An Appraisal Theoretic Approach to Modelling Affect Flow in Conversation Corpora

Anonymous ACL submission

Abstract

This paper presents a model of affect in conversations by leveraging Appraisal Theory as a generalizable framework. We propose that the multidimensional cognitive model of Appraisal Theory offers significant advantages for analyzing emotions in conversational contexts, addressing the current challenges of inconsistent annotation methodologies across corpora. To demonstrate this, we present AppraisePLM, a regression and classification model trained on the crowd-EnVent corpus that outperforms existing models in predicting 21 appraisal dimensions including *pleasantness*, *self-control*, and *alignment with social norms*. We apply AppraisePLM to diverse conversation datasets spanning task-oriented dialogues, general-domain chit-chat, affect-specific conversations, and domain-specific affect analysis. Our analysis reveals that AppraisePLM successfully extrapolates emotion labels across datasets, while capturing domain-specific patterns in affect flow – change in conversational emotion over the conversation. This work highlights the entangled nature of affective phenomena in conversation and positions affect flow as a promising model for holistic emotion analysis, offering a standardized approach to evaluate and benchmark affective capabilities in conversational agents.[1]

1 Introduction

Affect, which encompasses both emotion and mood, is crucial in conversations, influencing dynamics such as empathy, sarcasm, and naturalness (Ruusuvuori, 2012). In the domain of conversational agents (CAs), recognizing and responding to affective cues is essential (Skowron and Paltoglou, 2011; Yang et al., 2019). Various methodologies are employed for incorporating affect into CAs, including emotion classification, dimensional ratings, intent annotations, and vicarious emotion ratings such as empathy and condolence (Busso et al., 2008; Ma et al., 2020; Karna et al., 2020). While affect-annotated datasets exist across general and specialized domains, inconsistencies in annotation schemas and objectives pose challenges for standardizing affect modelling in conversational AI (Liu et al., 2021; Islam et al., 2022).

These inconsistencies arise due to variations in annotation methodologies, including differences in unit-level labelling (e.g., turn-wise versus full-conversation annotations) and dataset construction depending on the domain (Liu et al., 2024). Moreover, evaluation metrics for contextual affect interactions remain limited, coercing a generalization of findings across datasets. Many domain-specific models, such as diff-EPITOME (Lee et al., 2022), are trained within a specific domain but later applied broadly, highlighting the need for standardized affect evaluation (Schaaff et al., 2023). A generalizable framework for modelling affect in conversations could address these challenges, ensuring more consistent benchmarking for conversational agents.

This paper proposes that **Appraisal Theory** provides such a generalizable framework. Appraisal theory conceptualizes emotions as responses to an individual's evaluation of a stimulus along multiple cognitive dimensions (Ellsworth and Smith, 1988; Scherer, 2005). For example, anger can be characterized as an unpleasant, short-lived emotion with low self-control (Roseman and Smith, 2001). Such an approach not only allows for the modelling of emotional intensity and duration but also enables the analysis of *affect flow*, or how emotions evolve throughout a conversation (Hendriks et al., 2014; Poria et al., 2019b).

In this paper, we hypothesizes that: (**H1**) appraisal-theoretic emotion analysis aligns with existing emotion annotations; and that (**H2**) such a cognitive analysis captures affect flow: emotion

[1] Code is available here: `https://anonymous.4open.science/r/appraise-plm`

change over the course of a conversation. To test these hypotheses, the paper introduces **AppraisePLM**, a model for appraisal regression and emotion classification, trained on the crowd-EnVent corpus. Crowd-EnVENT is a benchmark emotion recognition and appraisal analysis corpus which provides fine-grained annotations of event descriptions on 21 appraisal dimensions including *pleasantness*, *self-control*, and *suddenness* (Troiano et al., 2023).[2]

Our model outperforms existing classifiers and regressors on this dataset and is subsequently applied to turn-wise appraisal annotation across four benchmark conversation corpora: EmoWOZ (Feng et al., 2022), EmpatheticDialogues (Rashkin et al., 2019), DailyDialog (Li et al., 2017), and Epitiome (Sharma et al., 2020). Our results show that AppraisePLM improves appraisal estimation performance on the crowd-EnVent corpus and can extrapolate categorical and emotion labels. Additionally, corpus domain influences affect flow, with distinct patterns emerging in specific domains (e.g., empathetic conversations improving *pleasantness*). Through this paper, we highlight the intertwined nature of affective phenomena and argue towards developing appraisal theory as an interpretable intradomain model of emotion in conversation.

2 Background and Motivation

2.1 Emotion Recognition in Conversation

Emotion recognition in conversation (ERC) often relies on Plutchik's wheel or Ekman's universal emotions for annotation (Plutchik, 1965; Ekman, 2000). Commonly used general-domain dialogue corpora, such as DailyDialog (Li et al., 2017), MELD (Poria et al., 2019a), and EmotionLines (Hsu et al., 2018), employ a set of basic emotions like joy, fear, sadness, anger, surprise, disgust, and neutral. However, some corpora use varying numbers of emotion categories, ranging from fine-grained annotations to broader affect labels (Qin et al., 2023). The veracity and similarity of emotions can differ significantly by domain, raising questions about the accessibility and identification of fine-grained emotions in conversation (Hancock et al., 2007; Machová et al., 2023).

The Valence-Arousal-Dominance (VAD) model is a prevalent dimensional model for emotion, with IEMOCAP serving as a reference corpus providing both dimensional and categorical emotion labels (Busso et al., 2008). The conversation corpus' domain heavily influences the taxonomy and distribution of emotion labels (Rajapakshe et al., 2024). For instance, mental health-focused corpora may prioritize certain emotions over others compared to general-domain corpora (Saha et al., 2022). Additionally, factors such as access to different modalities and the number of participants in the conversation can impact the emotion annotation methodology as well (Pereira et al., 2023, 2025). Appendix provides a table to show the inconsistencies across emotion annotations in conversation corpora.

2.2 Affective Phenomena in Conversation

The emotions expressed and perceived by interlocutors influence expected conversational behavior, though modeling "emotion shift" remains an open problem (Pereira et al., 2025). Corpora often use direct annotation methods to extract relevant affective features and behaviours. For instance, EmpatheticDialogues is a benchmark open-domain empathetic conversation corpus that uses 32 fine-grained emotion labels, also applied in EDOS (Rashkin et al., 2019).

Some domain-specific corpora, such as Epitome (Sharma et al., 2020), ALoE (Yang and Jurgens, 2024), PAIR (Pérez-Rosas et al., 2022), and ESConv (Liu et al., 2021) in the mental health domain, do not directly annotate emotion. Instead, they assess characteristics of empathetic interactions using direct annotator ratings, like *Emotional Interpretation* in Epitome. This approach allows models to access desirable interactional behaviours without relying solely on emotion (Lahnala et al., 2022). Metrics from PAIR and Epitome have been used to benchmark open-domain conversational agents, expecting these behaviours in general-domain contexts (Lee et al., 2024). For example, a general-purpose conversational agent should provide condolence, implying expected linguistic behaviour with an affective signal (Zhou and Jurgens, 2020). The manner and display of empathy vary with context, relationship, and personality, as noted in the PEC corpus (Zhong et al., 2020).

2.3 Appraisal Theory in Language and Conversation Analysis

Appraisal theory posits that experienced emotions result from cognitive appraisals of event stimuli,

[2] The list of appraisal dimensions are defined and described in Appendix B

such as pleasantness, suddenness, controllability, or alignment with social norms (Ellsworth and Smith, 1988). This theory offers a view of an experiencer's cognitive state by systematically choosing context-appropriate appraisals.

Appraisal theory has gained prominence in NLP and conversation analysis, enhancing emotion classification and interpersonal communication studies (Balahur et al., 2011; Hofmann et al., 2020). In NLP, it improves emotion classification accuracy through dimensional models and annotated corpora, aiding in understanding how emotions arise from event evaluations (Troiano et al., 2022; Resendiz and Klinger, 2023). In conversation analysis, appraisal theory reveals how speakers express attitudes and manage relationships through evaluative language.

The theory has also had use in analyzing motivational interviewing, with the ALOE dataset focusing on empathetic alignment in therapeutic conversations using appraisal theory (Yang and Jurgens, 2024). However, this work is domain-specific and lacks correspondence with other categorical or dimensional labels.

Troiano et al. (2023) introduced the crowd-enVENT dataset, consisting of 6,600 emotion-inducing event descriptions annotated with 21 appraisal dimensions, emotion labels, and author demographics. This dual-perspective annotation allows for comparing appraisal and emotion reconstruction by readers versus computational models, providing a human baseline for machine learning tasks. Unlike ISEAR, crowd-enVENT was compiled specifically for text analysis, studying the relationship between appraisals, emotions, and event descriptions.

3 The AppraisePLM Framework

In this section, we propose AppraisePLM, an appraisal-theoretic conversation analysis framework which estimates the aggregate change(s) and patterns in how the interlocutors appraise the conversation over time. First, we test the cross-comparability of appraisals with other representations (§3.1-3.3) then provide the methodology to do the same for conversations (§3.4).

3.1 Problem Definition

Given a dataset $D = \{(e_i, l_i, c_i)\}_{i=1}^{N}$ where e_i is the ith event description (text), $l_i = [l_i^1, l_i^2, ..., l_i^a]$ is a vector of a event description appraisals, and $c_i \in \mathbb{C}$ is a label from the set of n emotion class labels $\mathbb{C}$, we perform the following two tasks.

Appraisal Estimation Train a function $f_{app} : \mathbb{R}^d \rightarrow \mathbb{R}^a$ where d is the dimensionality of the encoded event description PLM(e_i) and a is the number of appraisals. The objective of this function is to find $\theta_{app}* = \arg\min_{\theta_{app}} \mathcal{L}_{app}$ such that:

$$\mathcal{L}_{app} = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{a}\sum_{j=1}^{a}(f_{app}(\text{PLM}(e_i))_j - l_i^j)^2$$

Emotion Classification Upon appraisal estimation, train a function $f_{emo} : \mathbb{R}^d \times \mathbb{R}^a \rightarrow \mathbb{C}$, where d is the dimensionality of the encoded event description PLM(e_i), $\mathbb{C}$ is a set of n emotion class labels, and a is the number of appraisals. The objective of this function is to find $\theta_{emo}\star = \arg\min_{\theta_{emo}} \mathcal{L}_{emo}$ such that:

$$\text{comb} = \text{PLM}(e_i) \oplus f_{app}(\text{PLM}(e_i))_k$$

$$\mathcal{L}_{emo}^{comb} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{k=1}^{n} 1_{[c_i=k]} * \ln f_{emo}(\text{comb})$$

3.2 Dataset Characteristics

The crowd-EnVENT dataset consists of 6,600 event descriptions (550 event descriptions for 13 emotion labels). Each event is annotated with 21 appraisal variables, which are cognitive evaluations of the event by the event's author. The fine-grained emotion labels allow us to analyze how experiencers appraise various emotions (including a no-emotion label). The distribution of appraisal values is skewed, more than 33% of the corpus being either 1 or 5. Their approach for appraisal classification involves a two-class classificaiton, which we do not use as the differences in appraisal values are a critical step in AppraisePLM.

3.3 Model Framework

The proposed AppraisePLM multitask framework jointly performs appraisal regression and emotion classification using attention-attenuated pretrained language models (PLMs) such as RoBERTa (Liu et al., 2019), DeBERTa (He et al., 2020), MP-Net (Song et al., 2020) and T5 (Raffel et al., 2020); with DeBERTa yielding the best performance. Figure 1 provides a simple schematic of the model architecture.

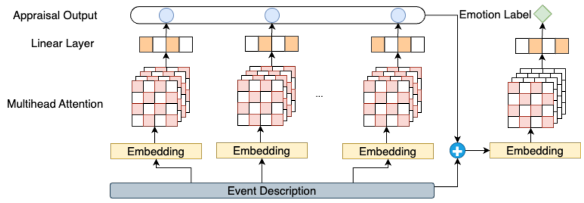


Figure 1: Model architecture for the AppraisePLM framework.

	T → A.	**T→ E**	**T + A → E**
Model	MSE ↓	F1 ↑	F1 ↑
Troiano et. al. (2022)	1.97	0.59	0.60
RoBERTa-large	1.62	0.59	0.67
T5-large	1.12	0.61	0.66
MPNet-base	**1.08**	0.64	0.70
DeBERTa-large	**1.08**	0.66	**0.71**

Table 1: Performance of the AppraisePLM architecture for appraisal estimation and emotion classification. Emotion classification is done in two modes; with only text (**T → E**) and both text and appraisals (**T + A → E**) on the crowd-EnVent corpus.

The event description is embedded using a PLM encoder and a multihead attention layer. Each appraisal dimension has a task-specific multihead attention layer and linear head. Regression is trained with individual MSE losses for all appraisal values.

Emotion classification utilizes both the PLM representation and predicted appraisal values. The encoded event description is concatenated with the predicted appraisal values, normalized and regularized before being decoded by another multihead attention layer and a linear classification head. Classification is trained on cross-entropy loss.

We use an AdamW optimizer with a weight decay 0.01 and a learning rate 2e−5. We use a standard grid search for hyperparameter tuning. Training employs Distributed Data Parallel (DDP)[3] on four RTX 2080 Ti GPUs, with a batch size of 16 and gradient checkpointing, early stopping within three epochs with a maximum training of ten epochs. Reproducibility report is provided in **??**.

3.4 AppraisePLM Results and Performance

Table 1 presents the test set performance of the AppraisePLM architecture on the crowd-enVENT dataset, compared to the baseline model. While attention attenuation marginally improves appraisal estimation, DeBERTa-large achieves the highest performance. However, the limited improvement in regression scores reflects the task's complexity (see Appendix D).

[3] https://pytorch.org/docs/stable/generated/torch.nn.parallel.DistributedDataParallel.html

Corpus	**P**	**R**	**F1**
EmpatheticDialogues	0.77	0.79	0.78
DailyDialog	0.63	0.66	0.62
EmoWOZ	0.62	0.56	0.59

Table 2: Zero-shot emotion classification performance of AppraisePLM (DeBERTa-large; T + A → E) model on conversation corpora with emotion labels. Comparisons are done after label folding, a smaller subset of the crowd-EnVENT emotion labels are mapped to the labelling schema of the corpus.

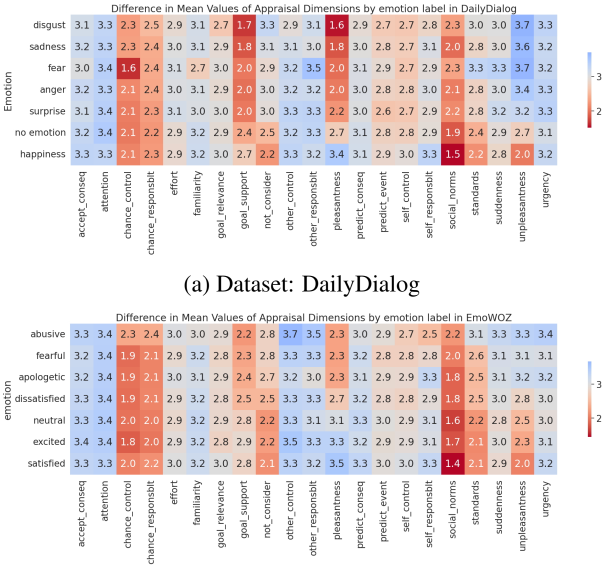
Appraisal representations enhance categorical emotion detection, with event descriptions appended with appraisal information yielding a 0.11 macro avg. F1 improvement over the baseline. Multi-head attention slightly improves standard emotion classification (**T → E**. in Table 6), but the AppraisePLM architecture shows a more substantial boost when integrating both text and appraisal data.

Figure 6 visualizes appraisal estimates across emotions using DeBERTa-large AppraisePLM, with emotions ordered by *pleasantness*. As expected, *no-emotion* separates positive and negative emotions, with *joy* being the most pleasant and *disgust* the least. Unpleasantness follows the inverse trend, while *urgency*, *attention*, and *other-control* exhibit minimal variation across emotions.

4 Affect Annotations in Dialogue Corpora

In this section, the applicability of AppraisePLM on conversational corpora is examined using four datasets: EmoWOZ, DailyDialog, EmpatheticDialogues, and Epitome. These datasets vary in emotion annotation schemes, label counts, and domains, so the analysis considers each corpus individually while maintaining methodological consistency.

AppraisePLM estimates the appraisal dimension for dialogue turn and concatenates them with utterance embeddings for zero-shot emotion classification. The DeBERTa-large AppraisePLM model is used for annotation. Due to differing labelling schemas for some datasets, label folding is applied, and a co-occurrence Emotion category similarities with crowd-EnVENT are assessed, and relevant labels are retained for weighted F1-score evaluation.



(a) Dataset: DailyDialog

(b) Dataset: EmoWOZ

Figure 2: Average estimate of each appraisal from the DailyDialog and EmoWOZ test corpora using the best performing AppraisePLM DeBERTa-large. The emotion labels are ordered by *pleasantness* from low (red) to high (blue).

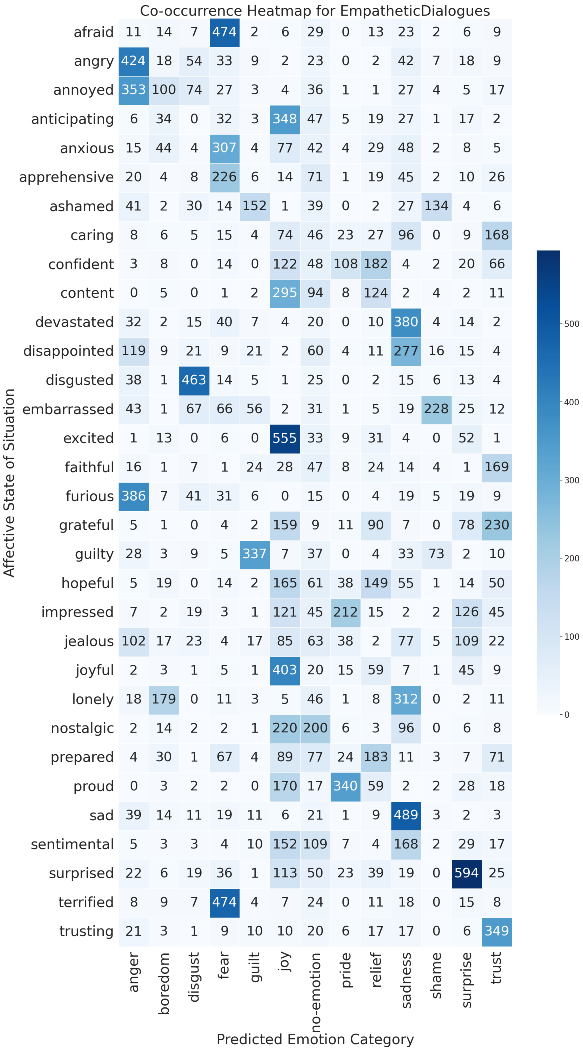


Figure 3: A co-occurrence heatmap of predicted emotion category and annotated emotion label for the EMPATHETICDIALOGUES corpus. Emotion categories are predicted for emotionally grounded situations.

4.1 DailyDialog

The DailyDialog dataset is a high-quality, manually labeled, multi-turn dialogue dataset designed to reflect everyday communication. It contains 13,118 dialogues, with an average of approximately 8 speaker turns per dialogue. The dataset covers various topics related to daily life, providing a diverse range of conversational context and includes manual annotations for topics, dialogue acts, and emotion.

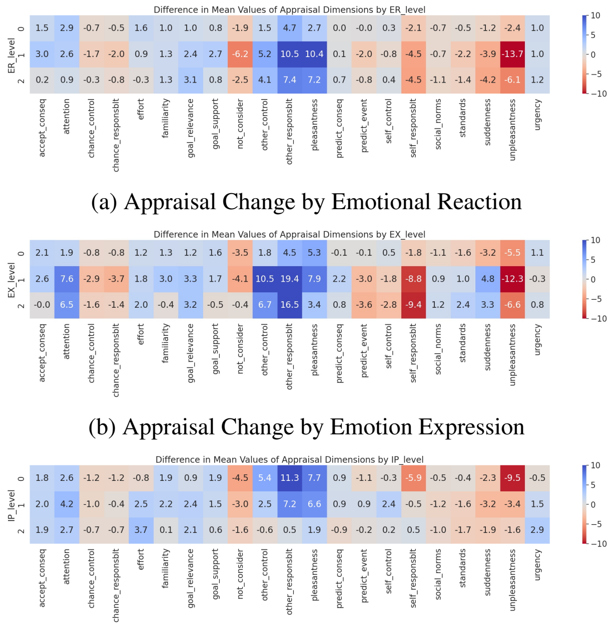
DailyDialog uses a six class emotion classification (*anger*, *fear*, *disgust*, *happiness*, *surprise*, *sadness*) along with a *no-emotion*. The latter is almost 80% of the corpus, while in the emotion labelled turns, 74% of them are labelled *happiness*. This label skew affected AppraisePLM's performance. Since the DailyDialog emotion categories are a subset of Plutchik's categories, no label folding or merging was performed, computing a strict macro weighted F1 score of 0.62 for emotion classification using AppraisePLM DeBERTa-large.

Figure 2a shows the average distribution of appraisal values across emotion labels for the DailyDialog corpus. We see that these appraisals are similar to the appraisal distribution by emotion label for crowd-EnVENT, except the average valence of the *no emotion* label and the slightly higher *pleasantness* and *unpleasantness* estimates of *disgust*- and *fear*-labelled conversation turns.

4.2 EMPATHETICDIALOGUES (ED)

The EmpatheticDialogues (ED) dataset comprises 24,850 one-to-one open-domain conversations, with 2,457 in the test set analyzed here. Each conversation features a speaker sharing a personal emotional experience and a listener responding empathetically. The dataset includes 32 fine-grained emotions, with 5.1% tagged as "surprised" and 1.9% as "faithful", and test set conversations averaging 4.2 turns.

Since ED uses a custom emotion list, AppraisePLM's emotion detection is evaluated using a coarser emotion set. Figure 3 shows that the model effectively distinguishes broad emotional



(a) Appraisal Change by Emotional Reaction

(b) Appraisal Change by Emotion Expression

(c) Appraisal Change by Interpretation

Figure 4: The change in appraisal estimate between the speaker and response posts of the EPITOME dataset. Change in appraisal estimates is computed as $a_r^3 - a_s^3$ where a_r is the average response appraisal and a_s is the seeker appraisal, scaled for trend analysis.

categories (e.g., "afraid," "anxious," "apprehensive," and "terrified" all align with "fear"). It also identifies theoretical correlations across annotation schemas (e.g., "lonely" and "annoyed" strongly correlate with "boredom"). Synonym-based label folding results are reported in Table 5.

Appendix Figure 7 presents appraisal estimates of emotions, ordered by pleasantness, showing similarities with the crowd-EnVENT corpus (Figure 6). The ordering of emotions reflects their perceived intensity or arousal (e.g., "furious" vs. "angry" and "disgusted" vs. "annoyed"). Notably, while "devastated" is among the most unpleasant, it is not the least pleasant and exhibits higher goal support than more negatively valenced emotions. Additionally, the range of appraisal estimates in ED is narrower than in crowd-EnVENT.

4.3 EPITOME

The EPITOME dataset is designed to examine empathy in text-based, asynchronous conversations, incorporating both emotional and cognitive aspects. It consists of 10,000 post-response pairs sourced from online platforms such as Reddit and TalkLife, annotated along three dimensions—Emotional Reaction (ER), Interpretation (IP), and Exploration (EX)—each rated on a 0-2 scale: ER demonstrates warmth, compassion, or concern, IP reflects an understanding of inferred feelings and experiences, and EX explores aspects of the seeker's experience not explicitly stated. Since these annotations rely on comparisons between posts, whereas Appraise-PLM annotates individual turns, we reinterpret the dimensions through differences in cognitive appraisals. Specifically: **High ER** corresponds to increased *pleasantness* and *other-responsibility* while decreasing *unpleasantness* and *self-responsibility*; **High IP** implies minimal change in appraisal values, ensuring emotional alignment with the seeker, and **High EX** suggests differences in *other-control*, *other-responsibility*, and *self-responsibility* between seeker and response, showing a distinct but similar affect.

The heatmap analysis (Fig. 4) highlights two key findings: (1) ER and EX ratings of 1 show greater shifts in appraisals than ratings of 2, and (2) IP ratings of 2 correspond to the lowest average appraisal shifts, indicating stronger alignment between seeker and response posts.

4.4 EmoWOZ

The EmoWOZ dataset is a large-scale, manually emotion-annotated corpus of task-oriented dialogues, derived from MultiWOZ. It is designed to examine how user emotions impact task-oriented dialogue systems. EmoWOZ contains 11,434 dialogues, including both human-human (MultiWOZ) and human-machine (DialMage) dialogues. The analysis focuses on the test set.

EmoWOZ employs a custom emotion labelling scheme for task-oriented dialogues, with seven labels: neutral, satisfied, dissatisfied, excited, apologetic, fearful, and abusive, adapted from the OCC emotion model. Due to differences in domain and classification intent, these labels do not directly correlate between corpora, with "neutral" being overwhelmingly dominant.

Figure 2b shows distinct appraisal profiles across emotion labels. The "neutral" category serves as a separator between positive and negative states. We can see that emotion ordering by pleasantness aligns with emotional valence. The range of appraisal values in EmoWOZ is lower than in other conversational datasets, likely due to the task-oriented nature of dialogues, which exhibit less emotional variability than chit-chat. Categorical labeling alone would not highlight such differences effectively. Table 2 indicates that emotion detection is more challenging in EmoWOZ, partly due to label imbalance, with notably fewer "abusive" and "dissatisfied" conversation turns.

5 Affect Flow in Conversation

So far, we have used appraisal theory to analyze categorical emotions in multi-turn dialogues. In this section, we extend this by modelling affect flow – how emotional appraisals evolve throughout a conversation. This is achieved by tracking shifts in appraisal dimensions between speakers and turns across multiple corpora. A key methodological refinement involves a power function transformation, which amplifies subtle but consistent variations in appraisal values, making it possible to discern meaningful patterns in conversational emotion shifts (as we did with EPITOME in §4.3).

To investigate affect flow, we model the gradient of appraisal values at each turn, differentiating between speaker and listener contributions to emotional evolution. Since appraisal shifts can be too subtle when averaged over an entire corpus, we categorize conversations by emotion to examine appraisal changes more precisely. The gradient analysis quantifies how a speaker's current appraisal estimate relates to the listener's next estimate, revealing that different corpora exhibit distinct patterns of emotional adaptation. We present our observations below. We refer to Figure 5 to examine the change in appraisals over time for a sample of the corpora.

EmoWOZ has the highest gradient and lowest central tendency for appraisal estimates, indicating large fluctuations in emotion appraisals over a conversation ((Figure 5a, 5b)). Conversations labelled satisfied exhibit strong positive valence shifts, with both speakers increasing appraisals of *pleasantness* and *goal support* over turns. In contrast, conversations labelled *dissatisfied* show an amplifying effect for positive appraisals and a dampening effect for negative appraisals by the second interlocutor, highlighting a different form of emotional adaptation compared to the other corpora.

EmpatheticDialogues (ED) contains the shortest conversations on average and shows low variation in appraisal shifts between turns (Figure 5e, 5b). Conversations in this corpus display empathic matching (Wondra and Ellsworth, 2015) for both positive and negative emotions: speakers and listeners tend to align their appraisals over time, leading to appraisal gradients closer to zero. The *happy/joyful* category exhibits strong alignment, consistent with theoretical expectations of interactional empathy, where interlocutors appraise events similarly over successive turns.

DailyDialog (DD) exhibits higher variation in appraisal gradients, particularly for negative emotions, suggesting that emotional shifts are more dynamic ((Figure 5c, 5d)). Unlike ED, where emotion directionality is clear (seeker vs. provider), DD does not enforce speaker roles. Either participant can elicit emotion, leading to non-uniform affect flow. Despite this variability, a general trend of appraisal convergence is observed over time, particularly for emotions like joy and sadness, although sadness shows a distinct decrease in unpleasantness near the end of conversations.

EPITOME Unlike the other corpora, EPITOME exhibits appraisal shifts where emotional convergence occurs but with different dominant appraisal dimensions. While *pleasantness* and *unpleasantness* remain key indicators, dimensions such as *self-responsibility*, *other-responsibility*, attention (for emotional expression), and *not consider* (for emotional reaction) play a larger role in distinguishing response quality. Higher quality responses, as measured by reaction, interpretation, and expression ratings, show distinct appraisal characteristics, reinforcing the importance of nuanced appraisal dynamics in emotion modelling.

6 Discussion

Appraisal theory, as a model of emotion realized in text, is based on post-hoc or simulated appraisals of cognitive dimensions correlated with universal emotion labels. We approximate conversation segments (situation, turn, or response) as event descriptions, assuming that post-hoc contextual rating of appraisals preserves the relationship between semantic and cognitive representations of affect. Our analysis of conversational corpora using appraisal estimation yields mixed quantitative results but offers promising qualitative insights. Label inconsistencies complicate the evaluation of emotion detection in AppraisePLM.

In this section, we examine the veracity of the hypotheses mentioned in §1.

6.1 H1: Aligning with Extant Affect Annotations

In **H1**, we hypothesized that appraisal-theoretic emotion analysis aligns with existing emotion annotations. We tested this at multiple levels by examining patterns of appraisal estimates for the overall corpus, characterized by its domain and annotation level (conversation, turn, or response).

We found that fully textual corpora, such as EMPATHETICDIALOGUES and DailyDialog, exhibit significantly higher alignment in categorical labels between the AppraisePLM emotion classification and existing annotations. This finding is notable, as both corpora have different approaches and goals for affect annotation. However, the domain of affect annotation poses challenges for quantitative analysis.

For instance, in the EPITOME corpus, changes in appraisal estimates between utterance and response align with the definitions of the annotated dimensions, while appraisal-informed emotion classification reflects the source of the conversation. A similar domain effect is observed in EmoWOZ, where emotion classification scores after label folding were baseline, but trends in appraisal by turn and speaker correspond to action states in the corpus.

In summary, appraisal theory shows reasonable alignment with existing affect annotations in conversational corpora, providing additional cognitive insights. Using appraisal theory as the grounding emotion annotation in general domain conversations would significantly improve the performance and reliability of the AppraisePLM approach.

6.2 H2: Appraisal Change as Affect Flow

In **H2**, we hypothesized that cognitive analysis captures affect flow, examined as emotion change over the course of a conversation. We observed that not all appraisals are relevant to a conversation or domain and may change minimally. However, those appraisals that do change exhibit a small but consistent gradient when aggregated over the conversation.

Section §5 details findings from one approach to examining affect flow using the power-amplified difference of appraisal estimates between conversation turns. Appraisal gradients differ by dataset: EMPATHETICDIALOGUES exhibits low appraisal shifts, DailyDialog shows greater variability in negative emotions, and EmoWOZ presents the highest appraisal gradient with distinct trends for satisfaction and dissatisfaction. Empathic matching, where interlocutors align appraisals over time, is evident in positive emotions across EMPATHETICDIALOGUES and DailyDialog. However, DailyDialog lacks directional speaker roles, leading to broader variability in emotional elicitation. The EPITOME corpus demonstrates distinct appraisal relevance, with dimensions like responsibility and attention influencing response quality.

In summary, changes in appraisal estimates represent emotion change in conversation. The multidimensionality and cognitive nature of appraisal theory reinforce its utility in emotion modeling and highlight corpus-specific affective dynamics.

7 Conclusion

In this paper, we introduced AppraisePLM, a multitask learning model designed to estimate appraisal dimensions and emotion categories using the crowd-EnVENT dataset. By applying our model to various conversational corpora, we leveraged its fine-grained dimensional representation of emotion to analyze affect flow—the subtle evolution of emotions within a conversation as it progresses.

Our findings demonstrate that appraisal theory provides a valuable framework for examining how emotions manifest in conversational data. While the crowd-EnVENT dataset is not a dialogue corpus, our results support the feasibility of using appraisal-based models to examine emotion dynamics in conversation. We observed not only improved appraisal estimation and emotion classification performance over baseline models but also reasonable success in appraisal-informed zero-shot emotion classification.

We identify two key applications for this approach. First, benchmarking affective conversational agents, such as those designed for emotional support or empathetic interaction, by assessing how their responses modulate appraisal dimensions. Second, informing agent response expectations when expressing specific emotions, offering insights into emotionally intelligent dialogue systems. These findings highlight the potential of AppraisePLM in advancing computational approaches to emotion modeling and affective dialogue analysis.

Limitations

This work has several important limitations that should be acknowledged. Firstly, we assume that appraisal annotation for conversations occurs in the same way as for statements or short-form text, which may not always be the case. Additionally, we presume that emotions are appraised similarly in human-human and human-machine interaction contexts, an assumption that requires further investigation. The granularity of our approach, while providing more detailed insights, also increases the potential for errors. We utilized 21 dimensions because it was possible, but future research should determine which of these dimensions are most applicable and relevant. Our current system employs power amplification of differences between values to identify interaction trends, which could be critiqued as potentially highlighting insignificant variations. A more robust approach would involve the development and use of conversationally defined and annotated corpora based on appraisal theory, given its relationship to and generality of emotion categorization systems. Lastly, the lack of longitudinal data prevents us from observing how appraisal patterns might change over time in ongoing human-machine interactions. Addressing these limitations in future research will be crucial for advancing our understanding of emotion appraisal in human-machine conversations.

References

Alexandra Balahur, Jesus M Hermida, and Andres Montoyo. 2011. Building and exploiting emotinet, a knowledge base for emotion detection based on the appraisal theory model. *IEEE transactions on affective computing*, 3(1):88–101.

Carlos Busso, Murtaza Bulut, Chi-Chun Lee, Abe Kazemzadeh, Emily Mower, Samuel Kim, Jeannette N Chang, Sungbok Lee, and Shrikanth S Narayanan. 2008. Iemocap: Interactive emotional dyadic motion capture database. *Language resources and evaluation*, 42:335–359.

Paul Ekman. 2000. Basic emotions. *Handbook of Cognition and Emotion*, page 45.

Phoebe C Ellsworth and Craig A Smith. 1988. From appraisal to emotion: Differences among unpleasant feelings. *Motivation and emotion*, 12(3):271–302.

Shutong Feng, Nurul Lubis, Christian Geishauser, Hsien-Chin Lin, Michael Heck, Carel van Niekerk, and Milica Gasic. 2022. Emowoz: A large-scale corpus and labelling scheme for emotion recognition in task-oriented dialogue systems. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 4096–4113.

Jeffrey T Hancock, Christopher Landrigan, and Courtney Silver. 2007. Expressing emotion in text-based communication. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 929–932.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2020. Deberta: Decoding-enhanced bert with disentangled attention. *arXiv preprint arXiv:2006.03654*.

Hanneke Hendriks, Bas van den Putte, and Gert-Jan de Bruijn. 2014. Changing the conversation: The influence of emotions on conversational valence and alcohol consumption. *Prevention Science*, 15:684–693.

Jan Hofmann, Enrica Troiano, Kai Sassenberg, and Roman Klinger. 2020. Appraisal theories for emotion classification in text. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 125–138.

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

Md Adnanul Islam, Md Saddam Hossain Mukta, Patrick Olivier, and Md Mahbubur Rahman. 2022. Comprehensive guidelines for emotion annotation. In *Proceedings of the 22nd ACM International Conference on Intelligent Virtual Agents*, pages 1–8.

Mounika Karna, D Sujitha Juliet, and R Catherine Joy. 2020. Deep learning based text emotion recognition for chatbot applications. In *2020 4th International Conference on Trends in Electronics and Informatics (ICOEI)(48184)*, pages 988–993. IEEE.

Allison Lahnala, Charles Welch, David Jurgens, and Lucie Flek. 2022. A critical reflection and forward perspective on empathy and natural language processing. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 2139–2158.

Andrew Lee, Jonathan Kummerfeld, Larry Ann, and Rada Mihalcea. 2024. A comparative multidimensional analysis of empathetic systems. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 179–189.

Young-Jun Lee, Chae-Gyun Lim, and Ho-Jin Choi. 2022. Does gpt-3 generate empathetic dialogues? a novel in-context example selection method and automatic evaluation metric for empathetic dialogue generation. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 669–683.

Yanran Li, Hui Su, Xiaoyu Shen, Wenjie Li, Ziqiang Cao, and Shuzi Niu. 2017. Dailydialog: A manually labelled multi-turn dialogue dataset. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 986–995.

Siyang Liu, Chujie Zheng, Orianna Demasi, Sahand Sabour, Yu Li, Zhou Yu, Yong Jiang, and Minlie Huang. 2021. Towards emotional support dialog systems. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3469–3483.

Tingting Liu, Salvatore Giorgi, Ankit Aich, Allison Lahnala, Brenda Curtis, Lyle Ungar, and João Sedoc. 2024. The illusion of empathy: How ai chatbots shape conversation perception. *arXiv preprint arXiv:2411.12877*.

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*.

Yukun Ma, Khanh Linh Nguyen, Frank Z Xing, and Erik Cambria. 2020. A survey on empathetic dialogue systems. *Information Fusion*, 64:50–70.

Kristína Machová, Martina Szabóova, Ján Paralič, and Ján Mičko. 2023. Detection of emotion by text analysis using machine learning. *Frontiers in Psychology*, 14:1190326.

Patrícia Pereira, Helena Moniz, and Joao Paulo Carvalho. 2025. Deep emotion recognition in textual conversations: A survey. *Artificial Intelligence Review*, 58(1):1–37.

Patrícia Pereira, Helena Moniz, Isabel Dias, and Joao Paulo Carvalho. 2023. Context-dependent embedding utterance representations for emotion recognition in conversations. In *Proceedings of the 13th Workshop on Computational Approaches to Subjectivity, Sentiment, & Social Media Analysis*, pages 228–236.

Verónica Pérez-Rosas, Kenneth Resnicow, Rada Mihalcea, et al. 2022. Pair: Prompt-aware margin ranking for counselor reflection scoring in motivational interviewing. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 148–158.

Robert Plutchik. 1965. What is an emotion? *The Journal of psychology*, 61(2):295–303.

Soujanya Poria, Devamanyu Hazarika, Navonil Majumder, Gautam Naik, Erik Cambria, and Rada Mihalcea. 2019a. Meld: A multimodal multi-party dataset for emotion recognition in conversations. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 527–536.

Soujanya Poria, Navonil Majumder, Rada Mihalcea, and Eduard Hovy. 2019b. Emotion recognition in conversation: Research challenges, datasets, and recent advances. *IEEE access*, 7:100943–100953.

Xiangyu Qin, Zhiyu Wu, Tingting Zhang, Yanran Li, Jian Luan, Bin Wang, Li Wang, and Jinshi Cui. 2023. Bert-erc: Fine-tuning bert is enough for emotion recognition in conversation. In *Proceedings of the AAAI conference on artificial intelligence*, volume 37, pages 13492–13500.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research*, 21(140):1–67.

Thejan Rajapakshe, Rajib Rana, Sara Khalifa, and Björn W Schuller. 2024. Domain adapting deep reinforcement learning for real-world speech emotion recognition. *IEEE Access*.

Hannah Rashkin, Eric Michael Smith, Margaret Li, and Y-Lan Boureau. 2019. Towards empathetic open-domain conversation models: A new benchmark and dataset. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5370–5381, Florence, Italy. Association for Computational Linguistics.

Yarik Menchaca Resendiz and Roman Klinger. 2023. Affective natural language generation of event descriptions through fine-grained appraisal conditions. In *Proceedings of the 16th International Natural Language Generation Conference*, pages 375–387.

Ira J Roseman and Craig A Smith. 2001. Appraisal theory. *Appraisal processes in emotion: Theory, methods, research*, pages 3–19.

Johanna Ruusuvuori. 2012. Emotion, affect and conversation. *The handbook of conversation analysis*, pages 330–349.

Tulika Saha, Vaibhav Gakhreja, Anindya Sundar Das, Souhitya Chakraborty, and Sriparna Saha. 2022. 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*, pages 2650–2656.

Kristina Schaaff, Caroline Reinig, and Tim Schlippe. 2023. Exploring chatgpt's empathic abilities. In *2023 11th international conference on affective computing and intelligent interaction (ACII)*, pages 1–8. IEEE.

Klaus R Scherer. 2005. What are emotions? and how can they be measured? *Social science information*, 44(4):695–729.

Ashish Sharma, Adam Miner, David Atkins, and Tim Althoff. 2020. A computational approach to understanding empathy expressed in text-based mental health support. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5263–5276.

Marcin Skowron and Georgios Paltoglou. 2011. Affect bartender—affective cues and their application in a conversational agent. In *2011 IEEE Workshop on Affective Computational Intelligence (WACI)*, pages 1–7. IEEE.

Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu. 2020. Mpnet: Masked and permuted pre-training for language understanding. *Advances in neural information processing systems*, 33:16857–16867.

Enrica Troiano, Laura Oberländer, and Roman Klinger. 2023. Dimensional modeling of emotions in text with appraisal theories: Corpus creation, annotation reliability, and prediction. *Computational Linguistics*, 49(1):1–72.

Enrica Troiano, Laura Oberländer, Maximilian Wegge, and Roman Klinger. 2022. x-envent: A corpus of event descriptions with experiencer-specific emotion and appraisal annotations. In *13th International Conference on Language Resources and Evaluation Conference, LREC 2022*, pages 1365–1375. European Language Resources Association (ELRA).

Joshua D. Wondra and Phoebe C. Ellsworth. 2015. An appraisal theory of empathy and other vicarious emotional experiences. *Psychological Review 122.3*, pages 4–11.

Jiamin Yang and David Jurgens. 2024. Modeling empathetic alignment in conversation. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 3127–3148.

Xi Yang, Marco Aurisicchio, and Weston Baxter. 2019. Understanding affective experiences with conversational agents. In *proceedings of the 2019 CHI conference on human factors in computing systems*, pages 1–12.

Peixiang Zhong, Chen Zhang, Hao Wang, Yong Liu, and Chunyan Miao. 2020. Towards persona-based empathetic conversational models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6556–6566.

Naitian Zhou and David Jurgens. 2020. Condolence and empathy in online communities. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 609–626.

A A Review of Emotion Annotations in Conversational Corpora

Table 3 shows the wide range of contemporary emotion classification and affect-annotated datasets. We see that there is little consistency in the emotion labelling, dimensionality, representation and expectation of emotion as a latent property of interaction. Standard

B Crowd-EnVent Dataset and Appraisal Definitions

B.1 Dataset Description

The crowd-EnVENT dataset consists of 6,600 instances of emotion-inducing event descriptions. Each event is annotated using 21 appraisals as well as the stable properties of text authors (demographics, personality traits). The dataset also captures categorical emotion. The data was collected from English native speakers from diverse backgrounds, not limited to college students. The dataset is annotated and validated by external crowdworkers who read the descriptions and inferred the original appraisals.

The distribution of labels for this corpus are provided in Table 4.

B.2 Appraisal Definitions

The crowd-EnVENT corpus highlights 21 appraisal dimensions, which can be categorized based on four affective state responses as established by Scherer (2005). These categories, which the paper and subsequent model treat as evaluation objectives, can be described as:

1. **Relevance**: Relevance may be determined as a combination of novelty, intrinsic pleasantness, and importance towards an experiencer's goal or objective; i.e. the relevance appraisal criterion determines the experiencer's familiarity with the event responsible for the emotion as well as linguistic cues about its alignment with the expected goals and outcomes.

2. **Implication**: Implication is seen as a combination of the causality of the agent, conduciveness of the situation towards the goal, anticipation of the consequence of the event, and the relative urgency of response to a given situation.

3. **Coping**: Coping as an appraisal objective examines how an experiencer handles the sit-

Dataset Name	Type	Annotation	Layer	Domain	Size
EmpatheticDialogues	Conversations	32 emo cat	dialogue	General Empathy	24,850
Epitome	Reddit	3 2pt dim	response	Mental Health	10,143
WASSA	Conversations	3 5pt dim	response	Empathic News Reactions	12,601
Condolence	Reddit	Distress/support	comment	Online Support	14.1M
ALoE	Reddit	Empathy levels	post	Mental Health	10,000+
ESConv	Conversations	Support strategies	turn	Mental Health	1,053
DailyDialog	Conversations	5 emo cat	turn	General Domain	13,118
MELD	Conversations (M.P)	3 senti + 7 emo cat	turn	Movie Dialogues	13,708
IEMOCAP	Conversations	3 5pt dim + 7 emo cat	turn	Multimodal Interaction	10,039
EmoWOZ	Conversations	3 senti + 7 emo cat	turn	Task Oriented	11,434
Twitter-Customer	Tweet-Response	3 5pt dim + 7 emo cat	turn	Customer Service	9,000+

Table 3: Overview of benchmark conversation corpora with emotion or affect annotations, highlighting the disparity between them. Corpora marked in **bold** are studied extensively in this paper. *cat* refers to categorical labels; *npt dim* refers to an n point dimensional Likert scale; *senti* refers to sentiment categories; *emo* refers to emotion categories. The disparity in emotion and affect annotations is apparent, depending on source and context. M.P refers to multi-party conversations. Size is measured in number of dialogues/conversations

Label	Frequency
5	1197
1	1034
4	859
2	627
3	603

Table 4: Distribution of labels from 1 to 5 in crowd-EnVENT, showing the label skew towards 1 and 5.

uation both in terms of their experience of control over the situation as well as the adjustment “felt necessary” by the experiencer. Some formalisms of the coping objective account for the experiencer’s “power” during the experience. Troiano et al. (2023) replaces this with the dimension of ‘effort’ instead.

4. **Normative Significance**: The normative significance of an event or situation is the degree of conformity that the response to that situation has to personal ideals as well as with external laws or norms, which may be based on the experiencer’s social or cultural environment.

These definitions are based on two critical underpinnings: that the person examining the event is also contextually involved in the event and outcome, and that this is a retrospective cognitive outcome of a given event. Given the methods adopted by Troiano et al. (2023) for curating the corpus, such an assumption is justified. However, in its applicability to dialogue, a principally reformulated set of appraisals would have to be determined. For example, the event in question could be the statement made by another conversation participant, or the scoping of *other responsibility* and *others’ control* would be limited to the other conversation participant, and any individual external to the conversation be treated as a part of the "situation". However, the suitability of appraisals is beyond the scope of a feasibility study and is a promising avenue for future work given that this work establishes the noticeable enrichment to dialogue done by an appraisal based approach.

C AppraisePLM: Implementation Details

C.1 Experimental Setup

All experiments were conducted using the PyTorch deep learning framework in conjunction with the Hugging Face `transformers` library. Model training was performed on a system equipped with four NVIDIA RTX 2080 Ti GPUs, employing mixed precision training (FP16) to enhance computational efficiency and memory utilization.

C.2 Dataset and Preprocessing

Text inputs were tokenized using a maximum sequence length of either 128 or 256 tokens, depending on the specific model configuration. No additional preprocessing steps, such as lowercasing, stopword removal, or normalization, were applied.

For the appraisal prediction task, appraisal values were directly used as regression labels. In the emotion classification task, when incorporating appraisal features, these values were projected through a linear transformation to ensure dimensional compatibility before classification.

Objectives	Dimensions	Definition
Relevance	Suddenness	The situation or event was sudden or abrupt to the experiencer.
	Familiarity	The situation or event was familiar to the experiencer.
	Predictability	The experiencer could have predicted that the event would occur or the situation would arise.
	Attention	The experiencer had to pay attention to the situation.
	Not Consider	The experiencer tried to shut the situation out of their mind.
	Pleasantness	The situation or event was a pleasant experience.
	Unpleasantness	The situation or event was an unpleasant experience.
	Goal Relevance	The experiencer expected the event to have important consequences for them.
Implication	Self Responsibility	The experiencer believes that the event occurred because of their behaviour.
	Other Responsibility	The experiencer believes that the event occurred because of somebody else's behaviour.
	Situational Responsibility	The experiencer believes that the event occurred because of circumstances external to them, such as chance, special circumstances, or natural forces.
	Goal Support	The experiencer expected a positive outcome of the event for them (this is different from goal relevance as an "important" event does not necessitate the belief of a positive outcome).
	Consequence Anticipation	The experiencer anticipated the consequences of the event.
	Urgency	The experiencer believes that the event requires an immediate response.
Coping	Own Control	The experiencer believes they can influence the ongoing of the event.
	Others' Control	The experiencer believes that someone other than them was influencing the ongoing event.
	Chance Control	The experiencer believes that the situation was the result of outside influences of which nobody had control.
	Anticipated Consequence	The experiencer anticipated the outcome of the event based on their past experiences.
	Effort	The experiencer believes that the event required additional ("a great deal of") effort to deal with.
N.S.	Standards	The event clashed with the experiencer's standards or ideals.
	Social Norms	The actions that produced the event violated laws or socially accepted norms.

Table 5: With appraisal objectives defined, each appraisal dimension can be examined based on the appraisal objective they contribute to. The appraisal dimensions can be defined based on the questions asked to annotators to examine a specific situation or event. **N.S.** refers to the Normative Significance objective.

	Text → Appr.			Text → Emo.			Text + Appr. → Emo.		
Model	MSE ↓	MAE ↓	RMSE ↓	P ↑	R ↑	F1 ↑	P ↑	R ↑	F1 ↑
Troiano et. al. (2022)	1.97	3.22	1.40	0.62	0.59	0.59	0.62	0.60	0.60
RoBERTa-large	1.62	2.96	1.11	0.62	0.59	0.60	0.66	0.67	0.67
T5-large	1.64	2.77	1.12	0.63	0.61	0.61	0.63	0.65	0.66
MPNet-base	1.49	2.68	**1.08**	0.66	0.64	0.62	0.64	**0.70**	0.70
DeBERTa-large	**1.44**	**2.60**	**1.08**	0.67	0.65	0.66	**0.73**	**0.71**	**0.71**

Table 6: Performance of the AppraisePLM architecture for the appraisal regression and emotion classification models on regression. Categorical emotion detection is done in two modes; with only text (**Text → Emo.**) and both text and appraisals (**Text + Appr. → Emo.**) on the crowd-ENVENT corpus. *Baseline* refers to the baseline RoBERTa-large regressor used in Troiano et. al. (2022). Per-appraisal performance and comparisons for **Text → Appr.** are provided in Appendix C.

C.3 Model Architectures and Training

C.3.1 Appraisal Prediction Model

Four pretrained language models (PLMs) were utilized: RoBERTa-large, DeBERTa-large, MPNet-base, and T5-large. Each PLM was augmented with a multihead attention layer comprising 8 attention heads and 2 layers, with a hidden size equal to that of the PLM embedding layer. The output of the attention mechanism was subsequently passed through a fully connected layer for final label prediction.

Optimization was performed using the AdamW optimizer with a linear learning rate decay schedule. The models were trained using Mean Squared Error (MSE) loss with balanced class weighting. An attention weight decay of 1×10^{-3} was applied, and a dropout rate of 0.01 was employed between sequential layers, except for T5, where a dropout rate of 0.001 yielded superior performance. To mitigate exploding gradients, gradient clipping was applied after the attention layer. Training was conducted for a maximum of 10 epochs, with early stopping enforced using a patience of 3 epochs. On average, model convergence was achieved in 4.6 epochs.

C.3.2 Emotion Detection Model

Two variations of the emotion detection model were developed: a text-only model and a text + appraisal model. The text-only model followed the architecture: PLM embeddings → attention layer → classification layer. The text + appraisal model incorporated appraisal features by concatenating them with text-based embeddings after passing them through a linear projection layer to ensure dimensional alignment before classification.

For classification, cross-entropy loss with balanced class weighting was utilized. Model performance was evaluated using Precision, Recall, and F1-score.

C.4 Hyperparameter Selection

A comprehensive grid search was conducted to determine optimal values for batch size, maximum sequence length, dropout rate, and attention weight decay. The final hyperparameter selections were as follows:

- **Batch size:** 16, except for RoBERTa, where a batch size of 8 was optimal.
- **Maximum sequence length:** 128, except for RoBERTa, where a length of 256 performed best.
- **Dropout rate:** 0.01, except for T5, where 0.001 was more effective.
- **Attention weight decay:** 0.01.
- **Learning Rate:** 2e-5

All models employed a linear decay learning rate schedule, with gradient clipping applied after the attention layer to prevent gradient explosion.

C.5 Evaluation and Baselines

For appraisal prediction, model performance was assessed using Mean Squared Error (MSE), Mean Absolute Error (MAE), and Root Mean Squared Error (RMSE). Emotion classification performance was evaluated using Precision, Recall, and F1-score.

As a baseline, our models were compared against a simple RoBERTa classifier released by the dataset authors. This baseline does not incorporate an attention mechanism and can be interpreted as an ensemble of single-task models rather than a fully integrated multitask model.

C.6 Statistical Analysis

To determine statistical significance, paired t-tests and ANOVA tests were conducted to compare model performance. These tests were performed both across different PLM architectures and before and after hyperparameter tuning. The results demonstrated statistically significant improvements in model performance following hyperparameter optimization.

C.7 Reproducibility Considerations

To ensure the reproducibility of our results, random seeds were set for model initialization, data shuffling, and optimizer state. Additionally, all hyperparameters, training procedures, and evaluation metrics are comprehensively documented in this report. All models were trained under controlled computational conditions to facilitate consistency and comparability across experimental runs.

D AppraisePLM: Performance Analysis Details

Since different datasets had a differing number of labels and we did not employ a semantic space implementation, we perform label folding in order to evaluate the AppraisePLM model. Here, we detail the emotion mapping used. Given the label skew in the EmoWOZ and DailyDialog datasets, the emotion detection metrics were computed *excluding* the neutral emotion label.

EMPATHETICDIALOGUES (ED) Since the ED corpus has 32 fine-grained emotions to the 13 (12 without no-emotion), we had to label fold from ED into crowd-EnVENT, i.e. predictions made by AppraisePLM would be considered true positive for more than one label of the ED corpus. We folded by synonymy, where each crowd-EnVENT emotion label was mapped as follows:

crowd-EnVENT	**ED**
anger	angry, annoyed, furious, disappointed
boredom	**None**
disgust	disgusted
fear	afraid, anxious, apprehensive, terrified
guilt	guilty
joy	joyful, excited, content
no-emotion	**None**
pride	proud
relief	prepared, hopeful
sadness	sad, devastated
shame	ashamed, embarrassed
surprise	surprised
trust	trust, grateful, faithful, caring
Removed	confident, nostalgic, sentimental

Table 7: Emotion categories and their associated terms from the crowd-EnVENT to the ED corpus

We do preserve the labels for qualitative tessting, as can be seen for Figure 3.

DailyDialog (DD) uses Plutchik's emotion labels: anger, disgust, fear, happiness, sadness, and surprise. However, from crowd-EnVENT, it is missing the labels *boredom*, *guilt*, *shame*, *trust*, *pride*, and *relief*. Therefore, we had to label fold from DD out of crowd-EnVENT, i.e. one or more predictions made by AppraisePLM would be considered true positive for the same label of the DD corpus. We folded here by affective synonymy, where each crowd-EnVENT emotion label was mapped as follows:

crowd-EnVENT	**DD**
anger	anger
boredom	no emotion
disgust	disgust
fear	fear
guilt	sadness
joy	happiness
no-emotion	no emotion
pride	happiness
relief	happiness
sadness	sadness
shame	sadness
surprise	surprise
trust	happiness

Table 8: Emotion categories and their corresponding mapped categories from crowd-EnVENT to the DD corpus.

EmoWOZ EmoWOZ uses a novel emotion labelling scheme tailored to task-oriented dialogues, with seven emotion labels: Neutral, Satisfied, Dissatisfied, Excited, Apologetic, Fearful, and Abusive. Interestingly, this system is adopted from the OCC emotion annotation schema (), which has its roots in early cognitive emotion theory. In fact, appraisal dimensions could theoretically be directly mapped to certain labels. However, practically, due to the presence of an overwhelmingly large category of no emotion, and the difference in source corpus of event descriptions and rarget corpus of textual instructional conversation, we do not use a semantic space representation of the OCC model, though we leave it up to future work. Instead, we follow an OCC mapping elicited by Steunebrink, Dastani, and Meyer (2009, Figure 2) .

The model performed worst on this dataset, partially because of the label skew, partially because the OCC mapping from crowd-EnVENT to EmoWOZ is less than satisfactory. The goal of an appraisal-based model is to have an interpretable semantic space adaptable to the affective lexicon

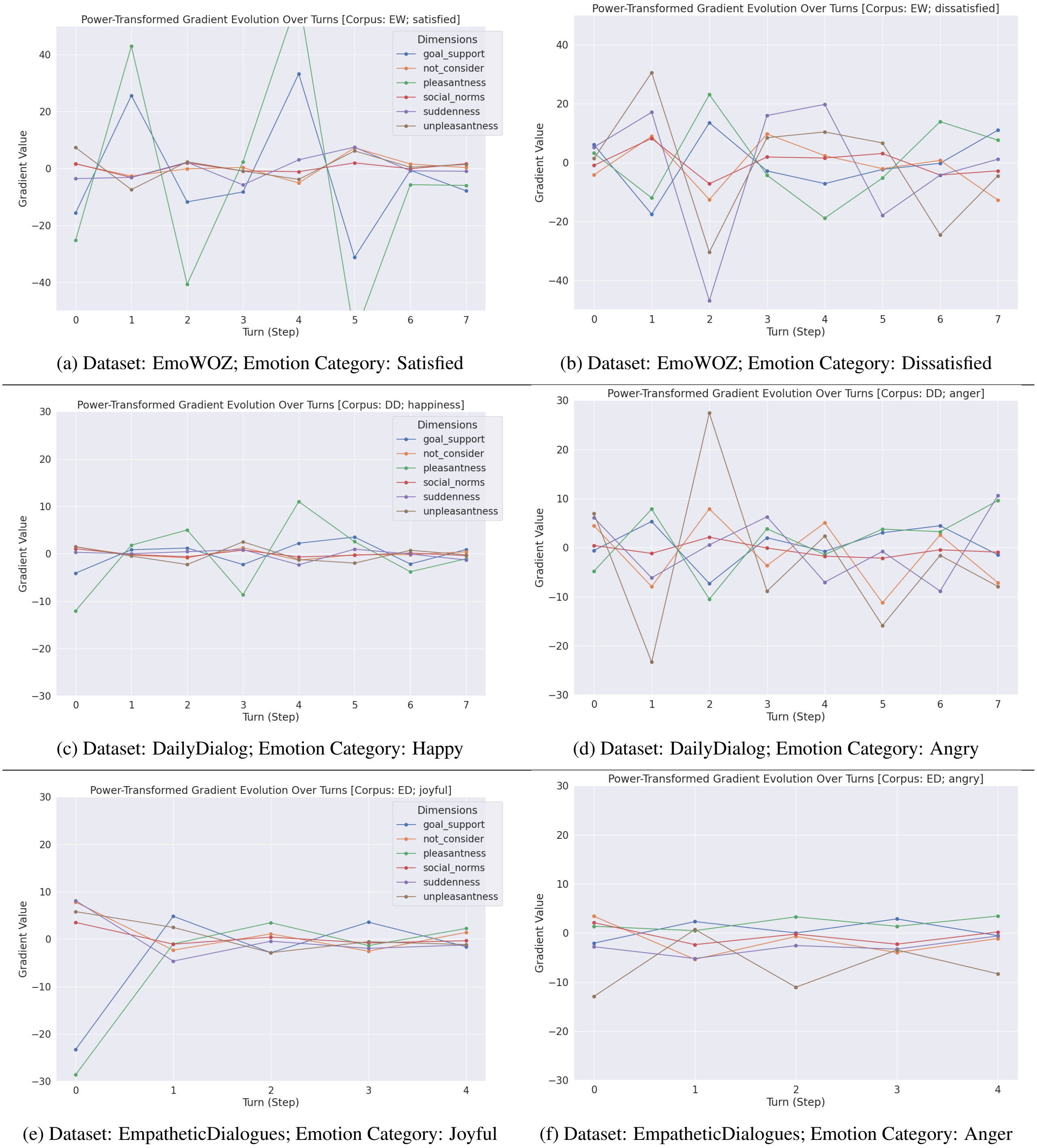
crowd-EnVENT	**EmoWOZ**
anger	dissatisfied
boredom	**None**
disgust	abusive
fear	fearful
guilt	apologetic
joy	satisfied
no-emotion	no emotion
pride	satisfied
relief	satisfied
sadness	fearful
shame	apologetic
surprise	**None**
trust	satisfied

Table 9: Emotion categories and their corresponding mapped categories from the crowd-EnVENT to the EmoWOZ corpus.

of a domain in order to avoid doing label mapping or using an uninterretible semantic space instead.

E Appraisal Distributions by Emotion Label for Conversational Corpora

In Section §4, we presented the mean appraisal estimates of emotion in the EmoWOZ and DailyDialog dataset. Figure 6 and 7 show the distribution of appraisal values by emotion category for crowd-EnVENT and estimates for ED respectively.



(a) Dataset: EmoWOZ; Emotion Category: Satisfied

(b) Dataset: EmoWOZ; Emotion Category: Dissatisfied

(c) Dataset: DailyDialog; Emotion Category: Happy

(d) Dataset: DailyDialog; Emotion Category: Angry

(e) Dataset: EmpatheticDialogues; Emotion Category: Joyful

(f) Dataset: EmpatheticDialogues; Emotion Category: Anger

Figure 5: The average gradient of change between appraisal estimates for an average number of turns isolated by emotion category. Each turn shows the gradient, i.e. the amplified power difference between the speaker and listener across conversational turns. We see that the way corpora expect models to handle the same emotion differs greatly based on the dataset and context. The legend is shared across all graphs.

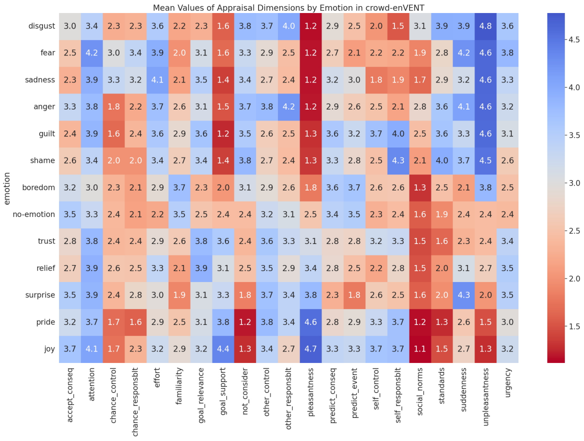


Figure 6: Average estimate of each appraisal from the crowd-enVENT test corpus. The emotion labels are ordered by *pleasantness* from low (red) to high (blue).

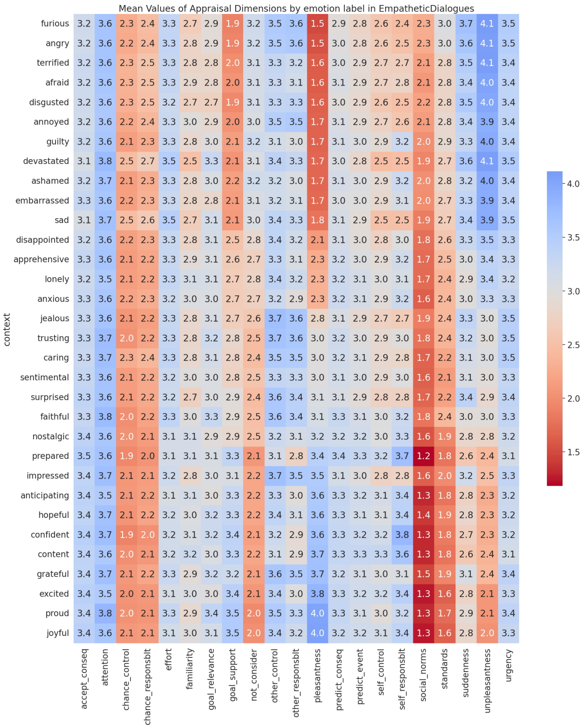
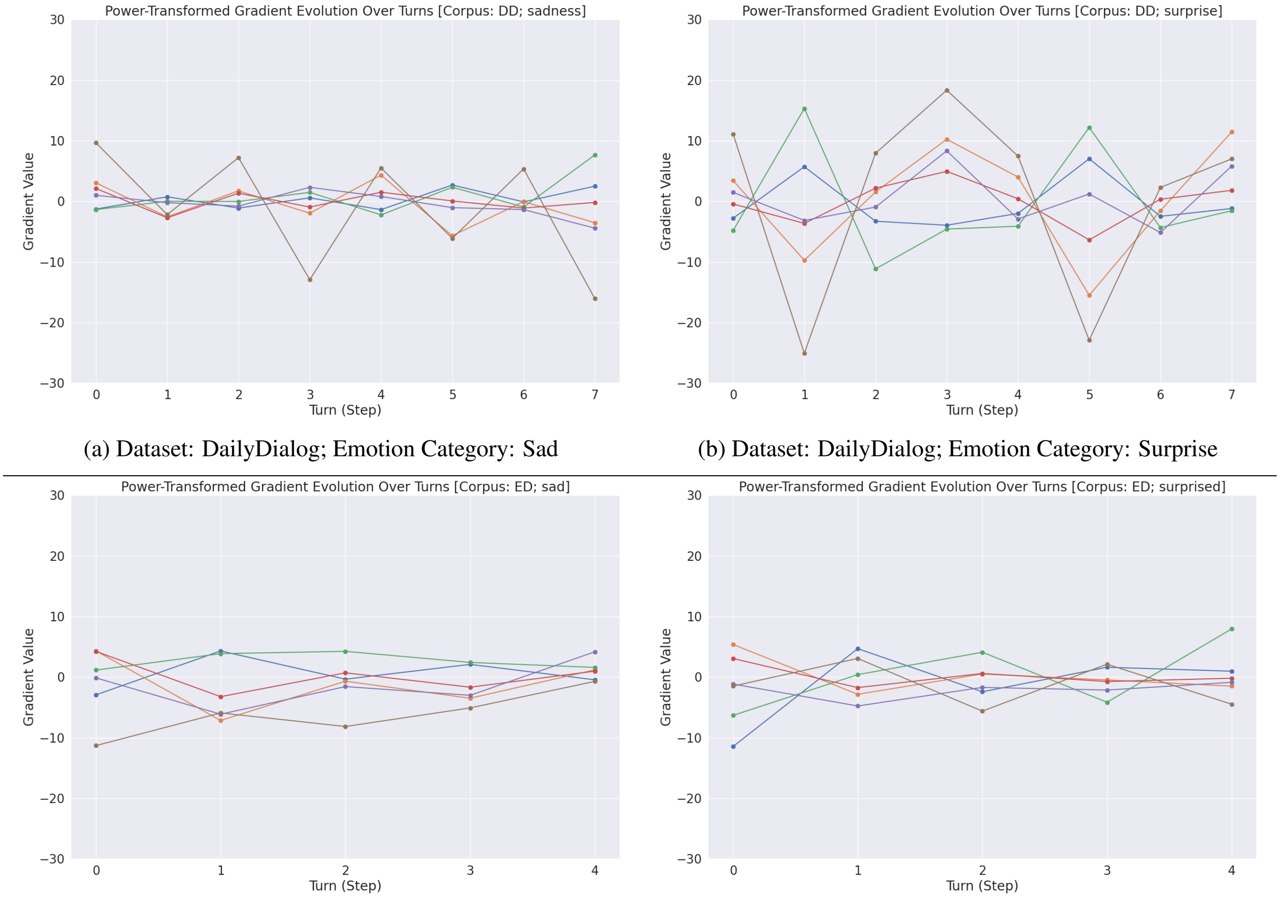


Figure 7: Average estimate of each appraisal from the EMPATHETICDIALOGUES test corpus using the best performing AppraisePLM DeBERTa-large. The emotion labels are ordered by *pleasantness* from low (red) to high (blue).



(a) Dataset: DailyDialog; Emotion Category: Sad

(b) Dataset: DailyDialog; Emotion Category: Surprise

(c) Dataset: EmpatheticDialogues; Emotion Category: Sad

(d) Dataset: EmpatheticDialogues; Emotion Category: Surprise

Figure 8: Comparing gradients from some other emotion labels in the EMPATHETICDIALOGUES and DailyDialog corpora.