Bringing Characters to Life - Emotion-Based Dialogue Generation

Charishma Tamarana, Kanchan Naik, Mrunali Katta, Soumya Challuru Sreenivas, Vaishnavi Nanduri, Yashasvi Kanchugantla
San Jose State University
Applied Data Intelligence

***Abstract*—This emotion-aware dialogue generation system aims to produce responses that are not only contextually relevant but also emotionally appropriate which makes conversational agents more human-like and engaging. Here in this project, we explore personality-affected emotion generation in dialogue systems which focus on how a speaker's current emotion, underlying mood, and stable personality traits influence the emotional tone of future utterances. We use the Personality-Affected Emotion Lines Dataset (PELD), which is derived from the Friends TV series, to train models capable of predicting the emotion of a speaker's response given the preceding dialogue and personality context. Emotions are represented in the Valence-Arousal-Dominance (VAD) space, and mood dynamics are modeled through VAD-based shifts influenced by prior emotion and personality. Our work demonstrates how integrating psychological factors—personality (OCEAN traits), mood, and emotion—into the generation pipeline can improve the emotional realism of generated dialogues.**

***Index Terms*—Emotion Prediction, Dialogue Systems, Personality Modeling, VAD, GRU, BERT, PELD Dataset, NLP**

I. Introduction

Recent advancements in conversational AI have enabled significant improvements in the syntactic and semantic quality of generated dialogue. However, emotional appropriateness and speaker-specific personality traits remain underexplored, yet critical, for natural and human-like interactions. In emotionally intelligent systems, understanding the interplay between a character's current emotion, underlying mood, and stable personality is essential for producing responses that feel authentic and relatable.

This project focuses on the task of emotion prediction in dyadic conversations, with a goal of extending it towards emotion-based dialogue generation. We use the Personality-Affected Emotion Lines Dataset (PELD), derived from scripted conversations in the Friends TV show, which contains utterances annotated with emotion labels and associated with personality profiles based on the OCEAN model. We model each speaker's emotion using a Valence-Arousal-Dominance (VAD) representation, and account for mood transitions influenced by both recent emotional states and inherent personality.

Our approach integrates BERT-based contextual embeddings, VAD representations, and gated fusion techniques to effectively model the nuanced dynamics of emotion evolution across conversations. This allows us to make emotion predictions that are not only context-aware but also character-consistent. By aligning with psychological insights, our model aims to bring characters to life with emotionally coherent dialogue, setting a foundation for future work in emotionally intelligent dialogue generation.

II. Motivation

Human-like dialogue needs more than just coherent language—it demands emotional intelligence as well. The traditional dialogue systems often focus on contextual relevance but neglect emotional consistency and personality-driven variation in emotional expression. These gaps lead to interactions that feel robotic, emotionally inconsistent, or insensitive to a speaker's character.

Psychological studies suggest that emotion is not only reactive but also influenced by long-term personality traits and evolving mood states. While many conversational AI systems predict emotions solely based on recent context, they fail to model how a speaker's personality (OCEAN traits) biases their emotional tendencies or how mood transitions evolve during a dialogue.

This project addresses these challenges by introducing personality-affected emotion prediction, using the PELD dataset, which annotates utterances from the Friends TV show with both emotion labels and personality vectors. By modeling emotion through Valence-Arousal-Dominance (VAD) vectors and incorporating personality into mood transition dynamics, we aim to predict speaker emotions that are not just appropriate in context—but also consistent with their psychological profile.

Therefore, our motivation is to bring emotionally believable characters to life in dialogue systems by integrating linguistic, affective, and personality cues into a unified framework.

III. Related Work

A. Emotion Recognition and Response Generation:

Dialogue systems have increasingly focused on understanding and generating emotional content. On the understanding side, emotion recognition in conversations is well-studied, with models leveraging conversational context to classify speaker emotions (Majumder et al., 2019). For example, recurrent and graph neural architectures capture inter-turn dependencies to

detect emotions in dialogue (e.g. DialogueRNN (Majumder et al., 2019), DialogueGCN (Hazarika et al., 2020)). On the generation side, researchers have developed methods to condition chatbot responses on a target emotion. The Emotional Chatting Machine introduced an encoder–decoder that conditions on specified emotion tags to produce emotionally colored responses (Zhou et al., 2018). Similarly, affect-controlled generation models use latent affective variables to guide response decoding (Zhou et al., 2018). These approaches successfully render a desired emotion (e.g. making a reply happy or sad) (Zhou et al., 2018), but they require deciding which emotion the system should express at each turn.

B. Empathetic and Emotional Dialogue Systems:

A prominent line of work is empathetic response generation, where the system responds in an emotionally appropriate way to user input. Rashkin et al. (Rashkin et al., 2019) released the EmpatheticDialogues dataset to train models that respond with empathy. Subsequent models predict an emotion for the next response based on the user's emotion and dialogue context (Rashkin et al., 2019). For instance, the MoEL model uses a mixture-of-experts to softly select among multiple "empathic listener" decoders tuned to different emotions (Rashkin et al., 2019). Other architectures (e.g. transformer-based EmpTransfo and the multi-factor CoMAE model) also generate responses conditioned on inferred emotions (Song et al., 2020). These systems improve conversational empathy, but they typically infer emotions only from context, without accounting for a speaker's inherent emotional tendencies (Rashkin et al., 2019; Song et al., 2020). As a result, the selected emotions may vary unpredictably, sometimes leading to inconsistent or inappropriate emotional tones across turns (Song et al., 2020). Recent extensions combine empathetic responding with predefined personas (Zhong et al., 2020), yet a stable trait-based emotional profile for the speaker is generally absent.

C. Personality-Aware NLP and Dialogue

Incorporating Personality into Dialogue: It is well established that a speaker's personality influences their communication style and emotional expressions (Mairesse & Walker, 2007; Mehrabian, 1996). Early work in embodied conversational agents showed that integrating personality models with emotion models yields more lifelike, coherent behavior (André et al., 2000; Ball & Breese, 2000). For example, André et al. (1999) implemented animated characters whose dialogue and affective responses were modulated by personality traits (André et al., 2000). Similarly, Ball and Breese (2000) explored how an agent's personality profile could affect its emotional reactions and dialogue decisions (Ball & Breese, 2000).

In the NLP community, personality has been introduced via "persona" profiles in open-domain chatbots. Li et al. (Li et al., 2016) pioneered a persona-based neural conversational model that conditions responses on speaker profile facts, ensuring more consistent behavior. This idea was extended with the Persona-Chat dataset (Zhang et al., 2018), enabling models to learn stylistic quirks of a given persona and avoid contradictory responses. While these persona-based systems improve consistency, they typically represent personality as textual profile information rather than psychological trait dimensions like Big Five.

Personality and Language Generation: Research has also examined controlling language style along psychological trait dimensions. Mairesse and Walker (Mairesse & Walker, 2007) developed a generation system that adjusts linguistic style (e.g. verbosity, formality) to reflect desired Big Five personality scores, demonstrating that subtle lexical and syntactic choices can project traits such as Extraversion or Neuroticism. Such style-control techniques highlight the feasibility of injecting personality into NLP systems.

Furthermore, large-scale analyses of social media confirm that language use strongly correlates with authors' personalities (Mohammad, 2018; C. Wen et al., 2021). For instance, certain words, sentiment patterns, or emoji usage can predict an individual's OCEAN traits (Mohammad, 2018). Leveraging these insights, recent models perform automatic personality recognition from text or dialogue (e.g. using deep contextual embeddings (Zhong et al., 2020)), which can provide a personality context for dialogue systems. Zhong et al. (Zhong et al., 2020) took a step toward integrating personality with empathy by personalizing empathetic responses to a given persona. However, even these approaches do not explicitly model how a fixed personality trait influences the dynamics of emotional state transitions over a conversation.

Our work addresses this gap by treating personality as a conditioning factor for emotion evolution, rather than only for static style or factual consistency.

D. Valence–Arousal–Dominance (VAD) and OCEAN Representations

Dimensional Emotion Representation: Emotions can be represented categorically (e.g. joy, anger, fear) or along continuous dimensions. Psychology research suggests three principal dimensions – Valence (pleasure vs. displeasure), Arousal (activation or intensity), and Dominance (sense of control) – that jointly characterize affective states (Scherer, 2005). This VAD space (also known as PAD) provides a graded representation of emotions, capturing nuances like mood intensity (Mehrabian, 1996).

NLP researchers have embraced dimensional models by creating resources such as EmotionBank, a text corpus annotated with valence, arousal, and dominance scores (Mohammad, 2018), and lexicons of words mapped to VAD values (Mohammad, 2018). Modeling emotion as a continuous vector enables treating emotion prediction as a regression task (Mohammad, 2018), often yielding richer predictions than discrete labels. Recent work has even transformed categorical emotion datasets into the VAD space to exploit correlations between emotions (Mohammad, 2018).

E. Linking Personality and Mood:

Crucially, personality and VAD representations can be unified in a common framework. Big Five (OCEAN) personal-

ity traits (Mehrabian, 1996) – Openness, Conscientiousness, Extraversion, Agreeableness, Neuroticism – have been analytically mapped into the VAD temperament model (Mehrabian, 1996). Mehrabian's seminal studies showed that an individual's personality scores predict their baseline position in the valence/arousal/dominance space (e.g. an extrovert tends toward higher arousal, a neurotic person toward lower valence) (Mehrabian, 1996).

This provides a psychological foundation for treating personality as a prior that biases emotional state. Prior works in affective computing and HCI have leveraged this idea: Itoh et al. (Han et al., 2018) and Han et al. (Han et al., 2018) designed robot emotion systems where a stable personality vector modulates a mood transition process. In these models, the robot maintains an internal mood (a point in VAD space) that drifts over time influenced by interactions, and the personality acts as a weight or constraint on these mood dynamics (Han et al., 2018). Similarly, Masuyama et al. (Masuyama et al., 2020) extend a VAD-based emotional model with personality factors and associative memory to personalize a robot's emotional reactions. These approaches demonstrate that mood can serve as an intermediate layer between long-term personality and momentary emotions: personality defines a characteristic range or tendency in mood, and mood in turn affects immediate emotional expressions (Masuyama et al., 2020; Mehrabian, 1996).

Our work draws inspiration from such architectures, bringing the concept of personality-conditioned mood transitions into the dialogue generation domain. By predicting a speaker's future emotion with a mood state that evolves under personality constraints, we build on the idea that emotional consistency arises from an interplay of trait and state.

F. Comparative Approaches:

To our knowledge, only a few recent studies explicitly combine personality with emotion prediction in conversational AI. Wen et al. (C. Wen et al., 2021) introduce a model that automatically selects a response emotion by simulating emotion transitions influenced by a given personality, showing initial gains in consistency. Our proposed method extends this line of research by more deeply integrating OCEAN-based personality context with VAD-based mood dynamics for emotion generation. In contrast to prior empathetic systems that choose emotions reactively from context alone (Rashkin et al., 2019), our approach provides a principled mechanism to maintain emotional coherence through a stable personality-conditioned mood trajectory.

IV. Methodology

A. Preliminaries

In any conversation three things play a major role - current emotion, mood, and personality. We can leverage the understand of interaction between these three to predict the emotion of the next dialogue in a conversation and potentially predict new dialogue.

1) Emotion: We are considering a basic set of emotions: Anger, Disgust, Fear, Joy, Neutral, Sadness, Surprise. Each emotion can be represented in terms of Valence, Arousal, and Dominance (VAD). In this project we are representing emotion in the VAD vector of dimension three. Each of these dimension represents if an emotion is positive or negative and intensity.(Z. Wen et al., 2024)

TABLE I
VAD meanings (Z. Wen et al., 2024)

Factor	***Description***
Valence	The degree of pleasure and displeasure
Arousal	Level of mental activity from low engagement to ecstasy
Dominance	Extent of control felt in situation

TABLE II
Emotion representation in VAD space (Z. Wen et al., 2024)

Emotion	***VAD***
Anger	(-0.51, 0.59, 0.25)
Disgust	(-0.60, 0.35, 0.11)
Fear	(-0.62, 0.82, -0.43)
Joy	(0.81, 0.51, 0.46)
Neutral	(0.00, 0.00, 0.00)
Sadness	(-0.63, -0.27, -0.33)
Surprise	(0.40, 0.67, -0.13)

2) Mood: In conversation, mood shifts based on emotion, previous mood, and personality. It is important to capture mood shift based on a new utterance. In our system, mood is represented in terms of M_1, M_2, M_3,M_4, and neutral mood, which is the origin of the VAD space. All moods can be represented with states of pleasantness and energy, eg, 1st - Positive, pleasant feelings and Energetic, 2nd - alert, unpleasant feelings with Energetic, alert, 3rd - Unpleasant with tired and calm, 4th - pleasant with calm (all by considering dominance as 0.0 at start). If we categorize mood in one of these categories it can be represented as follow. (Z. Wen et al., 2024)

TABLE III
Mood States Represented by Valence, Arousal, and Dominance (VAD) (Z. Wen et al., 2024)

Mood State	**VAD (Valence, Arousal, Dominance)**
M_1	(1.0, 1.0, 0.0)
M_2	(-1.0, 1.0, 0.0)
M_3	(-1.0, -1.0, 0.0)
M_4	(1.0, -1.0, 0.0)
Neutral	(0.0, 0.0, 0.0)

3) Personality: Personality of any character or individual is the same and can help in understanding long term pattern of thoughts and conversation behavior. Personality vector used in the project is defined in-terms of OCEAN factors. In order to compute mood shift, we need to bring personality into the same space as emotion and mood. To do so we need to project personality in-terms of VAD vector. This can be

TABLE IV
OCEAN PERSONALITY TRAITS (Z. WEN ET AL., 2024)

Factor	*Description*
Openness	Open-minded, imaginative, and sensitive.
Conscientiousness	Scrupulous, well-organized.
Extraversion	The tendency to experience positive emotions.
Agreeableness	Trusting, sympathetic, and cooperative.
Neuroticism	The tendency to experience psychological distress.

done by representing the VAD space using OCEAN factors. eg. (Z. Wen et al., 2024)

$$PV = 0.21E + 0.59A + 0.19N$$
$$PA = 0.15O + 0.30A - 0.57N$$
$$PD = 0.25O + 0.17C + 0.60E - 0.32A$$

B. Dataset

We used the Personality-Affected Emotion Lines Dataset (PELD) dataset to build this project. This data set is constructed from scripts of the Friends TV series. It is a dyadic conversational data set, which means it is a conversation between two people. PELD dataset contains a total of 6,510 samples, where each row contains characters and their utterances in a conversation. It contains 3 consecutive utterances (u1, u2, u3) of characters. u1 and u3 are utterances of character 1, and u2 is an utterance of the second character. Also, a fixed five-dimensional OCEAN personality vector for the speaker and an emotion label. The emotional label consists of seven classes. Classes: neutral, joy, surprise, anger, fear, sadness, disgust. This dataset is very unbalanced, having very large instances of neutral samples and very few representations for rare emotions like disgust and fear.

TABLE V
EMOTION DISTRIBUTION IN THE PELD DATASET (BASED ON EMOTION_3 COLUMN)

Emotion	Count
Neutral	2771
Joy	1123
Anger	858
Surprise	634
Sadness	493
Fear	487
Disgust	144

1) Speaker Selection and Personality Assignment: This project focus on six main characters. For each, there is a standard OCEAN profile vector associated, which is derived from physiological studies on these characters.

C. Train/Test Split

Samples are stratified and split 80:20 into training (5,208 triples) and test (1,302 triples) sets. Further training dataset is spitted in 90:10 split, representing training and validation set.

D. Pre-Processing

1) Sequence Formatting: we pretended [CLS] before each utterance, and terminate each with [SEP]. making it suitable for BERT input processing. BERT tokenier is used to tokenize each utterance in fix length tokens.

E. Model Architecture

1) Input: For input we are encoding utterances in VAD space ¡Tokens, V, A, D¿, which reflect affective content of the speech. Personality traits are also encoded in the VAD space, representing an individual's general affective disposition, which influences mood transitions and emotion classification.

2) BERT based context embeddings: This input sequence is fed through BERT's stack of Transformer layers, each comprising multi-head self-attention and a feed-forward network. BERT generated deep context sensitive hidden representations for every token.(Z. Wen et al., 2024)

3) Affective Embeddings: These embeddings integrate the affective content (VAD scores) of the utterances and emotions to produce a rich representation. These are forwarded to both the Mood Transition Model and Init mood module.(Z. Wen et al., 2024)

4) Init mood Module: This initializes the mood state using the emotional cues and context representation. It establishes a baseline mood vector for the subsequent modeling of mood dynamics. Init mood is initialized based on U_1 and E_1 (Z. Wen et al., 2024)

5) Mood Transition Model: The model takes as input Affective embeddings, Initial mood vector from Init mood, Personality traits. All of these inputs are encoded in VAD space. It models how mood changes dynamically over time in response to conversational and emotional cues. The output is a Mood vector that reflects the user's emotional state after accounting for transitions. (Z. Wen et al., 2024) It computes $Delta$VAD by subtracting the VAD of E_1 and U_1 from the VAD computed after considering E_2 and U_2.

6) Emotion Classification Model: It takes input, the Mood vector, Personality traits, Context Representation and predicts the Target Emotion based on the synthesized understanding of mood, personality, and context. The model is optimized to classify emotions from a predefined set using supervised learning. (Z. Wen et al., 2024)

7) Target Emotion:: The final predicted emotion label is derived from the classifier's interpretation of the mood vector and dialogue context.

F. Experiments

1) Multi-head Attention and Speaker-Enhanced BERT: As a baseline model, we implemented a Multi-head Attention mechanism on top of BERT, designed to capture emotional salience at the token level using VAD-guided attention. Each utterance pair is input in the format: `[speaker_1] [utt_1] [SEP] [speaker_2] [utt_2]`, explicitly encoding speaker identity to preserve conversational flow and character consistency.

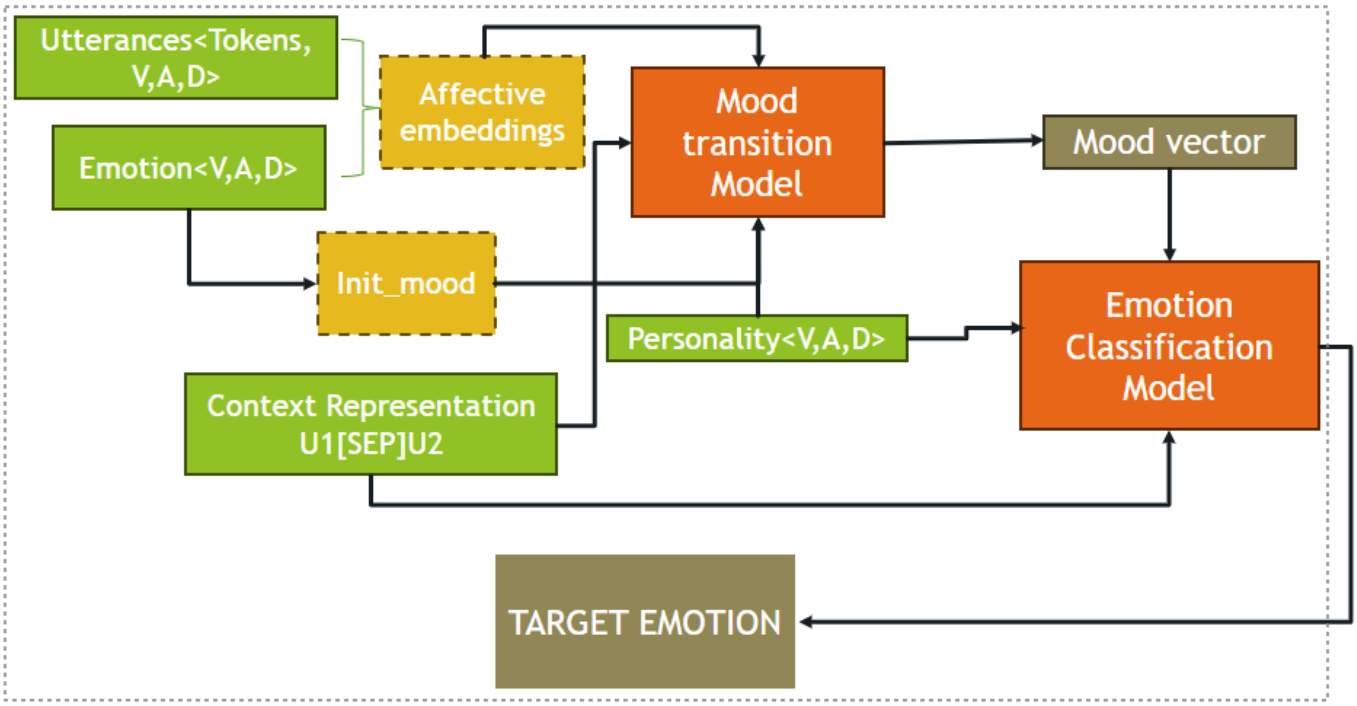

Fig. 1. High level Architecture

We apply three attention heads over the BERT token embeddings, each guided by a specific dimension of the target emotion's Valence-Arousal-Dominance (VAD) vector. For example, consider the utterance "I'm so excited", associated with the emotion *Joy*, which is represented in VAD space as (0.81, 0.51, 0.46). Each attention head focuses on how tokens align with one of the VAD axes:

TABLE VI
MULTI-HEAD VAD-GUIDED ATTENTION ON EXAMPLE UTTERANCE

Token	Valence Head	Arousal Head	Dominance Head
[CLS]	0.10	0.05	0.12
I'm	0.20	0.10	0.25
so	0.30	0.25	0.20
excited	**0.35**	**0.50**	**0.38**
[SEP]	0.05	0.10	0.05

As shown in Table VI, the token *"excited"* receives the highest attention across all heads due to its strong alignment with high valence and arousal, making it the most emotionally salient word in the sentence.

The model outputs are passed through a multi-layer perceptron (MLP) with dropout regularization. Training is performed with a dual-objective loss: Focal Loss for classifying emotions and Mean Squared Error (MSE) loss for regressing mood (VAD) vectors. While the model performs reasonably on high-frequency emotions such as *Neutral* and *Joy*, it struggles with low-resource classes like *Disgust* and *Surprise*, as reflected in their low F1 scores. This emphasizes the need for more advanced architectures that can model temporal and personality-conditioned affective shifts, which we address in subsequent sections.

*2) **GNN-Based Emotion and Mood Prediction**:* This experiment investigates the use of a Graph Neural Network (GNN) architecture to jointly model the speaker's personality and conversational flow for the task of emotion and mood prediction. Unlike sequential or transformer-based models that rely on temporal encodings, this approach models each sample as a graph, enabling the system to reason over utterance relationships and personality influence in a unified representation.

- **Graph Construction:** Each dialogue sample is represented as a directed graph with three nodes: U_1 (BERT embedding of Utterance_1), U_2 (BERT embedding of Utterance_2), and P (personality vector). The personality vector is projected from OCEAN to a 3-dimensional VAD space and then zero-padded to 768 dimensions. Directed edges include $U_1 \rightarrow U_2$ to represent conversational flow, and $P \rightarrow U_1$, $P \rightarrow U_2$ to inject personality context. The graph is stored using torch_geometric.data.Data with attributes for node features ($x \in \mathbb{R}^{3\times768}$), edge list ($edge\_index$), emotion label ($y$), and mood target ($m$, the BERT embedding of Utterance_3).
- **Data Split:** The dataset is stratified by emotion class and split into 80%
- **Utterance and Personality Embeddings:** Utterances are encoded using BERT-base-uncased to obtain 768-dimensional contextual embeddings. Personality vectors (originally 5-dimensional OCEAN) are mapped to 3-dimensional VAD vectors via a linear projection and then zero-padded to 768 dimensions to match the BERT output.

- **Graph Neural Network Architecture:** The graph is processed through three GCN layers:
  - GCNConv(768 $\rightarrow$ 256)
  - GCNConv(256 $\rightarrow$ 128)
  - GCNConv(128 $\rightarrow$ 64)

  Each layer is followed by Batch Normalization, ReLU activation, and Dropout (dropout rate = 0.3). Node embeddings are then aggregated using global_mean_pool() to produce a 64-dimensional graph-level embedding.
- **Prediction Heads:** The pooled graph embedding is passed through two linear heads:
  - **Emotion Head:** Linear(64 $\rightarrow$ 7) for Emotion_3 classification.
  - **Mood Head:** Linear(64 $\rightarrow$ 768) to regress the BERT embedding of Utterance_3.
- **Training Configuration:** The model is trained for 250 epochs using the Adam optimizer with a learning rate of 0.005 and batch size of 16. Dropout regularization is applied, and early stopping is used based on validation loss.
- **Loss Functions:** The model is trained using a multi-task objective:

$$\mathcal{L}_{\text{mood}} = \frac{1}{N}\sum_{i=1}^{N} \|\hat{m}_i - m_i\|^2 \tag{1}$$

$$\mathcal{L}_{\text{emotion}} = -\sum_{i=1}^{N} y_i \log(\hat{y}_i) \tag{2}$$

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{mood}} + \mathcal{L}_{\text{emotion}} \tag{3}$$

*3) **HTN + GRU**:* This experiment explores the model performance by adding one additional layer of Hierarchical Transformers on BERT context embeddings to understand the context of both the utterances. Only using BERT on utterances, encodes a single utterance at a time, and provides context from a single utterance. In a conversation, it is important to have a context of all previous utterances to predict the next utterance, the hierarchical transformer provides that context. It provides reach embeddings providing context between utterances.

- **Data Split:** For this experiment we split data into three parts with training set 80%, Validation split 10% and Test split 10% and seed value 42.
- **Base Encoder:** We used BERT-base-uncased as base encoder and HuggingFace AutoTokenizer to get contextual token embeddings.Helps encode nuanced meaning across utterances, which is critical for emotion recognition and makes downstream layers more effective at reasoning over text.
- **Affective Attention Module:** It takes BERT embeddings + VAD token embeddings and returns weighted VAD delta per utterance.
- **Context Transformer Layer:** It is two two-layer Transformer encoder with 8 heads, embedding size 768 and hidden size 3072. It is used to focus attention on tokens that are emotionally salient based on their Valence-Arousal-Dominance (VAD) scores. It computes a query (q) from BERT token embeddings. Performs element-wise product between query and token VAD embeddings. Applies softmax to get attention weights. Generates a weighted average VAD vector per utterance.
- **Context Transformer Layer:** It is two layer transformer used to model the temporal sequence of utterances and the flow of conversation. It models how contextual clues accumulate across turns and helps the model maintain a coherent dialogue state, essential for mood inference.
- **Mood Transition GRU:** This takes Contextual utterance representation + personality and returns predicted mood vector in VAD space by runing it through GRU. Then concatenate final GRU state with personality vector. Applies a linear transformation to predict the mood delta (ΔVAD). Adds this to the previous mood vector to get the new mood.
- **Emotion Classification Head:** It takes [CLS] embedding from last utterance + predicted mood + personality and applies a 2-layer MLP (Linear -¿ GELU -¿ Dropout -¿ Linear) to return 7 emotion logits (7-way classification)
- **Training Configuration:** Batch size use is 8 and epochs 5, optimizer Adam and schedular get_linear_schedule_with_warmup
- **Loss Functions:**

$$\text{Mood Loss} = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2 \tag{4}$$

$$\text{Emotion Loss} = -\alpha_t(1-p_t)^{\gamma}\log(p_t) \tag{5}$$

$$\alpha_t = \frac{1}{\text{class frequency normalization}} \tag{6}$$

$$\gamma = 2 \tag{7}$$

*4) **AutoEncoder**:* This experiment evaluates a multi-task autoencoder that simultaneously reconstructs 300-D text embeddings; classifies inputs into seven emotion categories; and predicts changes in mood (valence–arousal–dominance). The encoder compresses each example into a 64-D latent vector, which feeds two heads: an emotion classification head and a mood regression head. The emotion head is a two-layer MLP with dropout ($p = 0.3$) that outputs logits over seven emotion classes, while the mood head combines a base ΔVAD predictor with a personality-adjustment branch using a 5-D OCEAN vector.

- **Methodology:** A multi-task autoencoder that jointly reconstructs 300-D text embeddings, classifies into seven emotion categories, and regresses VAD mood values, leveraging a 5-D OCEAN personality vector for adaptive prediction.
- **Data Split:** Stratified 80% train, 10% validation, 10% test using a fixed random seed.

- **Class Balancing:** In the training set, each non-neutral emotion is sampled to 120% of the second-largest minority class, yielding a uniform 14.29% distribution per class.
- **Preprocessing:** Utterances are lowercased, punctuation is removed, and text is tokenized to length 50 (10 000-token vocabulary), then zero-padded or truncated. Personality OCEAN scores are parsed from metadata.
- **Model Architecture:**
  - *Base Encoder:* 300-D token embeddings; 5-D OCEAN $\rightarrow$ 32-D linear projection.
  - *Encoder MLP:* 332 $\rightarrow$ 256 $\rightarrow$ 128 $\rightarrow$ 64 layers with ReLU activations.
  - *Latent Space:* 64-D bottleneck vector capturing joint text/personality features.
  - *Decoder MLP:* 64 $\rightarrow$ 128 $\rightarrow$ 256 $\rightarrow$ 300 to reconstruct text embeddings.
  - *Emotion Head:* 64 $\rightarrow$ 128 $\rightarrow$ 64 $\rightarrow$ 7 with dropout ($p = 0.3$).
  - *Mood Head:* Two branches (64 $\rightarrow$ 64 $\rightarrow$ 3; 5 $\rightarrow$ 32 $\rightarrow$ 3) whose outputs are summed to predict ΔVAD, then added to the previous mood.
- **Training Configuration:** Trained for 30 epochs (batch size = 32) using Adam (learning rate 1×10^{-3}), dropout $p = 0.3$ in the emotion head, no weight decay or early stopping to ensure deterministic reproducibility. The 30-epoch choice was driven by convergence of the multi-task loss and validation metrics (plateauing with $< 1\%$ change).
- **Loss Functions:**
  - $\mathcal{L}_{\text{recon}} = \frac{1}{N}\sum_{i=1}^{N}\left\|\hat{\mathbf{x}}_i - \mathbf{x}_i\right\|^2$
  - $\mathcal{L}_{\text{emotion}} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{c=1}^{7} y_{c,i}\,\log \hat{y}_{c,i}$
  - $\mathcal{L}_{\text{mood}} = \frac{1}{N}\sum_{i=1}^{N}\left\|\hat{\mathbf{m}}_i - \mathbf{m}_i\right\|^2$

  Combined as

  $$\mathcal{L} = 0.4\,\mathcal{L}_{\text{recon}} + 0.4\,\mathcal{L}_{\text{emotion}} + 0.2\,\mathcal{L}_{\text{mood}}.$$

  The weights (0.4, 0.4, 0.2) balance the initial magnitudes of each component and were fine-tuned on validation to maximize macro-F_1.

*5) **Temporal Convolutional Network**:* Traditional feedforward models treat each utterance as an isolated input, failing to capture the natural flow of conversation. However, in human dialogue, the emotional context often evolves across utterances; a phenomenon crucial for accurate emotion prediction. To address this, we incorporate a Temporal Convolutional Network (TCN) to model temporal dependencies between utterances. TCN allows the model to understand how emotions shift and develop across a sequence of utterances, making it particularly well-suited for dialogue-based tasks. By learning patterns of emotional progression, the model can make more context-aware predictions for the final utterance.

- **Methodology:** Our model leverages a Temporal Convolutional Network (TCN) to capture mood dynamics across sequential utterances. Each of the three utterances (U1, U2, U3) is encoded using a pretrained BERT model into a 768-dimensional embedding. These are stacked into a (768, 3) tensor and passed through a TCN, which learns temporal dependencies and outputs a context vector capturing conversational flow. To incorporate personality, a fixed linear mapping projects a 5D OCEAN vector to a 3D mood estimate (VAD: Valence, Arousal, Dominance). A feedforward mood shift module predicts the delta mood (ΔVAD), which is added to the base mood vector. The final emotion prediction is made by concatenating the TCN context vector, the updated mood, and the personality, and passing them through a Feedforward Neural Network (FNN) classifier.
- **Class Balancing:** Adopted stratified sampling during data splitting to prevent bias toward majority classes and allows the model to learn balanced decision boundaries, thereby improving generalization, especially for minority emotion classes.
- **Data Split:** 80% train, 10% validation, 10% test using a fixed random seed.
- **Preprocessing:** Tokenized each utterance using the bert-base-uncased tokenizer and padded them to a fixed length of 64 tokens. Additionally, personality vectors were extracted from metadata and converted into initial VAD mood estimates.
- **Model Architecture:**
  - *Utterance Encoder: BERT-base (output: 768-D CLS embedding per utterance).*
  - *TCN Block: Two-layer Temporal Convolutional Network (768 input/output, kernel=2, dilations=[1,2]).*
  - *Mood Shift Head: (768 + 5 + 3) $\rightarrow$ 64 $\rightarrow$ 3 to predict ΔVAD.*
  - *Emotion Head: (768 + 5 + 3) $\rightarrow$ 128 $\rightarrow$ 7-class softmax.*
  - *Final Output: Emotion label of U3 and predicted VAD mood vector.*
- **Training Configuration:** We trained our model using an 80/10/10 stratified split for train, validation, and test sets. To address class imbalance, we used Focal Loss with $\gamma = 2.0$ and $\alpha = 1.0$. The model was optimized using AdamW with a learning rate of $2e$-5, trained for 3 epochs with a batch size of 16. Early stopping was applied based on test loss with a patience of 1.
- **Loss Functions:**

$$\text{Mood Loss} = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2 \tag{8}$$

$$\text{Emotion Loss} = -\alpha_t(1 - p_t)^{\gamma}\log(p_t) \tag{9}$$

$$\text{Total Loss} = \lambda_1 \cdot \text{Emotion Loss} + \lambda_2 \cdot \text{Mood Loss} \tag{10}$$

$$\lambda_1 = 0.7 and \lambda_2 = 0.3 \tag{11}$$

*6) **BERT + BiGRU**:* This experiment implements a BiGRU-based model layered on top of BERT embeddings, augmented with VAD-based attention and fused with both personality and mood signals. The goal is to capture emotionally salient patterns across dialogue while maintaining consistency with speaker-specific personality traits and mood dynamics.

- **Data Split:** We stratified the PELD dataset into 80% training, 10% validation, and 10% test splits using a fixed random seed (42) to preserve class distributions across splits.
- **Input Representation:** Each sample consists of Utterance_1 and Utterance_2 from the conversation history. These are concatenated with a `[SEP]` token and tokenized using the BERT tokenizer. A 5-dimensional OCEAN personality vector and a 3-dimensional VAD vector derived from the emotion of U_1 are also used as additional conditioning signals.
- **Model Architecture:**
  - *Utterance Encoder:* BERT-base-uncased is used to generate 768-dimensional token embeddings (Devlin et al., 2019).
  - *BiGRU Layer:* A bidirectional GRU (hidden size = 64) processes the BERT output sequence to capture both past and future emotional dependencies (Majumder et al., 2019).
  - *MLP Attention:* An MLP attention mechanism computes attention weights across BiGRU outputs, producing a 128-dimensional context vector.
  - *Fusion Layer:* The context vector is concatenated with the personality vector (5D) and VAD vector (3D), resulting in a 136-dimensional representation. This is passed through a linear projection with ReLU activation and dropout.
- **Prediction Heads:**
  - **Emotion Classifier:** Linear($128 \rightarrow 7$), followed by softmax activation to predict emotion class.
  - **Mood Regressor:** Linear($128 \rightarrow 3$) to output a predicted VAD vector.
- **Training Configuration:** The model is trained using the AdamW optimizer with a learning rate of 2×10^{-5} and a batch size of 16. Weighted random sampling is used to handle class imbalance. Training is conducted for up to 30 epochs with early stopping based on validation macro-F1. We progressively unfreeze the top 6 BERT layers every 5 epochs to allow gradual fine-tuning.
- **Loss Functions:** The model is trained using a dual-objective multi-task loss, as shown in Equation 12, combining emotion classification and mood regression:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{emotion}} + 0.3 \cdot \mathcal{L}_{\text{mood}} \tag{12}$$

where the emotion classification loss is defined as:

$$\mathcal{L}_{\text{emotion}} = -\sum_{i=1}^{N} y_i \log(\hat{y}_i) \tag{13}$$

and the mood regression loss is the mean squared error between predicted and true VAD values:

$$\mathcal{L}_{\text{mood}} = \frac{1}{N} \sum_{i=1}^{N} \|\hat{v}_i - v_i\|^2 \tag{14}$$

This multi-task learning formulation follows the psychological grounding and emotion modeling proposed in prior work (Z. Wen et al., 2024).

V. Results

TABLE VII
SUMMARY F_1 SCORES ON PELD TEST SET

Model	Micro	Macro	Weighted
Benchmark	0.41	0.33	0.33
Multi-head Attention	0.45	0.19	0.37
Graph Neural Network (GNN)	0.32	0.19	0.30
HTN + GRU	0.48	0.39	0.50
Autoencoder (AE)	0.41	0.38	0.42
Temporal Conv. Network (TCN)	0.55	0.43	0.54
BERT + BiGRU	0.43	0.34	0.39

Table VII summarizes the comparative F_1 performance of the six models on the PELD test set , along with benchmark. The Temporal Convolutional Network (TCN) outperforms all other architectures, achieving the highest micro-F_1 of 0.55 and macro-F_1 of 0.43. The Hierarchical Transformer with GRU (HTN+GRU) is the runner-up (micro-F_1=0.48, macro-F_1=0.39), demonstrating the benefit of multi-head VAD attention combined with recurrent fusion. The Autoencoder (AE) yields a balanced micro-F_1 of 0.41 and macro-F_1 of 0.38, indicating that its joint reconstruction and classification loss regularizes representations effectively. Simpler models—Multi-head Attention and Graph Neural Network (GNN)—achieve lower summary scores, highlighting the importance of explicit temporal modeling. BERT+BiGRU attains moderate gains (micro-F_1=0.43, macro-F_1=0.34) through pretrained embeddings but still falls short of specialized sequence learners.

Table VIII presents the per-class F_1 scores for each model. The Temporal Convolutional Network (TCN) consistently achieves the highest F_1 on most classes - Neutral (0.69), Surprise (0.53) and Anger (0.46) and remains competitive on Joy (0.51) and Fear (0.27), demonstrating its ability to model fine-grained temporal patterns. The HTN+GRU model excels on Joy (0.56) and maintains strong performance on Neutral (0.59) and Anger (0.44), but underperforms on Disgust (0.15) relative to the Autoencoder's 0.39. The Autoencoder delivers balanced results, notably on Disgust (0.39) and Surprise (0.46), while preserving moderate scores on Neutral (0.47) and Sadness (0.38). BERT+BiGRU shows solid contextual understanding

TABLE VIII
PER-CLASS F_1 SCORES ON PELD TEST SET

Model	Neutral	Joy	Surprise	Anger	Fear	Sadness	Disgust
Benchmark	0.54	0.25	0.27	0.27	0.24	0.27	0.25
Multi-head Attention	0.62	0.20	0.00	0.36	0.06	0.11	0.00
Graph Neural Network (GNN)	0.49	0.17	0.07	0.25	0.12	0.18	0.05
HTN + GRU	0.59	0.56	0.49	0.44	0.27	0.25	0.15
Autoencoder (AE)	0.47	0.40	0.46	0.40	0.23	0.38	0.39
Temporal Conv. Network (TCN)	0.69	0.51	0.53	0.46	0.27	0.30	0.29
BERT + BiGRU	0.54	0.48	0.40	0.42	0.20	0.26	0.15

with the second-best Joy F_1 (0.48) but lags on Fear (0.20) and Disgust (0.15). Simpler baselines - Multi-head Attention and GNN—struggle on low-frequency classes like Surprise and Disgust, highlighting the importance of both temporal modeling and multi-task regularization.

VI. CONCLUSION

This project presents a unified framework for modeling emotion and mood in conversations by integrating semantic, affective, and personality-driven features. Across all experiments, a common theme emerged: *emotion expression in dialogue is shaped not only by surface text but also by the interplay of affective state transitions and speaker traits*. By designing specialized architectures that account for this interplay—ranging from GNN-based graph reasoning to sequence-aware models like TCN and HTN-GRU—we demonstrate that incorporating *mood dynamics and personality context leads to more accurate and human-aligned emotion recognition*.

Affective embeddings, VAD-based attention, mood regression, and OCEAN-to-VAD projections all contributed complementary perspectives. The top-performing models (TCN and HTN-GRU) exemplified how modeling temporal mood shifts and personality-modulated context flow enables deeper emotional inference. Even architectures with lower overall scores, such as GNNs or multi-head attention, offered valuable insights—especially regarding structure-aware or token-level emotion cues.Meanwhile, the BERT+BiGRU model demonstrated that even without explicit temporal convolution or hierarchical attention, combining deep contextual embeddings with personality and mood fusion provides strong performance on common emotion classes.

Overall, this work provides a *modular and extensible design space* for affect-aware systems, combining traditional text encoding with psychologically grounded signals. It lays the groundwork for building *emotionally intelligent dialogue agents* capable of modeling both what is said and how the speaker feels across time.

VII. FUTURE WORK

Future efforts will focus on improving both architectural design and affective reasoning. Key directions include:

- **Enhanced Graph Modeling:** Extend the GNN structure using richer edge types (e.g., sentiment flow, temporal distance), relational GCNs, or Graph Transformers to better model discourse structure.
- **Multimodal Integration:** Incorporate speech prosody, visual signals (e.g., facial expressions), and physiological cues to improve emotion grounding and cross-modal affect detection.
- **Dynamic Mood Tracking:** Explore temporal graphs or memory-augmented models that explicitly track evolving mood states over extended dialogue sequences.
- **Explainability and Interpretability:** Integrate interpretable components (e.g., mood trajectory plots, attention visualizations) to build trust and transparency into affective models.
- **Real-World Deployment:** Fine-tune top-performing models on real conversational datasets, introduce noise tolerance, and evaluate generalization in interactive, user-facing systems.

REFERENCES

André, E., Rist, T., van Mulken, S., Klesen, M., & Baldes, S. (2000). The automated design of believable dialogues for animated presentation teams. In J. e. a. Cassell (Ed.), *Embodied conversational agents* (pp. 220–255). MIT Press.

Ball, G., & Breese, J. (2000). Emotion and personality in a conversational agent. In J. e. a. Cassell (Ed.), *Embodied conversational agents* (pp. 189–219). MIT Press.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). Bert: Pre-training of deep bidirectional transformers for language understanding. *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*, 4171–4186. https://doi.org/10.18653/v1/N19-1423

Han, Y., Yamashita, T., & Itoh, Y. (2018). Mood recognition and response generation for humanoid robots using valence-arousal-dominance emotion model. *2018 IEEE-RAS 18th International Conference on Humanoid Robots (Humanoids)*, 676–681. https://doi.org/10.1109/HUMANOIDS.2018.8624950

Hazarika, D., Zimmermann, R., & Poria, S. (2020). Conversational memory network for emotion recognition in dyadic dialogue. *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*, 2122–2132. https://doi.org/10.18653/v1/N18-1192

Li, J., Galley, M., Brockett, C., Spithourakis, G. P., Gao, J., & Dolan, B. (2016). A persona-based neural conversation model. *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (ACL)*, 94–104. https://doi.org/10.18653/v1/P16-1010

Mairesse, F., & Walker, M. A. (2007). Personage: Personality generation for dialogue. *Proceedings of the 45th Annual Meeting of the ACL*, 496–503. https://doi.org/10.3115/1557769.1557834

Majumder, N., Hazarika, D., Poria, S., & Mihalcea, R. (2019). Dialoguernn: An attentive rnn for emotion detection in conversations. *Proceedings of the AAAI Conference on Artificial Intelligence*, *33*(01), 6818–6825. https://doi.org/10.1609/aaai.v33i01.33016818

Masuyama, H., Nagai, T., & Asada, M. (2020). A model of personality traits based on valence–arousal–dominance space and its application to dialogue control in a robot. *Frontiers in Robotics and AI*, *7*, 84. https://doi.org/10.3389/frobt.2020.00084

Mehrabian, A. (1996). Pleasure-arousal-dominance: A general framework for describing and measuring individual differences in temperament. *Current Psychology*, *14*(4), 261–292. https://doi.org/10.1007/BF02686918

Mohammad, S. M. (2018). Obtaining reliable human ratings of valence, arousal, and dominance for 20,000 english words. *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (ACL)*, 174–184. https://doi.org/10.18653/v1/P18-1017

Rashkin, H., Smith, E. M., Li, M., & Boureau, Y.-L. (2019). Empatheticdialogues: A dataset for empathetic response generation. *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 5370–5381. https://doi.org/10.18653/v1/D19-1543

Scherer, K. R. (2005). What are emotions? and how can they be measured? *Social Science Information*, *44*(4), 695–729. https://doi.org/10.1177/0539018405058216

Song, Y., Zhang, Z., Wu, L., & Yu, M. (2020). Constrained multi-aspect empathetic dialogue generation. *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2917–2927. https://doi.org/10.18653/v1/2020.emnlp-main.233

Wen, C., Yao, Y., Yu, Z., & Yu, Z. (2021). You sound like someone who…: Persona based grounded emotion recognition. *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*, 4559–4571. https://doi.org/10.18653/v1/2021.naacl-main.361

Wen, Z., Cao, J., Shen, J., Yang, R., Liu, S., & Sun, M. (2024). Personality-affected emotion generation in dialog systems. *ACM Transactions on Information Systems*, *42*(5), 1–27. https://doi.org/10.1145/3655616

Zhang, S., Dinan, E., Urbanek, J., Szlam, A., Kiela, D., & Weston, J. (2018). Personalizing dialogue agents: I have a dog, do you have pets too? *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (ACL)*, 2204–2213. https://doi.org/10.18653/v1/P18-1205

Zhong, Z., Wang, D., Muresan, S., & Ostendorf, M. (2020). Towards persona-based empathetic dialogue generation. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL)*, 6556–6566. https://doi.org/10.18653/v1/2020.acl-main.585

Zhou, H., Huang, M., Zhang, T., Zhu, X., & Liu, B. (2018). Emotional chatting machine: Emotional conversation generation with internal and external memory. *AAAI Conference on Artificial Intelligence (AAAI)*, *32*(1), 730–739.