The Thirty-Fourth AAAI Conference on Artificial Intelligence (AAAI-20)

Sentiment Classification in Customer Service Dialogue with Topic-Aware Multi-Task Learning

Jiancheng Wang,[1] Jingjing Wang,[1*] Changlong Sun,[2] Shoushan Li,[1]
Xiaozhong Liu,[2] Luo Si,[2] Min Zhang,[1] Guodong Zhou[1]
[1]School of Computer Science and Technology, Soochow University, China
[2]Alibaba Group, China
jcwang@stu.suda.edu.cn, {djingwang, lishoushan, minzhang, gdzhou}@suda.edu.cn
{changlong.scl, xiaozhong.lxz, luo.si}@alibaba-inc.com

Abstract

Sentiment analysis in dialogues plays a critical role in dialogue data analysis. However, previous studies on sentiment classification in dialogues largely ignore topic information, which is important for capturing overall information in some types of dialogues. In this study, we focus on the sentiment classification task in an important type of dialogue, namely customer service dialogue, and propose a novel approach which captures overall information to enhance the classification performance. Specifically, we propose a topic-aware multi-task learning (TML) approach which learns topic-enriched utterance representations in customer service dialogue by capturing various kinds of topic information. In the experiment, we propose a large-scale and high-quality annotated corpus for the sentiment classification task in customer service dialogue and empirical studies on the proposed corpus show that our approach significantly outperforms several strong baselines.

1 Introduction

In an active traditional/online business environment, customer service is of critical importance, and customer service dialogue mining can be an essential piece of business intelligence. In this background, sentiment analysis in customer service dialogue plays a critical role in various applications, such as service satisfaction analysis (Geetha 2017) and intelligent agent (Cui et al. 2017). In this study, we focus on sentiment classification in customer service dialogue, which aims to assign a proper sentiment label to each utterance in a customer service dialogue.

Different from in the dialogues of chit-chat in daily life, there exist two inherent challenges to perform sentiment classification in customer service dialogue: **(1) Overall Dialogue Motivation:** a *customer* or *service agent* normally initiates a dialogue with a strong and clear motivation and tends to talk around some specific topics which can be critical to detecting the utterance's sentiment in the dialogue context. Take the exemplar dialogue as shown in Figure 1, both the fourth and sixth utterances seem to express the *neutral* sentiment if only considering their content, i.e. ignoring the dialogue context. However, both the utterances actually express the *negative* sentiment if they are considered in the context of the whole dialogue under the goods delivery topic since the motivation of both the utterances is to inquire the reason for slow shipping and delivery. In this case, topic-related semantic clues from the whole dialogue context may help determine their sentiment, such as "*shipped*", "*tracking information*", "*last week*", "*warehouse*". **(2) Different Roles:** The *customer* (speaker **C**) and the *service agent* (speaker **A**) play different roles in a customer service dialogue. While the *customer* (**C**) tends to inquire questions or express dissatisfaction with the service, the *agent* (**A**) needs to address the issue for better service or promoting products. Such role-based motivations can be speculated from their utterances in the dialogue context, such as "*shipped*" for the customer and "*patient*" for the agent, and can further help determine their utterance sentiments.

C: *Hello, are you there?* [*neutral*]
A: *Hello, I'm here. It's my pleasure to help.*
What can I do for you? [*positive*]
C: *Why hasn't my order shipped yet?* [*negative*]
C: *My tracking information isn't available yet.* [*negative*]
A: *We need several days for processing. Please be patient.* [*neutral*]
C: *I placed my order last week.* [*negative*]
C: *What a business! Your service is poor.* [*very negative*]
A: *I'm sorry to keep you waiting, I just contacted the warehouse*
and your order will be on the way soon. [*neutral*]

Figure 1: A dialogue example in E-commerce customer service, where mark **C** denotes a *customer* and mark **A** denotes a *service agent*.

In order to overcome above two challenges, we propose a new topic-aware multi-task learning approach to sentiment classification in customer service dialogue by capturing various kinds of topic information, i.e., overall topic inference, customer-role topic inference, and agent-role topic inference. On one hand, overall topic inference captures the global topic information for the entire dialogue to model the overall dialogue motivation. On the other hand, customer-role topic inference and agent-role topic inference capture

[*] Corresponding Author: Jingjing Wang.

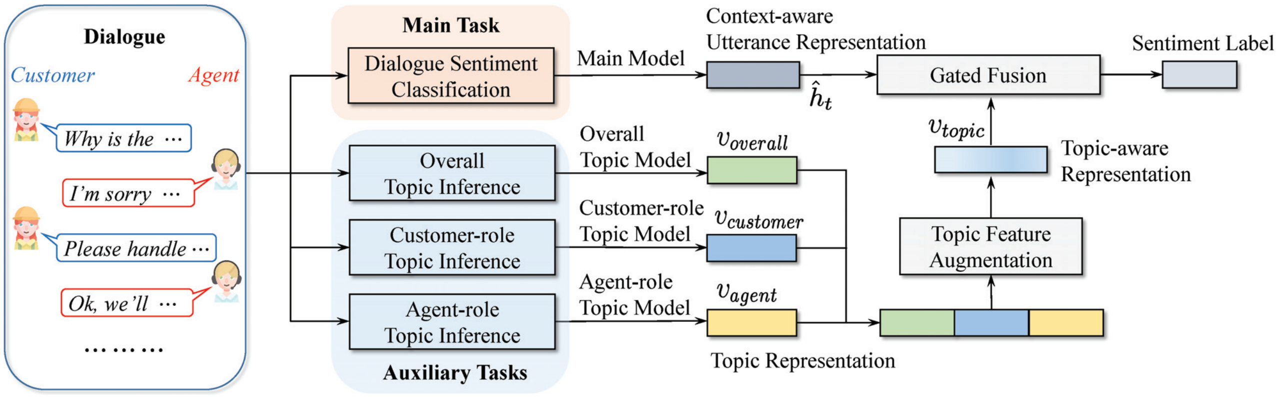

Figure 2: Overall architecture of our proposed Topic-aware Multi-task Learning (TML) approach.

role-based topic information to characterize different roles played by the *customer* and the *service agent* respectively. On this basis, we adopt a feature augmentation method to obtain the topic-aware customer representation and agent representation, which are further incorporated into the representation of each utterance using a gated fusion method.

Furthermore, to facilitate the research, we annotate a large-scale corpus[1] of high quality on a real-world dataset in E-commerce customer service. Experimentation on the proposed corpus shows that our topic-aware multi-task learning approach performs significantly better than several strong baselines.

It is worthwhile to note that, although the proposed approach is specifically designed for sentiment classification in customer service dialogue, our approach is believed to be applicable in many other types of dialogues, such as business meeting dialogue and court debate dialogue, where the dialogue motivation and multi-role information can share the same topic-aware multi-task learning approach.

2 Related Work

In the last decade, many researchers have devoted their efforts to sentiment analysis in NLP. Related studies could be divided into two groups, i.e., sentiment analysis in dialogue and sentiment analysis with topic model.

2.1 Sentiment Analysis in Dialogue

Early studies on sentiment analysis in dialogue often ignore contextual dependencies in dialogue and each utterance is regarded as an independent instance. In this scenario, text classification approaches like SVM and CNN can be readily applied to solve sentiment classification in dialogue. More recently, to model the contextual dependencies, some studies use LSTM to learn context-dependent representation from surrounding utterances (Hsu et al. 2018; Cerisara et al. 2018). Alternatively, Hazarika et al. (2018b) utilize memory network (Sukhbaatar et al. 2015) to model the contextual history. Hazarika et al. (2018a) consider inter-personal dependencies in contextual history and Majumder et al. (2019) define the information flow in several GRU (Chung et al. 2014) cells, and incorporate contextual extraction in the processing of each utterance. Especially, Shen et al. (2018) and Wang et al. (2019) employ a bidirectional attention network to capture the semantic matching information inside the single-turn dialogue, i.e., the question and answer pair.

Existing efforts in sentiment analysis in dialogue mainly focus on modeling contextual connections between utterances, while they neither address the characteristics of the dialogue as a whole, nor do they consider characteristics of specific roles. In contrast, our work attempts to capture various kinds of topic information for characterizing dialogue motivation and different roles in customer service dialogue.

2.2 Sentiment Analysis with Topic Model

Traditional topic models, such as probabilistic latent semantic analysis (pLSA) (Hofmann 1999) and latent Dirichlet allocation (LDA) (Blei, Ng, and Jordan 2003) have been widely used for inferring a low dimensional representation that captures the latent semantics of a document. Recently, neural topic models (Miao, Grefenstette, and Blunsom 2017; Srivastava and Sutton 2017) have been proposed to infer latent distribution due to the success of neural variational inference (Kingma and Welling 2014). In early studies, topic models have been employed to help text classification. For instance, Lin and He (2009) and Li, Huang, and Zhu (2010) propose LDA-based models to detect sentiment and topic simultaneously. Zeng et al. (2018) attempt to use topic model to address data sparsity problem in short text categorization while Huang et al. (2018) use a topic representation layer to get LDA features for helping sentence encoding.

However, all above methods use the topic model as a sentence encoder, similar to RNN or Transformer (Vaswani et al. 2017). Different from them, we utilize topic-aware multi-task learning with multiple topic models to mine various kinds of topic information in dialogues for sentiment classification.

3 Topic-aware Multi-task Learning (TML)

Figure 2 shows the overall architecture of our approach which consists of one main task and three auxiliary tasks.

[1] https://github.com/jc-wang/TML

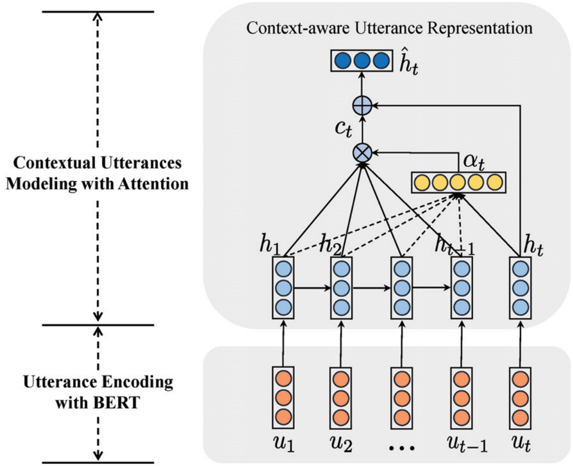
Context-aware Utterance Representation
$\hat{h}_t$
c_t
α_t
Contextual Utterances Modeling with Attention
h_1
h_2
h_{t-1}
h_t
Utterance Encoding with BERT
u_1
u_2
...
u_{t-1}
u_t

Figure 3: The architecture of the main task.

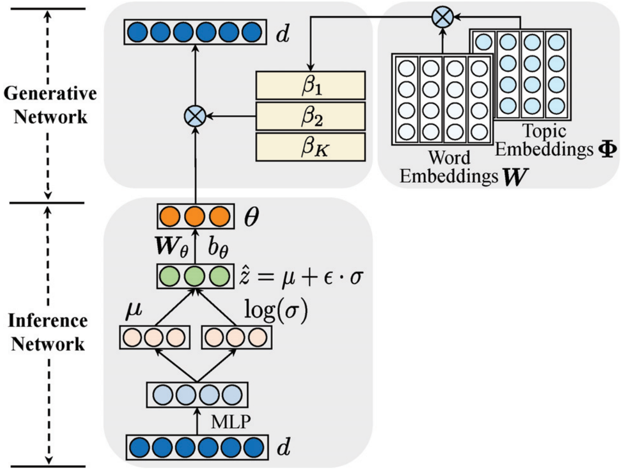
d
β_1
β_2
β_K
Generative Network
Word Embeddings $\boldsymbol{W}$
Topic Embeddings $\boldsymbol{\Phi}$
θ
$\boldsymbol{W}_\theta$ b_θ
$\hat{z} = \mu + \epsilon \cdot \sigma$
μ
$\log(\sigma)$
Inference Network
MLP
d

Figure 4: The architecture of the neural topic model.

Formally, assume that we have a dialogue dataset, where each dialogue d is composed of continuous utterances $\{x_t\}_{t=1}^{T}$ with corresponding sentiment labels $\{y_t\}_{t=1}^{T}$. In Section 3.1, we will introduce the main task of dialogue sentiment classification, which assigns a sentiment label to each utterance in the dialogue. In Section 3.2, we will introduce the three auxiliary tasks, i.e. overall topic inference, customer-role topic inference, and agent-role topic inference. In Section 3.3, we will give the strategy to combine the main and auxiliary tasks.

3.1 Main Task: Dialogue Sentiment Classification

Figure 3 shows the architecture of the main task, which contains two main components: utterance encoding with BERT and contextual utterance modeling with attention.

Utterance Encoding with BERT. Pre-trained language representations have been shown to improve many downstream NLP tasks such as question answering and natural language inference. In our approach, we apply the pre-trained model BERT (Devlin et al. 2019) as the shared utterance encoding model. In BERT, a special [CLS] token is added to the head of each utterance x_t. The final hidden state corresponding to this token is denoted as u_t and it is used as the aggregate utterance representation. Moreover, we fine-tune BERT and update the utterance representation u_t.

Contextual Utterances Modeling with Attention. In a dialogue, the sentiment of each utterance depends on the dialogue context. Thus, within a dialogue, there is a high probability of inter-dependency with respect to their sentimental clues. To characterize the information flow in the dialogue, we feed the utterance representation into an LSTM which sequentially connects utterances in a dialogue context. Formally, given the t-th utterance representation u_t of the utterance x_t, we update u_t as follows:

$$h_t = \text{LSTM}(u_t, h_{t-1}, m_{t-1}) \tag{1}$$

where $h_t \in \mathbb{R}^D$ is the hidden state of the LSTM for the utterance representation u_t and m_{t-1} is the memory cell state at the time-step $t-1$.

Inspired by Majumder et al. (2019), to estimate a long-range dependency, we employ an attention mechanism to compute the context representation $c_t \in \mathbb{R}^D$ relevant to utterance x_t:

$$\alpha_t = \text{softmax}(h_t^\top \boldsymbol{W}_\alpha [h_1, h_2, \cdots, h_{t-1}]) \tag{2}$$

$$c_t = \alpha_t [h_1, h_2, \cdots, h_{t-1}]^\top \tag{3}$$

where $h_1, h_2, \cdots, h_{t-1}$ are the hidden states for the preceding utterances of x_t. $\boldsymbol{W}_\alpha \in \mathbb{R}^{D \times D}$ is a model parameter to be learned. $\text{softmax}(x_i) = e^{x_i} / \sum_j e^{x_j}$. Then, we employ a fully connected layer to compute the context-aware utterance representation $\hat{h}_t \in \mathbb{R}^D$ as follows:

$$\hat{h}_t = \text{ReLU}(\boldsymbol{W}_c(h_t \oplus c_t) + b_c) \tag{4}$$

where ReLU denotes the rectified linear unit activation (Nair and Hinton 2010). $\oplus$ denotes vector concatenation operation. $\boldsymbol{W}_c \in \mathbb{R}^{D \times 2D}$ and $b_c \in \mathbb{R}^D$ are trainable parameters.

3.2 Auxiliary Tasks: Topic Modeling

In our approach, we employ three neural topic models as the auxiliary tasks to construct three different topic representations, i.e., overall topic representation, customer-role topic representation, and agent-role topic representation. Figure 4 shows the architecture of the neural topic model. In principle, the neural topic model consists of two components, i.e., the inference network and the generative network (Miao, Grefenstette, and Blunsom 2017).

• **Inference Network** is used to infer the topic distribution θ from document d. Formally, let $d \in \mathbb{R}^{|V|}$ be the bag-of-words representation (with stop words excluded) of a document where V is the vocabulary. We first construct an inference network to generate $\mu(d)$ and $\sigma(d)$ to parameterize $q(z|d) = \mathcal{N}(\mu(d), \sigma^2(d))$, where $z \in \mathbb{R}^K$ is a latent variable in the topic model. $q(z|d)$ is a diagonal Gaussian distribution. $\mu(d)$ and $\sigma(d)$ are functions of d which are implemented by multilayer perceptron (MLP). Then, we sample $\hat{z}$ from $q(z|d)$ using a reparameterization trick as described in Kingma and Welling (2014):

$$\hat{z} = \mu(d) + \epsilon \cdot \sigma(d) \tag{5}$$

where ϵ is sampled from $\mathcal{N}(\mathbf{0}, \boldsymbol{I}^2)$. Finally, we obtain topic distribution $\theta \in \mathbb{R}^K$ as follows:

$$\theta = \text{softmax}(\boldsymbol{W}_\theta \hat{z} + b_\theta) \tag{6}$$

where $\boldsymbol{W}_\theta \in \mathbb{R}^{K \times K}$ and $b_\theta \in \mathbb{R}^K$ are trainable parameters.

- **Generative Network** is used to parameterize $p(d|\theta, \beta)$ which is the probability distribution of the data given the latent topic distribution θ and the topic-word distribution β. Here, each row β_k in β is a distribution over words for the k-th topic. In the generative network, word embeddings $\boldsymbol{W} \in \mathbb{R}^{V \times M}$ and topic embeddings $\mathbf{\Phi} \in \mathbb{R}^{K \times M}$ (where M is the embedding dimension and K is the number of topics) are used to construct the topic-word distribution as follows:

$$\beta_k = \text{softmax}\Big(\frac{\boldsymbol{W} \cdot \phi_k}{\sqrt{M}}\Big) \tag{7}$$

where ϕ_k is k-th topic embedding in $\mathbf{\Phi}$.

The loss function for the topic model is defined as follows:

$$\mathcal{L} = \text{KL}[q(z|d)||p(z)] - \mathbb{E}_{q(z|d)}\big[\text{log}p(d|\theta, \beta)\big] \tag{8}$$

where $p(z)$ is a standard Normal prior $\mathcal{N}(\mathbf{0}, \boldsymbol{I}^2)$. The first term ensures that our learned distribution $q(z|d)$ is similar to the true prior distribution $p(z)$, measured by the KL divergence. The second term represents the reconstruction document likelihood obtained from the generative network. For more derivation details, we refer the readers to Miao, Grefenstette, and Blunsom (2017) due to the space limitation.

In our approach, the topic embeddings learned in the generative network and the topic distribution learned in the inference network will be further used to construct topic representation for each dialogue. On this basis, we propose three auxiliary tasks, i.e., overall topic inference, customer-role topic inference, and agent-role topic inference.

- **Auxiliary Task 1: Overall Topic Inference.** Each dialogue is represented as a bag-of-words representation $d_{overall}$, which is fed into the overall topic model to get the topic distribution $\theta_{overall}$. Then, the overall topic representation $\upsilon_{overall} \in \mathbb{R}^M$ is obtained as follows:

$$\upsilon_{overall} = \mathbf{\Phi}_{overall}^\top \cdot \theta_{overall} \tag{9}$$

where $\mathbf{\Phi}_{overall} \in \mathbb{R}^{K \times M}$ and $\theta_{overall} \in \mathbb{R}^K$ denote the topic embeddings and topic distribution in the overall topic model.

- **Auxiliary Task 2: Customer-Role Topic Inference.** In a customer service dialogue, utterances are divided into customer's utterances and agent's utterances. All customer's utterances are represented as a bag-of-words representation $d_{customer}$ which is fed into the customer-role topic model to get the topic distribution $\theta_{customer}$. Then, the customer-role topic representation $\upsilon_{customer} \in \mathbb{R}^M$ is obtained as follows:

$$\upsilon_{customer} = \mathbf{\Phi}_{customer}^\top \cdot \theta_{customer} \tag{10}$$

where $\mathbf{\Phi}_{customer} \in \mathbb{R}^{K \times M}$ and $\theta_{customer} \in \mathbb{R}^K$ denote the topic embeddings and topic distribution in the customer-role topic model.

- **Auxiliary Task 3: Agent-Role Topic Inference.** Similar to the process of customer-role topic inference, all agent's utterances are represented as a bag-of-words representation d_{agent} which is fed to the corresponding agent-role topic model to get the topic distribution θ_{agent}. Then, the agent-role topic representation $\upsilon_{agent} \in \mathbb{R}^M$ is obtained as follows:

$$\upsilon_{agent} = \mathbf{\Phi}_{agent}^\top \cdot \theta_{agent} \tag{11}$$

where $\mathbf{\Phi}_{agent} \in \mathbb{R}^{K \times M}$ and $\theta_{agent} \in \mathbb{R}^K$ denote the topic embeddings and topic distribution in the agent-role topic model.

Topic Feature Augmentation. Due to different roles in a customer service dialogue, we employ different feature augmentation operations for the customer and agent utterances.

For a customer utterance, we compute the topic-aware customer representation as follows:

$$\upsilon_{topic} = \upsilon_{overall} \oplus \upsilon_{customer} \oplus \mathbf{0} \tag{12}$$

where $\oplus$ denotes vector concatenation and $\mathbf{0} \in \mathbb{R}^M$ denotes a $\boldsymbol{M}$-dimensional vector with all elements set to 0.

For an agent utterance, we compute the topic-aware agent representation as follows:

$$\upsilon_{topic} = \upsilon_{overall} \oplus \mathbf{0} \oplus \upsilon_{agent} \tag{13}$$

3.3 Combination Strategy

Gated Fusion. To distinguish the main task representation from those of auxiliary tasks, a fusion gate $g \in \mathbb{R}^D$ is used to combine utterance representations from both the main and auxiliary tasks. The final utterance representation $r_t \in \mathbb{R}^D$ is obtained as follows:

$$g = \text{sigmoid}\Big(\boldsymbol{W}_g(\hat{h}_t \oplus \upsilon_{topic}) + b_g\Big) \tag{14}$$

$$r_t = (1 - g) \odot \hat{h}_t + g \odot (\boldsymbol{W}_\upsilon \upsilon_{topic}) \tag{15}$$

where $\odot$ denotes element-wise multiplication. $\boldsymbol{W}_g \in \mathbb{R}^{D \times (D+3M)}$, $b_g \in \mathbb{R}^D$ and $\boldsymbol{W}_\upsilon \in \mathbb{R}^{D \times 3M}$ are trainable parameters. $\hat{h}_t$ is the context-aware utterance representation (see Eq.(4)). Then, the utterance representation r_t is fed to a softmax classifier for performing sentiment classification:

$$p_\Theta = \text{softmax}(\boldsymbol{W}_\gamma r_t + b_\gamma) \tag{16}$$

where $\boldsymbol{W}_\gamma \in \mathbb{R}^{C \times D}$ and $b_\gamma \in \mathbb{R}^C$ are trainable parameters. C is the number of sentiment categories. p_Θ is the predicted probability distribution of all sentiment labels.

Joint Learning. We employ the joint loss function to optimize all the main and auxiliary tasks simultaneously. Here, the joint loss consists of two parts. One is the supervised loss for the main task of dialogue sentiment classification, and the other contains the unsupervised losses for the three auxiliary tasks of neural topic modeling.

Specifically, the loss $\mathcal{L}_{main}$ for the main task of dialogue sentiment classification is computed as follow:

$$\mathcal{L}_{main} = -\frac{1}{N}\sum_{i=1}^{N} \log p_\Theta(y_i|x_i) + \frac{\delta}{2}||\Theta||_2^2 \tag{17}$$

where N is the number of all utterances across all dialogues. y_i is the ground-truth label for utterance x_i. δ is an L_2 regularization weight. Θ denotes all trainable parameters in the model.

	# of dialogues	# of utterances	# of each sentiment label				
			very negative	*negative*	*neutral*	*positive*	*very positive*
Train	2555	49710	541	4789	23187	19395	1798
Dev.	320	6357	85	805	2798	2471	198
Test	621	12149	131	1391	5210	4934	483

Table 1: Data distributions of the annotated data.

In addition, the loss function for neural topic model has been shown in Eq.(8). For clarity, the losses for the auxiliary tasks, i.e., overall topic inference, customer-role topic inference, agent-role topic inference, are denoted as $\mathcal{L}_{overall}$, $\mathcal{L}_{customer}$, $\mathcal{L}_{agent}$ respectively.

Finally, the joint loss $\mathcal{L}$ is obtained as follows:

$$\mathcal{L} = \mathcal{L}_{main} + \lambda(\mathcal{L}_{overall} + \mathcal{L}_{customer} + \mathcal{L}_{agent}) \quad (18)$$

where λ is a weight parameter to balance the losses for the main task and auxiliary tasks.

4 Experimentation

4.1 Experimental Settings

Data Collection and Annotation. We collect a dialogue dataset from an online customer service system in a top E-commerce company in China. For the annotation, we define five sentiment labels, i.e., very negative, negative, neutral, positive, and very positive. Specifically, for the utterances containing obvious negative words, such like dirty words, we annotate them as *very negative*; for the utterances which express dissatisfaction/complaint with the product or service but without strong negative words, we annotate them as *negative*; for the utterances that are commonly used to express gratitude or the utterances with one or two positive sentiment words, we annotate them as *positive*; for the utterances with positive sentiment words modified by intensity adverbs or with a number of strong positive sentiment words or phrases, we annotate them as *very positive*; for the other utterances without a clear sentiment polarity, we annotate them as *neutral*. During the annotation, we assign two annotators to annotate each utterance in a dialogue and the final Kappa consistency check value of this annotation is 0.79. When two annotators cannot reach an agreement, an expert is invoked to make the final decision, ensuring the quality of data annotation. Table 1 shows the detail statistics of the final dataset.

Implementation Details. We use the pre-trained BERT-Base model[2] to initialize the BERT model in our TML approach. The dimension of LSTM hidden state is set to be 256. The dimension M of embeddings in topic model is set to be 100 and the number of topics (K) is set to be 20 for all topic models. The word embeddings in neural topic model are pre-trained using *Glove* (Pennington, Socher, and Manning 2014). All weights of other layers are initialized by the Glorot uniform initializer (Glorot and Bengio 2010). Batch size is set to be 32. In addition, other hyper-parameters are fine-tuned with the development data. Specifically, λ in Eq.(18) is set to be 0.01. The dropout rate (Srivastava et al. 2014) is 0.3. The L_2 regularization weight of parameters is 10^{-5}. Finally, we use Adam optimizer (Kingma and Ba 2014) for training our TML approach with the initial learning rate of 0.001.

Evaluation Metrics. F1 score is used to measure the performance of the proposed approach for each label and Macro-F1 score is used to evaluate the overall performance by averaging the performances of all labels. In addition, t-test is used to evaluate the significance of the performance difference between two approaches (Yang, Liu, and others 1999).

4.2 Experimental Results

Overall Performance. For thorough comparison, we implement following baselines to sentiment classification in customer service dialogue.

- **BERT (Devlin et al. 2019),** where a BERT model is used to construct the utterance representations which are sent to a two-layer perceptron with a final softmax layer for sentiment classification.
- **LDA-LSTM (Huang et al. 2018),** where a BiLSTM model is first used to get the sentence representation. Then, a topic representation layer is used to get topic features from LDA. Finally, all features are concatenated for sentiment classification.
- **LDA-BERT (Huang et al. 2018),** where LSTM is replaced with BERT in LDA-LSTM.
- **CMN (Hazarika et al. 2018b),** where one utterance and previous utterances of each speaker are fed to two memory networks for obtaining the final representation. For a fair comparison, we use BERT as the utterance encoder to generate the utterance representations.
- **ICON (Hazarika et al. 2018a),** where one utterance with previous utterances is first provided as input and then a memory-based GRU is used to obtain the final representation. We also use BERT as the utterance encoder.
- **DialogueRNN (Majumder et al. 2019),** where a variant of RNN combined with attention mechanism is used to model the information flow during the dialogue. We also use BERT as the utterance encoder.
- **ConGCN (Zhang et al. 2019),** where a graph is used to model the connections between utterances and speakers. Each utterance and speaker in the corpus is represented as a node in the graph for performing sentiment classification. We also use BERT as the utterance encoder.
- **TML (Our Approach).** This is exactly our topic-aware multi-task learning approach, where a main task on dialogue sentiment classification along with three auxiliary tasks are jointly learned.

[2] https://github.com/jc-wang/TML

Models	*very negative*	*negative*	*neutral*	*positive*	*very positive*	*Macro-F1*
BERT (Devlin et al. 2019)	49.8	56.4	87.4	92.6	75.5	72.3
LDA-LSTM (Huang et al. 2018)	48.6	54.3	87.1	92.6	75.4	71.6
LDA-BERT (Huang et al. 2018)	50.1	57.2	87.3	92.7	75.3	72.5
CMN (Hazarika et al. 2018b)	52.9	57.4	87.8	92.9	75.6	73.3
ICON (Hazarika et al. 2018a)	53.6	57.5	88.0	93.0	75.8	73.6
DialogueRNN (Majumder et al. 2019)	55.1	58.2	87.9	92.9	75.6	73.9
ConGCN (Zhang et al. 2019)	52.7	57.9	87.9	92.6	75.2	73.3
TML (Our Approach)	**57.2**	**63.4**	**88.4**	**93.2**	**77.4**	**75.9**

Table 2: Comparison of our approach and other state-of-the-art baselines.

Models	*very negative*	*negative*	*neutral*	*positive*	*very positive*	*Macro-F1*
Main Task	54.3	57.0	87.9	92.7	75.2	73.4
+ Auxiliary Task 1	55.5	60.5	88.2	93.2	76.7	74.8
+ Auxiliary Task 1,2	**57.5**	62.6	88.3	93.1	76.5	75.6
+ Auxiliary Task 1,3	55.7	60.9	**88.7**	93.0	**77.6**	75.2
+ Auxiliary Task 1,2,3	57.2	**63.4**	88.4	**93.2**	77.4	**75.9**

Table 3: Performances of the main task with different auxiliary tasks.

Table 2 shows the comparison results of all above approaches. From this table, we can see that:

1) BERT without using any topic information performs better than **LDA-LSTM**. This result indicates the effectiveness of using **BERT** to generate utterance representations in our approach.

2) CMN, **ICON**, **DialogueRNN**, and **ConGCN** are all superior to a single **BERT** model. This result highlights the importance of considering contextual information in sentiment classification in dialogues.

3) Among all approaches, our **TML** approach performs best and the significance test shows that these improvements are all significant (p-value < 0.05). This result verifies the effectiveness of using topic information to perform sentiment classification in customer service dialogue.

Contribution of Various Kinds of Topic Information. This is done by augmenting the main task with different kinds of topic information, where

Main Task only adopts the main task, i.e., dialogue sentiment classification,

+ Auxiliary Task 1 only incorporates the auxiliary task of overall topic inference with the main task,

+ Auxiliary Task 1,2 incorporates two auxiliary tasks, i.e., overall topic inference and customer-role topic inference, with the main task,

+ Auxiliary Task 1,3 incorporates two auxiliary tasks, i.e., overall topic inference and agent-role topic inference, with the main task,

+ Auxiliary Task 1,2,3 is exactly our TML approach.

Table 3 shows the performances of the main task with various auxiliary tasks. From this table, we can see that:

1) + Auxiliary Task 1 significantly outperforms **Main Task** with the improvement of 1.4% in terms of *Macro-F1* (p-value < 0.05). This result indicates that the overall topic information is helpful for sentiment classification in customer service dialogue.

2) + Auxiliary Task 1,2 and **+ Auxiliary Task 1,3** both perform slightly better than **+ Auxiliary Task 1**. This result indicates that further incorporating customer-role topic or agent-role topic besides overall topic is also helpful for sentiment classification in customer service dialogue.

3) + Auxiliary Task 1,2 is superior to **Main Task** in two categories, *negative* and *very negative*. This is reasonable since the Auxiliary Task 2, i.e., customer-role topic inference, aims to mine the information in the customer utterances. Compared to a *service agent*, a *customer* is more likely to express the *negative* sentiment in his/her utterances. Therefore, mining more information in customer utterances is more helpful for detecting *negative* utterances.

4) + Auxiliary Task 1,2,3 performs best among all approaches and significantly outperforms **Main Task** by 2.5% (p-value < 0.05). This confirms the effectiveness of considering both overall topic and different role topic information in sentiment classification in customer service dialogue.

4.3 Analysis and Discussion

To better understand the auxiliary tasks and corresponding topic models, we qualitatively evaluate the semantic information learned by topic models on the customer service dialogue dataset and give a case study of our TML approach.

Topic Analysis for Auxiliary Tasks. In the topic model, a particular topic is a probability distribution over the vocabulary calculated by Eq.(7), in which we selected the 10 words with the highest probability to represent the topic. Figure 5 gives 2 selected topics with top 10 words for each topic model. From the table, we can see that, in the overall topic model, Topic 1 often exists in pre-sale dialogues. For instance, in some pre-sale dialogues, some *customers* would like to ask detailed information about products. In this sce-

Overall Topic Model		Customer-role Topic Model		Agent-role Topic Model	
Topic 1	Topic 2	Topic 1	Topic 2	Topic 1	Topic 2
穿*(wear)*	退货*(return)*	发货*(ship)*	退*(return)*	拍*(order)*	退款*(refund)*
建议*(suggestion)*	运费*(shipping cost)*	几天*(a few days)*	运费*(shipping cost)*	付款*(payment)*	申请*(application)*
尺码*(size)*	申请*(application)*	催*(reminder)*	问题*(problem)*	活动*(activity)*	原因*(reason)*
身高*(height)*	退款*(refund)*	多久*(how long)*	质量*(quality)*	喜欢*(like)*	退货*(return)*
发货*(ship)*	运费险*(freight insurance)*	慢*(slow)*	退款*(refund)*	下单*(order)*	包裹*(package)*
体重*(weight)*	原因*(reason)*	没货*(out of stock)*	差评*(bad review)*	优惠*(discount)*	销售*(sale)*
下单*(order)*	承担*(bear)*	今天*(today)*	感觉*(feel)*	尺码*(size)*	售后*(after sales)*
谢谢*(thanks)*	问题*(problem)*	快点*(hurry)*	不好*(bad)*	放心*(don't worry)*	提交*(submit)*
推荐*(recommend)*	质量*(quality)*	快递*(delivery)*	怎么办*(how to do)*	优惠券*(coupon)*	理由*(reason)*
款式*(style)*	理由*(reason)*	时候*(when)*	麻烦*(trouble)*	购买*(buy)*	处理*(deal with)*

Figure 5: Top-10 words of topics in three topic models.

<table>
<tr><th colspan="3">Dialogue Example 1</th><th colspan="3">Dialogue Example 2</th></tr>
<tr><td colspan="3">………
C: 为什么还不发货？ [*negative*]
(*Why hasn't my order shipped yet?*)
C: 快递信息还查不到。 [*negative*]
(*My tracking information isn't available yet.*)
A: 我们需要几天处理一下，麻烦您耐心等待下。 [*neutral*]
(*We need several days for processing. Please be patient.*)
C: 我上个星期就买了。 [To classify]
(*I placed my order last week.*)</td><td colspan="3">C: 我要退货，裤子太小了。 [*negative*]
(*I want to return the goods, the pants are too small.*)
………
C: 那运费怎么算? [*neutral*]
(*Who will pay for the shipping?*)
A: 非质量问题运费自理。 [*neutral*]
(*You should pay for it if the item has no quality problems.*)
C: 我信你的推荐，买小了，运费也要我付? [To classify]
(*I trusted the size you recommended, but found pants too small, should I pay for the shipping?*)</td></tr>
<tr><td>BERT
✗(*neutral*)
P(*negative*)=0.41</td><td>DialogueRNN
✓(*negative*)
P(*negative*)=0.59</td><td>TML(Our Approach)
✓(*negative*)
P(*negative*)=0.81</td><td>BERT
✗(*neutral*)
P(*negative*)=0.25</td><td>DialogueRNN
✗(*neutral*)
P(*negative*)=0.37</td><td>TML(Our Approach)
✓(*negative*)
P(*negative*)=0.74</td></tr>
</table>

Figure 6: Examples from the test data with their categories predicted by different approaches (i.e., BERT, DialogueRNN and our approach). ✓(or ✗) denotes that the predicted category is correct (or wrong).

nario, the utterances often contain the words, such as ***"size"*** and ***"weight"***. Topic 2 is more likely to exist in after-sale dialogues. For instance, in some after-sale dialogues, some *customers* want to return and discuss the shipping costs with the *agent*. In this scenario, the utterances often contain the words, such as ***"return"*** and ***"refund"***.

In the role-based topic model, some topics are largely related to specific roles, which is difficult to observe in the overall topic model. For instance, we can find that some words, such as ***"a few days"*** and ***"slow"*** occur in Topic 1 in the customer topic model, and these words describe the customer's true feelings about the delivery. From the topic words of the Topic 1 in the agent topic model, we can find some words, such as ***"activity"*** and ***"discount"***, which often appear in some sale promotion activities.

Case Study. Figure 6 shows two dialogue examples, along with their predicted categories and probabilities of the ground-truth label by different approaches. From this table, we can see that: **1)** For dialogue example 1 about the delivery, both the predictions of **DialogueRNN** and **TML (Our Approach)** are *negative* and correct, while **BERT** modeling each utterance independently is wrong. However, the probability for ground-truth label *negative* by **TML** is much higher than that of **DialogueRNN** (0.81 vs. 0.59). This indicates that **TML** considering the additional overall and role topic information (i.e., shipping and delivery slowness) is superior to **DialogueRNN** which only models contextual connections between utterances. **2)** For dialogue example 2 about the return of goods, the utterance (to classify) is used to express complaints by the *customer* for that the *agent* recommended the wrong size of pants. We find that both **BERT** and **DialogueRNN** give wrong predictions, while **TML** can still give the correct prediction. This is reasonable because in a customer service dialogue about the return of goods, the utterance which is used to express complaints about shipping costs by the *customer*, is more likely to express the *negative* sentiment.

5 Conclusion

In this paper, we conduct our research on sentiment classification in customer service dialogue. In particular, a large-scale and high-quality corpus is constructed for this task. On this basis, we propose a Topic-aware Multi-task Learning (TML) approach to solving the challenges therein. Specifically, we propose three auxiliary topic inference tasks to learn the overall, customer-role, and agent-role topic information, so as to improve the main task of dialogue sentiment classification. Empirical studies show that our TML approach significantly outperforms several strong baselines.

In our future work, we would like to handle the class imbalance problem in sentiment classification in customer service dialogue. As shown in the data distribution of dialogue dataset, there are much less *negative* samples than *positive* or *neutral* samples but recognizing *negative* samples is some-times more important in customer service data mining.

Acknowledgments

We thank our anonymous reviewers for their helpful comments. This work was supported by three NSFC grants, i.e., No.61702149, No.61672366, and No.61525205. This work was also supported by a Project funded by China Postdoctoral Science Foundation No.2019M661930, a Project Funded by the Priority Academic Program Development of Jiangsu Higher Education Institutions (PAPD) and a joint research project of Alibaba Group and Soochow University.

References

Blei, D. M.; Ng, A. Y.; and Jordan, M. I. 2003. Latent dirichlet allocation. *Journal of machine Learning research* 3(Jan):993–1022.

Cerisara, C.; Jafaritazehjani, S.; Oluokun, A.; and Le, H. T. 2018. Multi-task dialog act and sentiment recognition on mastodon. In *Proceedings of COLING-2018*, 745–754.

Chung, J.; Gülçehre, Ç.; Cho, K.; and Bengio, Y. 2014. Empirical evaluation of gated recurrent neural networks on sequence modeling. *CoRR* abs/1412.3555.

Cui, L.; Huang, S.; Wei, F.; Tan, C.; Duan, C.; and Zhou, M. 2017. Superagent: A customer service chatbot for e-commerce websites. In *Proceedings of ACL-2017*, 97–102.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-2019*, 4171–4186.

Geetha, M. e. a. 2017. Relationship between customer sentiment and online customer ratings for hotels-an empirical analysis. *Tourism Management* 61:43–54.

Glorot, X., and Bengio, Y. 2010. Understanding the difficulty of training deep feedforward neural networks. In *Proceedings of AISTATS-2010*, 249–256.

Hazarika, D.; Poria, S.; Mihalcea, R.; Cambria, E.; and Zimmermann, R. 2018a. Icon: Interactive conversational memory network for multimodal emotion detection. In *Proceedings of EMNLP-2018*, 2594–2604.

Hazarika, D.; Poria, S.; Zadeh, A.; Cambria, E.; Morency, L.-P.; and Zimmermann, R. 2018b. Conversational memory network for emotion recognition in dyadic dialogue videos. In *Proceedings of NAACL-2018*, volume 1, 2122–2132.

Hofmann, T. 1999. Probabilistic latent semantic indexing. In *Proceedings of SIGIR-1999*, 50–57. ACM.

Hsu, C.-C.; Chen, S.-Y.; Kuo, C.-C.; Huang, T.-H.; and Ku, L.-W. 2018. Emotionlines: An emotion corpus of multiparty conversations. In *Proceedings of LREC-2018*.

Huang, Y.; Jiang, Y.; Hasan, T.; Jiang, Q.; and Li, C. 2018. A topic bilstm model for sentiment classification. In *Proceedings of ICIAI-2018*, 143–147. ACM.

Kingma, D. P., and Ba, J. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Kingma, D. P., and Welling, M. 2014. Auto-encoding variational bayes. In *Proceedings of ICLR-2014*.

Li, F.; Huang, M.; and Zhu, X. 2010. Sentiment analysis with global topics and local dependency. In *Proceedings of AAAI-2010*.

Lin, C., and He, Y. 2009. Joint sentiment/topic model for sentiment analysis. In *Proceedings of CIKM-2009*, 375–384. ACM.

Majumder, N.; Poria, S.; Hazarika, D.; Mihalcea, R.; Gelbukh, A.; and Cambria, E. 2019. Dialoguernn: An attentive rnn for emotion detection in conversations. In *Proceedings of AAAI-2019*, volume 33, 6818–6825.

Miao, Y.; Grefenstette, E.; and Blunsom, P. 2017. Discovering discrete latent topics with neural variational inference. In *Proceedings of ICML-2017*.

Nair, V., and Hinton, G. E. 2010. Rectified linear units improve restricted boltzmann machines. In *Proceedings of ICML-2010*, 807–814.

Pennington, J.; Socher, R.; and Manning, C. 2014. Glove: Global vectors for word representation. In *Proceedings of EMNLP-2014*, 1532–1543.

Shen, C.; Sun, C.; Wang, J.; Kang, Y.; Li, S.; Liu, X.; Si, L.; Zhang, M.; and Zhou, G. 2018. Sentiment classification towards question-answering with hierarchical matching network. In *Proceedings of EMNLP-2018*, 3654–3663.

Srivastava, A., and Sutton, C. 2017. Autoencoding variational inference for topic models. In *Proceedings of ICLR-2017*.

Srivastava, N.; Hinton, G.; Krizhevsky, A.; Sutskever, I.; and Salakhutdinov, R. 2014. Dropout: a simple way to prevent neural networks from overfitting. *The journal of machine learning research* 15(1):1929–1958.

Sukhbaatar, S.; Weston, J.; Fergus, R.; et al. 2015. End-to-end memory networks. In *Proceedings of NIPS-2015*, 2440–2448.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. In *Proceedings of NIPS-2017*, 5998–6008.

Wang, J.; Sun, C.; Li, S.; Liu, X.; Si, L.; Zhang, M.; and Zhou, G. 2019. Aspect sentiment classification towards question-answering with reinforced bidirectional attention network. In *Proceedings of ACL-2019*, 3548–3557.

Yang, Y.; Liu, X.; et al. 1999. A re-examination of text categorization methods. In *Proceedings of SIGIR-1999*, 42–49.

Zeng, J.; Li, J.; Song, Y.; Gao, C.; Lyu, M. R.; and King, I. 2018. Topic memory networks for short text classification. In *Proceedings of EMNLP-2018*, 3120–3131.

Zhang, D.; Wu, L.; Sun, C.; Li, S.; Zhu, Q.; and Zhou, G. 2019. Modeling both context- and speaker-sensitive dependence for emotion detection in multi-speaker conversations. In *Proceedings of IJCAI-2019*, 5415–5421.