Adding SPICE to Life: Speaker Profiling in Multiparty Conversations

Shivani Kumar[1], **Rishabh Gupta**[1], **Md Shad Akhtar**[1], **Tanmoy Chakraborty**[2]
[1]IIIT Delhi, [2]IIT Delhi
{shivaniku, rishabh19089, shad.akhtar}@iiitd.ac.in, tanchak@iitd.ac.in

Abstract

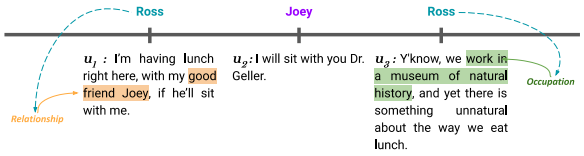
In the realm of conversational dynamics, individual idiosyncrasies challenge the suitability of a one-size-fits-all approach for dialogue agent responses. Prior studies often assumed the speaker's persona's immediate availability, a premise not universally applicable. To address this gap, we explore the Speaker Profiling in Conversations (SPC) task, aiming to synthesize persona attributes for each dialogue participant. SPC comprises three core subtasks: persona discovery, persona-type identification, and persona-value extraction. The first subtask identifies persona-related utterances, the second classifies specific attributes, and the third extracts precise values for the persona. To confront this multifaceted challenge, we've diligently compiled SPICE, an annotated dataset, underpinning our thorough evaluation of diverse baseline models. Additionally, we benchmark these findings against our innovative neural model, SPOT, presenting an exhaustive analysis encompassing a nuanced assessment of quantitative and qualitative merits and limitations.

Keywords: Speaker profiling, personalisation, dialogue systems, dialogue understanding, natural language processing.

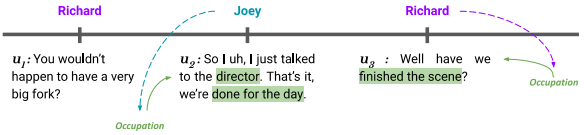
1. Introduction

Understanding natural language inputs is crucial for effective processing, as evidenced by a substantial body of work dedicated to the analysis of standalone textual content (Schank, 1972; Pruksachatkun et al., 2020; Zhou et al., 2013; Badjatiya et al., 2017). However, recent research has shifted towards contextual conversational data, emphasizing the need for mutual understanding among speakers and leading to extensive investigations in emotional analysis (Poria et al., 2019; Jiao et al., 2020; Shen et al., 2020), intent discernment (Larson et al., 2019; Gangadharaiah and Narayanaswamy, 2019), and dialogue act detection (Qin et al., 2020; Liu et al., 2017). This change is driven by the growing prevalence of dialogue agents, necessitating contextually appropriate response generation. In this context, research has explored the engagement of participants, including empathetic (Lin et al., 2020; Shin et al., 2019; Rashkin et al., 2018) and stylistic dialogue generation (Su et al., 2020; Akama et al., 2017; Danescu-Niculescu-Mizil and Lee, 2011). While such agents enhance system appeal, there is a need to address personalized dialogue generation, incorporating users' personas as essential inputs (Zhang et al., 2018; Weston et al., 2018; Roller et al., 2020; Dinan et al., 2018; Chen et al., 2020). Although persona details improve response intuitiveness and engagement (Zhang et al., 2018; Weston et al., 2018; Roller et al., 2020; Dinan et al., 2018), the studies in this domain assume prior persona provision, a rarity in practical applications.

To tackle the challenge of persona information unavailability within chatbots, we embark on the task of **Speaker Profiling in Conversations (SPC)**. SPC is geared towards the creation of comprehensive profiles for all participants engaged in a conversation, encompassing various speaker-centric attributes, including traits, likes, and occupation. To elucidate the intricacies of SPC, we offer an illustrative depiction through two dialogues, as showcased in Figure 1. Within the first dialogue (Figure 1a), Ross characterizes Joey as his good friend, thereby establishing a friendly rapport between them. Furthermore, Ross explicitly discloses his occupation, citing his role in a natural history museum within his third utterance (u_3). While the identification of overt relationship and occupational details appears straightforward in meticulously curated dialogues, authentic conversations frequently involve implicit information that poses a formidable identification challenge. The second dialogue depicted in Figure 1b underscores the complexity of extracting persona-related data concerning one's occupation, often demanding specialized



(a) Explicit Persona



(b) Implicit Persona

Figure 1: Persona information in dialogues.

LREC-COLING 2024, pages 415–425
20-25 May, 2024.

knowledge to discern phrases such as 'director' and 'scene' as indicative of a profession within the movie industry. This paper undertakes the dual responsibility of addressing both explicit and implicit manifestations of persona identification.

The intricate task of speaker profiling unfolds into a triad of subtasks: *persona discovery*, *persona-type identification*, and *persona-value extraction*. In the first subtask, the objective is to discern which utterances within the conversation harbor persona-related insights. Subsequently, the second subtask entails the discernment of the specific persona information type within each identified utterance. Finally, the last subtask involves the meticulous extraction of precise values associated with each recognized persona type. To bolster research efforts in this domain, we present SPICE[1], a novel dataset teeming with multi-party conversations, thoughtfully adorned with annotated labels for all three subtasks. Complementing this, we introduce SPOT[2], a neural methodology that amalgamates RoBERTa (Liu et al., 2020), Transformer (Vaswani et al., 2017), and attention based methods, adept at capturing both the minutiae of dialogue-level context and the nuances of speaker-specific context for persona discovery. In our rigorous evaluation, SPOT outshines four baseline approaches, both in standalone and pipeline configurations, excelling in both subtasks. To gain deeper insights into its efficacy, we conduct a comprehensive analysis of the discrete components of SPOT, thereby affording a more nuanced understanding of its strengths and limitations.

In a nutshell, our contributions are four-fold:

1. We explore the problem of **Speaker Profiling in Conversations** from a new angle, where given a dialogue as input, the task is to extract the speaker-centric personality information of all speakers present in the dialogue.
2. We curate SPICE, a **multi-party conversation dataset** with human annotated SPC labels.
3. We benchmark it with a RoBERTa and attention based **novel model**, SPOT for the SPC task.
4. We perform a **comparative analysis** of our model with several baselines and establish the superiority of SPOT.

Reproducibility: The source code for SPOT and the SPICE dataset (partial) along with the execution instructions can be found here: bit.ly/3o4sWIU.

2. Related Work

Various studies have focused on natural language understanding in conversations, including intent classification (Larson et al., 2019; Gangadharaiah and Narayanaswamy, 2019), dialogue act recognition (Qin et al., 2020; Liu et al., 2017), and emotion analysis (Poria et al., 2019; Jiao et al., 2020; Shen et al., 2020). The primary aim of comprehending conversations is to develop more engaging dialogues. One way to achieve this is by catering to the interests of the users. This can be illustrated by the following example: Suppose Andrew wants to go on a date with Lisa, and he already knows her likes and dislikes. With this knowledge, Andrew can not only arrange an outstanding date but also engage in captivating conversations with Lisa. Similarly, for online dialogue agents, such additional information can enhance dialogue generation.

Personalised dialogue systems. It has been widely recognized that personalization improves the performance of dialogue systems (Weston et al., 2018; Roller et al., 2020; Dinan et al., 2018; Chen et al., 2020; Lucas et al., 2009; Joshi et al., 2017). While certain investigations have concentrated on user profiles to customize goal-oriented dialogue systems (Lucas et al., 2009; Joshi et al., 2017), recent research has shifted towards informal chit-chat settings. Historically, personalization, in vector format, has been applied, exemplified by Li et al. (2016), wherein distributed embeddings were acquired for individual Twitter users, encapsulating unique attributes like writing style and prior experience. Subsequently, Zhang et al. (2018) introduced the Persona-chat dataset, encompassing $10,907$ dyadic dialogues wherein speakers were endowed with fictional personas, dictating their conversational demeanor. Several studies subsequently underscored the Persona-chat dataset's efficacy in enhancing personalization when generating responses for users (Weston et al., 2018; Roller et al., 2020; Dinan et al., 2018). Nevertheless, the utility of persona data for dialogue generation necessitates careful consideration, as it should not be assumed to be readily accessible.

Persona identification in dialogues. Many studies have dealt with identifying most fitting persona for a speaker using classification mechanism, where given a dialogue, the task is to classify each speaker into one of the persona categories (Chu et al., 2018; Gu et al., 2021). Other studies try to extract speaker qualities from the given dialogues (Tigunova et al., 2019; Wu et al., 2019).

How is Our Task Different? Several studies have attempted to extract speaker characteristics from dialogues, as previously noted. To illustrate the differences between these studies, Figure 2 presents a sample dialogue. Tigunova et al. (2019) extracted four types of persona information, including profession, gender, age, and family status, while Wu et al. (2019) extracted information in the form of triplets, both of which used heuristics to collect ground-truth labels without the use of

[1] SPICE: **S**peaker **P**rofiling **I**n **C**onv**E**rsation
[2] SPOT: **S**peaker **Pr**O**f**iling using **T**ransformers

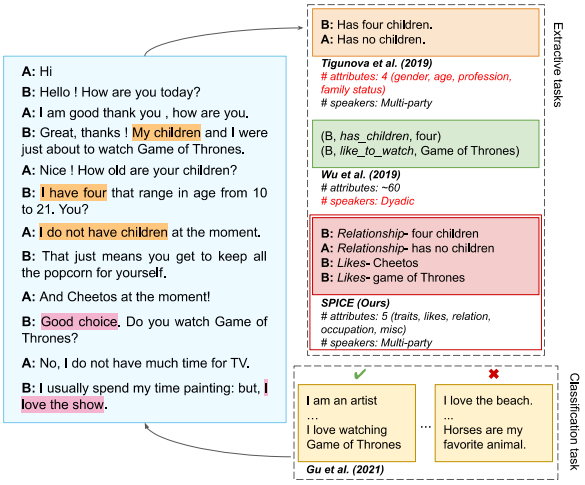


Figure 2: Difference between speaker profiling in conversation (proposed) and related works.

human-level gold labels. In contrast, our solution is based on human-annotated ground-truth labelling, which ensures cleaner and less noisy data. It is important to note that while Wu et al. (2019) is only able to handle dyadic conversations, our task can also handle multi-party scenarios. Gu et al. (2021) projected the task of speaker profiling as a classification task, determining the ranking among the available personalities and assigning the most appropriate identity to the dialogue where a speaker can only be assigned a pre-existing persona. In contrast, our proposed work extracts persona on the fly in a generative way.

3. Problem Statement

The task of SPC can be conceptualized as a synthesis of three subtasks. Formally, we define the subtasks below:

- ***Persona discovery***: Given a sequence of n utterances as a dialogue, $D = \{u_1, u_2, \ldots, u_n\}$, in a multiparty conversation where u_i represents the i^{th} utterance of the dialogue, we associate a binary label with u_i, $i \in [1, n]$, signifying whether u_i contains a persona information of the speaker articulating that utterance.
- ***Persona-type identification***: Given an instance, $I_p = \{u_1, u_2, ..., u_m\}$, where u_m is identified as an utterance containing persona information, we associate a label p, which represents what type of persona is present in the concerned utterance. Here, p belongs to a set of carefully curated possible persona types P. Section 4 elaborates on the types of persona we consider.
- ***Persona-value extraction***: Given an instance, $I_v = \{u_1, u_2, ..., u_m\}$, where u_m contains information about the presence of persona information and persona type p, the task is to extract the exact persona value v associated with p from I_v.

Set	#Dlg	#Utt	#Sp/Dlg	#P Utt	#P Utt/Dlg
Train	1039	9989	2.70	1005	0.96
Dev	114	1109	3.01	109	0.96
Test	280	1983	2.66	305	1.09

Set	#Persona Slot				
	Trait	Likes	Relation	Occ	Misc
Train	389	244	107	89	179
Dev	32	36	10	10	24
Test	120	88	28	18	53

Table 1: Statistics of SPICE. (Dlg: Dialogue; Utt: Utterance; Sp: Speaker; P Utt: Persona Utterance)

4. Dataset

We introduce a new dataset, SPICE, tailored for speaker profiling in multi-party dialogues. Leveraging conversations extracted from the MELD dataset (Poria et al., 2018), we meticulously annotate each utterance for our designated task. Following MELD's original train-dev-test distribution, we undertake three subtasks for annotation.

- *Persona discovery*: We identify the presence of persona information in each utterance of the dialogue by marking it as 'yes' in this subtask.
- *Persona-type identification*: We associate a type of persona with each utterance marked as 'yes' in the previous phase, within this subtask. Following a comprehensive analysis of each conversation in the dataset, we define five persona types - *trait*, *likes*, *relation*, *occupation*, and *misc* - to encapsulate various personality characteristics of the speakers.
- *Persona-value extraction*: In this subtask, we extract persona values from the given instance for each identified persona type. These values may include a span from the input (e.g., for *occupation*), reference to another speaker present in the conversation (e.g., for *relationship*), or something inferred from the context (e.g., for *trait*).

Three annotators[3] were engaged in annotating SPICE. The initial two annotators assigned relevant persona labels to dialogue utterances, with any discrepancies resolved by the third annotator. Inter-annotator agreement was assessed using Krippendorff's Alpha (Krippendorff, 2011). For persona discovery, an inter-annotator agreement score of 0.83 was attained, while persona-type identification achieved an agreement score of 0.71. Refer to Table 1 for dataset statistics, including the persona type distribution within SPICE.

5. Proposed Methodology

In this section, we illustrate SPOT, our proposed method to benchmark the task and the dataset.

[3]They were NLP researchers or linguistics by profession; and their age ranges between $20 - 45$ years.

SPOT constitutes three subtasks – persona discovery, persona-type identification, and persona-value extraction.

5.1. Persona Discovery

In this stage, we employ a RoBERTa encoder (Liu et al., 2020) to capture dialogue-level contextual information, as illustrated in Figure 3. The model input comprises a sequence of utterances forming the dialogue, denoted as $D = \{u_1, u_2, ..., u_m\}$. Subsequently, the dialogue-level representations are fed into fully-connected layers for classification. It's noteworthy that SPICE, resembling a real world scenario, exhibits an inherent skew towards utterances lacking persona information. To address this imbalance, we apply the SMOTE upsampling technique (Chawla et al., 2002) to boost the representation of persona-related utterances.

5.2. Persona-type Identification

With a clear identification of persona-bearing utterances within the dialogue, the focus shifts to the subsequent subtask. This phase processes a sequence of utterances denoted as $I = \{u_1, u_2, ..., u_k\}$, where u_k signifies the target utterance, i.e., the one containing persona information, while $u_1, u_2, ..., u_{k-1}$ encompass the contextual utterances. For this task, our proposition integrates a fusion of RoBERTa and Transformer (Vaswani et al., 2017) elements, as depicted in Figure 3.

Dialogue representation. Every utterance u_j within the input sequence I undergoes processing via a Transformer layer to yield its representation.

Speaker-specific representation. In our quest to proficiently encapsulate the speaker sequence within a dialogue, we implement distinct Transformer encoder layers, one for each participating speaker in the discourse, thus yielding contextually tailored speaker-specific representations. Each of these speaker-specific encoders receives input from the utterance representations that are specific to the respective speaker i.

Attention. The creation of a speaker-aware representation, denoted as H_{SAR}, hinges on the judicious application of attention mechanisms (Bahdanau et al., 2014) that facilitate the integration of speaker-specific representations with the target representation. In a parallel fashion, the derivation of a context-aware representation, designated as H_{CAR}, transpires seamlessly. Building upon this foundation, we embark on the synthesis of a global attention representation, denoted as H_{GAR}, serving as the cohesive fusion point for the speaker-aware and context-aware representations.

Adaptive decision boundary. Additionally, to capture the dialogue-level context effectively, we forward the RoBERTa embeddings to the model as a skip connection. Finally, we acquire the adaptive decision boundary for the persona classes through a k-means (Lloyd, 1982; MacQueen et al., 1967) methodology. Moreover, for the optimization of decision boundaries specific to each class and to achieve maximal class separation, we employ the innovative boundary loss (Zhang et al., 2021). The loss is computed using the following equation:

$$L_b = \frac{1}{N}\sum_{i=1}^{N}[\Delta_i(||z_i - c_{yi}||_2 - \delta_{yi}) + (1 - \Delta_i)(\delta_{yi} - ||z_i - c_{yi}||_2)]$$

where N is the total number of samples in our set, z_i is the representation of the i^{th} instance, c_{yi} is the centroid for class yi, and δ_{yi} is the radius for class yi. Here,

$$\Delta_i = \begin{cases} 1, if ||z_i - c_{yi}||_2 > \delta_{yi} \\ 0, if ||z_i - c_{yi}||_2 \leq \delta_{yi} \end{cases}$$

5.3. Persona-value Extraction

In the culmination of shaping a speaker profile, the extraction of persona values for designated persona types within a dialogue emerges as the ultimate stride. Notably, these persona values can be conjectured from the input, often lacking specific constraints. Consequently, we adopt an encoder-decoder framework aligned with a generative objective to undertake this task effectively. This endeavor hinges upon adeptly encompassing the entirety of the conversation to grasp its essence, the contextual utterances to assimilate contextual knowledge, and the focal utterance, as it constitutes the primary wellspring for persona attributes.

In our approach, we employ a BART encoder (Lewis et al., 2020) to meticulously encode the context, target, and dialogue utterances, resulting in c, t, and d, respectively. These representations undergo a pivotal phase where an attention mechanism amalgamates the key k_c and value v_c extracted from the context representation with the query q_t derived from the target utterance. This sophisticated interplay encapsulates the dynamic interaction between the target utterance and contextual utterances, thereby adeptly capturing the context-driven persona nuances embedded in the target utterance. The resultant representation is seamlessly integrated with the dialogue representation, and subsequently channeled into the BART decoder for the generation of output.

6. Experiments and Results

6.1. Experimental Setup

We perform experiments for all three subtasks of speaker profiling in two settings – standalone and pipeline. Following sections present both settings.

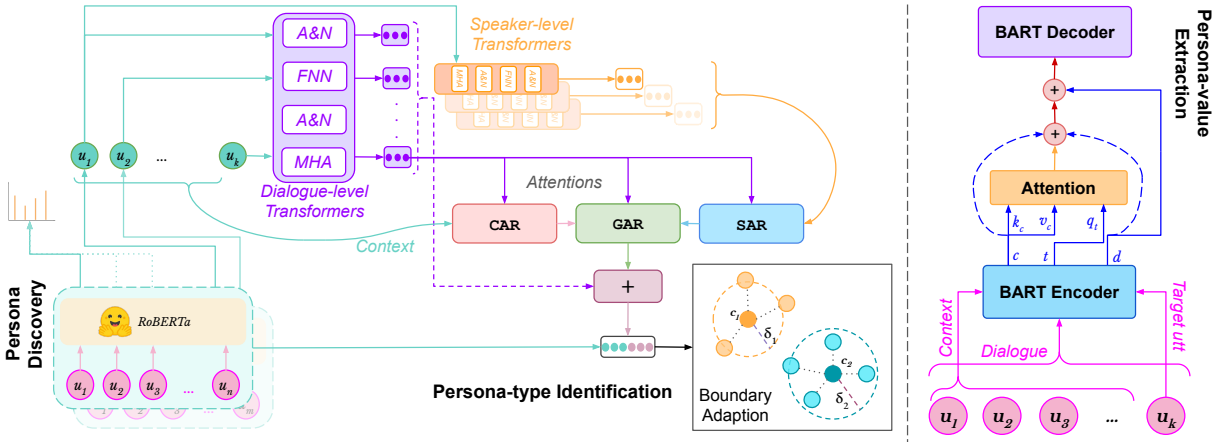


Figure 3: *Persona discovery*: The utterance representations obtained from dialogue-level RoBERTa are used for classification. *Persona-type identification*: Utterance representations are obtained from Dialogue-level Transformers and the speaker-specific Transformers. After receiving the representation from context, speaker, and global attention mechanism, the final representation is used to obtain adaptive decision boundary. We initialize the centroids $\{c_i\}_{i=1}^{K}$ and the radius of decision boundaries $\{\delta\}_{i=1}^{K}$ for each persona type and use the boundary loss for optimisation. *Persona-value extraction*: The context, target utterance, and the complete dialogue is transformed using a BART encoder following which attention is applied to get target attended vectors. Finally a concatenated vector is sent to the BART decoder for output generation.

Standalone Evaluation. In this configuration, the distinct phases' models are individually trained and assessed. To elucidate, in the persona discovery phase, all dialogues undergo processing by SPOT, meticulously scrutinizing each utterance for persona-related cues. Transitioning to persona-type identification, we strictly adhere to the ground-truth, selectively forwarding solely the persona-laden utterances, alongside their contextual information, to the model, as visually depicted in Figure 3. Ultimately, when it comes to persona-value extraction, we furnish SPOT with the ground truth persona categories, together with the persona-imbued utterances and their contextual backdrop, thereby enabling the precise extraction of persona values.

Pipeline evaluation. In this context, the persona discovery process aligns with the standalone setup. Yet, in the persona-type identification phase, we exclusively supply the model with the utterances pinpointed as persona-bearing in the preceding subtask, without adhering to the ground truth. Subsequently, the results yielded from the second phase serve as input for the ultimate persona-value extraction task, devoid of any ground-truth reference.

Baseline methods. Given that persona discovery and persona-type identification are classification-oriented tasks, we have leveraged four classification baselines that are originally designed for akin tasks like emotion detection and dialogue-act identification. **BERT:** BERT (Bidirectional Encoder Representations from Transformers) (Devlin et al., 2018) is encoder stack of transformer architecture (Vaswani et al., 2017). We use pre-trained BERT base and fine-tune it for our tasks. **RoBERTa:** RoBERTa (Robust BERT) (Liu et al., 2020) extends upon BERT by adjusting critical hyperparameters, eliminating the next-sentence pretraining task, and utilizing significantly larger mini-batches and learning rates during training. **DialogXL:** Shen et al. (Shen et al., 2020) modified XLNet by changing the segment-level recurrence mechanism to an utterance-level recurrence mechanism so that XLNet could be mapped to a dialogue setting. They also incorporated dialogue-aware self-attention to capture the intra- and inter-speaker dependencies in a conversation. **Co-GAT:** Qin et al. (Qin et al., 2020) proposed a co-interactive graph interaction layer with cross-utterance and cross-tasks connections. **AGHMN:** Jiao et al. (Jiao et al., 2020) used an attention-based GRU to monitor the flow of information through a hierarchical memory network. The attention weights are calculated over the contextual utterances in the conversation and combined for the final classification.

For the third subtask of persona-value extraction, we consider sequence-to-sequence models for comparison. **RNN:** OpenNMT5 provides with an implementation of the RNN seq-to-seq architecture which we use in our study. **Transformers:** We use the standard encoder and decoder stack to generate the output (Vaswani et al., 2017). **Pointer Generator Network (PGN):** The standard seq-to-seq architecture supporting both generation of new words as well as copying words from input (See et al., 2017). **BART:** BART (Lewis et al., 2020) contains a bidirectional encoder and an auto-

Systems	*Persona Discovery*			*Persona-type Identification*					
	P	R	F1	Trait	Likes	Rel	Misc	Occ	Wtd
BERT	0.17	0.72	0.27	0.48	0.0	0.09	0.24	0.05	0.24
RoBERTa	0.20	0.56	0.29	0.51	0.17	0.11	0.26	0.04	0.27
DialogXL	0.45	0.23	0.31	0.52	0.0	0.0	0.0	0.0	0.18
Co-GAT	0.15	**0.94**	0.27	0.50	0.35	0.06	0.14	0.05	0.33
AGHMN	0.43	0.14	0.21	0.56	0.58	0.38	0.26	0.25	0.48
SPOT	**0.47**	0.58	**0.52**	**0.61**	**0.63**	**0.49**	**0.35**	**0.31**	**0.56**

Table 2: Comparative results for standalone evaluation. (P: Precison; R: Recall; Rel: Relationship; Occ: Occupation; Wtd: Weighted F1 score.)

Systems	*Persona Discovery*			*Persona-slot Identification*					
	P	R	F1	Trait	Likes	Rel	Misc	Occ	Wtd
BERT	0.17	0.72	0.27	0.12	0.0	0.05	0.07	0.0	0.07
RoBERTa	0.20	0.56	0.29	0.14	0.02	0.07	0.09	0.02	0.09
DialogXL	0.45	0.23	0.31	0.33	0.0	0.0	0.0	0.0	0.07
Co-GAT	0.15	**0.94**	0.27	0.31	0.26	0.0	0.07	0.0	0.15
AGHMN	0.43	0.14	0.21	0.48	0.43	**0.41**	0.21	0.16	0.40
SPOT	**0.47**	0.58	**0.52**	**0.57**	**0.52**	0.38	**0.28**	**0.24**	**0.46**

Table 3: Comparative results for pipeline evaluation. (P: Precison; R: Recall; Rel: Relationship; Occ: Occupation; Wtd: Weighted F1 score.)

regressive decoder to create a denoising autoencoder model. **T5:** T5 (Raffel et al., 2020) is a seq-to-seq model trained on a mixture of unsupervised and supervised tasks.

Evaluation metrics. Since the first two subtasks are multi-class classification problems, we use F1 score as our choice of evaluation metric. We consider F1 score of the positive class for the task of *persona discovery*, while weighted F1 score is used for *persona type identification*. On the other hand, since the task of *persona-value extraction* follows a generative objective, we use the ROUGE (Lin, 2004) and the BLEU (Papineni et al., 2002) scores to gauge the performance of the systems.

6.2. Results

Standalone evaluation. We evaluate SPOT for all subtasks of SPC separately and show the results in Table 2 and Table 4.

- *Persona discovery*: We train SPOT as a binary classifier using cross-entropy loss. We obtain 52% F1 score, which is $\sim 21\%$ better than the best baseline, DialogXL, as can be seen in Table 2. The gain in performance can be attributed to the efficient way we use different modules in our architecture to capture different essence of a conversation. It is interesting to observe that Co-GAT produce the best performance in terms of recall scores while SPOT holds a balance between precision and recall to obtain the highest F1 score, which is our metric of choice due to the class imbalance present in our data.
- *Persona-type identification*: We use boundary loss (Zhang et al., 2021) to train SPOT for this task. Table 2 shows that SPOT yields a weighted average of 56% F1-score with the maximum score for persona-type *likes*. AGHMN, the best baseline, results in a weighted average of 48% F1-score, which is $\sim 8\%$ less than SPOT. We observe that SPOT achieves the best F1-score for all persona slots showcasing a global dominance of our system. It is interesting to observe that our model performs quite well across the persona slots that are dominantly present in our data and consistently decreases for the slots based on their availability in the data.
- *Persona-value extraction*: Using a generative objective, we obtain the results by SPOT for this subtask. Table 4 demonstrates that SPOT outperforms the baselines by around 1% for all metrics except BLEU 1 and BLEU 3.

Models	Standalone					Pipeline				
	R1	R2	B1	B2	B3	R1	R2	B1	B2	B3
RNN	26.85	2.28	24.78	1.48	0.36	19.64	0.37	18.98	0.36	1.12
Transformer	26.02	2.37	23.93	1.54	0.58	19.48	0.70	18.82	0.69	2.05
PGN	24.40	1.59	23.12	1.08	0.36	17.11	0.49	16.13	0.33	**9.28**
BART	28.93	2.16	**27.23**	1.51	0.36	22.41	**1.01**	21.37	0.80	0.08
T5	15.07	0.0	14.90	2.25	**1.20**	11.62	0.37	11.51	0.36	1.12
SPOT	**29.51**	**2.97**	27.16	**2.27**	0.60	**23.40**	0.60	**22.12**	**1.12**	0.08

Table 4: Comparative results for standalone and pipeline evaluation for *persona-value extraction*. (R1/2: ROUGE1/2; B1/2/3: BLEU1/2/3)

Systems	Trait	Likes	Relation	Misc	Occ	Weighted
SPOT_{Base}	0.36	0.27	0.22	0.23	0.07	0.28
$\text{SPOT}_{Base+RoBERTA}$	0.56	0.53	0.45	0.40	0.26	0.49
SPOT	**0.61**	**0.63**	**0.49**	**0.35**	**0.31**	**0.56**

Table 5: Ablation results for *persona-type identification*. (Misc: Miscellaneous; Occ: Occupation)

Pipeline evaluation. Tables 3 and 4 show the performance obtained by our model along with the baseline scores. For the task of persona discovery, we obtain the same results as standalone due to the same type of input and evaluation strategies. However, we observe a performance drop of $\sim 10\%$ for the persona-type identification task and a drop of at most $\sim 6\%$ for the persona-value extraction task when compared to the standalone results. This is expected as the erroneous predictions from the previous stage may propagate to the next stage. Nevertheless, when compared to the baseline systems, our proposed mechanism gives the best score, with an increase of $\sim 6\%$ over the best baseline and $\sim 39\%$ over the worst performing baseline for the former task (c.f. Table 3). Apart from *relation*, SPOT performs the best for all persona slots. While for the last subtask, we obtain an improvement of $\sim 1\%$ over the baselines. Consequently, we establish that SPOT is able to capture the essence of persona more clearly when compared with the baseline systems.

6.3. Ablation Study

SPOT captures two primary aspects of a dialogue – the dialogue context and the speaker semantics. To capture the dialogue-level context, we use SPOT_{Base}, containing the dialogue-level Transformer followed by context H_{CAR}, and global attention representation H_{GAR}. We reinforce the dialogue context by using the RoBERTa represen-

True \ Predicted	No	Yes
No	1191	487
Yes	121	**184**

(a) SPOT

True \ Predicted	No	Yes
No	1594	84
Yes	235	**70**

(b) DialogXL

Table 6: Confusion matrices for SPOT and DialogXL (best baseline) for the persona discovery task.

True \ Predicted	Trait	Occ	Misc	Likes	Relation
Trait	72/65	8/11	10/10	22/23	6/9
Occ	4/5	8/6	3/3	2/3	1/1
Misc	16/16	7/8	17/12	5/7	8/10
Likes	24/25	2/5	3/4	55/51	4/3
Relation	3/3	0/0	7/8	3/5	17/12

Table 7: Confusion matrices for the *persona-type identification* task. Each cell represents value like {SPOT/AGHM}.

tation as a skip connection in this architecture, $\text{SPOT}_{Base+RoBERTa}$. Speaker semantics are captured by the speaker-specific Transformers and attention representation H_{SAR}. We add these modules in our final model, SPOT. We observe that the addition of the RoBERTa representation improves the performance of our model considerably (21%) while the addition of speaker module improves it further (7%) verifying the use of each module.

7. Error Analysis

In this section, we present a detailed analysis of the results obtained for SPOT. We first show the quantitative analysis by analysing the confusion matrices obtained. After this, we show a qualitative analysis by observing a few test samples and their predicted persona slots and values. We also pick some predicted examples to illustrate the shortcomings of our approach and give a possible direction for future research.

7.1. Quantitative Analysis

Persona discovery. Table 6 presents the confusion matrices for SPOT and the best performing baseline, DialogXL. SPOT correctly predicts 184 out of 305 positive instances (60.3%) while DialogXL is only able to predict 70 (22.9%). Although DialogXL performs poorly while identifying the true positives, it does a better job in identifying the true negatives. It is able to correctly classify 1594 instances as negative (94.9%), whereas SPOT predicts only 1191 true negative instances (70.9%).

Persona-type identification. We compare the confusion matrices obtained by SPOT and the best baseline for this subtask, AGHMN in Table 7. Both the models produce a comparable performance with maximum accurate predictions for the persona-type *trait*. Moreover, we observe that the classes

#	Speaker	Utterance	PD True	PD Pred
1	Chandler	We're in a relationship?	0	1
2	Danny	So you like the short hair better?	0	1
3	Rachel	Yeah. Oh! Was how you invented the cotton gin?!	0	1
4	Phoebe	Well, so, umm, anyway umm, I've been, I've been looking for my Father, and umm, have you heard from him, or seen him?	0	1
5	Janice	So, I hear, you hate me?	0	1

(a) Persona does not lie in questions.

#	Speaker	Utterance	PD True	PD Pred
1	Monica	Oh my god, I am losing my mind.	0	1
2	Phoebe	Because we're girls.	0	1
3	Monica	No, Phoebe, I'll tell you what, if you get ready now I'll let you play it at the wedding.	0	1
4	Leslie	My best shoes, so good to me.	0	1
5	Chandler	Okay uh, for now, temporarily, you can call me, Clint.	0	1

(b) Persona is not temporary/trivial attributes.

Table 8: Examples of false positives by SPOTfor the Persona Discovery (PD) task.

likes and *trait* are most confused, followed by *likes*, *relation*, *misc* and *occupation* for SPOT as well as for AGHMN, while *occupation* and *relation* are least confused among all classes.

7.2. Common Errors by SPOT

False positives. While attaining a decent value for true positives, SPOT obtains a significant value of false positives (487) for *persona discovery* (c.f. Table 6). We analyse the type of misclassified instances and observe that SPOT often identifies utterances containing questions as having persona information. For example, the utterance *'We're in a relationship?'* is marked as a positive instance for persona discovery when in true sense, its answer was the one carrying persona. Table 8 presents similar examples. This phenomenon can be attributed to the presence of words such as 'relationship' (instance 1), 'like' (instance 2), or 'father' (instance 4) in the utterances as these words may hint towards the presence of explicit persona information in a statement. In addition, SPOT frequently predicts the utterances expressing a temporary/trivial state for the speaker as containing persona information. For instance, the utterance *'Oh my god, I am losing my mind.'* is marked as the one containing persona information. We show more such instances in Table 8. Future work could be done to handle such cases of false positives.

False negatives. In addition to falsely identifying utterances containing no persona information as positive instances, SPOT identifies 121 true positive instances as belonging to the negative class. We analyse the misclassified positive instances and

#	Speaker	Utterance	*Persona Discovery*			*Persona Type Identification*		
			True	Predicted		True	Predicted	
				SPOT	DialogXL		SPOT	AGHMN
u_1	Rachel	Everybody, this is Paolo, Paolo, I want you to meet my friends. This is Monica	Yes	Yes	No	relationship	relationship	likes
u_2	Monica	Hi!	No	No	No	-	-	
u_3	Rachel	And Joey...	Yes	No	Yes	relationship	relationship	relationship
u_4	Monica	Hi!	No	No	No	-	-	
u_5	Rachel	And Ross...	Yes	Yes	Yes	relationship	trait	likes
u_6	Monica	Hi!	No	No	Yes	-	-	

Table 9: Actual and predicted labels for the *persona discovery* and *persona type identification* tasks. DialogXL and AGHM are the best performing baseline for the respective tasks.

#	#	Speaker	Utterance	PD	
				True	Pred
1	u_1	Ross	Okay! All right! Now, Chandler you-you wanna live with Monica, right?	0	1
	u_2	Chandler	Yeah, I do.	1	0
2	u_1	Judge	So based on your petition you are seeking an annulment on the grounds that Mr. Geller is mentally unstable?	1	1
	u_2	Ross	Fine, I'm mentally unstable.	1	0
3	u_1	Ross	Are you intrigued?	0	0
	u_2	Chandler	You're flingin'-flangin' right I am!	1	0
4	u_1	Rachel	Why, does she have a bad personality?	1	1
	u_2	Phoebe	Oh no, Bonnie's the best!	1	0
5	u_1	Chandler	Soo, ah, Eric, what kind of photography do ya do?	0	0
	u_2	Eric	Oh, mostly fashion, so there may be models here from time to time, I hope that's cool.	1	0

(a) Persona lies in answer to a question.

#	Speaker	Utterance	PD	
			True	Pred
1	Joey	Ya see, it's just, see I was a regular on a soap opera y'know?	1	0
2	Joey	Awww, one of my students got an audition. I'm so proud.	1	0
3	Joey	Yeah but we won't be able to like get up in the middle of the night and have those long talks about our feelings and the future.	1	0
4	Janice	Oh, Chandler, look. You and Monica are meant to have children. I am sure it's gonna be just fine.	1	0
5	Steve	Umm, see, I was thinking maybe you two could switch apartments because Phoebe's more our kind of people.	1	0

(b) Persona is implicit.

Table 10: Examples of false negatives by SPOTfor the Persona Discovery (PD) task.

identify two situations where such misclassifications happen. When the persona information is present in the answer to a question, it is often misclassified by SPOT. For example, in the dialogue *'Ross: Okay! All right! Now, Chandler you-you wanna live with Monica, right? Chandler: Yeah, I do.'*, Chandler's utterance contains information about his persona (*relationship* with Monica), but SPOT is unable to identify this instance correctly. Table 10 highlights similar examples from SPICE. Furthermore, SPOT often misclassifies instances where the persona information is implicit in nature. For instance, the utterance *'Ya see, it's just, see I was a regular on a soap opera y'know?'* contains persona information (that the speaker's *occupation* is an actor); however, SPOT is not able to relate the phrase 'soap opera' to *occupation* and thus does not mark the instance as having persona information. Supporting examples are shown in Table 10.

7.3. Qualitative Analysis

This section presents a subjective analysis of the quality of predictions made by SPOT and the best baselines, based on a sample dialogue from the test set. The dialogue contains six utterances, where utterances u_1, u_3, and u_5 are identified as having the persona type *relationship*, as shown in Table 9. In the first subtask of persona discovery, SPOT correctly identifies two positive instances out of the total three, while the best baseline, DialogXL, only identifies one such instance. However, both SPOT and DialogXL misclassify one utterance as false negative.

Moving on to the second subtask of persona-type identification, SPOT correctly classifies two instances of persona-type *relationship* but misclassifies one instance as *trait*. On the other hand, the best baseline, AGHMN, only predicts one correct class and misclassifies the others as *likes*.

8. Conclusion

In this study, we delved into the intricate task of speaker profiling within conversations, with the goal of unearthing persona information linked to each participant. This undertaking was subdivided into three distinct subtasks: persona discovery, persona-type identification, and persona-value extraction. To facilitate our research, we introduced a novel dataset named SPICE, meticulously designed to serve as a benchmark for this purpose. Moreover, we introduced SPOT, a sophisticated neural approach that harnessed RoBERTa, Transformer, and specialised attention modules to adeptly capture both the conversational context and speaker-specific semantics. Concurrently, we adapted several cutting-edge models for comparative evaluation. Our rigorous experimentation unveiled the supremacy of SPOTover these alternatives. In addition, we conducted a comprehensive

ablation study, justifying the utility of each component within SPOT. Strengthening our findings, we presented error analyses, including confusion matrices and qualitative predictions. Furthermore, we candidly acknowledged the limitations of SPOT, exemplified through test set samples from SPICE.

Acknowledgements

The authors acknowledge the support of the ihub-Anubhuti-iiitd Foundation, set up under the NM-ICPS scheme of the DST.

Bibliographical References

Reina Akama, Kazuaki Inada, Naoya Inoue, Sosuke Kobayashi, and Kentaro Inui. 2017. Generating stylistically consistent dialog responses with transfer learning. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 408–412.

Pinkesh Badjatiya, Shashank Gupta, Manish Gupta, and Vasudeva Varma. 2017. Deep learning for hate speech detection in tweets. In *Proceedings of the 26th international conference on World Wide Web companion*, pages 759–760.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*.

Nitesh V Chawla, Kevin W Bowyer, Lawrence O Hall, and W Philip Kegelmeyer. 2002. Smote: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, 16:321–357.

Guanyi Chen, Yinhe Zheng, and Yupei Du. 2020. Listener's social identity matters in personalised response generation. *arXiv preprint arXiv:2010.14342*.

Eric Chu, Prashanth Vijayaraghavan, and Deb Roy. 2018. Learning personas from dialogue with attentive memory networks. *arXiv preprint arXiv:1810.08717*.

Cristian Danescu-Niculescu-Mizil and Lillian Lee. 2011. Chameleons in imagined conversations: A new approach to understanding coordination of linguistic style in dialogs. *arXiv preprint arXiv:1106.3077*.

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*.

Emily Dinan, Stephen Roller, Kurt Shuster, Angela Fan, Michael Auli, and Jason Weston. 2018. Wizard of wikipedia: Knowledge-powered conversational agents. *arXiv preprint arXiv:1811.01241*.

Rashmi Gangadharaiah and Balakrishnan Narayanaswamy. 2019. Joint multiple intent detection and slot labeling for goal-oriented dialog. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 564–569.

Jia-Chen Gu, Zhen-Hua Ling, Yu Wu, Quan Liu, Zhigang Chen, and Xiaodan Zhu. 2021. Detecting speaker personas from conversational texts. *arXiv preprint arXiv:2109.01330*.

Wenxiang Jiao, Michael Lyu, and Irwin King. 2020. Real-time emotion recognition via attention gated hierarchical memory network. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 8002–8009.

Chaitanya K Joshi, Fei Mi, and Boi Faltings. 2017. Personalization in goal-oriented dialog. *arXiv preprint arXiv:1706.07503*.

Klaus Krippendorff. 2011. Computing krippendorff's alpha-reliability.

Stefan Larson, Anish Mahendran, Joseph J Peper, Christopher Clarke, Andrew Lee, Parker Hill, Jonathan K Kummerfeld, Kevin Leach, Michael A Laurenzano, Lingjia Tang, et al. 2019. An evaluation dataset for intent classification and out-of-scope prediction. *arXiv preprint arXiv:1909.02027*.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.

Jiwei Li, Michel Galley, Chris Brockett, Georgios P Spithourakis, Jianfeng Gao, and Bill Dolan. 2016. A persona-based neural conversation model. *arXiv preprint arXiv:1603.06155*.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Zhaojiang Lin, Peng Xu, Genta Indra Winata, Farhad Bin Siddique, Zihan Liu, Jamin Shin, and Pascale Fung. 2020. Caire: An end-to-end empathetic chatbot. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 13622–13623.

Yang Liu, Kun Han, Zhao Tan, and Yun Lei. 2017. Using context information for dialog act classification in dnn framework. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2170–2178.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Ro{bert}a: A robustly optimized {bert} pretraining approach.

Stuart Lloyd. 1982. Least squares quantization in pcm. *IEEE transactions on information theory*, 28(2):129–137.

JM Lucas, F Fernández, J Salazar, J Ferreiros, and R San Segundo. 2009. Managing speaker identity and user profiles in a spoken dialogue system. *Procesamiento del lenguaje natural*, (43):77–84.

James MacQueen et al. 1967. Some methods for classification and analysis of multivariate observations. In *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, volume 1, pages 281–297. Oakland, CA, USA.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Soujanya Poria, Devamanyu Hazarika, Navonil Majumder, Gautam Naik, Erik Cambria, and Rada Mihalcea. 2018. Meld: A multimodal multi-party dataset for emotion recognition in conversations. *arXiv preprint arXiv:1810.02508*.

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

Yada Pruksachatkun, Jason Phang, Haokun Liu, Phu Mon Htut, Xiaoyi Zhang, Richard Yuanzhe Pang, Clara Vania, Katharina Kann, and Samuel R Bowman. 2020. Intermediate-task transfer learning with pretrained models for natural language understanding: When and why does it work? *arXiv preprint arXiv:2005.00628*.

Libo Qin, Zhouyang Li, Wanxiang Che, Minheng Ni, and Ting Liu. 2020. Co-gat: A co-interactive graph attention network for joint dialog act recognition and sentiment classification. *arXiv preprint arXiv:2012.13260*.

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

Hannah Rashkin, Eric Michael Smith, Margaret Li, and Y-Lan Boureau. 2018. Towards empathetic open-domain conversation models: A new benchmark and dataset. *arXiv preprint arXiv:1811.00207*.

Stephen Roller, Emily Dinan, Naman Goyal, Da Ju, Mary Williamson, Yinhan Liu, Jing Xu, Myle Ott, Kurt Shuster, Eric M Smith, et al. 2020. Recipes for building an open-domain chatbot. *arXiv preprint arXiv:2004.13637*.

Roger C Schank. 1972. Conceptual dependency: A theory of natural language understanding. *Cognitive psychology*, 3(4):552–631.

Abigail See, Peter J. Liu, and Christopher D. Manning. 2017. Get to the point: Summarization with pointer-generator networks. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1073–1083, Vancouver, Canada. Association for Computational Linguistics.

Weizhou Shen, Junqing Chen, Xiaojun Quan, and Zhixian Xie. 2020. Dialogxl: All-in-one xlnet for multi-party conversation emotion recognition. *arXiv preprint arXiv:2012.08695*.

Jamin Shin, Peng Xu, Andrea Madotto, and Pascale Fung. 2019. Happybot: Generating empathetic dialogue responses by improving user experience look-ahead. *arXiv preprint arXiv:1906.08487*.

Yixuan Su, Deng Cai, Yan Wang, Simon Baker, Anna Korhonen, Nigel Collier, and Xiaojiang Liu. 2020. Stylistic dialogue generation via information-guided reinforcement learning strategy. *arXiv preprint arXiv:2004.02202*.

Anna Tigunova, Andrew Yates, Paramita Mirza, and Gerhard Weikum. 2019. Listening between the lines: Learning personal attributes from conversations. In *The World Wide Web Conference*, pages 1818–1828.

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.

Jason Weston, Emily Dinan, and Alexander H Miller. 2018. Retrieve and refine: Improved sequence generation models for dialogue. *arXiv preprint arXiv:1808.04776*.

Chien-Sheng Wu, Andrea Madotto, Zhaojiang Lin, Peng Xu, and Pascale Fung. 2019. Getting to know you: User attribute extraction from dialogues. *arXiv preprint arXiv:1908.04621*.

Hanlei Zhang, Hua Xu, and Ting-En Lin. 2021. Deep open intent classification with adaptive decision boundary. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 14374–14382.

Saizheng Zhang, Emily Dinan, Jack Urbanek, Arthur Szlam, Douwe Kiela, and Jason Weston. 2018. Personalizing dialogue agents: I have a dog, do you have pets too? *arXiv preprint arXiv:1801.07243*.

Xujuan Zhou, Xiaohui Tao, Jianming Yong, and Zhenyu Yang. 2013. Sentiment analysis on tweets for social events. In *Proceedings of the 2013 IEEE 17th international conference on computer supported cooperative work in design (CSCWD)*, pages 557–562. IEEE.