Successful New-entry Prediction for Multi-Party Online Conversations via Latent Topics and Discourse Modeling

Lingzhi Wang
lzwang@se.cuhk.edu.hk
The Chinese University of Hong Kong
Hong Kong, China

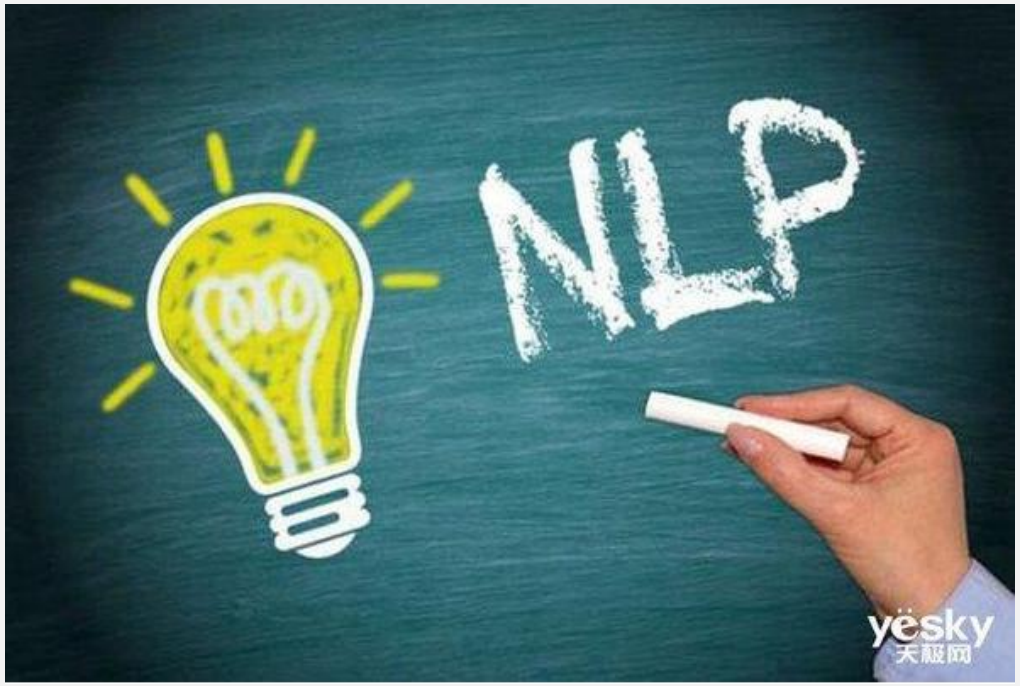
Jing Li
jing-amelia.li@polyu.edu.hk
The Hong Kong Polytechnic University
Hong Kong, China

Xingshan Zeng
zxshamson@gmail.com
Huawei Noah's Ark Lab
Hong Kong, China

Kam-Fai Wong
kfwong@se.cuhk.edu.hk
The Chinese University of Hong Kong
Hong Kong, China

(WWW-2022)

Reported by Jia Wang



1. Introduction
2. Approach
3. Experiments

Introduction

History Conversation 1 for U_5:
U: ***Pizza***. Make it at home. …
U_5 : Try ***deep dish***! Layer it with maple smoked ***ham***, …
…
History Conversation 2 for U_5 :
U: … The U.S. and China have had a hard-on for this ***tech*** …
U_5 : … considering no international ***laws*** … destructive ***weapons*** …
…
History Conversation 3 for U_5 :
U: OP didn't see [Devil.](<URL>)
U_5 : This was such a good ***movie***, though five years later …
…

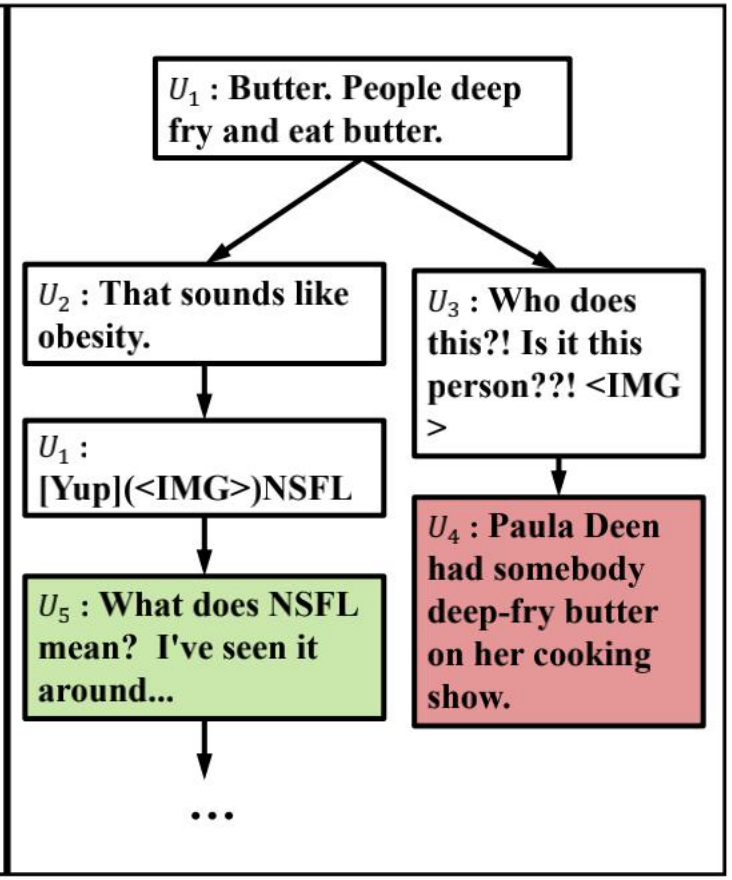


Figure 1: A Reddit conversation on the left part (U_i: the i-th user). U_3 made a successful engagement (i.e. receiving a reply, we omit here to save space). The right column shows U_3's chatting history, where the topic words are in *bold and italic*.

Introduction

The main contributions of this work can be summarized as follows:

• We first formulate the task of successful new-entry prediction and contribute two large-scale datasets, Twitter and Reddit. The SNP task can benefit the development of online assistants and early socialization strategies.

• We propose a novel framework combining unsupervised and supervised neural networks. VAE and RNN-based modules are incorporated for the personalized user engagement prediction via learning latent topic and discourse representation.

• Experimental results on both Twitter and Reddit show that the proposed model significantly outperforms the baselines. For example, we achieve 34.6 F1 on Reddit compared with 32.5 achieved by a BERT-based method.

Approach

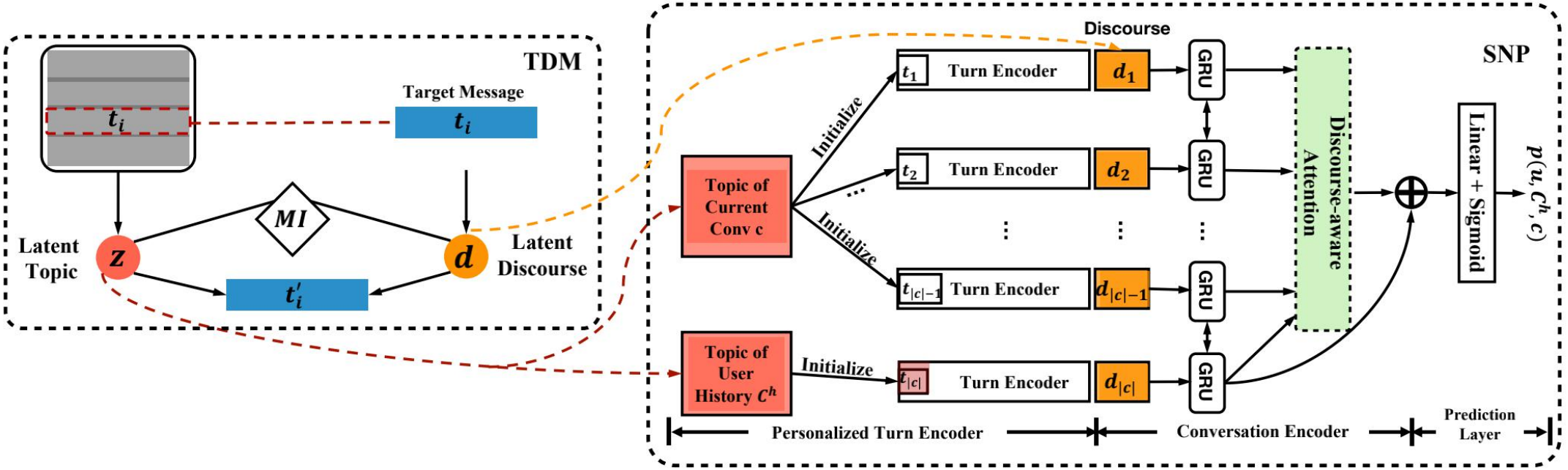
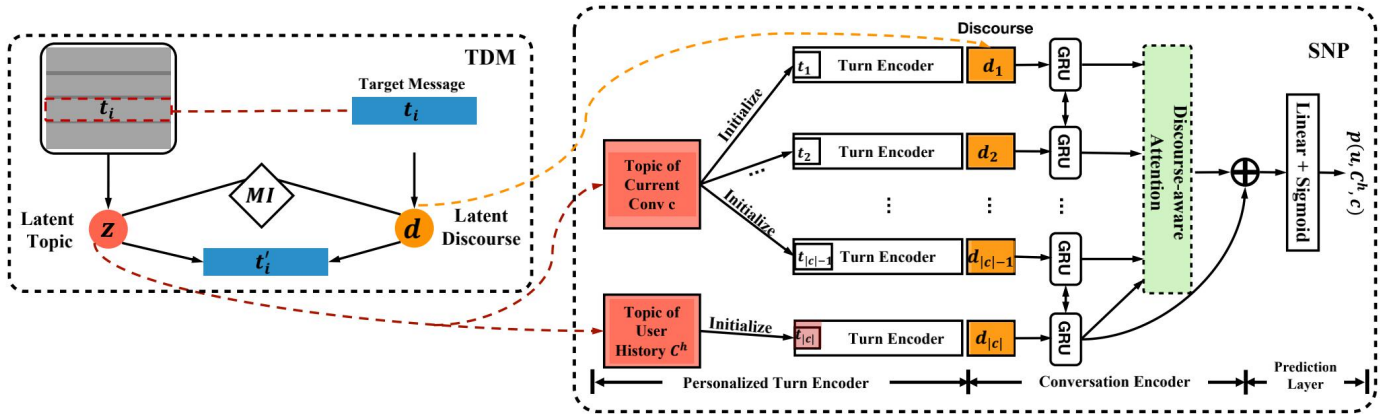


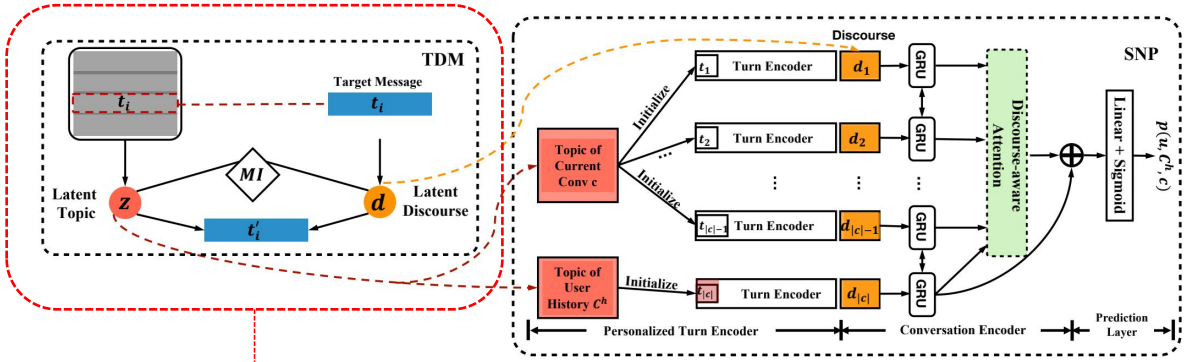
Figure 2: Our generic framework for successful new-entry prediction. It contains two modules: Topic and Discourse Modeling (TDM) and Successful New-entry Prediction (SNP). SNP consists of three parts: turn encoder, conversation encoder and prediction layer.

Approach



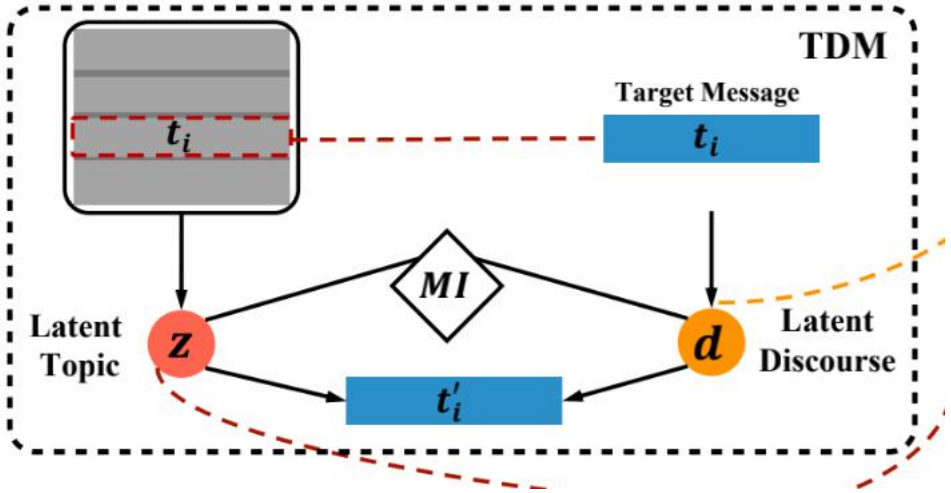
The input for our model can be divided into two parts: the observed conversation c and the history conversation set $C^h = \{c_1^h, c_2^h, ..., c_k^h\}$ of the newcomer u, where k is the number of history conversations obtained from training set. The conversation c is formalized as a sequence of turns (e.g., posts or tweets) $\{t_1, t_2, ..., t_{|c|}\}$, and the $|c|^{th}$ turn is posted by the newcomer u (we predict whether u can get others' response afterwards). The conversations in user's history conversation set C^h are organized similarly into the sequences of turns. For output, we yield a Bernoulli distribution $p(u, C^h, c)$ to indicate the estimated likelihood of whether u gets responses from other participants (successful new-entries).

Topic and Discourse Modeling (TDM)



conversation c, $\begin{cases} \boldsymbol{e}_c \\ \boldsymbol{d} = \langle \boldsymbol{d}_1, ..., \boldsymbol{d}_{|c|} \rangle. \end{cases}$

C^h of the newcomer $u \longrightarrow \boldsymbol{e}_u$.

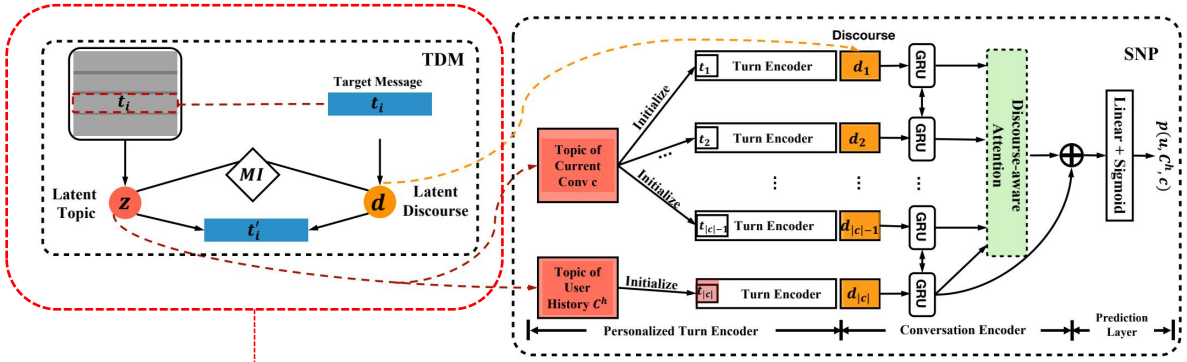


Encode step

$$\mu = f_\mu(f_e(\boldsymbol{c}_{bow})),\ log\,\sigma = f_\sigma(f_e(\boldsymbol{c}_{bow}))$$
$$\pi = \text{softmax}(f_\pi(\boldsymbol{t}_{i_{bow}})) \qquad (1)$$

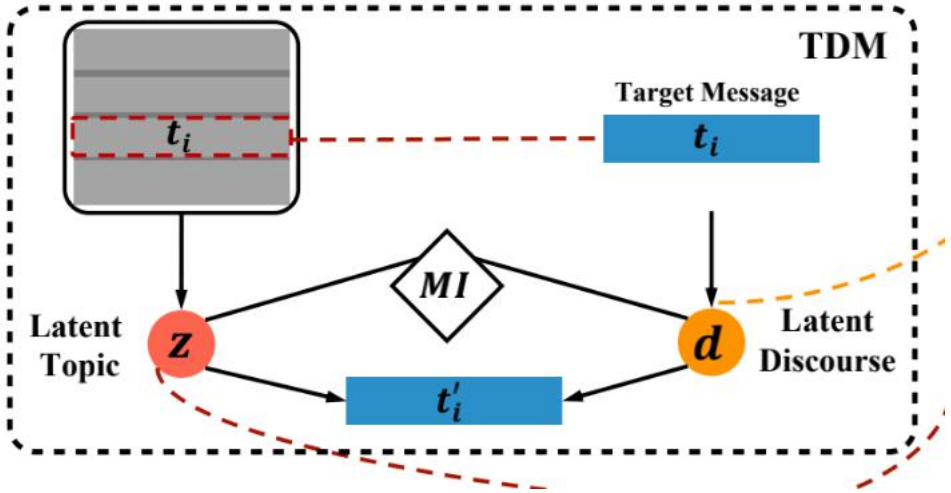
where $f_*(\cdot)$ is neural perceptrons performing linear transformations activated with an ReLU function [33].

Topic and Discourse Modeling (TDM)



conversation c, $\begin{cases} \boldsymbol{e}_c \\ \boldsymbol{d} = \langle \boldsymbol{d}_1, ..., \boldsymbol{d}_{|c|} \rangle. \end{cases}$

C^h of the newcomer $u \longrightarrow \boldsymbol{e}_u$.

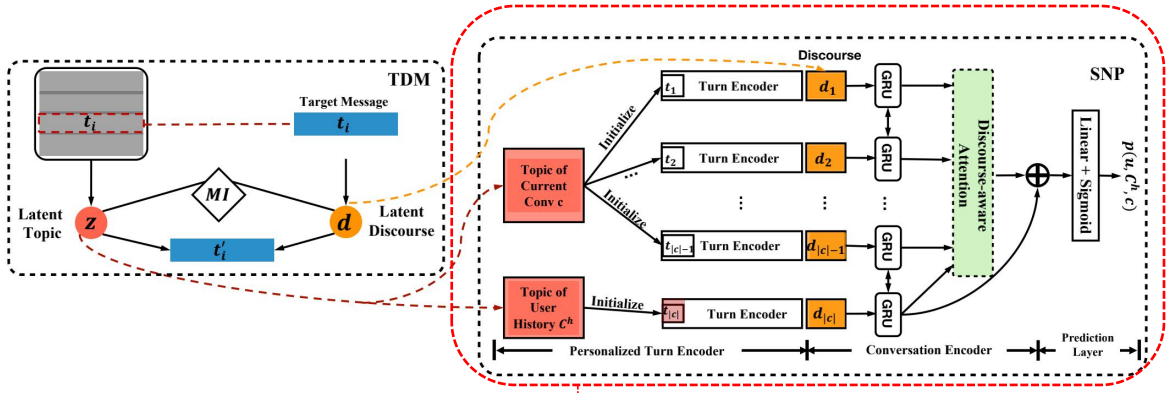


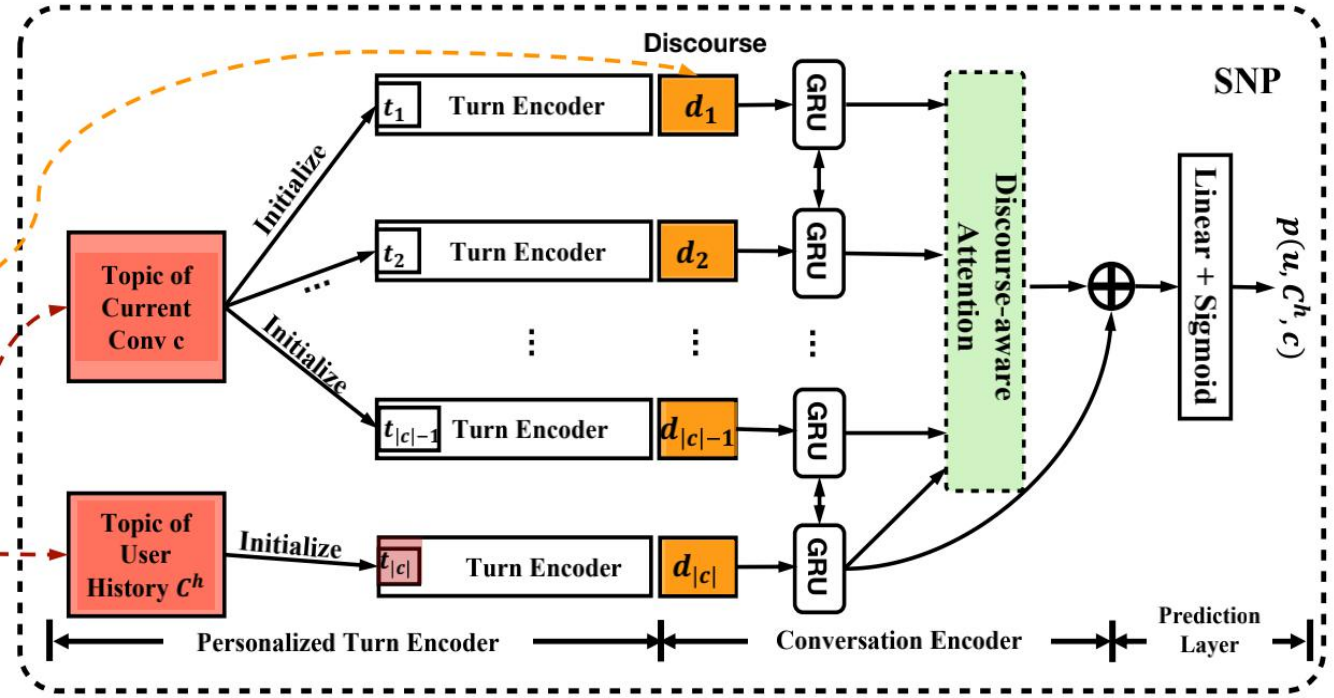
Decode step

- Draw latent topic $\mathbf{z} \sim \mathcal{N}(\mu, \sigma^2)$.
- Topic mixture $\boldsymbol{\theta} = \text{softmax}(f_\theta(\mathbf{z}))$.
- Draw the latent discourse $\boldsymbol{d} \sim \text{Multi}(\pi)$.
- For the n-th word in the conversation:
  - $\beta_n = \text{softmax}(f_{\phi^T}(\boldsymbol{\theta}) + f_{\phi^D}(\boldsymbol{d}))$
  - Draw the word $w_n \sim \text{Multi}(\beta_n)$.

In particular, the weight matrix of $f_{\phi^T}(\cdot)$ (after the softmax normalization) is considered as the topic-word distribution ϕ^T. We can also get the discourse-word distribution ϕ^D in a similar way.

Successful New-entry Prediction (SNP)





Personalized Turn Encoder

For each turn t_i in conversation c,

$$w_j$$

$$\langle r_{i1}, r_{i2}, ..., r_{i,n_i} \rangle$$

We divide the observed conversation turns into context turns (turns before the last turn) and query turn (last turn, posted by newcomer u). For query turn, we use u's topic representation $\boldsymbol{e}_u$ (produced by TDM module in Section 3.2) to initialize the aforementioned Bi-GRU.

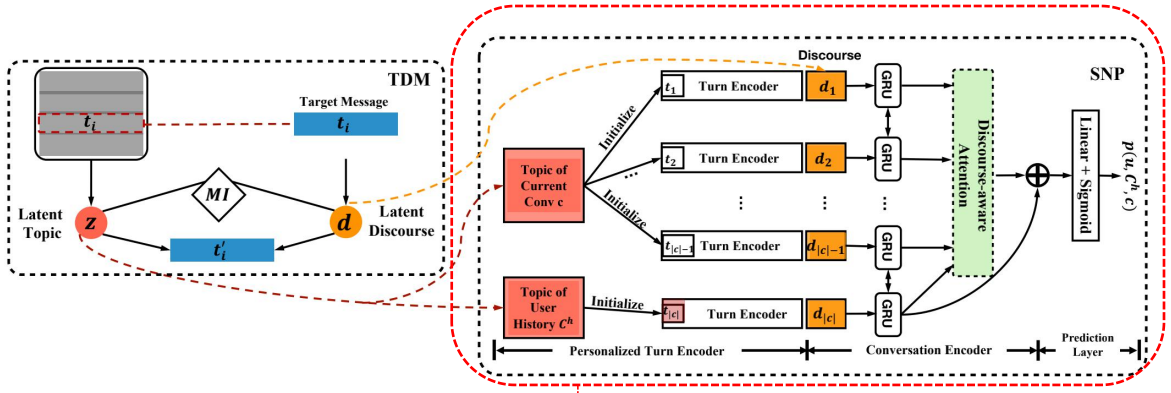
$$\overrightarrow{\boldsymbol{h}_{i,0}} = \overleftarrow{\boldsymbol{h}_{i,n_i}} = \boldsymbol{W}^P \boldsymbol{e} + b^P \qquad \boldsymbol{e} \text{ is } \boldsymbol{e}_u \text{ or } \boldsymbol{e}_c$$

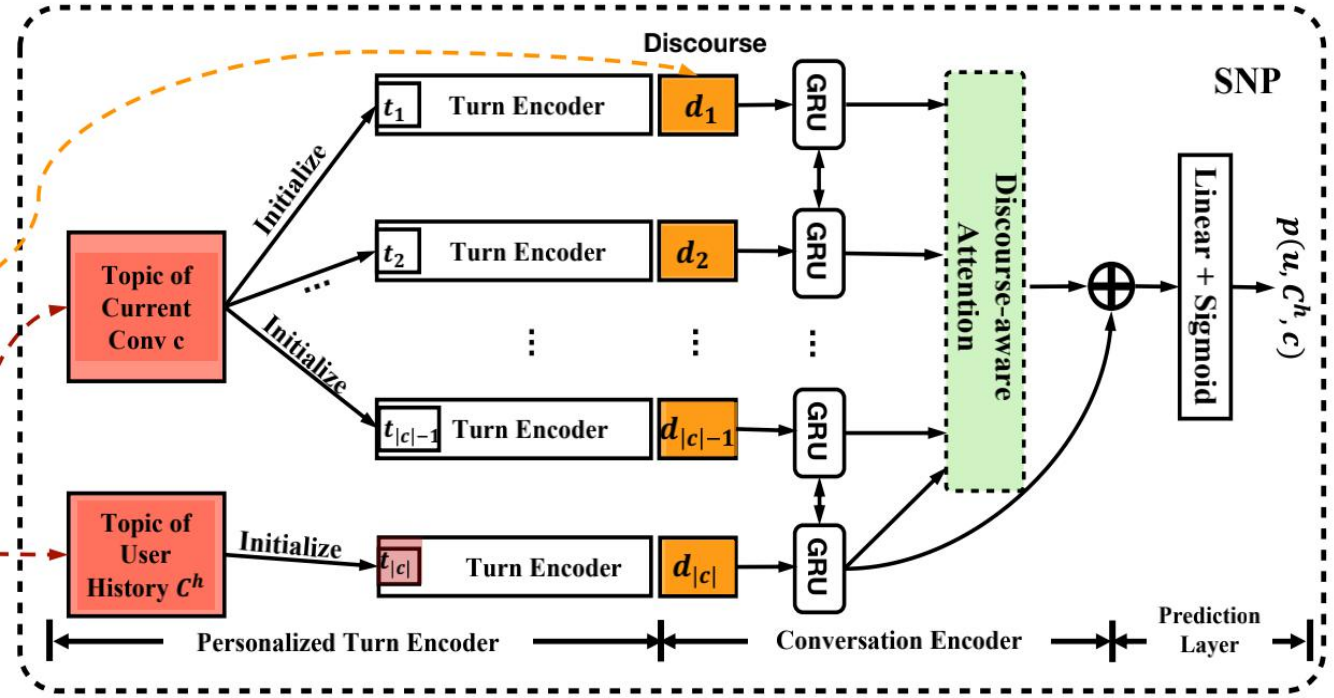
$$\overrightarrow{\boldsymbol{h}_{ij}} = f_{GRU}(\boldsymbol{r}_{ij}, \boldsymbol{h}_{i,j-1}), \ \overleftarrow{\boldsymbol{h}_{ij}} = f_{GRU}(\boldsymbol{r}_{ij}, \boldsymbol{h}_{i,j+1}) \qquad (2)$$

$$\boldsymbol{h}_i = [\overrightarrow{\boldsymbol{h}_{i,n_i}}; \overleftarrow{\boldsymbol{h}_{i,0}}]$$

$$\langle \boldsymbol{h}_1, \boldsymbol{h}_2, .., \boldsymbol{h}_{|c|} \rangle.$$

Successful New-entry Prediction (SNP)





Discourse-aware Conversation Encoder

$$\overrightarrow{\boldsymbol{h}_j^d} = f_{GRU}(\boldsymbol{s}_j, \boldsymbol{h}_{j-1}^d),\ \overleftarrow{\boldsymbol{h}_j^d} = f_{GRU}(\boldsymbol{s}_j, \boldsymbol{h}_{j+1}) \tag{3}$$

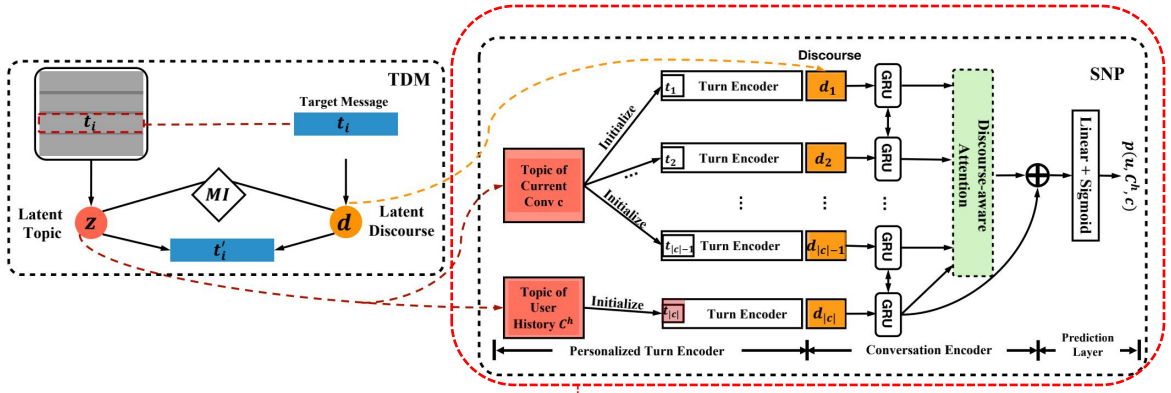
where $\boldsymbol{s}_j = [\boldsymbol{h}_j; \boldsymbol{d}_j]$ and the representation of each turn after GRU is $\boldsymbol{h}_j^d = [\overrightarrow{\boldsymbol{h}_j^d}; \overleftarrow{\boldsymbol{h}_j^d}]$.

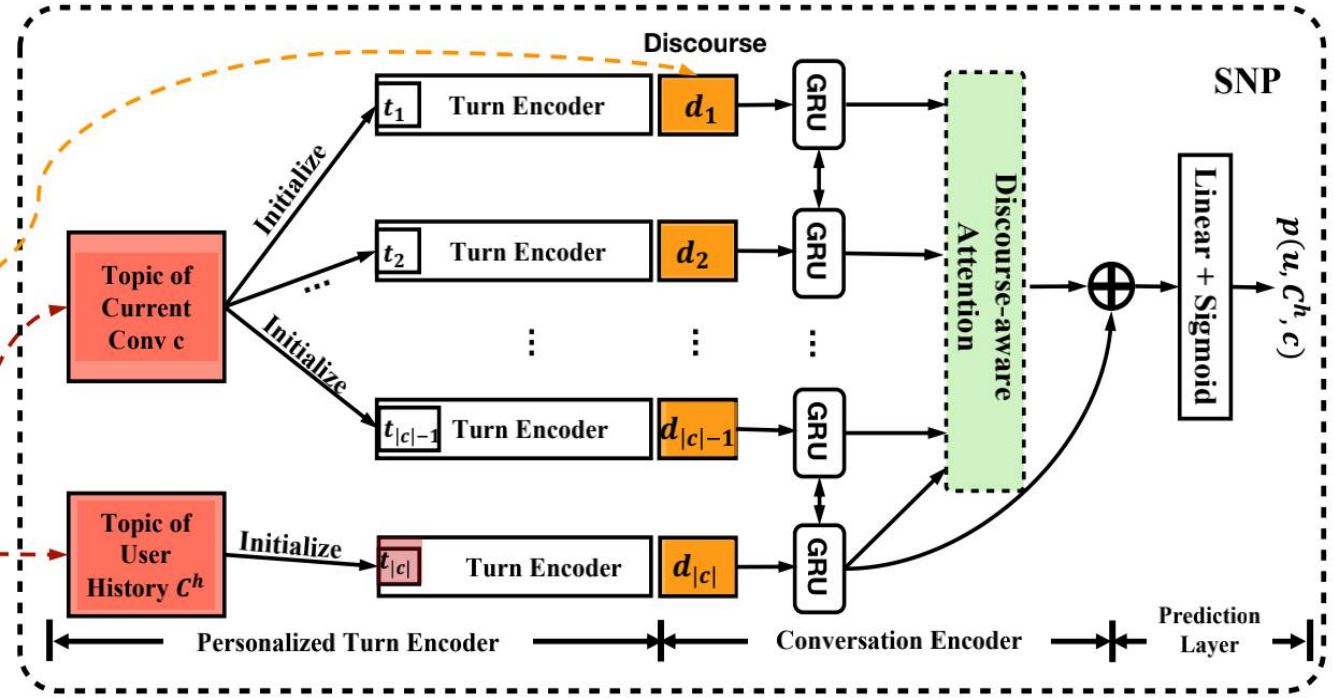
$$a_j = f_d(\text{argmax}(\boldsymbol{d}_j)) \tag{4}$$

where $\text{argmax}(\boldsymbol{d}_j)$ means the learned discourse behavior to turn j, and $f_d(\cdot)$ maps the discourse behaviors to different weight values.

$$\boldsymbol{h}^c = [\boldsymbol{h}_{|c|}^d; \sum_j \text{softmax}(a_j)\boldsymbol{h}_j^d] \tag{5}$$

Successful New-entry Prediction (SNP)



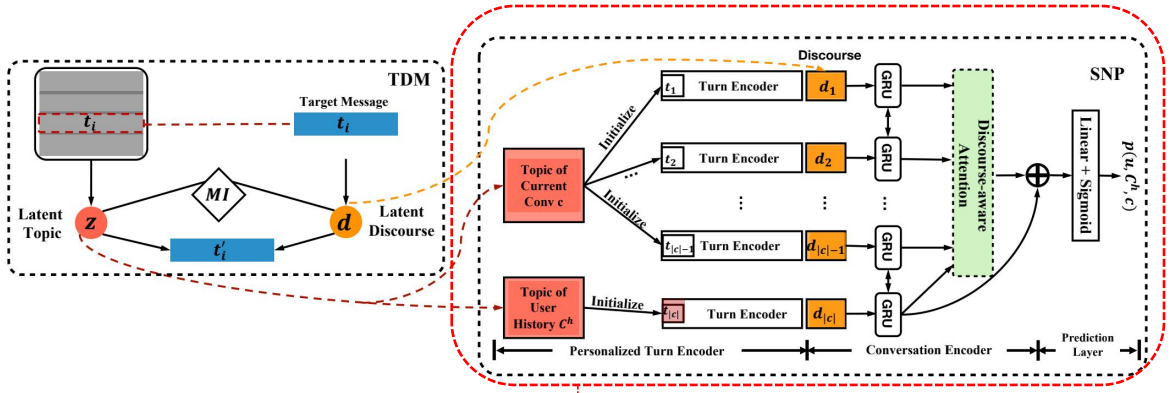


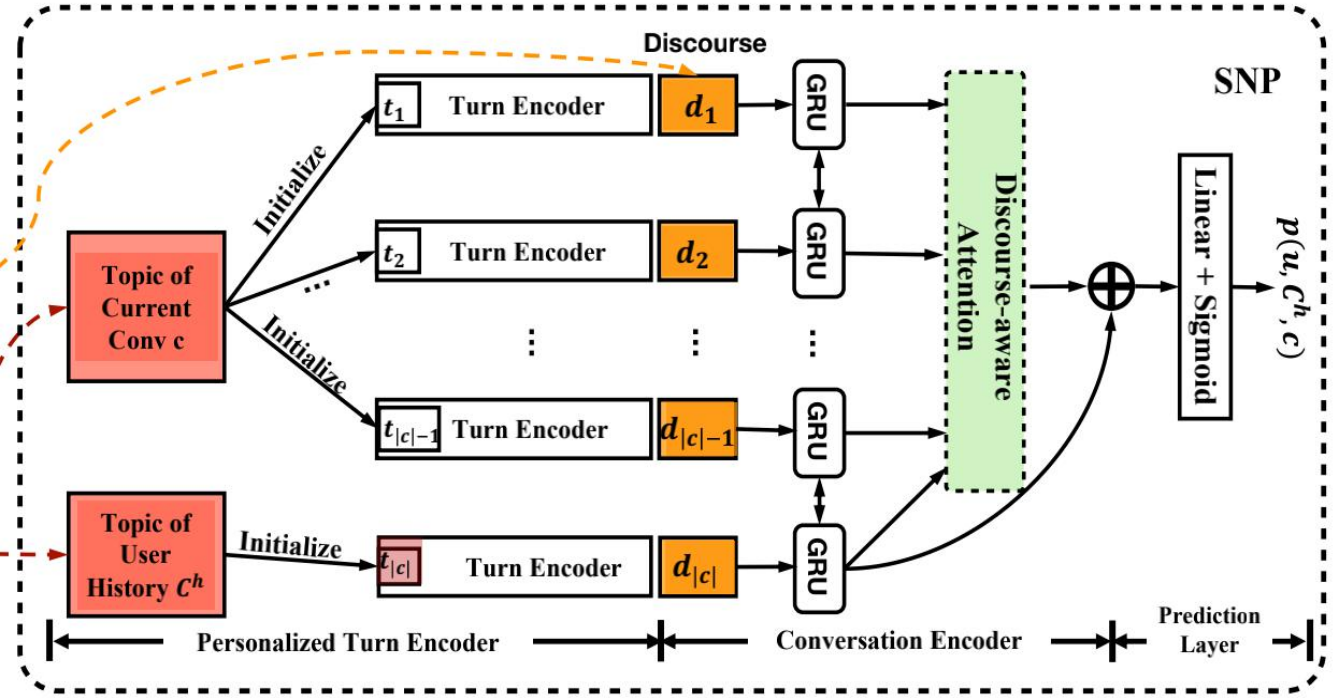
Prediction Layer

$$p(u, C^h, c) = \sigma(\boldsymbol{w}^T \boldsymbol{h}^c + b) \quad (6)$$

where $\boldsymbol{w}^T$ and b are trainable, and $\sigma(\cdot)$ is the sigmoid activation function.

Learning Objective





$$\mathcal{L} = \mathcal{L}_{TDM} + \mathcal{L}_{SNP} \tag{7}$$

$$\mathcal{L}_{TDM} = \mathcal{L}_z + \mathcal{L}_d + \mathcal{L}_t - \lambda \mathcal{L}_{MI} \tag{8}$$

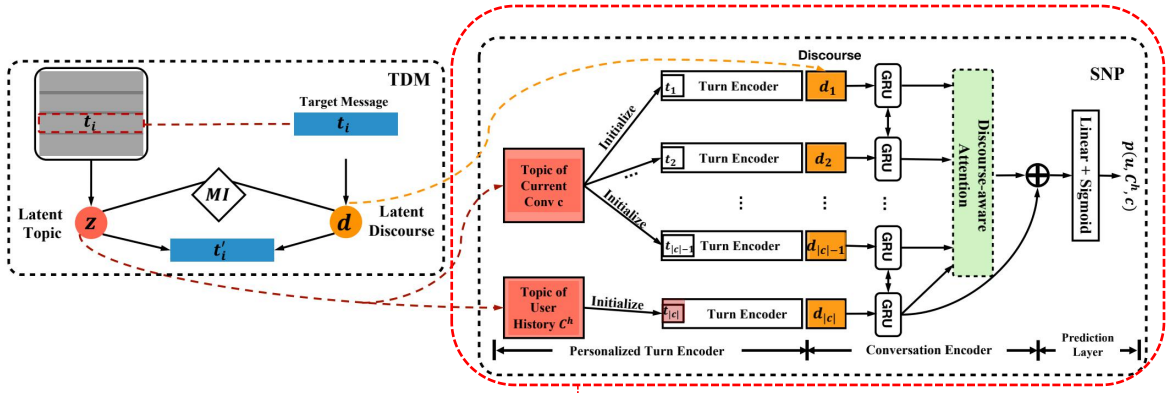
where $\mathcal{L}_z$ and $\mathcal{L}_d$ are objectives about learning topics and discourse, $\mathcal{L}_t$ is the loss for target message reconstruction, and $\mathcal{L}_{MI}$ ensures that topics and discourse learn differently.

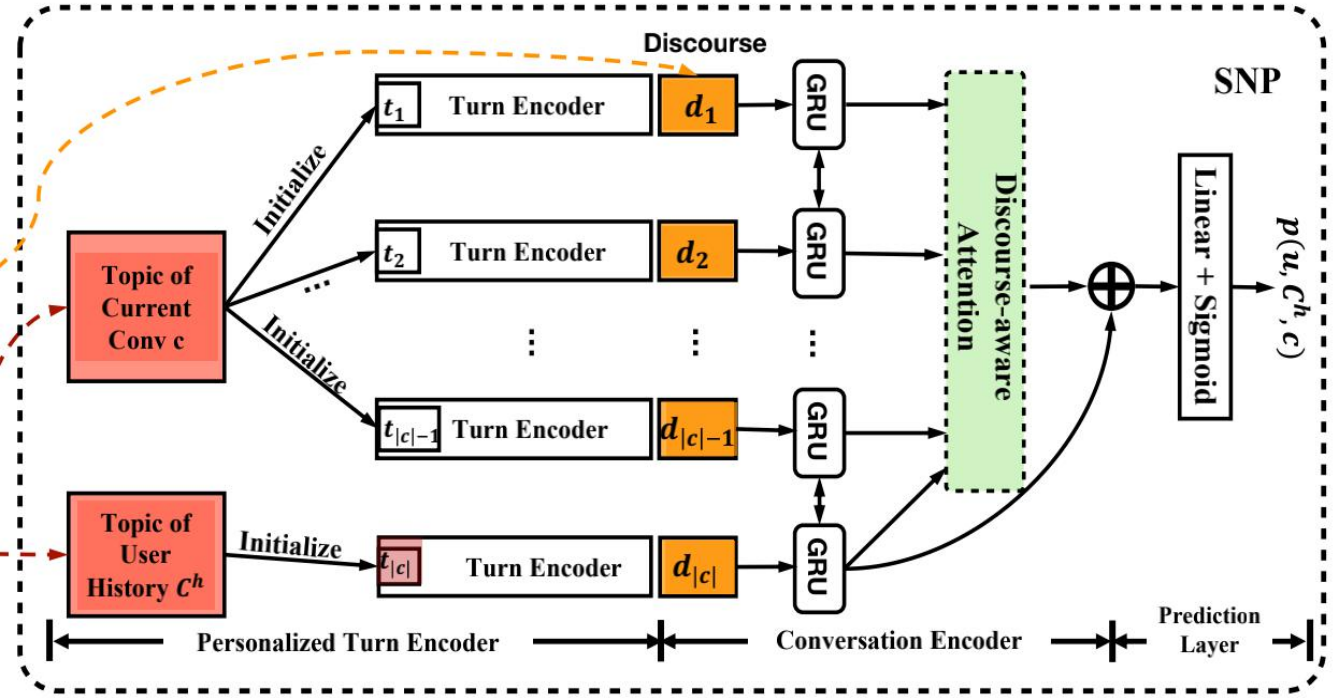
$$\mathcal{L}_z = \mathbb{E}_{q(z|c)}\left[p(c \mid z)\right] - D_{KL}(q(z \mid c)||p(z)) \tag{9}$$

$$\mathcal{L}_d = \mathbb{E}_{q(\boldsymbol{d}|t)}\left[p(t \mid \boldsymbol{d})\right] - D_{KL}(q(\boldsymbol{d} \mid t)||p(\mathbf{d})) \tag{10}$$

where $q(z|c)$ and $q(\boldsymbol{d}|t)$ are approximated posterior probabilities describing how the latent topic z and the latent discourse $\boldsymbol{d}$ are generated from the conversations and message turns. $p(c \mid z)$ and $p(t \mid \boldsymbol{d})$ represent the corpus likelihoods conditioned on the latent variables. $p(z)$ follows the standard normal prior $\mathcal{N}(\mathbf{0}, \mathbf{I})$ and $p(\boldsymbol{d})$ is the uniform distribution $Unif(0, 1)$. D_{KL} refers to the Kullback-Leibler divergence that ensures the approximated posteriors to be close to the true ones.

Learning Objective





$$\mathcal{L}_t = \mathbb{E}_{q(z|c)q(\boldsymbol{d}|t)}\left[\log p(t \mid z, \boldsymbol{d})\right] \tag{11}$$

$$\mathcal{L}_{MI} = \mathbb{E}_{q(z)} D_{KL}(p(\boldsymbol{d} \mid z) || p(\boldsymbol{d})) \tag{12}$$

$$\mathcal{L}_{SNP} = -\sum_i \mu\, y_i \log(\hat{y}_i) + (1 - y_i)\log(1 - \hat{y}_i) \tag{13}$$

where $\hat{y}_i$ denotes the probability estimated from $p(u, C^h, c)$ for the i-th instance, and y_i is the corresponding binary ground truth label (1 for successful entries and 0 for the opposite). To take the potential data imbalance into account, we also adopt a trade-off weight μ to give more weight to the minority class. μ is set based on the proportion of positive and negative instances in the training set.

Experiments

Table 1: Statistics of Twitter and Reddit datasets.

	Twitter	Reddit
# of users	53,488	96,001
# of convs	37,339	69,428
# of conv turns	179,265	236,764
# of successful entries	29,340	12,199
# of failed entries	7,999	57,229
Avg turn number per conv	4.8	3.4
Avg token number per turn	20.5	20.7
Ratio of newcomer with history	0.59	0.62
Avg # of history for newcomers	2.5	6.3

Experiments

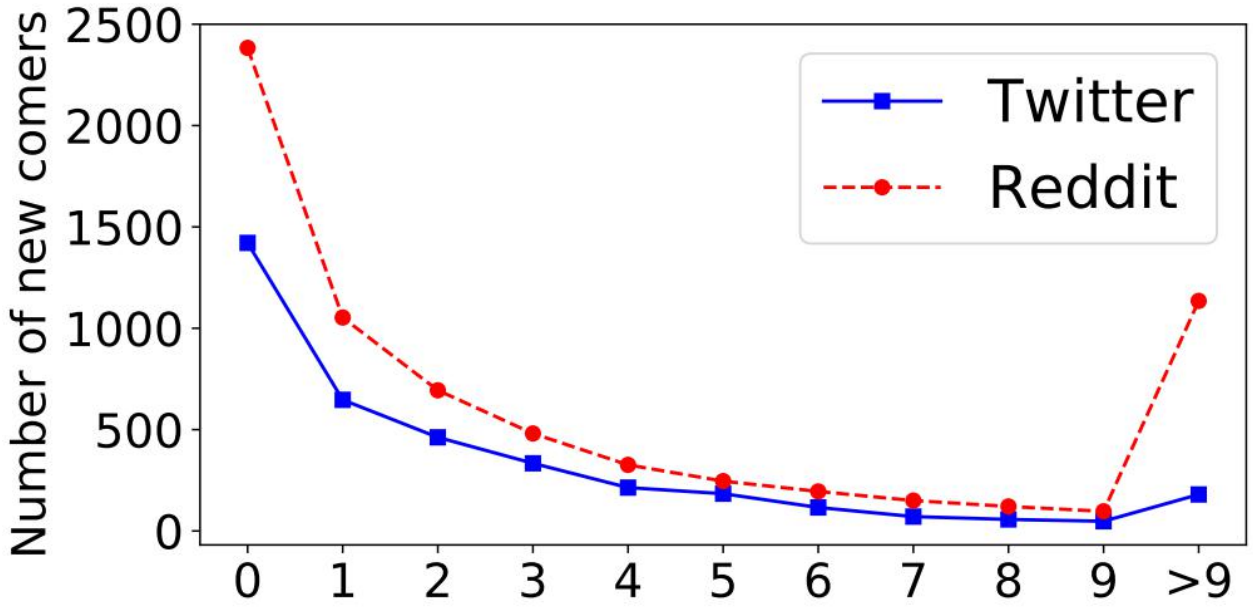


Figure 3: The distribution over the number of history conversations (X-axis). Y-axis: number of newcomers.

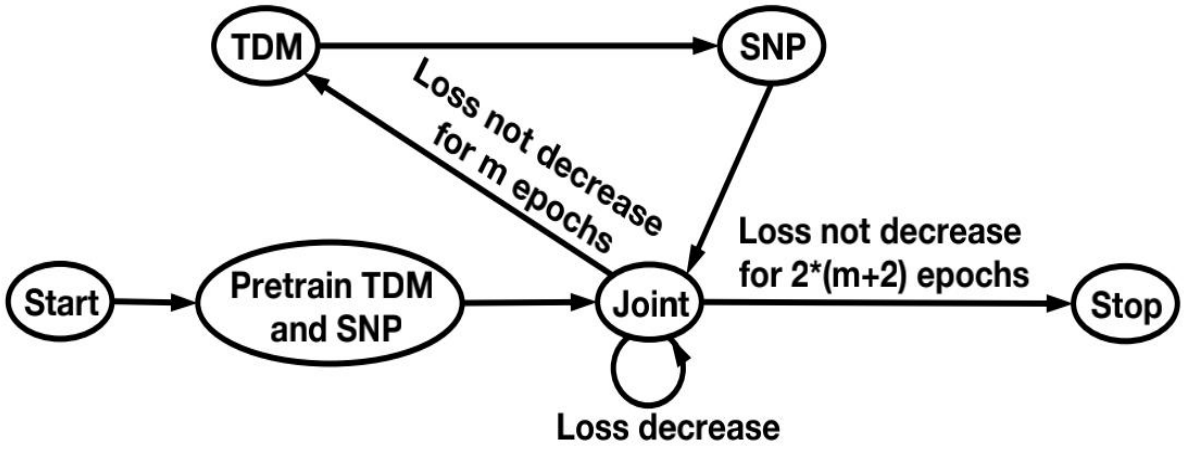


Figure 4: State transition diagram of training process.

Experiments

Table 2: Comparison results on Twitter and Reddit datasets (in %). Higher scores indicate better performance. The best results in each column are in bold. Our model gets significantly better scores than all other comparisons for all metrics ($p<0.01$, paired t-test).

Models	Twitter				Reddit			
	AUC	Accuracy	Precision	F1	AUC	Accuracy	Precision	F1
Simple Baselines								
RANDOM	50.1	49.5	78.0	60.1	49.9	50.2	15.4	24.1
HISTORY	44.1	41.2	73.2	50.6	53.6	46.5	18.1	27.5
Comparisons								
SVM	51.5	56.3	75.7	74.2	54.3	50.1	18.9	29.1
BILSTM	52.5	77.4	78.2	87.2	59.3	54.1	22.2	31.7
BERT	70.5	80.2	80.4	89.0	63.2	51.1	21.6	32.5
CONVERNET	73.6	79.2	78.9	88.2	60.6	55.3	21.6	31.2
JECUH	75.2	80.1	80.3	88.4	60.7	57.6	22.6	31.9
Our Model	**83.2**	**82.9**	**84.7**	**90.2**	**64.8**	**62.7**	**24.9**	**34.6**

Experiments

Table 3: Comparison results with ablations(in %). Higher scores indicate better performance.

Models	Twitter		Reddit	
	Accuracy	F1	Accuracy	F1
w/o Topic Init	79.5	88.1	60.5	32.7
w/o Disc Concat	80.8	88.3	61.4	33.2
w/o Disc Att	81.1	88.6	60.2	33.5
Full Model	**82.9**	**90.2**	**62.7**	**34.6**

Experiments

Table 4: C_v scores for Top 5 and 10 words of leaned topics. The values range from 0.0 to 1.0, and higher scores indicate better topic coherence.

Models	Twitter		Reddit	
	5	10	5	10
LDA	0.498	0.393	0.483	0.377
NTM	0.499	0.425	0.492	0.397
Our	**0.504**	**0.431**	**0.495**	**0.412**

Experiments

Table 5: 5 sample latent discourse behaviors discovered from Reddit (The top 5 terms by likelihood are shown here). Names in the first column are our interpretation of the discourse behaviors according to the learned clusters. Discourse words indicating the behavior are highlighted in *blue and italic*.

Discourse	Top 5 representative terms
Disagreement	*but* have ask *different* see
Explanation	*because* stil when of that
Opinion	*think* my here ! *never*
Doubt	*n't* always want like *why*
Question	*?* For ! *where what*

Experiments

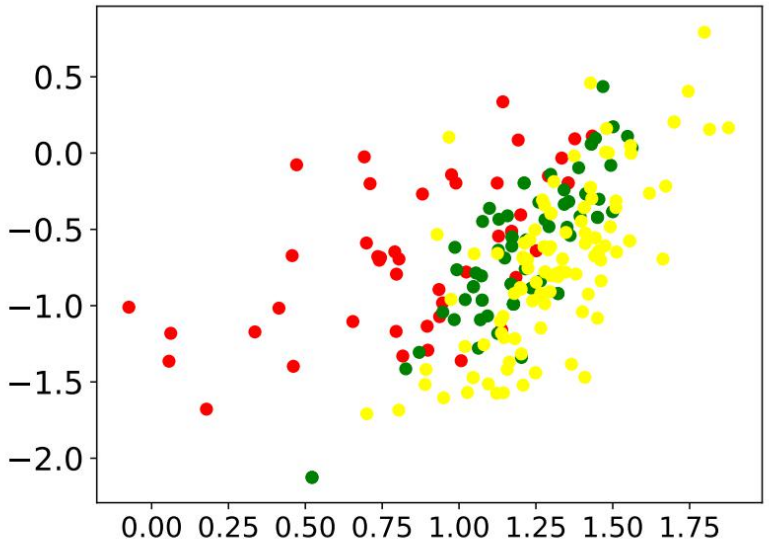
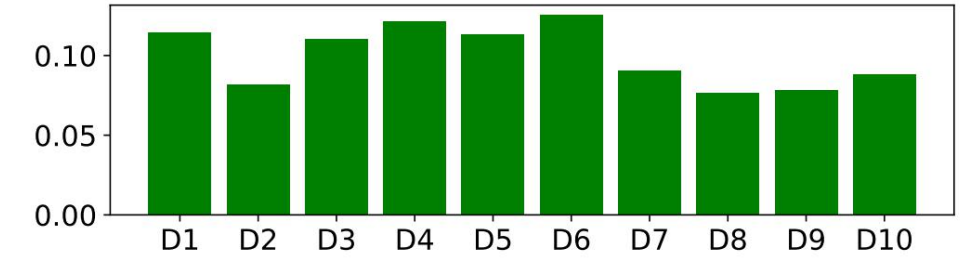
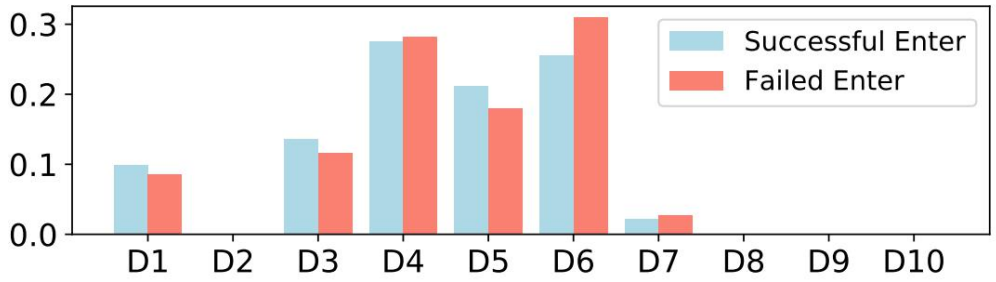


Figure 5: Topic mixture visualization (before softmax normalization) of three users' history conversations. A point refers to a conversation while different users are in distinguished colors.



(a) Disc distribution for newcomers



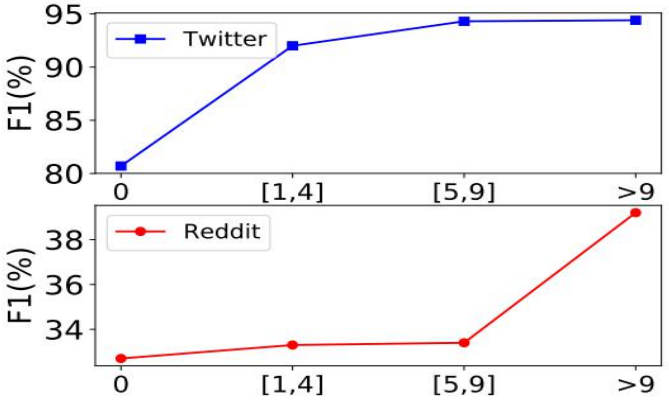
(b) Distribution for successful and failed new-entries

Figure 6: 6(a) is the distribution over discourse behaviors used in new-entries. 6(b) is the distribution of discourse behaviors for successful and failed new-entries. For both, X-axis: the 10 discourse behavior learned by our model; Y-axis: corresponding probability.

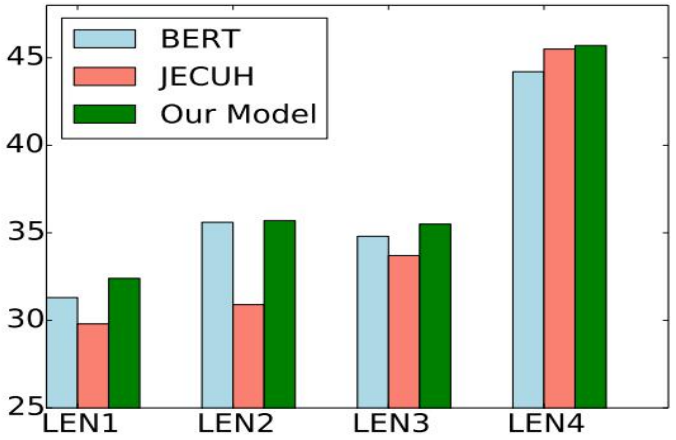
Experiments

Table 6: Human evaluation results (%). The overall inter-rater agreement achieved Krippendorff's α of 0.74, which indicates reliable results [30].

Models	Twitter				Reddit			
	OT	AQ	CL	CS	OT	AQ	CL	CS
SUCCESSFUL	100	28	0	54	98	48	14	38
FAILED	96	4	0	6	88	20	22	20



(a) F1 with Varying History #



(b) F1 with Varying Turn #

Figure 7: Y-axis: F1 score. In 7(a), X-axis: user history conversation numbers. In 7(b), LENi in X-axis: the i-th quantile by turn numbers (smaller i, shorter length).

Thanks !