HADE: Hierarchical Affective Dialog Encoder for Personality Recognition in Conversation

Anonymous ACL submission

Abstract

Personality recognition in conversation aims to determine the personality traits of speakers through the dialogue content, which is of great importance in designing personalized conversational AI. Existing methods that use only linguistic patterns in utterances limit their performance. To fill in the gap, we investigate the effectiveness of incorporating affective information and modeling the interactions among speakers in conversations for personality recognition. However, available corpus with personality and explicit affective annotations is rare. Besides, modeling the dialog flow with multiple speakers is difficult. Faced with the issues, we proposed Hierarchical Affective Dialog Encoder (HADE) for effective personality recognition in conversation. HADE utilizes manual annotated Valance-Arousal-Dominance (VAD) vectors of single words and implicitly extracts affective information from utterances. Then, it introduces a hierarchical architecture with the dialog state embeddings to identify the speakers and encode the whole dialog flow. Finally, the affective information is integrated by an auxiliary VAD regression task to enhance personality recognition. Extensive experiments on a well-known dataset, **FriendsPersona**, demonstrate the effectiveness of our method compared with state-of-the-art models. Besides, we conduct an ablation study to discuss different approaches for integrating affective information and dialog flow modeling; the design of both parts in HADE is also verified to be effective for personality recognition in conversation[1].

1 Introduction

Personality is relatively permanent traits and unique characteristics that give both consistency and individuality to a person's behavior (Feist and Feist, 2012). In the conversation scenario, personality influences both semantic content and emotional expressions. Therefore, recognizing the personality of speakers is critical for understanding the conversation content so that the dialog systems are able to provide appropriate and personalized responses to users.

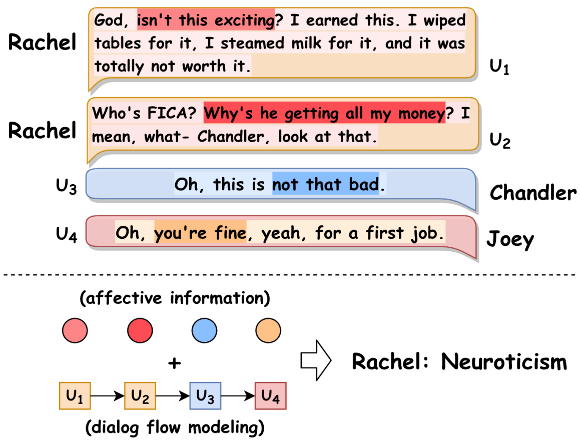

Figure 1: A toy example for personality recognition in conversation. In this example, we first analyze the affective information in utterances from Rachel: excited and nervous, while for Chandler and Joey, the affective information is quite positive. Besides, the dialog flow contains the interaction between Rachel (U_1, U_2), Chandler (U_3), and Joey (U_4), showing that others are comforting her. So, we infer that the current personality exhibited by Rachel is *Neuroticism*.

Existing researches (Rissola et al., 2019; Jiang et al., 2020) simply focused on extracting linguistic patterns in utterance to recognize certain personality, which limited their performance. The main reason is that they fails to model complicated yet effective factors (e.g., the affective information in utterances or the interactions among the speakers) of personality recognition in conversation intentionally in their approaches.

Psychology studies (Watson and Clark, 1992; Mehrabian, 1995) find that there is a strong correlation between personalities and affective information in expression. Besides, by observing the conversation data, we found that in addition to the

[1]Our code will be released at github.com.

semantics in utterances, the interactions among different speakers in the dialog flow also helps to recognize the personality.

However, implementing the insights above meets two major challenges. The first one is the lack of explicit affective annotations in the personality analysis corpus. Personality analysis datasets in the conversation scenario are already rare because collecting such data may cause privacy concerns. Nevertheless, almost none of them incorporates explicit affective annotations. The second challenge arises in modeling the dialog flow to analyze the specified speakers. The data shortage tends us to use general pre-train language models. However, it is difficult to indicate specific speakers efficiently with existing conversational models (*e.g.*, DialoGPT (Zhang et al., 2019), PLATO (Bao et al., 2019), and EVA (Zhou et al., 2021)).

To tackle the issues mentioned above, we propose the Hierarchical Affective Dialog Encoder (HADE) to implicitly extract the affective information from the dialog content and design a hierarchical architecture to encode the dialog flow for personality recognition. First, to alleviate the lack of explicit affective annotations in the personality analysis corpus, HADE uses the pre-annotated Valence-Arousal-Dominance (VAD) vectors for single words to represent the implicit affective factors in utterances. Then, we design HADE based on BERT (Devlin et al., 2018), and a transformer (Vaswani et al., 2017) encoder hierarchically. BERT in the bottom layer encodes all the utterances, respectively. After that, the transformer encoder receives the output from the bottom layer and the dialog state embeddings designed to identify different speakers for personality recognition. To incorporate the affective information to enhance personality recognition, we integrate an auxiliary VAD regression task in the upper layer of HADE through a regression head of BERT.

To show the effectiveness of our method, we conduct extensive experiments on **FriendsPersona** constructed by (Jiang et al., 2020). It is the dialog script with personality annotations in 711 different dialogues, including 8,157 utterances from the famous TV Series *Friends*[2]. Adequate results validate that our model outperforms the state-of-the-art methods. We also design an ablation study to evaluate different modules in our model. The utilization of affective information in personality recognition is also verified to be effective in HADE. The contributions of this work are summarized as follows:

- We investigate the effectiveness of incorporating affective information and modeling the interactions among speakers and proposed HADE for personality recognition in conversation.
- In HADE, we utilize pre-annotated VAD vectors of single words and introduce a hierarchical architecture with the dialog state embeddings, which solves the challenges of affective annotation shortage and the dialog flow modeling.
- HADE outperforms state-of-the-art methods on a public conversation dataset, **FriendsPersona**. Besides, through ablation study, the modules in HADE are validated effective to integrate affective information and model the dialog flow.

2 Related Works

In this section, we review existing researches that related to personality analysis in conversation. These researches are categorized into two aspects: Text-based Personality Analysis and Dialog Modelling in Conversation.

2.1 Text-based Personality Analysis

Most existing researches in text-based personality recognition are limited to analyzing self-reported essays (Pennebaker and King, 1999; Tighe et al., 2016) or behaviors on social media (Golbeck et al., 2011; Schwartz et al., 2013). (Schwartz et al., 2013) analyzed 700 million words, phrases, and topic instances collected from the Facebook messages of 75,000 volunteers and found striking variations in language with personality, gender, and age. The Facebook data is also studied in (Lynn et al., 2020). They hierarchically encode all posts from one user with attention-based GRU (Cho et al., 2014) to produce the whole contextual representations for personality identification.(Moreno et al., 2019) adopted a feature-engineering approach to extract text-based features from Twitter blogs to identify the personality of Twitter users. Only a few works (Mehl et al., 2006; Rissola et al., 2019; Jiang et al., 2020) focus on the conversation scenario due to the shortage of available data: (1) The number of conversation datasets with personality

[2]https://www.imdb.com/title/tt0108778/

labels is insufficient as collecting such kinds of data may cause privacy concerns; and (2) The length of the dialog flow is short compared with self-reports, essays, and multiple posts on social media.

2.2 Dialog Flow Modeling in Conversation

Modeling the dialog flow is also helps to understand the personalities of speakers in conversation. In the early stage, (Serban et al., 2017) regards the tokens in utterances and utterances in a dialog flow as two kinds of sequences and proposes the classic hierarchical RNN encoder for dialog data. (Mehri et al., 2019) proposes two novel pre-training objectives: masked-utterance retrieval and inconsistency identification to better capture both the utterance-level and context-level information. Similarly, (Gu et al., 2020) employs a hierarchical BERT architecture to encode the utterances and the dialog context separately to enable the model to capture multi-level coherences. Furthermore, (Wolf et al., 2019b) adds the dialog state embeddings during utterance encoding so that the model can identify the utterances from different speakers.

3 Preliminaries

Before introducing our method, we first present the development of the Big-five personality traits and the affective information for personality analysis. This part inspires the design of HADE and helps to understand our method as preliminary knowledge.

3.1 The Big-five Personality Traits

The Big-five trait theory presents a discrete taxonomy of personality as shown in Table 1[3], which is naturally suitable for personality analysis as a classification problem. This theory was defined by several independent sets of researchers who used factor analysis of verbal descriptors of human behavior. It is developed from the trait theory and the lexical hypothesis and in psychology.

Factor	Description
Openness	Openminded, 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.

Table 1: The OCEAN personality traits and description (Costa and McCrae, 1992)

[3]https://en.wikipedia.org/wiki/Big_Five_personality_traits

In the trait theory, personality is the set of psychological traits and mechanisms within the individual that are organized and relatively enduring and that influence their interactions with, and adaptations to, the intrapsychic, physical, and social environments (Larsen and Buss, 2008). The lexical hypothesis first states that those personality characteristics that are important to a group of people will eventually become a part of that group's language (Cattell, 1943). It second states that more important personality characteristics are more likely to be encoded into language as a single word (John et al., 1988), which also explains the principles of existing personality analysis researches based on linguistic patterns.

Therefore, the big-five personality traits are widely applied as personality recognitionclassification labels in social medias (Iacobelli et al., 2011; Souri et al., 2018) and conversations (Mairesse and Walker, 2006; Mairesse et al., 2007).

3.2 Affective Information for Personality Analysis

Besides linguistic patterns, affective information in expressions is important for personality analysis. Affect, in psychology, refers to the underlying experience of feeling, emotion, or mood (Fiske and Taylor, 1991). Affective states vary along three principal dimensions: valence, arousal, and motivational intensity (Harmon-Jones et al., 2013) (also interpreted as dominance in some works (Bradley and Lang, 1999; Mohammad, 2018)).

(Watson and Clark, 1992) pointed out that there are strong relations between the *Extraversion* and *Conscientiousness* traits and the positive affects, and between *Neuroticism* and *disagreeableness* and various negative affects. (Mehrabian, 1995) analyzed the relationship between the big-five personality with the PAD[4] scales as follows: *Extraversion* includes pleasant and dominant characteristics: *Agreeableness* consists of pleasant and submissive qualities; *Conscientiousness* relates positively to trait pleasure; *Neuroticism* includes pleasant and arousable qualities; and *Openness* is comprised of pleasant, arousable, and dominant characteristics. Based on the analysis above, (Mehrabian, 1996) further estimates the relationship into a set of re-

[4]It is Pleasure-Arousability-Dominance (PAD) in the original paper, PAD and VAD share the same meaning in the context of verbal text, we will use VAD for consistency henceforth.

gression equations. These theories are also adopt to design human-like robots (Han et al., 2012; Masuyama et al., 2018), and empathetic dialog systems (Ball and Breese, 2000; Wen et al., 2021).

The following section will introduce the studied problem and the HADE model in detail.

4 Methodology

4.1 Problem Statement

The studied problem is stated as follows. Given a dialog flow $C = \{U_1, U_2, ..., U_n\}$ including n utterances from multiple speakers, the objective is to recognize the personality trait P of the analyzed speaker s through the semantic content and the affective information in C.

The personality trait p is represented as a 5-d binary vector $[O, C, E, A, N]$ indicating the Openness, Conscientiousness, Extraversion, Agreeableness, and Neuroticism respectively referring to the big-five personality theory. The affective information is indicated by the manual-annotated VAD vectors of words. Therefore, following the problem settings in some similar personality analysis works (Rissola et al., 2019; Jiang et al., 2020), we model the personality recognition as a binary classification problem over the five personality traits, respectively.

4.2 The HADE Model

To solve the challenges mentioned earlier of affective annotation shortage and effective speaker identification in dialog flow encoding, we design the HADE model as shown in Figure 2. HADE includes three modules: Utterance Encoding, Dialog Flow Encoding, and Utterance VAD regression. We will introduce these modules in detail.

4.2.1 Utterance Encoding

In conversation, utterances convey the personality trait of the speaker in addition to their semantic content (Mairesse et al., 2007). We choose BERT in the bottom layer of HADE to encode all the utterances, respectively. Pre-trained on the massive corpus, BERT does not rely on training with a large dataset to extract the semantics in utterances, which meets the challenge of data shortage.

For each utterance U_i in the dialog flow, we add a [CLS] and a [SEP] token in the beginning and the last position during tokenization. Hereafter, the $U_1, U_2, ..., U_n$ are separately encoded by the BERT encoder as a list of hidden representations $E_1, ..., E_n$, where the E_i is the embedding of the [CLS] token in U_i from the last pooling layer output in BERT.

4.2.2 Dialog Flow Encoding

By observing the dialog data, we found that the sentence-level interaction among the speakers (i.e., what are the current speaker talks to others and how others respond to the current speaker) is also essential to analyze the personality traits. Therefore, in the upper layer, we design the dialog flow encoding module based on a vanilla transformer encoder, as shown in the upper left of Figure 2. The transformer encoder receives the output of the bottom layer and the dialog state embeddings designed to identify the speakers for personality recognition.

First, $\{E_1, ..., E_n\}$ are the utterance embeddings from the BERT encoder. Inspired by (Wolf et al., 2019b) and (Lin et al., 2019), we then construct the dialog state embedding $\{D_1, ..., D_n\}$ to identify the utterance from the analyzed speaker s and the context. To be more specific, we use 1 to indicate the utterances from s, and 0 for utterances from other speakers. To feed the indicaters into the model, we obtain the dialog state embedding by $D_i = MLP(is\_uttr(U_i))$, where $is\_uttr()$ outputs 1 and 0 as mentioned above. We also follow the original setting in (Vaswani et al., 2017) and construct the positional encodings $\{P_1, ..., P_n\}$ to help the model understand the dialog flow:

$$\begin{aligned} P_i(2j) &= sin(\frac{i}{10000^{\frac{2j}{d_{model}}}}) \\ P_i(2j+1) &= cos(\frac{i}{10000^{\frac{2j}{d_{model}}}}) \end{aligned} \tag{1}$$

where i is the token position in the utterance, d_{model} is the size of the positional encodings, $j = 0, 1, ..., d_{model}/2 - 1$.

After we get the three embeddings/encodings, we sum them together and feed them into the transformer model. We use all the last layer output of the transformer as the utterance representations $R_1, ..., R_n$ containing the sentence-level interactions through the self-attention mechanism. Then, we adopt the average pooling on the utterance representations for the personality classification minimizing the cross-entropy loss $\mathcal{L}_{ce}$ during training:

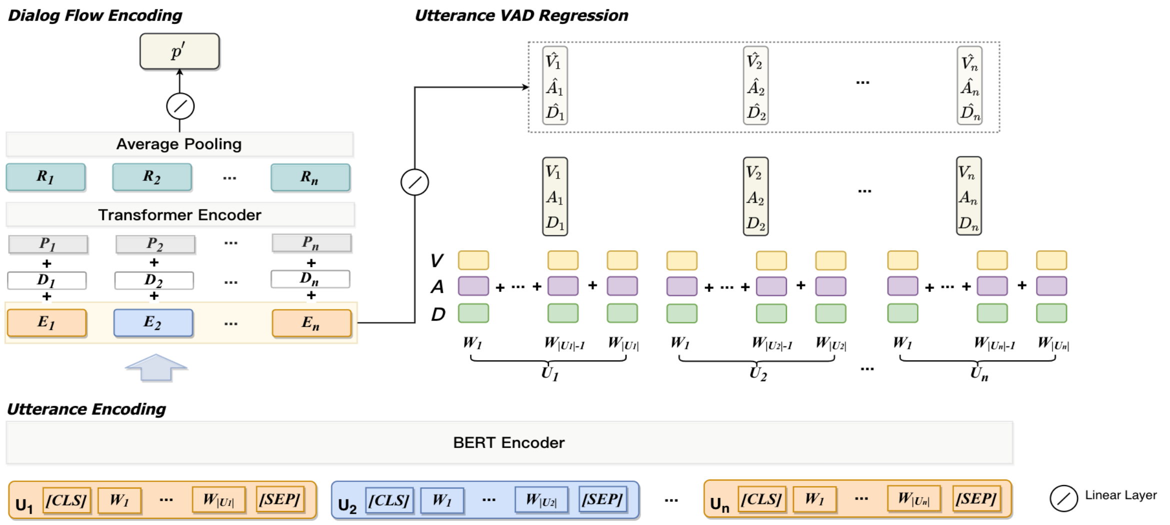

Figure 2: The model illustration of HADE. We use the same color to represent the utterances from the same speaker. e.g., U_1 and U_n.

$$
\begin{aligned}
R_i &= f_t(E_i + D_i + P_i) \\
p' &= MLP(\sum_{i=1}^{n} \frac{R_i}{n}) \\
\mathcal{L}_{ce} &= plog(p') + (1-p)log(1-p')
\end{aligned} \tag{2}
$$

where f_t is the transformer encoder, p' is the predicted personality label, and p is the ground truth personality label.

HADE first extracts the token-level semantic information in the bottom layer and then models the sentence-level interactions among speakers in the upper layer to facilitate personality recognition. The hierarchical modeling is verified as an efficient way to extract semantics in text with different granularities (Nawrot et al., 2021). It is also widely adopted in the conversation scenarios (Serban et al., 2017; Lynn et al., 2020; Gu et al., 2020).

4.2.3 Utterance VAD Regression

Although plenty of researches (Rank et al., 2013; Skowron et al., 2013; Asghar et al., 2018) work on the affective dialog systems, few works (Bauerhenne et al., 2020; Wen et al., 2021) combine it with personality analysis. One of the reasons is the lack of datasets with both emotion and personality annotations. Therefore, HADE extracts the affective information implicitly from utterances with VAD annotations for all the words in the utterances. Not only does this approach not need explicit emotion annotations, but it also can present the strength of emotions with numeric VAD vectors rather than discrete emotion labels.

Specifically, to preserve the encoding ability of BERT in HADE, we design an utterance VAD regression task with a regression head for the affective information extraction. The utterance VAD regression task supervises the model to capture affective information from the utterances.

For each utterance U_i in the input, we obtain the VAD vectors annotated by (Mohammad, 2018) of each word, which is also commonly utilized to represent affective information in conversation (Zhong et al., 2019; Colombo et al., 2019; Wen et al., 2021; Lee and Lee, 2021). The VAD vectors are numeric values ranging in $[0, 1]$ that indicate emotion intensity in three different dimensions. The valence measures positivity/negativity, arousal is for the excitement/calmness, and dominance is for the powerfulness/weakness.

$$
\begin{aligned}
V_i, A_i, D_i &= \sum_{j=1}^{|U_i|} V_j, A_j, D_j \\
\hat{V}_i, \hat{A}_i, \hat{D}_i &= f(E_i) \\
\mathcal{L}_{mse} &= \frac{1}{n} \sum_{i=1}^{n} (\sqrt{(V_i - \hat{V}_i)^2} \\
&+ \sqrt{(A_i - \hat{A}_i)^2} + \sqrt{(D_i - \hat{D}_i)^2})
\end{aligned} \tag{3}
$$

We sum the VAD vectors of all the words in each utterance as the regression objectives $\{V_j, A_j, D_j\}$ for U_i. Then, each E_i obtained from the bottom layer is fed into a linear function f to regression the objective by generating $\hat{V}_i, \hat{A}_i, \hat{D}_i$. Finally, the regression loss $\mathcal{L}_{mse}$ is calculated by averaging the regression loss for all the utterances. This procedure is formulated in Formular 3.

4.2.4 Training Strategy

Our model is based on the bert-base-uncased model implemented by Huggingface Transformer repository (Wolf et al., 2019a). With 110 million parameters pre-trained on the massive corpus, we found that it is challenging to incorporate such a big model with the modules we designed in HATE. Therefore, we fixed the look-up embeddings and the parameters in the first 11 encoder layers in the BERT encoder during training, only to fine-tune the last encoder layer and the pooler layers in BERT and train other modules designed by us at the same time.

Although there are two optimization objectives ($\mathcal{L}_{ce}, \mathcal{L}_{mse}$) for HADE, it is still designed to focus on personality recognition. So, we conduct a two-stage training approach by first minimize the overall loss function $\mathcal{L} = \mathcal{L}_{ce} + \mathcal{L}_{mse}$, and then remove the gradients in the auxiliary utterance VAD regression task and only train HADE on $\mathcal{L}_{ce}$ in the second stage.

5 Experiment Settings

5.1 Dataset

Most personality recognition datasets focus on the posts on social media (Schwartz et al., 2013) or essays (Pennebaker and King, 1999). Recording daily conversation for analysis, especially including multiple speakers in the conversation, is privacy-intrusive. So, we use the **FriendsPersona** constructed by (Jiang et al., 2020) to evaluate our method. It is a dialog script dataset with personality annotations in 711 different dialogues, including 8,157 utterances. These dialogues are from the famous TV Series *Friends*. In **FriendsPersona**, the average length of the dialog flows is 11.47 utterances, while the average number of tokens for the utterances is 16.27.

The personality in **FriendsPersona** is represented as 5-d binary vectors for the big-five traits. The distribution of the personality annotations is shown in Figure 3. The **AGR**, **CON**, **EXT**, **OPN** and **NEU** indicate the big-five personality traits respectively.

To facilitate the utterance VAD regression module in our method, we also calculate the number of tokens that have accurate VAD annotations from (Mohammad, 2018) in the dataset. It suggests that among 5,346 unique tokens, 2,796 of them have

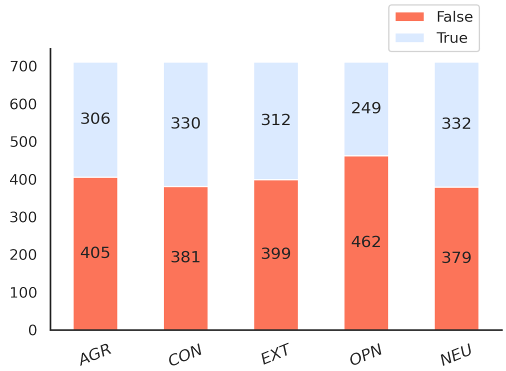

Figure 3: Personality annotations in **FriendsPersona**.

valid VAD annotations, the coverage is around 52.3%. As for the overall tokens, the corresponding number is 28.6% (27,669/96,801).

5.2 Baseline Models

To show the effectiveness of our method, we compare HADE with three state-of-the-art models as below with a personality classification task on **FriendsPersona**:

HAN: Hierarchical Attention Network (HAN) is proposed in (Yang et al., 2016). It encodes dialogue on both utterance and token levels by RNN encoders with attention layers for personality classification.

RoBERTa(S) and **RoBERTa(F)** are proposed in (Jiang et al., 2020). They use the RoBERTa (Liu et al., 2019) as the dialog encoder and try different input for personality classification. **RoBERTa(S)** only use the utterances from the analyzed speaker as input; while **RoBERTa(F)** input all the utterances within the whole dialog flow in their natural order for classification.

5.3 Ablation Study Settings

To further investigate the effectiveness of different modules in HADE and the methods we process the input, we adopt an ablation study to compare the performances of the following sub-models:

Uttr: We only use the BERT to encode the utterances from the speaker s for personality classification through a classification head.

Uttr VAD: Based on the **Uttr**, we add an auxiliary VAD regression head beside the original classification head. The additional VAD regression task is to supervise the model to extract affective information through a multi-task learning scheme.

VAD Embedding: We obtain the affective embeddings by inputting the VAD vectors of all the single words in the utterance into a linear layer. Then, we add the affective embeddings on the pre-trained look-up embeddings in BERT as the model input. This sub-model is to compare the way to utilize affective information with **Uttr VAD**.

Flow (Dialog State): We concatenate all the utterances in the whole dialog flow and feed it into the BERT encoder for personality classification. Simultaneously, we indicated the utterances from the analyzed speaker and the context with the segment embeddings in the BERT inspired by (Wolf et al., 2019b): **1** for utterances and **0** for the rest dialog context.

Hierarchical Flow: We first use the BERT model to encode each utterance in the bottom layer, and then in the second layer, we model the dialog flow as described in Section 4.2.2.

To sum up, **Uttr VAD** and **VAD Embedding** show the different ways to process the affective information; while **Flow (Dialog State)** and **Hierarchical Flow** are different approches to model the dialog flow.

5.4 Implementation Details

During implementation, we pad all the utterances with [PAD] to a MAX_LEN of 64; besides, each dialog flow is padded to 20 utterances according to the dataset statistics. The dialog flows are fed into the models in batches of 16. As for the transformer model for the dialog flow encoding in HADE, we choose four heads and 512 as the d_model according to the best performance.

Due to the limited data, we do not conduct the warm-up training. Besides, we set the drop-out rate as 0.1 to avoid overfitting in training. We use the Adam (Kingma and Ba, 2014) as the optimization algorithm in training. The learning rate for each model is selected to refer to the best performance in evaluation.

6 Results Analysis

In this section, we describe the results of the evaluation of our method through experiments with the settings above. We analyze the result by answering the following two research questions (RQs):

- RQ1: What is the performance of HADE in personality recognition in conversation?
- RQ2: How do the affective information and the dialog flow encoding influence the personality recognition HADE, respectively?

RQ1: What is the performance of our method in personality recognition in conversation?

We compare HADE with **HAN**, **RoBERTa(S)**, and **RoBERTa(F)** on binary personality classification. Following the settings in (Jiang et al., 2020), we conduct the 10-folds cross validation on **FriendsPersona**, and calculate the average classification accuracy of the test sets over the 10 splits. The results are shown in Table 2.

Model	AGR	CON	EXT	OPN	NEU	Avg
HAN	0.619	0.578	0.584	0.664	0.584	0.606
RoBERTa (S)	0.656	0.568	**0.642**	0.685	0.601	0.630
RoBERTa (F)	0.645	0.574	0.601	0.672	0.593	0.617
HADE	**0.659**	**0.627**	0.639	**0.689**	**0.643**	**0.651**

Table 2: Accuracy of binary personality classification.

We first focus on the performance of HADE. It achieves the highest accuracy (0.689) when predicting the Openness of the speakers. The lowest accuracy (0.627) occurs when indicating *Conscientiousness*. The average accuracy among the five personality traits is 0.651, and the standard deviation is around 0.021.

HADE outperforms other baseline models in four (**AGR**, **CON**, **OPN**, and **NEU**) over five personality traits with a considerable improvement. Besides, the average accuracy among the five personality traits of our model is also higher than the best baseline **RoBERTa(S)** over 3.3%. Although for **EXT**, our model does not outperform the **RoBERTa(S)**, the result is also close to the best. The results show that with our model design, the affective information and the dialog flow modeling can effectively help the personality recognition in conversation.

We also conclude that methods based on pre-trained language models are more competitive than those (e.g., **HAN**) with the traditional RNN encoders. Moreover, **RoBERTa(S)** beats **RoBERTa(F)** on overall performance, which indicates that even if input information is more, pure pre-trained language models are not appropriate to model the dialog flow data without modification.

Model	AGR	CON	EXT	OPN	NEU	Avg
Uttr	0.675 ± 0.023	0.613 ± 0.075	0.613 ± 0.134	**0.791** ± 0.002	0.632 ± 0.087	0.665
Uttr VAD	0.700 ± 0.099	**0.632** ± 0.047	0.625 ± 0.047	**0.791** ± 0.003	0.621 ± 0.089	0.674
VAD Embedding	0.642 ± 0.084	0.588 ± 0.125	0.469 ± 0.103	0.716 ± 0.052	0.602 ± 0.120	0.603
Flow (Dialog State)	0.672 ± 0.066	0.625 ± 0.098	0.614 ± 0.033	0.656 ± 0.104	0.609 ± 0.021	0.641
Hierarchical Flow	0.710 ± 0.035	0.625 ± 0.109	0.623 ± 0.023	0.780 ± 0.030	0.612 ± 0.044	0.670
HADE	**0.719** ± 0.100	0.627 ± 0.072	**0.625** ± 0.062	0.787 ± 0.017	**0.643** ± 0.091	**0.680**

Table 3: F1 scores for binary classification of personality traits.

RQ2: How do the affective information and the dialog flow encoding influence the personality recognition HADE, respectively?

After we verify the effectiveness of HADE, we are still curious about how and to what extent the modules in HATE influence the performance. Hence, we conduct an ablation study as the setting above. To better describe the personality classification performances, we use F-score (considers both precision and recall) rather than merely accuracy as the metric in the ablation study. Moreover, we run each experiment 10 times with ten different random seeds for dataset partition and model parameter initialization (except for parameters in BERT). We also record the standard deviations. The results are shown in Table 3.

In general, by integrating all the modules, HADE does outperform the **Uttr** in most of the traits, which verifies the benefit of our model design. By comparing **Uttr** and **Uttr VAD**, we can see that adding the VAD regression task improves the accuracy in **AGR** and **CON**, but slightly reduce the performance in **EXT** and **NEU**. Consequently, the average performance is still improved. Nevertheless, when we focus on the **VAD Embedding**, which modifies the look-up embeddings in the pre-trained language model by VAD vectors, the accuracy decrease in all the traits compared with both **Uttr** and **Uttr VAD**. The reason is that VAD vectors damage the original semantics stored in the look-up embeddings pre-trained in the massive corpus. However, the training dataset is too small to supervise the model to learn to process the VAD vectors in the input. Therefore, even both methods integrate the affective information in the model; only the appropriate way can preserve the strength of BERT and improve the performance.

Then, we turn to the dialog flow modeling. We compare the results between the **Uttr** and **Flow (Dialog State)** and found that although incorporating the dialog flow improves the performance in **CON** and **EXT**, it decreases the performance in other traits, especially in predicting **OPN**. It shows that similar to **VAD Embedding**, directly incorporating with the dialog state embeddings in the pre-trained language model fails to make it learn to process such information appropriately with such a small training set. However, if we focus on the performance of **Hierarchical Flow**, we can see the results are much better. So, hierarchically and separately modeling the utterances (in token level) and the dialog flow (in sentence-level) is a better approach to utilize pre-trained language models in our problem.

Combining **Uttr VAD** and **Hierarchical Flow** forms HADE and improves both sub-models. Nevertheless, we can also see that the average performance of **Uttr VAD** is slightly higher than **Hierarchical Flow**, even they are calculated on ten different random seeds. So, we conclude that affective information is more important in personality recognition under the design of HADE.

7 Conclusion and Future Work

We propose HADE to extract affective information implicitly and model the dialog flow for personality recognition in conversation. We utilize pre-defined VAD vectors of single words and design a hierarchical architecture to model the dialog flow, which solves the challenging issues met in existing works. Our model outperforms state-of-the-art methods on a public conversation dataset. Through ablation study, our approach is validated as an effective way to apply affective information into the model design with pre-trained language models.

HADE outperforms state-of-the-art models on **FriendsPersona**; we also want to verify the generality of HADE in other conversation scenarios. One significant barrier is that conversation datasets with personality annotations are rare due to privacy concerns. So, in future work, we will investigate the conversational dataset construction in a privacy-nonintrusive manner so that HADE, and even more approaches can be evaluated.

References

Nabiha Asghar, Pascal Poupart, Jesse Hoey, Xin Jiang, and Lili Mou. 2018. Affective neural response generation. In *European Conference on Information Retrieval*, pages 154–166. Springer.

Gene Ball and Jack Breese. 2000. Emotion and personality in a conversational agent. *Embodied conversational agents*, 189.

Siqi Bao, Huang He, Fan Wang, Hua Wu, and Haifeng Wang. 2019. Plato: Pre-trained dialogue generation model with discrete latent variable. *arXiv preprint arXiv:1910.07931*.

C Bauerhenne, A Gammoudi, M Moussaoui, J Reichelt, and J Wang. 2020. Emotional states and personality profiles in conversational ai.

Margaret M Bradley and Peter J Lang. 1999. Affective norms for english words (anew): Instruction manual and affective ratings. Technical report, Technical report C-1, the center for research in psychophysiology ....

Raymond B Cattell. 1943. The description of personality: Basic traits resolved into clusters. *The journal of abnormal and social psychology*, 38(4):476.

Kyunghyun Cho, Bart Van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning phrase representations using rnn encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078*.

Pierre Colombo, Wojciech Witon, Ashutosh Modi, James Kennedy, and Mubbasir Kapadia. 2019. Affect-driven dialog generation. *arXiv preprint arXiv:1904.02793*.

Paul T Costa and Robert R McCrae. 1992. Normal personality assessment in clinical practice: The neo personality inventory. *Psychological assessment*, 4(1):5.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Jess Feist and Gregory Feist. 2012. J.(2008). theories of personality.

Susan T Fiske and Shelley E Taylor. 1991. *Social cognition*. Mcgraw-Hill Book Company.

Jennifer Golbeck, Cristina Robles, and Karen Turner. 2011. Predicting personality with social media. In *CHI'11 extended abstracts on human factors in computing systems*, pages 253–262.

Xiaodong Gu, Kang Min Yoo, and Jung-Woo Ha. 2020. Dialogbert: Discourse-aware response generation via learning to recover and rank utterances. *arXiv preprint arXiv:2012.01775*.

Meng-Ju Han, Chia-How Lin, and Kai-Tai Song. 2012. Robotic emotional expression generation based on mood transition and personality model. *IEEE transactions on cybernetics*, 43(4):1290–1303.

Eddie Harmon-Jones, Philip A Gable, and Tom F Price. 2013. Does negative affect always narrow and positive affect always broaden the mind? considering the influence of motivational intensity on cognitive scope. *Current Directions in Psychological Science*, 22(4):301–307.

Francisco Iacobelli, Alastair J Gill, Scott Nowson, and Jon Oberlander. 2011. Large scale personality classification of bloggers. In *international conference on affective computing and intelligent interaction*, pages 568–577. Springer.

Hang Jiang, Xianzhe Zhang, and Jinho D Choi. 2020. Automatic text-based personality recognition on monologues and multiparty dialogues using attentive networks and contextual embeddings (student abstract). In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 13821–13822.

Oliver P John, Alois Angleitner, and Fritz Ostendorf. 1988. The lexical approach to personality: A historical review of trait taxonomic research. *European journal of Personality*, 2(3):171–203.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Randall J Larsen and David M Buss. 2008. Psihologija ličnosti. *Naklada Slap, Jastrebarsko*.

Joosung Lee and Wooin Lee. 2021. Compm: Context modeling with speaker's pre-trained memory tracking for emotion recognition in conversation. *arXiv preprint arXiv:2108.11626*.

Zhaojiang Lin, Andrea Madotto, Jamin Shin, Peng Xu, and Pascale Fung. 2019. Moel: Mixture of empathetic listeners. *arXiv preprint arXiv:1908.07687*.

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*.

Veronica Lynn, Niranjan Balasubramanian, and H Andrew Schwartz. 2020. Hierarchical modeling for user personality prediction: The role of message-level attention. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5306–5316.

François Mairesse and Marilyn Walker. 2006. Automatic recognition of personality in conversation. In *Proceedings of the Human Language Technology Conference of the NAACL, Companion Volume: Short Papers*, pages 85–88.

François Mairesse, Marilyn A Walker, Matthias R Mehl, and Roger K Moore. 2007. Using linguistic cues for the automatic recognition of personality in conversation and text. *Journal of artificial intelligence research*, 30:457–500.

Naoki Masuyama, Chu Kiong Loo, and Manjeevan Seera. 2018. Personality affected robotic emotional model with associative memory for human-robot interaction. *Neurocomputing*, 272:213–225.

Matthias R Mehl, Samuel D Gosling, and James W Pennebaker. 2006. Personality in its natural habitat: Manifestations and implicit folk theories of personality in daily life. *Journal of personality and social psychology*, 90(5):862.

Albert Mehrabian. 1995. Relationships among three general approaches to personality description. *The journal of Psychology*, 129(5):565–581.

Albert Mehrabian. 1996. Analysis of the big-five personality factors in terms of the pad temperament model. *Australian journal of Psychology*, 48(2):86–92.

Shikib Mehri, Evgeniia Razumovskaia, Tiancheng Zhao, and Maxine Eskenazi. 2019. Pretraining methods for dialog context representation learning. *arXiv preprint arXiv:1906.00414*.

Saif Mohammad. 2018. Obtaining reliable human ratings of valence, arousal, and dominance for 20,000 english words. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 174–184.

Daniel Ricardo Jaimes Moreno, Juan Carlos Gomez, Dora-Luz Almanza-Ojeda, and Mario-Alberto Ibarra-Manzano. 2019. Prediction of personality traits in twitter users with latent features. In *2019 International Conference on Electronics, Communications and Computers (CONIELECOMP)*, pages 176–181. IEEE.

Piotr Nawrot, Szymon Tworkowski, Michał Tyrolski, Łukasz Kaiser, Yuhuai Wu, Christian Szegedy, and Henryk Michalewski. 2021. Hierarchical transformers are more efficient language models. *arXiv preprint arXiv:2110.13711*.

James W Pennebaker and Laura A King. 1999. Linguistic styles: language use as an individual difference. *Journal of personality and social psychology*, 77(6):1296.

Stefan Rank, Marcin Skowron, and David Garcia. 2013. Dyads to groups: modeling interactions with affective dialog systems. *International Journal of Computational Linguistics*, 4(1):22–37.

Esteban Andres Rissola, Seyed Ali Bahrainian, and Fabio Crestani. 2019. Personality recognition in conversations using capsule neural networks. In *IEEE/WIC/ACM International Conference on Web Intelligence*, pages 180–187.

H Andrew Schwartz, Johannes C Eichstaedt, Margaret L Kern, Lukasz Dziurzynski, Stephanie M Ramones, Megha Agrawal, Achal Shah, Michal Kosinski, David Stillwell, Martin EP Seligman, et al. 2013. Personality, gender, and age in the language of social media: The open-vocabulary approach. *PloS one*, 8(9):e73791.

Iulian Serban, Alessandro Sordoni, Ryan Lowe, Laurent Charlin, Joelle Pineau, Aaron Courville, and Yoshua Bengio. 2017. A hierarchical latent variable encoder-decoder model for generating dialogues. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 31.

Marcin Skowron, Mathias Theunis, Stefan Rank, and Arvid Kappas. 2013. Affect and social processes in online communication–experiments with an affective dialog system. *IEEE Transactions on Affective Computing*, 4(3):267–279.

Alireza Souri, Shafigheh Hosseinpour, and Amir Masoud Rahmani. 2018. Personality classification based on profiles of social networks' users and the five-factor model of personality. *Human-centric Computing and Information Sciences*, 8(1):1–15.

Edward P Tighe, Jennifer C Ureta, Bernard Andrei L Pollo, Charibeth K Cheng, and Remedios de Dios Bulos. 2016. Personality trait classification of essays with the application of feature reduction. In *SAAIP@ IJCAI*, pages 22–28.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.

David Watson and Lee Anna Clark. 1992. On traits and temperament: General and specific factors of emotional experience and their relation to the five-factor model. *Journal of personality*, 60(2):441–476.

Zhiyuan Wen, Jiannong Cao, Ruosong Yang, Shuaiqi Liu, and Jiaxing Shen. 2021. Automatically select emotion for response via personality-affected emotion transition. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 5010–5020.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. 2019a. Huggingface's transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771*.

Thomas Wolf, Victor Sanh, Julien Chaumond, and Clement Delangue. 2019b. Transfertransfo: A transfer learning approach for neural network based conversational agents. *arXiv preprint arXiv:1901.08149*.

Zichao Yang, Diyi Yang, Chris Dyer, Xiaodong He, Alex Smola, and Eduard Hovy. 2016. Hierarchical attention networks for document classification. In *Proceedings of the 2016 conference of the North American chapter of the association for computational linguistics: human language technologies*, pages 1480–1489.

Yizhe Zhang, Siqi Sun, Michel Galley, Yen-Chun Chen, Chris Brockett, Xiang Gao, Jianfeng Gao, Jingjing Liu, and Bill Dolan. 2019. Dialogpt: Large-scale generative pre-training for conversational response generation. *arXiv preprint arXiv:1911.00536*.

Peixiang Zhong, Di Wang, and Chunyan Miao. 2019. Knowledge-enriched transformer for emotion detection in textual conversations. *arXiv preprint arXiv:1909.10681*.

Hao Zhou, Pei Ke, Zheng Zhang, Yuxian Gu, Yinhe Zheng, Chujie Zheng, Yida Wang, Chen Henry Wu, Hao Sun, Xiaocong Yang, et al. 2021. Eva: An open-domain chinese dialogue system with large-scale generative pre-training. *arXiv preprint arXiv:2108.01547*.