integrate textual and class semantic knowledge into the ANP representation, an adversarial loss (Zhu et al., 2018) is employed, described in the "Training and Inference" section. This adversarial loss aims to disentangle syntax (captured by ANP) from semantics (captured by contextual character embedding), which could enhance interpretability and control over the learned representations.

Penalty. We introduce a penalty value into the system to enhance token prediction. This is intended to improve the model's ability to grasp the relationship between various labels and the input utterance. Incorporating this penalty into the loss function is driven by the complexity of defining a clear decision boundary for token markers in tasks involving information extraction. The uncertainty surrounding this boundary can make it challenging for a standard softmax/sigmoid classifier to precisely distinguish between different classes, which might result in misclassification of certain instances. The original equations representing softmax and sigmoid are as follows:

$$\mathcal{L}_{softmax} = -\frac{1}{b_s}\sum_{i=1}^{b_s}\log\frac{\exp^{\mathcal{W}l_i+b_i}}{\sum_{j=1}^{N}\exp^{\mathcal{W}l_j+b_j}} \quad (4)$$

$$\mathcal{L}_{sigmoid} = -\frac{1}{b_s}\sum_{i=1}^{b_s}\frac{1}{\exp^{\mathcal{W}l_i+b_i}} \quad (5)$$

Here, $l_i \in \mathbf{R}^d$ denotes the feature of the i^{th} sample, while b_s indicates the batch size. Moreover, b_i and b_j correspond to the biases, and $\mathcal{W} \in \mathbf{R}^{d\times N}$ stands for the weight matrix.

To tackle the challenge of establishing a decision boundary for token markers in information extraction tasks, the Insightface loss technique (Deng et al., 2019) offers a solution by normalizing the feature l_i and weight matrix $\mathcal{W}$. It evaluates the similarity of features based on the angle difference between them. To expedite feature convergence, a penalty value v is introduced to the angle θ in the loss function. This adjustment applies to both softmax and sigmoid and is expressed in the following manner:

$$\mathcal{L}_{softmax} = -\frac{1}{b_s}\sum_{i=1}^{b_s}\log\frac{\exp^{a(cos(\theta+v))}}{\exp^{a(cos(\theta+v))}+\sum_{j=1}^{N}\exp^{a(cos(\theta))}} \quad (6)$$

$$\mathcal{L}_{sigmoid} = -\frac{1}{b_s}\sum_{i=1}^{b_s}\frac{1}{\exp^{a(cos(\theta+v))}+\exp^{a(cos(\theta))}} \quad (7)$$

In the context provided, θ denotes the angle between the weight matrix $\mathcal{W}$ and the feature l_i, while a represents the amplifier function. The equation $\exp^{a(cos(\theta+v))}$ is employed to compute the similarity score for the positive sample, and $\exp^{a(cos(\theta))}$ is used for the negative samples' similarity score. The inclusion of the penalty value v introduces a margin to the classification boundary, enhancing the feature's convergence rate.

Training and inference. We outline the process of training our model and explain how to make predictions for politeness cause and intensity. Our model is trained in an end-to-end fashion using four different loss functions.

Models	PI Task		PIT Task		PCE Task				
	F1 (%)	ACC. (%)	F1 (%)	ACC. (%)	FM	PM	HD	JF	ROS
BiRNN-Attn (Liu and Lane, 2016)	66.32	67.59	61.43	63.32	25.86	29.32	0.49	0.66	0.72
CNN-GRU (Zhang et al., 2018)	67.34	69.43	61.19	63.93	26.77	30.65	0.47	0.65	0.74
BERT (Liu et al., 2019)	70.54	72.53	64.63	66.18	32.66	34.51	0.56	0.68	0.76
SpanBERT (Joshi et al., 2020)	71.42	73.65	66.75	68.17	34.65	36.55	0.59	0.72	0.78
BiRNN-HateXplain (Mathew et al., 2021)	68.55	69.47	65.17	66.43	29.77	31.43	0.51	0.70	0.73
BERT-HateXplain (Mathew et al., 2021)	72.63	74.32	68.11	69.54	32.65	36.32	0.62	0.76	0.78
CMSEKI (Ghosh et al., 2022a)	74.33	76.74	69.63	70.93	35.62	37.24	0.60	0.74	0.80
PAANTH *(Proposed)*	**77.12**	**78.77**	**71.93**	**73.31**	**37.59**	**39.41**	**0.67**	**0.81**	**0.83**

Table 2: Results from the **PAANTH** model and the various baselines. Here, the bolded values indicate maximum scores. Here, PI: Politeness Identification, PCE: Politeness Cause Elicitation, PIT Politeness Intensity Tagging. The results are statistically significant. The statistical significance test, Welch's t-test (Welch, 1947) is conducted at 5% (0.05) significance level.

Reconstruction Loss. The aim is to bring the structures of label features and adjective-noun pair features into alignment within the learned latent space using an autoencoder. This autoencoder is responsible for reconstructing adjective-noun pair features and generating latent features while retaining politeness-related information. The optimization of autoencoder parameters involves minimizing a loss function that quantifies the similarity between the autoencoder's input and output. This loss function is defined as $\mathcal{L}_{re} = ||\hat{A}((t)) - A((t))||_2^2$, where $\hat{A}$ and A represents the input and output embedding features of the autoencoder, respectively.

Alignment loss. Our objective is to synchronize the latent space and label semantic spaces within an autoencoder, ensuring that the generated label representations are closely associated with latent polite concepts. This goal is pursued by optimizing the loss function $\mathcal{L}_{al} = ||h(x) - \phi(l_{po})||_2^2$, where the function $h(x)$ represents the latent space embedding produced by the autoencoder, and l_{po} signifies the politeness embedding. The comprehensive objective function is achieved by merging the alignment loss and the reconstruction loss: $\mathcal{L}_{re} + \mathcal{L}_{al}$.

Zero-shot loss. To assess the effect of emotion on the proposed PCEIT task, we feed the emotional information to the model in a zero-shot fashion. The objective of this loss function is to minimize the difference between the feature of text represented by $\theta(x)$, and the semantic feature of the emotion label[8] computed using pre-trained BERT (Devlin et al., 2019), represented by $\phi(l_{emo})$, through optimization. This loss is defined as $\mathcal{L}_{zl} = ||\theta(SA(x) - \phi(l_{emo})||_2^2$.

Adversarial loss. Our goal is to reduce the gap between the discriminative capability of the text ($\theta(x)$ representing $SA(t)$) and the intricate politeness structural information encapsulated in the feature $\phi(l_{po})$. This is accomplished by employing an adversarial constraint designed to deceive the discriminator network $\mathcal{D}$, thereby making the output features of $A(\theta(x))$ as similar to the ANP features as feasible. It is defined as $\mathcal{L}_{ad} = \mathcal{E}_y(\log \mathcal{D}(h(y)) - \mathcal{E}_y(\log \mathcal{D}(\theta(y))$. In this context, $\theta(y)$ represents the feature of the text, while $h(y)$ signifies the latent feature space.

Joint Loss. We train our model by incorporating a blend of the diverse loss functions as follows: $\mathcal{L}_{joint} = \mathcal{L}_{ad} + \mathcal{L}_{zl} + (\mathcal{L}_{re} + \mathcal{L}_{al})$. We sum up these loss functions by assigning equal weights to each for effective joint optimization during training. The equal weights of the different loss components ensure that the proposed model treats all tasks equally. All tasks are intricately related and hold significant relevance for a dialogue system.

Experimental Setup. The details about baselines, implementation process, and evaluation metrics are given in Section A.2 of the Appendix.

5 Results and Analysis

Table 2 displays the outcomes of the proposed **PAANTH** framework in comparison to various baselines using the newly introduced **HING-POEM** dataset. The results in the table reveal that CMSEKI, which taps into common-sense knowledge from external sources to comprehend input data, stands out as the top-performing baseline. Nevertheless, the proposed **PAANTH** model consistently exhibits even better performance than CMSEKI across all evaluation metrics. Notably, **PAANTH** achieves a substantial improvement of 2.79% in F1 for the *PI* task, 3 points in ROS for the *PCE* task and 2.3% in F1 for *PIT* task. Among the baselines that do not rely on external information,

[8]Ekman's (Ekman, 1992) basic emotion classes (*Anger*, *Disgust*, *Sad*, *Joy*, *Surprise*, *Fear*, *Fear*). Additionally, we consider the *Neutral* class to accommodate instances that do not fall in the scope of Ekman's categorization.

Model	Text	Label
1. Human Annotator	*Kripya shaant ho jaiye aur dhairya rakhiye* , *hum yahan aapki har tarah se help karne ke liye hai.* *Kya aap* bata sakte hain, ki hum kisse interact kar rahe hai?	
	(Please calm down and have patience. We are here to help you in every possible way. Can you tell with whom we are interacting?)	**Polite**
BERT-HateXplain	*Kripya shaant* ho jaiye aur dhairya rakhiye, hum yahan aapki har tarah se help karne ke liye hai. Kya *aap* bata sakte hain, ki hum kisse interact kar rahe hai?	**Polite**
SpanBERT	*Kripya shaant* ho jaiye aur dhairya rakhiye, hum yahan aapki har tarah se help karne ke liye hai. Kya aap bata sakte hain , ki hum kisse interact kar rahe hai?	**Impolite**
CMSEKI	Kripya shaant ho jaiye aur dhairya rakhiye, hum yahan aapki har tarah se help karne ke liye hai. Kya *aap bata sakte hain* , ki hum kisse interact kar rahe hai?	**Polite**
PAANTH (Proposed)	*Kripya shaant ho jaiye aur dhairya rakhiye* , *hum yahan aapki har tarah se help karne ke liye hai* . *Kya aap* bata sakte hain , ki hum kisse interact kar rahe hai?	**Polite**
2. Human Annotator .	online complaint mode chunane ke liye dhanyavad. please www.cybercrime.gov.in par log on kare and diye gaye instructions ke sath aage badhe.	
	(Thanks for choosing the online complaint mode. Please log on to www.cybercrime.gov.in and proceed with the given instruction.)	**Polite**
BERT-HateXplain	online complaint mode chunane ke liye dhanyavad. please www.cybercrime.gov.in par log on kare and diye gaye instructions ke sath aage badhe.	**Polite**
SpanBERT	online complaint mode chunane ke liye dhanyavad. please www.cybercrime.gov.in par log on kare and diye gaye instructions ke sath aage badhe.	**Polite**
CMSEKI	online complaint mode chunane ke liye dhanyavad. please www.cybercrime.gov.in par log on kare and diye gaye instructions ke sath aage badhe.	**Polite**
PAANTH (Proposed)	online complaint mode chunane ke liye dhanyavad. please www.cybercrime.gov.in par log on kare and diye gaye instructions ke sath aage badhe.	**Polite**

Table 3: Sample predictions from the various systems.

SpanBERT emerges as the most effective, surpassing other comparable systems. Yet, when compared to **PAANTH**, SpanBERT falls short by 5.7% in F1 for the *PI* task, 5 points in ROS for the *PCE* task and 5.18% in F1 for *PIT* task.The relatively lower performance of BERT, SpanBERT, and BERT-HateXplain underlines the challenges that powerful language models face in comprehending intricate tasks like politeness cause elicitation and intensity tagging, particularly in scenarios involving mental health and legal counseling, where training data is limited.

Qualitative Analysis. We conduct a comprehensive analysis of the predictions made by different systems. Consider the examples presented in Table 3. In the top row, we can see the tokens (referred to as 'causes') identified by human annotators as the representations of the causes for the utterance labeled as *polite*. The subsequent four rows present the tokens extracted by various models. It is evident that the proposed **PAANTH** model accurately identifies the examples as instances of politeness and provides high-quality causal spans. While the SpanBERT model correctly captures a partial causal span, it misclassifies the label as *impolite*. We also delve into cases where the proposed model exhibits lower performance.

Ablation Study. As shown in Table 4, we perform an ablation study on the **HING-POEM** dataset to analyze the performance of the different components in our proposed framework. The values of the metrics for the PCEIT task are shown to drop when either the penalty factor (**PAANTH**$_{\text{-Penalty}}$), enhanced attention (**PAANTH**$_{\text{-EA}}$), adjective-noun pair (**PAANTH**$_{\text{-ANP}}$) or contextual embedding (**PAANTH**$_{\text{-CCE}}$) is omitted. The performance drop is more profound when either two, three or all the components are removed. This affirms that the involvement of the penalty factor, enhanced attention, adjective-noun pair and contextual embedding of the utterances significantly contributes to the effectiveness of the proposed PCEIT task.

Setup	$F1^{PI}$ (%)	$F1^{PIT}$ (%)	JF^{PCE} (%)	ROS^{PCE} (%)
[**PAANTH**]$_{\text{-Penality}}$	75.14(-1.98)	70.50(-1.43)	0.79(-0.020)	0.82(-0.15)
[**PAANTH**]$_{\text{-EA}}$	74.23(-2.89)	69.62(-2.31)	0.79(-0.025)	0.81(-0.019)
[**PAANTH**]$_{\text{-EA+ Penality}}$	72.39(-4.73)	67.60(-4.33)	0.78(-0.037)	0.79(-0.040)
[**PAANTH**]$_{\text{-ANP}}$	75.69(-1.43)	70.88(-1.05)	0.80(-0.009)	0.82(-0.012)
[**PAANTH**]$_{\text{-EA+ANP+Penality}}$	71.02(-6.10)	66.54(-5.39)	0.76(-0.050)	0.77(-0.057)
[**PAANTH**]$_{\text{-CEm}}$	74.69(-2.43	69.82(-2.11)	0.79(-0.016)	0.80(-0.023))
[**PAANTH**]$_{\text{-CEm+EA+ANP+Penality}}$	68.91(-8.21)	63.95(-7.98)	0.73(-0.07)	0.75(-0.07)
PAANTH *(Proposed)*	**77.12**	**71.93**	**0.81**	**0.83**

Table 4: Results of ablation experiments. The % fall in scores are shown in brackets. EA: Enhanced Attention, ANP: Adjective-Noun Pair, CCE: Contextual Character Embedding

Additional Analysis. Due to space limitation, we have included more analyses such as (1) Analysis of Embeddings; (2) Comparison with ChatGPT; (3) Varying Context Length; (4) Emotion analysis for Politeness tasks; (5) Loss Function Analysis; (6) Analysis of Task Setting; and (7) Error Analysis under the Additional Analysis section A.3 of the Appendix.

6 Conclusion

This study introduces a novel task titled "*Politeness Cause Elicitation and Intensity Tagging*" (PCEIT) in *Hinglish* conversations. To address this, we present the **HING-POEM** dataset, a novel code-mixed conversational data for mental health and legal counseling involving crime victims. Leveraging the capabilities of the BLOOMZ, we generate code-mixed dialogues with in-context few-shot examples. We annotate the dataset at the utterance level to include politeness, the causes of politeness, and the intensity of politeness. To identify politeness along with its underlying causal span(s) and intensity, we design **PAANTH**, a framework built upon a Contextual Enhanced Attentive Convolution Transformer. Notably, the **PAANTH** is the first task-specific system tailored to address the PCEIT task within conversational settings. To underscore the effectiveness of our approach, we benchmark it against various state-of-the-art baselines.

Limitations

Our current study focuses on identifying politeness, politeness cause elicitation, and intensity tagging in code-mixed *Hinglish* conversations focused on mental health and legal counseling of crime victims. The primary limitation lies in the scarcity of labeled data for modeling politeness cause and intensity in conversations. Nevertheless, we opted for the meticulous process of annotating data with the assistance of human annotators, recognizing its reliability. As a result of the absence of a dataset specifically dedicated to politeness cause and intensity annotation, we conducted experiments exclusively with the newly constructed **HING-POEM** dataset. Nevertheless, the applicability of our proposed conversational code-mixed Hinglish dataset, **HING-POEM** is not limited to the proposed task. It can be used for several other downstream tasks like code-mixed sentiment analysis, emotion recognition, emotion cause extraction, conversational agents capable of conversing in the Hinglish language, and dialogue summarization, to mention a few. However, we would like to highlight that in this work, we did not assess the extent to which our semi-automatically synthetically generated code-mixed data enhances the proficiency of language models in processing code-mixed text for downstream NLP tasks. While earlier studies have demonstrated that refining models with synthetic code-mixed data results in fewer performance improvements compared to naturally existing code-mixed data (Santy et al., 2021), we anticipate that this performance gap will lessen as the quality of data generation improves with more powerful future multilingual LLMs.

In the future, we plan to expand our experiments to encompass more task-oriented datasets. Additionally, due to constraints in computational resources within academic settings, we could not perform experiments utilizing advanced language models such as GPT3 (Brown et al., 2020), PaLM (Chowdhery et al., 2022), LLaMa (Touvron et al., 2023), and others.

Ethics Statement

This study has been evaluated and approved by our Institutional Review Board (IRB). In this study, we utilized the POEM, a collection of dialogues centered on mental health and legal counseling of crime victims. We obtained permission to use this dataset for our research, adhering to the copyright guidelines provided by the copyright holder. Considering the severity of the research area, we make sure that at each step, we maintain the privacy of the personal data of victims. The dataset proposed in the study will be made available for the sole purpose of facilitating research and educational intentions upon acceptance, accompanied by appropriate copyright provisions.

Acknowledgements

Priyanshu Priya acknowledges the financial support provided by the Department of Science and Technology, Ministry of Science and Technology, Government of India, through the Innovation in Science Pursuit for Inspired Research (INSPIRE) Fellowship.

References

2003. *Using tf-idf to determine word relevance in document queries*. Citeseer.

Emily Ahn, Cecilia Jimenez, Yulia Tsvetkov, and Alan W Black. 2020a. What code-switching strategies are effective in dialog systems? In *Proceedings of the Society for Computation in Linguistics 2020*, pages 254–264.

Yuna Ahn, Yilin Zhang, Yujin Park, and Joonhwan Lee. 2020b. A chatbot solution to chat app problems: Envisioning a chatbot counseling system for teenage victims of online sexual exploitation. In *Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems*, pages 1–7.

Malika Aubakirova and Mohit Bansal. 2016. Interpreting neural networks to improve politeness comprehension. *arXiv preprint arXiv:1610.02683*.

Anshul Bawa, Pranav Khadpe, Pratik Joshi, Kalika Bali, and Monojit Choudhury. 2020. Do multilingual users prefer chat-bots that code-mix? let's nudge and find out! *Proceedings of the ACM on Human-Computer Interaction*, 4(CSCW1):1–23.

Manjot Bedi, Shivani Kumar, Md Shad Akhtar, and Tanmoy Chakraborty. 2021. Multi-modal sarcasm detection and humor classification in code-mixed conversations. *IEEE Transactions on Affective Computing*.

Chandrakant Bothe and Stefan Wermter. 2022. Conversational analysis of daily dialog data using polite emotional dialogue acts. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 2395–2400.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda

Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Niyati Chhaya, Kushal Chawla, Tanya Goyal, Projjal Chanda, and Jaya Singh. 2018. Frustrated, polite, or formal: Quantifying feelings and tone in email. In *Proceedings of the Second Workshop on Computational Modeling of People's Opinions, Personality, and Emotions in Social Media*, pages 76–86.

Jason PC Chiu and Eric Nichols. 2016. Named entity recognition with bidirectional lstm-cnns. *Transactions of the Association for Computational Linguistics*, 4:357–370.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. 2022. Palm: Scaling language modeling with pathways. *arXiv preprint arXiv:2204.02311*.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Édouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *ACL*, pages 8440–8451.

Cristian Danescu-Niculescu-Mizil, Moritz Sudhof, Dan Jurafsky, Jure Leskovec, and Christopher Potts. 2013. A computational approach to politeness with application to social factors. *arXiv preprint arXiv:1306.6078*.

Amitava Das and Björn Gambäck. 2014. Identifying languages at the word level in code-mixed indian social media text. In *Proceedings of the 11th International Conference on Natural Language Processing*, pages 378–387.

Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. 2019. Arcface: Additive angular margin loss for deep face recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4690–4699.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT*.

Suman Dowlagar and Radhika Mamidi. 2021. Cmsaone@ dravidian-codemix-fire2020: A meta embedding and transformer model for code-mixed sentiment analysis on social media text. *arXiv preprint arXiv:2101.09004*.

Paul Ekman. 1992. An argument for basic emotions. *Cognition and Emotion*, pages 169–200.

Mattia Falduti and Sergio Tessaris. 2022. On the use of chatbots to report non-consensual intimate images abuses: The legal expert perspective. In *Proceedings of the 2022 ACM Conference on Information Technology for Social Good*, pages 96–102.

Mauajama Firdaus, Asif Ekbal, and Pushpak Bhattacharyya. 2020. Incorporating politeness across languages in customer care responses: Towards building a multi-lingual empathetic dialogue agent. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4172–4182.

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.

Claudia García-Moreno and Avni Amin. 2016. The sustainable development goals, violence and women's and children's health. *Bulletin of the World Health Organization*, 94(5):396.

Soumitra Ghosh, Asif Ekbal, and Pushpak Bhattacharyya. 2022a. A multitask framework to detect depression, sentiment and multi-label emotion from suicide notes. *Cogn. Comput.*, 14(1):110–129.

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.

Soumitra Ghosh, Swarup Roy, Asif Ekbal, and Pushpak Bhattacharyya. 2022b. Cares: Cause recognition for emotion in suicide notes. In *European Conference on Information Retrieval*, pages 128–136. Springer.

Mandar Joshi, Danqi Chen, Yinhan Liu, Daniel S Weld, Luke Zettlemoyer, and Omer Levy. 2020. Spanbert: Improving pre-training by representing and predicting spans. *Transactions of the association for computational linguistics*, 8:64–77.

E Kasthuri and S Balaji. 2021. A chatbot for changing lifestyle in education. In *2021 Third International Conference on Intelligent Communication Technologies and Virtual Mobile Networks (ICICV)*, pages 1317–1322. IEEE.

Selma Yilmazyildiz Kayaarma, Sherik Lehal, and Hichem Sahli. 2019. Politeness detection in speech for human-computer interaction. In *BNAIC/BENELEARN*.

John F Kelley. 1984. An iterative design methodology for user-friendly natural language office information applications. *ACM Transactions on Information Systems (TOIS)*, 2(1):26–41.

Shakir Khan, Mohd Fazil, Agbotiname Lucky Imoize, Bayan Ibrahimm Alabduallah, Bader M Albahlal, Saad Abdullah Alajlan, Abrar Almjally, and Tamanna Siddiqui. 2023. Transformer architecture-based transfer learning for politeness prediction in conversation. *Sustainability*, 15(14):10828.

Dean G Kilpatrick and Ron Acierno. 2003. Mental health needs of crime victims: Epidemiology and

outcomes. *Journal of Traumatic Stress: Official Publication of The International Society for Traumatic Stress Studies*, 16(2):119–132.

Hyeok Kim, Youjin Hwang, Jieun Lee, Youngjin Kwon, Yujin Park, and Joonhwan Lee. 2022. Personalization trade-offs in designing a dialogue-based information system for support-seeking of sexual violence survivors. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, pages 1–15.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Taku Kudo and John Richardson. 2018. Sentencepiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. *arXiv preprint arXiv:1808.06226*.

Ritesh Kumar. 2014. Developing politeness annotated corpus of hindi blogs. In *LREC*, pages 1275–1280.

Ritesh Kumar. 2021. Towards automatic identification of linguistic politeness in hindi texts. *arXiv preprint arXiv:2111.15268*.

Mingyang Li, Louis Hickman, Louis Tay, Lyle Ungar, and Sharath Chandra Guntuku. 2020. Studying politeness across cultures using english twitter and mandarin weibo. *Proceedings of the ACM on Human-Computer Interaction*, 4(CSCW2):1–15.

Bing Liu and Ian Lane. 2016. Attention-based recurrent neural network models for joint intent detection and slot filling. *arXiv preprint arXiv:1609.01454*.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Gale M Lucas, Jonathan Gratch, Aisha King, and Louis-Philippe Morency. 2014. It's only a computer: Virtual humans increase willingness to disclose. *Computers in Human Behavior*, 37:94–100.

Aman Madaan, Amrith Setlur, Tanmay Parekh, Barnabas Poczos, Graham Neubig, Yiming Yang, Ruslan Salakhutdinov, Alan W Black, and Shrimai Prabhumoye. 2020. Politeness transfer: A tag and generate approach. *arXiv preprint arXiv:2004.14257*.

Wookjae Maeng and Joonhwan Lee. 2022. Designing and evaluating a chatbot for survivors of image-based sexual abuse. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, pages 1–21.

Binny Mathew, Punyajoy Saha, Seid Muhie Yimam, Chris Biemann, Pawan Goyal, and Animesh Mukherjee. 2021. Hatexplain: A benchmark dataset for explainable hate speech detection. In *Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Thirty-Third Conference on Innovative Applications of Artificial Intelligence, IAAI 2021, The Eleventh Symposium on Educational Advances in Artificial Intelligence, EAAI 2021, Virtual Event, February 2-9, 2021*, pages 14867–14875. AAAI Press.

Mary L McHugh. 2012. Interrater reliability: the kappa statistic. *Biochemia medica*, 22(3):276–282.

Golden Millar, Lana Stermac, and Mary Addison. 2002. Immediate and delayed treatment seeking among adult sexual assault victims. *Women & health*, 35(1):53–64.

Kshitij Mishra, Mauajama Firdaus, and Asif Ekbal. 2022. Predicting politeness variations in goal-oriented conversations. *IEEE Transactions on Computational Social Systems*.

Kshitij Mishra, Priyanshu Priya, Manisha Burja, and Asif Ekbal. 2023a. e-therapist: I suggest you to cultivate a mindset of positivity and nurture uplifting thoughts. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 13952–13967.

Kshitij Mishra, Priyanshu Priya, and Asif Ekbal. 2023b. Help me heal: A reinforced polite and empathetic mental health and legal counseling dialogue system for crime victims. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 14408–14416.

Kshitij Mishra, Priyanshu Priya, and Asif Ekbal. 2023c. Pal to lend a helping hand: Towards building an emotion adaptive polite and empathetic counseling conversational agent. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12254–12271.

Niklas Muennighoff, Thomas Wang, Lintang Sutawika, Adam Roberts, Stella Biderman, Teven Le Scao, M Saiful Bari, Sheng Shen, Zheng-Xin Yong, Hailey Schoelkopf, et al. 2022. Crosslingual generalization through multitask finetuning. *arXiv preprint arXiv:2211.01786*.

Carol Myers-Scotton. 1993. Common and uncommon ground: Social and structural factors in codeswitching. *Language in society*, 22(4):475–503.

Joseph Newbold, Gavin Doherty, Sean Rintel, and Anja Thieme. 2019. Politeness strategies in the design of voice agents for mental health.

Sumit Pandey, Srishti Sharma, and Samar Wazir. 2022. Mental healthcare chatbot based on natural language processing and deep learning approaches: ted the therapist. *International Journal of Information Technology*, 14(7):3757–3766.

Tanmay Parekh, Emily Ahn, Yulia Tsvetkov, and Alan W Black. 2020. Understanding linguistic accommodation in code-switched human-machine dialogues. In *Proceedings of the 24th Conference on*

Computational Natural Language Learning, pages 565–577.

Matthew E Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. *arXiv preprint arXiv:1802.05365*.

Michael Planty, Lynn Langton, Christopher Krebs, Marcus Berzofsky, and Hope Smiley-McDonald. 2013. *Female victims of sexual violence, 1994-2010*. US Department of Justice, Office of Justice Programs, Bureau of Justice ....

Soujanya Poria, Navonil Majumder, Devamanyu Hazarika, Deepanway Ghosal, Rishabh Bhardwaj, Samson Yu Bai Jian, Pengfei Hong, Romila Ghosh, Abhinaba Roy, Niyati Chhaya, et al. 2021. Recognizing emotion cause in conversations. *Cognitive Computation*, 13(5):1317–1332.

Adithya Pratapa, Monojit Choudhury, and Sunayana Sitaram. 2018. Word embeddings for code-mixed language processing. In *Proceedings of the 2018 conference on empirical methods in natural language processing*, pages 3067–3072.

Priyanshu Priya, Mauajama Firdaus, and Asif Ekbal. 2023a. A multi-task learning framework for politeness and emotion detection in dialogues for mental health counselling and legal aid. *Expert Systems with Applications*, 224:120025.

Priyanshu Priya, Kshitij Mishra, Palak Totala, and Asif Ekbal. 2023b. Partner: A persuasive mental health and legal counselling dialogue system for women and children crime victims. In *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence, IJCAI-23*, pages 6183–6191.

Antonia Quadara. 2008. *Responding to young people disclosing sexual assault: A resource for schools*. Australian Institute of Family Studies.

Vikram Ramanarayanan and David Suendermann-Oeft. 2017. Jee haan, i'd like both, por favor: Elicitation of a code-switched corpus of hindi-english and spanish-english human-machine dialog. In *Interspeech*, pages 47–51.

Azlaan Mustafa Samad, Kshitij Mishra, Mauajama Firdaus, and Asif Ekbal. 2022. Empathetic persuasion: reinforcing empathy and persuasiveness in dialogue systems. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 844–856.

Sebastin Santy, Anirudh Srinivasan, and Monojit Choudhury. 2021. Bertologicomix: How does code-mixing interact with multilingual bert? In *Proceedings of the Second Workshop on Domain Adaptation for NLP*, pages 111–121.

Gopendra Vikram Singh, Mauajama Firdaus, Shruti Mishra, Asif Ekbal, et al. 2022a. Knowing what to say: Towards knowledge grounded code-mixed response generation for open-domain conversations. *Knowledge-Based Systems*, 249:108900.

Gopendra Vikram Singh, Priyanshu Priya, Mauajama Firdaus, Asif Ekbal, and Pushpak Bhattacharyya. 2022b. Emoinhindi: A multi-label emotion and intensity annotated dataset in hindi for emotion recognition in dialogues. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 5829–5837.

Vorada Socatiyanurak, Nittayapa Klangpornkun, Adirek Munthuli, Phongphan Phienphanich, Lalin Kovudhikulrungsri, Nantawat Saksakulkunakorn, Phonkanok Chairaungsri, and Charturong Tantibundhit. 2021. Law-u: Legal guidance through artificial intelligence chatbot for sexual violence victims and survivors. *IEEE Access*, 9:131440–131461.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 5998–6008.

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.

Bernard L Welch. 1947. The generalization ofstudent's' problem when several different population variances are involved. *Biometrika*, 34(1/2):28–35.

World Health Organization WHO. 2023a. Crime against Children. https://www.who.int/news-room/fact-sheets/detail/violence-against-children.

World Health Organization WHO. 2023b. Crime against Women. https://www.who.int/news-room/fact-sheets/detail/violence-against-women.

Felix Wu, Amauri Souza, Tianyi Zhang, Christopher Fifty, Tao Yu, and Kilian Weinberger. 2019. Simplifying graph convolutional networks. In *International conference on machine learning*, pages 6861–6871. PMLR.

Zhanghao Wu, Zhijian Liu, Ji Lin, Yujun Lin, and Song Han. 2020. Lite transformer with long-short range attention. *arXiv preprint arXiv:2004.11886*.

Ziqi Zhang, David Robinson, and Jonathan Tepper. 2018. Detecting hate speech on twitter using a

convolution-gru based deep neural network. In *European semantic web conference*, pages 745–760. Springer.

Yizhe Zhu, Mohamed Elhoseiny, Bingchen Liu, Xi Peng, and Ahmed Elgammal. 2018. A generative adversarial approach for zero-shot learning from noisy texts. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1004–1013.

A Appendix

A.1 Dataset

In this section, we provide the details pertaining to the dataset.

A.1.1 POEM Dataset Description

The POEM dataset consists of 5K dialogues between a dialogue agent and a crime victim. These dialogues are primarily concentrated to address the mental health and/or legal counseling needs of women and children who have faced violence in any form. The dialogues encompass 16 different categories of crimes, including conventional and cyber-crimes committed against women and children, namely domestic violence, rape, acid attacks, physical/cyber-stalking, workplace harassment, online harassment, impersonation, trolling, matrimonial fraud, financial fraud, child pornography, women/child trafficking, non-consensual sexting, doxing/outing, and exclusion.

The dialogues in this dataset are created using real-life stories of crimes against women and children crawled from news articles and related case studies. Further, during the POEM corpus creation, several authentic websites, *viz.* National Cybercrime Reporting Portal, National Commission for Women, etc. are referred to ensure the authenticity of mental health counseling and legal information. The dialogues are prepared in the Wizard-of-Oz fashion (Kelley, 1984) involving a pair of participants, where one participant plays the role of the agent/counselor, and the other acts as a victim who needs either mental health counseling, legal counseling, or both to recover from the victimization. The dataset creation is characterized by comprehensive domain expert supervision to ensure diverse, informative, engaging, and realistic conversations. The dataset is available in English. We reconstruct the dataset by converting the English utterances in dialogues into *Hinglish* code-mixed versions.

A.1.2 Prompt Evaluation

The synthetic code-mixed dialogues are evaluated by the same three human translators for (i). *Fluency* (***F***): Assess if the utterances are syntactically and grammatically correct; (ii). *Adequacy* (***A***): Assess if the utterances are semantically equivalent to the original English utterance; and (iii). *Colloquialism* (***C***): Assess if the code-mixed utterances

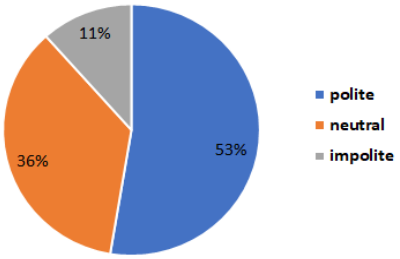

Figure 3: Politeness distribution in **HING-POEM** dataset.

are colloquial rather than forced, on a scale of 1-5[9]. The prompt which gives the best average scores of 3.42, 3.20, and 3.92 for ***F***, ***A***, and ***C***, respectively, is selected as the final prompt. An example prompt for the generation of code-mixed *Hinglish* utterance is shown in Figure 4.

The following is a conversation snippet between the Agent and the Victim of a crime. The utterances in English and their code-mixed Hinglish (a blend of Hindi and English languages) equivalent are given as context. You are required to provide a code-mixed Hinglish version of the target utterance.

Context: [***Agent:*** Rakshak sends his greetings. What can I do now to help you? **::** Rakshak apani best wishes deta hai. Main aapki help ke lie kya kar sakata hoon?,
Victim: I'm not sure why this sort of thing constantly occurs to me. We live in a culture that is murderous. **::** main sure nahi hoon ki is tarah ki cheejen mere sath lagatar kyon hoti hain. hum ek aise culture men rahate hain jo janleva hai.,
Agent: I'm sorry you had to experience it. May I please know your good name before proceeding? **::** I am sorry ki aapko ye sab experience karna pad raha hai. Aage proceed karne se pehle kya main aapka shubh name jaan sakta hoon, please?,
Victim: My name is <person_name>, and I'm <person_age> years old.:: Mera naam <person_name> hai aur main <person_age> years old hoon.,
Agent: Good day, <person_name>. Could you further elaborate on your difficulty so that I may better assist you? **::** Good day, <person_name>. Kya aap please apani difficulty aur elaborate kar sakati hain taki main aapko achche se assist kar sakoon?,
Victim: Despite the fact that my boss expects favours in exchange for a promotion, he has not increased my compensation. That jerk beats me up and occasionally insults me when I go inside his cabin. **::** Is fact ke bavajood ki mera boss promotion ke badale men favours ki ummeed karta hai, usne mera compensation nahi badhaya hai. jab main uske cabin ke andar jati hoon to vah jerk mujhe marata hai aur kabhi kabhi mera insult karata hai.]

Target Utterance: ***Agent:*** So, have you filed an official complaint with the appropriate authorities? **::**

Prompting BLOOMZ Model

So, kya aapne appropriate authorities ke pas official complaint file karai hai?

Figure 4: Example of the six-shot version of prompt

A.1.3 Challenges in Dataset Preparation

We encounter the following challenges during *Hinglish* translation process:

(i) Procuring precise and meaningful translations of idioms and phrases from English to Hindi. For instance, for the utterance from a dialogue in POEM, "Seriously, you're now **driving me up the wall**. I work in a bank in Dhanbad, Jharkhand, and my account number is xxxxxxxxx.", the speaker is intended to convey annoyance. However, translators may interpret it literally, leading to a distortion in the intended meaning of the utterance.

(ii) Translating homographs, for example, "I feel so tired, I can't bear the weight of this burden anymore. I have **tear** in my eyes most of the time.", here '*tear*' means a drop of liquid from crying. In contrast, in "She was insane. She could not **tear** herself away from her husband.", '*tear*' means to move away.

(iii) Translating sarcastic utterances, for instance, in the utterance, "An FIR was filed, and one of the four scumbags was arrested, but he was freed two days later, and the cops have now awarded the rascals **a clean bill of health**.", the victim is expressing anger and disappointment over the clean chit given to the accused. However, a translator might not grasp the speaker's sarcastic intent and interpret it literally. In such instances, the translator might translate the sentence word-for-word and render it as "Ek FIR file ki gayi, aur chaaron sumbugs mein se ek ko arrest kiya gaya, lekin woh do din baad free kar diya gaya, aur ab police ne in badmashon ko **health ka ek clean bill de diya hai**.". Such translations can result in unnatural and contextually incorrect *Hinglish* dialogues.

(iv) Translating polite/impolite markers in utterances. For example, "**Please relax**, I am here to help you. May I know to whom I am talking?", here, the polite marker '*Please relax*' is conveying a request for the victim to calm down and become less anxious or stressed. However, it might be taken literally and translated into *Hinglish* as "**Please aaram karen**, main aapko help karne ke liye yahan hoon. May I know ki meri baat kisse ho rahi hai?". Such translations can make *Hinglish* conversations sound unnatural.

(v) Capturing the appropriate code-switching patterns and maintaining a smooth flow between languages, particularly while translating longer or complex utterances like "You have the option of filing a complaint with the Ministry of Women and Child Development. The Ministry was established with the primary

[9] ***Fluency*** - 5: Flawless, 4: Good, 3: Non-native, 2: Disfluent, 1: Incomprehensible; ***Adequacy*** - 5: All, 4: Most, 3: Much, 2: Little, 1: None; ***Colloquialism*** - 5: Very Good, 4: Good, 3: Average, 2: Poor, 1: Very Poor

goal of filling gaps in state action for women and children by promoting inter-ministerial and inter-sectoral convergence in order to develop gender-equitable and child-centered legislation, policies, and programmes. Do you want to go ahead with this option?"

A.2 Experiments

In this section, we provide the implementation details, a comprehensive description of the evaluation metrics, and the baselines used in the present work.

A.2.1 Implementation Details

We use PyTorch[10], a Python-based deep learning package, to develop our proposed model. We conduct experiments with the BERT import from the huggingface transformers [11] package. To establish the ideal value of the additive angle x, which affects performance, five values ranging from 0.1 to 0.5 were examined. The default value for x is 0.30. We set amplification value a as 64. All experiments are carried out on an NVIDIA GeForce RTX 2080 Ti GPU. We perform a grid search across 200 epochs.

We find empirically that our Embedding size is 812 bytes. We use Adam (Kingma and Ba, 2015) for optimization. The learning rate is 0.05, and the dropout is 0.5. The auto-latent encoder's dimension is fixed at 812. The discriminator $\mathcal{D}$ consists of two completely linked layers and a ReLU layer and accepts 812-D input features. Stochastic gradient descent has a learning rate of 1e-4 and a weight decay of 1e-3. with a momentum of 0.5. We perform 5 cross-validations on the **HING-POEM** dataset for training and testing purposes. We run our experiments for 200 epochs and report the averaged scores after 5 runs of the experiments to account for the non-determinism of Tensorflow GPU operations.

A.2.2 Evaluation Metrics

Since the proposed dataset has skewed class proportion, hence, to better assess the competency of our proposed method against the various baselines, we conduct 5-fold cross-validation. Finally, we report both Accuracy and macro-F1 scores for *Politeness Identification* (*PI*) and *Politeness Intensity Tagging* (*PIT*) tasks, F-Measure-Modified (FM), Precision-Modified (PM), Hamming Distance (HD), Jaccard F1 (JF) and Recall-Oriented Score (ROS) scores to evaluate the *Politeness Cause Elicitation* (*PCE*) task.

A.2.3 Baselines

We evaluate the efficacy of our proposed approach on the **HING-POEM** dataset against five state-of-the-art baseline models: BiRNN-Attn (Liu and Lane, 2016), CNN-GRU (Zhang et al., 2018), BiRNN-HateXplain (Mathew et al., 2021), BERT (Liu et al., 2019), BERT-HateXplain (Mathew et al., 2021), SpanBERT (Joshi et al., 2020) and Cascaded Multitask System with External Knowledge Infusion (CMSEKI) (Ghosh et al., 2022a). To adapt the RNN-based baselines to our code-mixed scenario, we use utterances' meta-embeddings formed from GloVe and fastText.

BiRNN-Attention. The only difference between this model and the BiRNN model is the addition of an attention layer (Liu and Lane, 2016) after the sequential layer. In order to further train the attention layer outputs, we calculate the cross entropy loss between the attention layer output and the ground truth attention.

CNN-GRU. We employ CNN-GRU (Zhang et al., 2018) on our proposed dataset. We add convolutional 1D filters of window sizes 2, 3, 4, with 100 filters per size, to the existing architecture. We employ the GRU layer for the RNN component and max-pool the hidden layer output representation. This hidden layer is routed via a fully connected layer to yield prediction logits.

BERT. We fine-tune BERT (Liu et al., 2019) by adding a fully connected layer, with the output corresponding to the CLS token in the input. Next, to add attention supervision, we try to match the attention values corresponding to the CLS token in the final layer to the ground truth attention. This is calculated using a cross-entropy between the attention values and the ground truth attention vector, as detailed in (Mathew et al., 2021).

BiRNN-HateXplain and BERT-HateXplain. We fine-tune the models[12] made available by (Mathew et al., 2021) on our **HING-POEM** dataset by changing the output layers as described earlier to suit our task's objective.

SpanBERT. SpanBERT (Joshi et al., 2020) follows a different pre-training objective compared to traditional BERT system (e.g. predicting masked contiguous spans instead of tokens) and performs better on question-answering tasks. Following the

[10] https://pytorch.org/

[11] https://huggingface.co/docs/transformers/index

[12] https://github.com/punyajoy/HateXplain

Setup	Politness Identification (PI)		Politness Intensity Tagging (PIT)		Politeness Cause Elicitation (PCE)				
	F1 (%)	ACC. (%)	F1 (%)	ACC. (%)	FM	PM	HD	JF	ROS
PAANTH-*EMotion*	75.43	76.77	67.93	70.29	36.31	38.91	0.65	0.80	0.81
PAANTH+*EMotion (Proposed)*	**77.12**	**78.77**	**71.93**	**73.31**	**38.39**	**39.41**	**0.67**	**0.81**	**0.83**

Table 5: Results from the PAANTH model with zero-shot emotion and without zero-shot emotion. Here, the bolded values indicate maximum scores.

Setup	Politness Identification (PI)		Politness Intensity Tagging (PIT)		Politeness Cause Elicitation (PCE)				
	F1 (%)	ACC. (%)	F1 (%)	ACC. (%)	FM	PM	HD	JF	ROS
PAANTH - ELMo	75.32	76.31	69.88	71.65	35.11	37.32	0.65	0.79	0.80
PAANTH - Tf-Idf	76.05	76.29	70.31	70.43	35.04	38.11	0.66	0.78	0.81
PAANTH - (ELMo+Tf-Idf)	75.03	75.11	67.98	69.20	34.84	37.71	0.65	0.77	0.78
PAANTH (Proposed)	77.12	78.77	71.93	73.31	37.59	39.41	0.67	0.81	0.83

Table 6: Effect of different embeddings in the PAANTH model.

work in (Ghosh et al., 2022b) where SpanBERT is used to solve a mix of classification and cause extraction tasks, we fine-tune the SpanBERT base model on our HING-POEM dataset to meet our objective.

Cascaded Multitask System with External Knowledge Infusion (CMSEKI). We contrast the performance of our model with the state-of-the-art CMSEKI system presented in (Ghosh et al., 2022a). CMSEKI leverages common-sense knowledge in the learning process to address multiple tasks simultaneously.

A.3 Additional Analysis

This section delineates additional analysis for our proposed PAANTH framework.

A.3.1 Analysis of Embeddings

We investigate the importance of the different embeddings on the performance of the proposed PAANTH framework by ablating the different embedding types. In the first ablated model, we remove the embeddings generated by the ELMo and observe a drop in F1-score for the PI task by 2.46%, F1-score for the PIT task by 1.66%, and ROS for the PCE task by 0.03 points. Similarly, we notice a performance drop when Tf-Idf embeddings are removed. The performance degrades significantly when both the embeddings are ablated (3.66% in the F1 PI task, 4.11% in F1 for the PIT task, and 0.05 ROS points for the PCE task).

A.3.2 Comparison with ChatGPT

We perform a pilot study using ChatGPT[13] to demonstrate the effectiveness of our proposed framework. We notice that PAANTH has an overwhelming performance advantage over ChatGPT; one possible reason is that the few-shot prompt setting may not be enough to achieve satisfactory performance for complex tasks like politeness identification, politeness cause elicitation and politeness intensity tagging. A few sample predictions from ChatGPT on the PCEIT task are shown below:

- **Utterance:** *Kripya shaant ho jaiye aur dhairya rakhiye , hum yahan aapki har tarah se help karne ke liye hai. Kya aap bata sakte hain , ki hum kisse interact kar rahe hai?* (Please calm down and have patience. We are here to help you in every possible way. Can you tell with whom we are interacting?);
  Human Annotators: Politeness Label: *Polite*, Politeness Causal Span: *Kripya shaant ho jaiye aur dhairya rakhiye , hum yahan aapki har tarah se help karne ke liye hai. Kya aap bata sakte hain , ki hum kisse interact kar rahe hai?*, Politeness Intensity Value: *2*
  ChatGPT: Politeness Label: *Polite*, Politeness Causal Span: *Kripya shaant ho jaiye aur dhairya rakhiye , hum yahan aapki har tarah se help karne ke liye hai. Kya aap bata sakte hain, ki hum kisse interact kar rahe hai?* , Politeness Intensity Value: *2*.
- **Utterance:** *What a load of nonsense, and yet another inquiry. main ek house visit lena chahungi.* (What a load of nonsense, and yet another inquiry. I want to opt for a house visit.);
  Human Annotators: Politeness Label: *Impolite*, Politeness Causal Word: *What a load of nonsense, and yet another inquiry. maiN ek house visit karna chahungi.*,

[13] https://chat.openai.com/

Basic Model	$\mathcal{L}_{ad}$	$\mathcal{L}_{re}$	$F1^{PI}(\%)$	$F1^{PIT}(\%)$	$JF^{PCE}(\%)$	$ROS^{PCE}(\%)$
✓			74.23 (-2.89)	70.24 (-1.69)	0.81(-0.089)	0.82(-0.013)
✓	✓		75.79 (-1.33)	70.64 (-1.29)	0.80(-.07)	0.82(-0.011)
✓		✓	75.38 (-1.74)	70.42 (-1.51)	0.80(-0.012)	0.82(-0.013)
✓	✓	✓	**77.12**	**71.93**	**0.81**	**0.83**

Table 7: Effect of different loss functions. The basic model combines semantic features via zero-shot loss function

utterance	Extracted Span	Predicted Label
Partially extracted causal spans		
1. kya aap hamse share karenge ki vah kaise aapke credentials ko harm karne ki activity kar raha hai?	**kya aap hamse**	**Impolite**
2. Mujhe khushi hai ki is tough situation men aap itne positive hain. Kya aap mujhe batayenge ki vah aapko blakemail karta hai ya kuch aur bhi?	**Kya aap mujhe batayenge**	**Polite**
Causal spans not extracted		
3. aapki information ke liye dhnyawaad . main ispar aapas me discuss karke aapse baat karungi. Bye.	**No Cause**	**Neutral**

Table 8: Error analysis from the proposed **PAANTH** framework. Color Coding: Blue- Correct, Red: Incorrect; Teal: Incomplete. Highlighted text in pink shows the human annotated causal spans.

Politeness Intensity Value: *1*;
ChatGPT: Politeness Label: *Impolite*, Politeness Causal Word: *What a load of nonsense , and yet another inquiry. maiN ek house visit karna chahungi.*, Politeness Intensity Value: *1*.

A.3.3 Varying Context Length

By changing context sizes(ψ), we examine the role that context plays in PCEIT task. The following context lengths were trained for by **PAANTH**: 1, 3, 5, 7, 9, 10 and 12. The results are represented in Figure 5. 1 means there is no context, and the model merely receives the target utterance as input. We observe a steady improvement in performance as the number of previous utterances increases. When the ψ is set to 7, we get the best results. More context does not provide useful information, resulting in model confusion and poor performance.

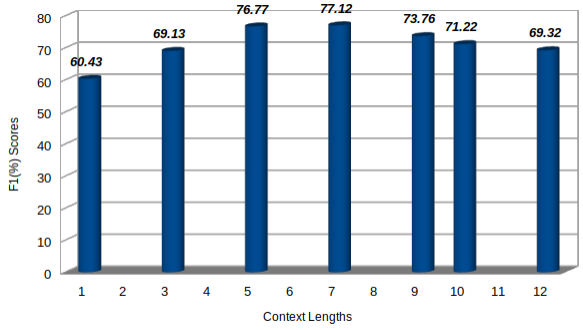

Figure 5: Graphical depiction of results of **PAANTH** on varying context length.

A.3.4 Emotion analysis for Politeness task

As politeness and emotion are interconnected (Priya et al., 2023a), we also attempt to investigate the relationship between them in code-mixed setting. However, annotating emotions proved to be a significant challenge for our annotators, leading us to incorporate zero-shot emotion into our model. The results presented in Table 5 demonstrate that our hypothesis was indeed correct. By utilizing zero-shot emotion, we observe a notable improvement in our model's performance, with a 1.69% increase in the F1 score for *PI* task, 4% increase in F1-score for *PIT* task and a 2-point increase in the ROS score for *PCE* task. Hence, by mitigating the burden of explicit emotion annotation, our approach yielded positive outcomes.

A.3.5 Loss Function Analysis

We further investigate the significance of the loss functions in **PAANTH** by removing one of them one by one. We report the ablation analysis for various loss functions in Table 7. In the first ablated model, we remove all two loss functions (i.e., $\mathcal{L}_{ad}$, and $\mathcal{L}_{re}$. We remove the $\mathcal{L}_{re}$ loss function in the second model, and the $\mathcal{L}_{ad}$ adversarial function in the third. In the fourth model, we remove $\mathcal{L}_{ad}$ and $\mathcal{L}_{re}$. When any of these losses are eliminated from **PAANTH**, we see a performance decline when compared to the proposed method. The performance decline is the largest (2.89% in F1 for *PI* task, 1.69% in F1 for *PIT* task and 0.013 POS points for *PCE* task) when all the losses are eliminated. Clearly, loss functions play a crucial role in training the entire model end-to-end.

A.3.6 Analysis on Task Setting

We also perform an ablation study on our proposed approach to analyze the importance of various tasks in three different task settings (uni-task: *PI* or *PCE* task, bi-task: *PI+PIT* or *PI+PCE*, multi-task (proposed task setting): *PI+PIT+PCE*). As shown in Figure 6, we observe that when either *PIT* or *PCE* task is ablated, the performance drops significantly in comparison to the tri-task setting. Specifically, *PI* accuracy, *PCE* JI and ROS drops by 1.65, 1.0, and 2.0 points, respectively, when *PIT* task is omitted. On removal of *PCE* task, the *PI* and *PIT* accuracy drops by 2.2 and 5.5 points, respectively. We notice a further decline in performance in unitask settings in terms of all the metrics. This confirms that all the tasks are interrelated and help each other in order to achieve the best overall performance.

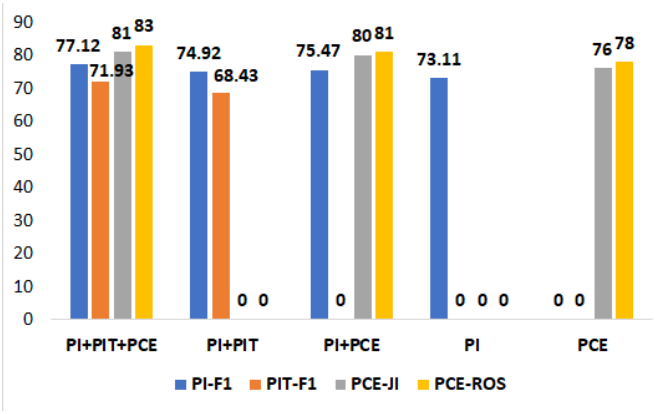

Figure 6: Graphical depiction of multi-task vs uni-task/bi-task comparison

A.3.7 Error Analysis

Although our proposed **PAANTH** framework performs well in majority of the test cases, still there are certain scenarios where it fails to make the correct predictions. We show some sample predictions from the test set in Table 8. In the first two instances, our model is able to partially predict the causal spans; however, in the first example, it fails to categorize the utterance as *Politeness*. It is also to be noted that the model extracted span in the second example seems to be more appropriate than the actual annotation by the human annotator. The model rightfully ignores some information but focuses on the other relevant action part of the utterance. This illustrates our model's strong ability to comprehend politeness reasoning among diverse test cases. In the third and fourth examples, our model fails to extract any relevant cause from the given input. Moreover, in the third example, the model wrongly categorizes the utterance as Non-polite. This can be due to the lack of sufficient context that hindered our model's comprehension ability for the given input.