A Two-stage Attention-based Model for Customer Satisfaction Prediction in E-commerce Customer Service

Anonymous ACL submission

Abstract

Nowadays, customer satisfaction prediction (CSP) on e-commerce platforms has become a hot research topic for both intelligent customer service and artificial customer service. CSP aims to discover customer satisfaction according to the dialogue content of customer and customer service, for the purpose of improving service quality and customer experience. In this paper, we focus on CSP for intelligent customer service chatbots. Although previous works have made some progress in many aspects, they mostly ignore the huge differences of expressions between customer and customer service, and fail to adequately consider the internal relations of those two kinds of personalized expressions. Thus, for emphasizing the importance of modeling customer part and service part separately, in this work we propose a two-stage dialogue-level classification model, which contains an intra-stage and an inter-stage to handle the issues above. In the intra-stage, we model customer part and service part separately by using attention mechanism combined with personalized context to obtain *customer state* and *service state*. Then we interact those two states with each other in the inter-stage to capture the final satisfaction representation of the whole dialogue. Experiment results demonstrate that our model achieves better performance than several competitive baselines on our in-house dataset and four public datasets.

1 Introduction

With the development of e-commerce platforms in recent years, a large number of companies use customer service chatbots, for the reasons that they could answer to customers' questions quickly and save labor cost. Customer satisfaction prediction (CSP) for the dialogue of customer and customer service chatbot has become an important problem in industry. For one thing, customers' satisfaction is a crucial indicator to evaluate the quality of service, which can help improve the ability of chatbots.

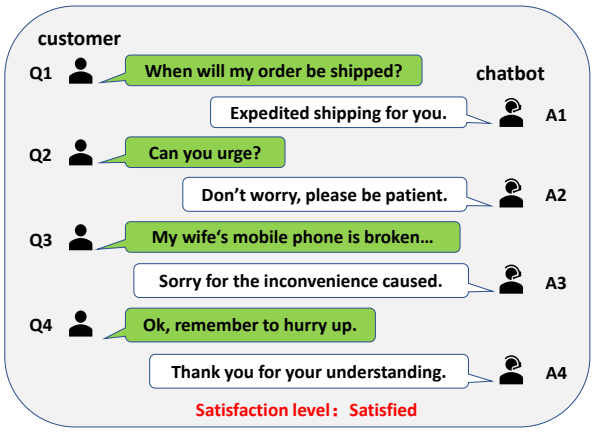

Figure 1: A dialogue of customer and chatbot on e-commerce platform.

For another, predicting customers' satisfaction in real time helps platforms handle problematic dialogues by transferring customer service chatbots to staffs timely, which can improve the customers' experience.

CSP is a multi-class classification task. Existing researches on CSP is mainly divided into two parts, one is turn-level CSP, the other is dialogue-level CSP. The former task concerns satisfaction prediction in every customer-service turn, while the latter one predicts satisfaction level of the whole dialogue. On e-commerce platforms, customer service aims to provide information for customers and solve their problems. Customer's satisfaction of the whole dialogue is the key point to evaluate the quality of the service and whether the customer's problem has been solved. In this study, we concentrate on dialogue-level CSP with five satisfaction levels (*strongly satisfied*, *satisfied*, *neutral*, *dissatisfied*, or *strongly dissatisfied*). As shown in figure 1, the customer expressed his anxiety and displeasure at the beginning, then turned into satisfied after the answers of the chatbot.

To address the dialogue-level CSP task, many

approaches extracted features from dialogue content and built models to fully utilize the interaction between customer questions and customer service answers. Some earlier studies use manual features to present conversational context (Hakkani-Tür and Ostendorf, 2010; Gangemi et al., 2015), while recent studies concerned more on how questions and answers interact each other (Song et al., 2019a; Yao et al., 2020). Although these works have made great progress in CSP task, two issues still remain. Firstly, existing studies ignore the huge differences of expressions between customers and customer service chatbots, in terms of the emotion intensity, language habits, language richness, and sentence length etc. Secondly, most prior studies fail to adequately consider the internal relations of personalized expressions for customers and staffs/chatbots respectively.

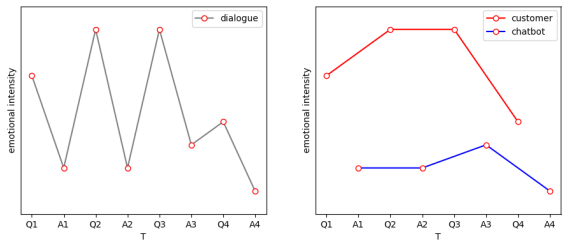

Figure 2: The emotional intensity trends are obviously after split

According to the above analysis, we figure that besides handling the interaction of customer questions and customer service answers, modeling customer part and service part separately should also be taken into consideration due to their expression differences in many aspects. For example, customers' questioning emotion is volatile and the intensity is high, while the answering emotion of service is relatively stable and the intensity is low. Figure 2 shows the emotional intensity trend of the case in figure 1, in which the customer's emotional intensity is higher with greater fluctuation, while chatbot is the opposite. After splitting the dialogue into customer part and service part, we are able to catch the emotional intensity trends of both intuitively. For the similar reason, other aspects of expression differences also matter.

Thus, we propose a two-stage classification model for CSP in E-commerce service. Our model consists of four sub-modules. Firstly, we adopt convolutional neural networks (CNNs) as the encoding module to extract features in dialogue content. Next comes the intra-stage, which consists of customer part and service part. We split customer questions and customer service answers as two independent sequences and send them into two parts separately. Specifically, each part exploits Long Short-Term Memory (LSTM) networks and Gated Recurrent Units (GRUs) to fully extract the internal relations of the sequence. After an attention layer combined with personalized context and a GRU, we get *customer state* and *service state* as the results of intra-stage. Then, the inter-stage apply an interactive attention mechanism to capture satisfaction representations of the whole dialogue from *customer state* and *service state*. In the end, a decoder module contributes to predict the final satisfaction classification.

To summarize, our contributions are as follows:

- We propose a dialogue-level classification model for CSP in E-commerce customer service chatbots.
- By bringing forward a two-stage architecture, we split the dialogue content into customer part and service part to model them separately. With the results of *customer state* and *service state*, we construct interaction to capture final satisfaction representation. This architecture handles two issues well while absorbing merits from existing works.
- Experimental results indicate that our proposed model outperforms all the baselines on our in-house dataset and four public datasets.

2 Related Work

In recent years, people pay much more attention to CSP and similar tasks. Some earlier works aim to predict sentiment levels for subjective texts in different granularities, such as words (Song et al., 2016), sentences (Ma et al., 2017), short texts (Song et al., 2015) and documents (Yang et al., 2018). More recently, mainstream research direction concentrates on turn-level and dialogue-level CSP.

Some researchers explore the turn-level structure, such as modeling dialogues via a hierarchical RNN (Cerisara et al., 2018), keeping track of satisfaction states of dialogue participants (Majumder et al., 2019), exploring contrastive learning (Toutanova et al., 2021) and so on. But, due to the labels of turn-level satisfaction is difficult to obtain and dialogue-level CSP appears to re-

flect service quality more realistically, we focus on dialogue-level CSP in this paper.

To study the dialogue-level CSP, earlier methods used manual features (Hakkani-Tür and Ostendorf, 2010; Gangemi et al., 2015), while recent studies prefer deep neural networks and attention mechanism to explore how questions and answers interact with each other. Some researchers adopt a Bi-directional LSTM network to capture the contextual information of conversational services and use the hidden vector of the last utterance for satisfaction prediction (Hashemi et al., 2018), some researchers uses each question to capture information from all answers to model customer-service interaction (Song et al., 2019a), while another study focusing on dialogue-level CSP uses LSTM networks to capture contextual features and computes the semantic similarity scores between customer questions and customer service answers across different turns to model customer-service interaction (Yao et al., 2020). However, these works didn't consider about the differences of expressions between customer and customer service. Morever, they failed to excavate the internal relations of personalized expression sequences. In this work, we work on addressing the two existing issues, thus proposing a two-stage classification model for dialogue-level CSP.

3 Methodology

3.1 Problem Definition

In the real scenario, customers ask questions and chatbots will provide the corresponding answers in turn, so the dialogue content is defined as a sequence of utterances $C = \{q_1, a_1, q_2, a_2, ..., q_n, a_n\}$. Each question q_i is followed by an answer a_i, and the length of conversation is $2n$. The goal of our task is to predict the satisfaction level y based on dialogue content C, while the satisfaction level is divided into five classes: *strongly satisfied, satisfied, neutral, dissatisfied, strongly dissatisfied*.

3.2 Proposed Model

As shown in Figure 3, we propose a two-stage classification model for dialogue-level CSP. Our model consists of four sub-modules: session-encoder, intra-stage, inter-stage and session-decoder. The session-encoder is a dialogue encoding module to process the raw conversation content. Intra-stage is comprised of customer part and service part, which helps extract sufficient internal features of question sequences and answer sequences separately. For both parts in inter-stage, we utilize attention mechanisms to adequately discover the sentence characteristics at each time step from their personalized context, served as *customer state* and *service state*. Next, inter-stage applies an interactive attention mechanism to fully capture satisfaction representations of the whole dialogue from *customer state* and *service state*. Finally, the session-decoder contributes to predict the final satisfaction classification. In the following sections, we will introduce the details of the model structure in order.

3.3 Session-encoder

Session-encoder aims to encode natural language dialogues into semantic representations. Our input is the whole dialogue text, in which words are separately transformed into 300 dimensional vectors by using pre-trained GloVe model (Pennington et al., 2014).

$$E = \text{GloVe}\,(C) \tag{1}$$

Then, inspired by previous study (Kim, 2014), we leverage a CNN layer with max-pooling to extract context independent features of each utterance. Concretely, we apply three filters of size 1,2,3 with 50 feature maps each, and employ ReLU activation (Nair and Hinton, 2010) and max-pooling to deal with these feature maps.

$$fm_{1,2,3} = \text{ReLU}\left(\text{CNN}_{1,2,3}\left(E\right)\right) \tag{2}$$

$$fm'_{1,2,3} = \text{max-pooling}\left(fm_{1,2,3}\right) \tag{3}$$

Then, we concatenate these features and send them into a fully connected layer, which produces the context representations cr as follow.

$$fm' = \text{concat}\left(fm'_{1,2,3}\right) \tag{4}$$

$$cr = \text{ReLU}\left(W_0 fm' + b_0\right) \tag{5}$$

3.4 Intra-stage

Intra-stage is a core module of our two-stage model, which consists of the customer part and service part. We can alternately divide cr into question representations $qr = \{qr_1, qr_2, ..., qr_n\}$ and answer representations $ar = \{ar_1, ar_2, ..., ar_n\}$ as the input of customer part and service part. In the following, we will illustrate how these two parts of intra-stage adequately exploit the inside relations of their own utterance sequences.

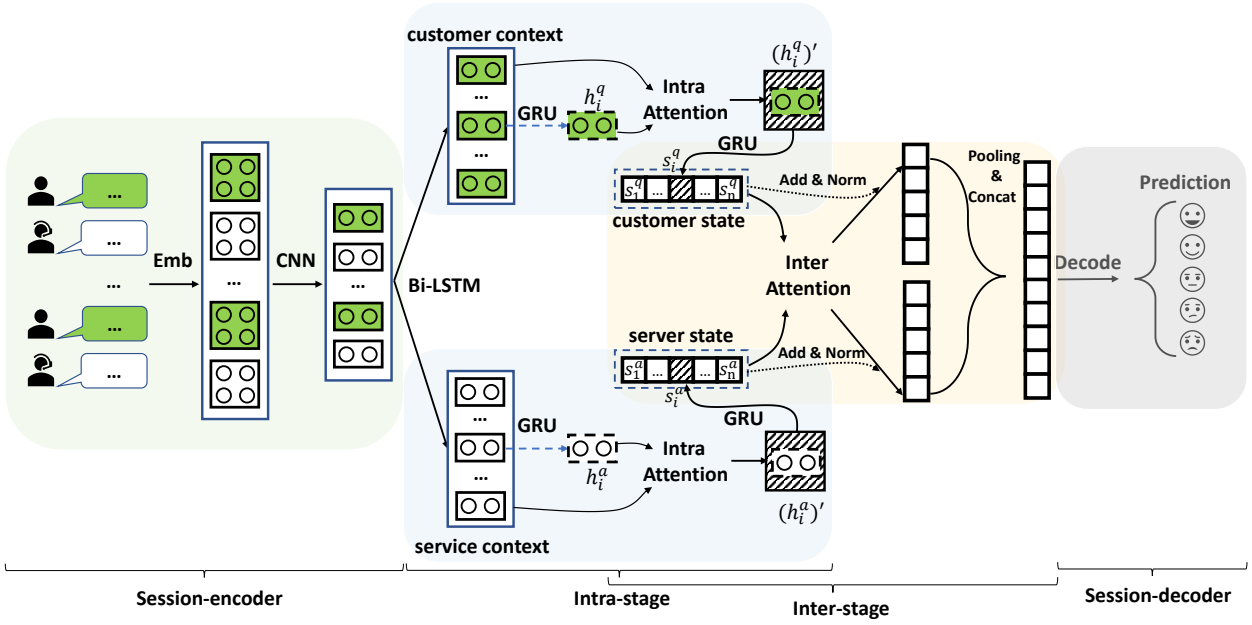

Figure 3: Framework of the two-stage classification model for dialogue-level CSP.

3.4.1 customer part

LSTM has a special unit called memory cell, which is similar to an accumulator or a gated neuron. We adapt a Bi-directional LSTM to capture long-term dependencies of qr,

$$m_i^q = \mathrm{BiLSTM}^{\mathrm{q}}\left(m_{i\pm1}^q, qr_i\right) \tag{6}$$

where $i = 1, 2, ..., n$. m_i is the output of Bi-directional LSTM at time step i, the whole context representation of question sequence is $m^q = \{m_1^q, m_2^q, ..., m_n^q\}$.

To better explore the internal relations of question sequence, we capture the satisfaction representation of each time step iteratively by adequately interacting current features with context information. Firstly, an GRU encoder is used to process the sequence,

$$h_i^q = \mathrm{GRU}_{\mathrm{encode}}^{\mathrm{q}}\left(m_i^q, h_{i-1}^q\right) \tag{7}$$

where $h^q = \{h_1^q, h_2^q, ..., h_n^q\}$, h^q is the hidden state of GRU. Secondly, we use an attention mechanism to match h_i^q with the masked personalized context,

$$masked\left(m_j^q\right) = \begin{cases} m_j^q, & j \in \{1, 2, ..., i\} \\ 0, & \text{Otherwise} \end{cases} \tag{8}$$

$$q, k, v = h_i^q, masked\left(m^q\right), masked\left(m^q\right) \tag{9}$$

$$h_i^{q\prime} = \mathrm{IntraAtt}^{\mathrm{q}}\left(q, k, v\right) \tag{10}$$

where $h^{q\prime} = \{h_1^{q\prime}, h_2^{q\prime}, ..., h_n^{q\prime}\}$, $h^{q\prime}$ is the result of this attention layer.

Up to now, we have adequately obtained the internal relations of question sequence. Then, a GRU is used to decode the result from the intra attention layer,

$$s_i^q = \mathrm{GRU}_{\mathrm{decode}}^{\mathrm{q}}\left(h_i^{q\prime}, s_{i-1}^q\right) \tag{11}$$

$s^q = \{s_1^q, s_2^q, ..., s_n^q\}$, where s^q is *customer state* after the complete process of customer part.

3.4.2 service part

Service part is the other part in intra-stage, which contributes to the satisfaction state of service. The whole structure of service part is similiar to customer part,

$$m_i^a = \mathrm{BiLSTM}^{\mathrm{a}}\left(m_{i\pm1}^a, ar_i\right) \tag{12}$$

$$h_i^a = \mathrm{GRU}_{\mathrm{encode}}^{\mathrm{a}}\left(m_i^a, h_{i-1}^a\right) \tag{13}$$

$$masked\left(m_j^a\right) = \begin{cases} m_j^a, & j \in \{1, 2, ..., i\} \\ 0, & \text{Otherwise} \end{cases} \tag{14}$$

$$q, k, v = h_i^a, masked\left(m^a\right), masked\left(m^a\right) \tag{15}$$

$$h_i^{a\prime} = \mathrm{IntraAtt}^{\mathrm{a}}\left(q, k, v\right) \tag{16}$$

$$s_i^a = \mathrm{GRU}_{\mathrm{decode}}^{\mathrm{a}}\left(h_i^{a\prime}, s_{i-1}^a\right) \tag{17}$$

where s^a is the *service state*.

3.5 Inter-stage

Inter-stage aims to fully interact s^q with s^a. Zhou et al. (2018) utilize attention mechanisms to capture the most relevant information and construct interaction between two sequences on natural language processing tasks. Inspired by this work, we use an attention mechanism to interact s^q with s^a.

$$\tilde{s}^q = \mathrm{InterAtt}^{\mathrm{q}}\left(s^q, s^a, s^a\right) \tag{18}$$

$$\tilde{s}^a = \mathrm{InterAtt}^{\mathrm{a}}\left(s^a, s^q, s^q\right) \tag{19}$$

In order to make the learning process smoother, we adopt a layer of add & normalization (Vaswani et al., 2017).

$$\tilde{s}^{q'} = \mathrm{Normalization}\left(\mathrm{Add}\left(\tilde{s}^q, s^q\right)\right) \tag{20}$$

$$\tilde{s}^{a'} = \mathrm{Normalization}\left(\mathrm{Add}\left(\tilde{s}^a, s^a\right)\right) \tag{21}$$

In the end of the inter-stage, by using average pooling, we transform $\tilde{s}^{q'}$ and $\tilde{s}^{a'}$ into vectors and concatenate them together as follow.

$$s = \mathrm{concat}\left(\mathrm{pooling}\left(\tilde{s}^{q'}\right), \mathrm{pooling}\left(\tilde{s}^{a'}\right)\right) \tag{22}$$

where s is the final satisfaction representation of the whole dialogue.

3.6 Session-decoder

Session-decoder module is used to decode the satisfaction state s to predict the customer satisfaction. We use two layers of fully connected network, $\hat{y}$ is the prediction of satisfaction level.

$$H = \mathrm{ReLU}\left(W_1 s + b_1\right) \tag{23}$$

$$P = \mathrm{softmax}\left(W_2 H + b_2\right) \tag{24}$$

$$\hat{y} = \underset{k}{\mathrm{argmax}}\left(P[k]\right) \tag{25}$$

As for the loss function, we choose cross-entropy:

$$\mathcal{L}(\theta) = -\sum_{v \in y_V} \sum_{z=1}^{Z} Y_{vz} \ln P_{vz} \tag{26}$$

where $y_{\mathcal{V}}$ is the set of dialogue indexes that have real labels. Y is the label indicator matrix, and θ is the collection of trainable parameters in this two-stage classification model.

4 Experimental Settings

This section mainly introduces datasets, hyper parameters and baselines used in our experiments.

4.1 Datasets

We evaluate our two-stage classification model on our in-house dataset *(Five-classification task)* and four released public datasets *(Three-classification task)*.

4.1.1 CECSP

This is our in-house Chinese E-commerce CSP dataset collected from one of the largest E-commerce platforms. We use real customer feedback as the dialogue-level satisfaction labels which include *strongly satisfied*, *satisfied*, *neutral*, *dissatisfied* and *strongly dissatisfied*.

4.1.2 Clothes & Makeup

These are two CSP datasets in clothes and makeup domain collected from a top E-commerce platform (Song et al., 2019b). Each dialogue is annotated as one of the three satisfaction classes: *satisfied*, *neutral* and *dissatisfied*.

4.1.3 MELD

This is a multi-party conversation corpus collected from the TV show Friends (Poria et al., 2019). Each utterance is annotated as one of the three sentiment classes: *negative*, *neutral* and *positive*. While *negative* and *positive* are considered as *dissatisfied* and *satisfied* respectively, *neutral* is kept unchanged.

4.1.4 EmoryNLP

This is also a multi-party conversation corpus collected from Friends, but varies from MELD in the choice of scenes and emotion labels (Zahiri and Choi, 2018). The emotion labels include *neutral*, *joyful*, *peaceful*, *powerful*, *scared*, *mad* and *sad*. To create three satisfaction classes: *joyful*, *peaceful* and *powerful* are grouped together to form the *satisfied* class; *scared*, *mad* and *sad* are grouped together to form the *dissatisfied* class; and *neutral* is kept unchanged.

4.1.5 Transforming rules for MELD & EmoryNLP

Original MELD and EmoryNLP are two released conversational emotion recognition (CER) datasets. We transform them into the conversational service scenario following three rules: (1) We consider the first speaker of a dialogue as the customer and

Datasets	Train	Val	Test	Avg-turns
CECSP	22576	2822	2801	3.67
Clothes	8000	1000	1000	8.14
Makeup	2832	354	354	8.01
MELD	1037	113	279	3.19
EmoryNLP	685	88	78	3.86

Table 1: The statistics of the five datasets. While **CECSP** is our constructed Chinese E-commerce CSP dataset, **Clothes** and **Makeup** are two released corpora in different domains. **MELD** and **EmoryNLP** are two CER datasets.

map all the emotion labels of his/her utterances into a new dialogue-level satisfaction label; (2) We concatenate consecutive utterances from the same person as a long utterance; (3) If a dialogue is ended by the first speaker, we extra use utterance "NULL" as the answer of the last turn.

4.2 Hyper parameters

The batch sizes are set to be {256,64,64,64,64} for CECSP, Clothes, Makeup, MELD and EmoryNLP. We adopt Adam (Kingma and Ba, 2015) as the optimizer with initial learning rates of {1e-3,1e-4,1e-4,1e-4,1e-4} and L2 weight decay rates of {1e-4, 1e-5, 1e-5, 1e-5, 1e-5}, respectively. The dropout is set to be 0.5 (Srivastava et al., 2014). We train all models for a maximum of 200 epochs and stop training if the validation loss does not decrease for 30 consecutive epochs. The total number of parameters in this model is 59.84 million. We use a piece of Tesla P40 24GB. Each epoch of these experiments costs around 400 seconds.

4.3 Baselines

We compared our model with the following baselines in our experiments:

- **LSTMCSP** (Hashemi et al., 2018):This model uses a Bi-directional LSTM network to capture the user's intent and identify user's satisfaction.
- **CMN** (Hazarika et al., 2018):It is an end-to-end memory network which updates contextual memories in a multi-hop fashion for conversational emotion recognition.
- **DialogueGCN** (Ghosal et al., 2019): It is a graph-based approach which leverages inter-speakers' dependency of the interlocutors to model conversational context for emotion recognition.
- **CAMIL** (Song et al., 2019a): This Context-Assisted Multiple Instance Learning model predicts the sentiments of all the customer utterances and then aggregates those sentiments into service satisfaction polarity.
- **LSTM-Cross** (Yao et al., 2020):This model uses LSTM networks to capture contextual features .Then, these features are concatenated with the cross matching scores to predict the satisfaction.
- **DialogueDAG** (Shen et al., 2021):This model uses directed graphs to collect nearby and distant historical informative cues. We aggregate the node representations to capture dialogue-level representations for CSP.

5 Results and Analysis

5.1 Overall Results

The overall results of all the models on five datasets are shown in Table 2. We can learn from the results that our proposed two-stage model achieves better performance than all the baselines on five datasets.

LSTMCSP, CMN, and DialogueGCN achieve similar performance on CECSP, MELD and EmoryNLP. CMN is capable of capturing the emotional cues in context, thus achieving better F1 scores than LSTMCSP on Clothes and Makeup. However, chatbot answers are always neutral in conversational service, which narrow the gap between CMN and LSTMCSP on CECSP. In our scene, customer questions and chatbot answers are alternating, so the related positions between them cannot provide additional information. Thus, the position method in DialogueGCN does not have better performance here.

CAMIL takes turn-level sentiment information into account and outperforms previous strategies on four datasets except MELD. Due to the customer-service interaction modeling method, LSTM-Cross has made further improvement on all datasets, which implies the importance of interactions in single turn. DiologueDAG uses graphical structure to effectively collect nearby and distant information, so it performs well on datasets with shorter average turns, such as MELD and EmoryNLP. But when the average turns become longer, it doesn't work well.

Model	CECSP		Clothes		Makeup		MELD		EmoryNLP	
	Acc.	F1	Acc.	F1	Acc.	F1	Acc.	F1	Acc.	F1
LSTMCSP	51.55	50.10	75.59	75.78	76.31	76.56	42.29	43.08	50.01	47.56
CMN	52.09	50.32	78.5	78.1	81.07	80.88	45.52	44.08	52.56	48.52
DialogueGCN	52.69	50.25	76.89	76.82	77.72	77.78	46.39	44.99	52.72	48.78
CAMIL	55.43	52.92	78.30$^{\#}$	78.40	78.50$^{\#}$	78.64	44.44	39.02	55.13	49.52
LSTM-Cross	55.51	53.11	78.91	79.33	79.88	79.58	48.03	47.28	55.28	51.00
DialogueDAG	55.12	51.97	75.4	75.04	73.73	73.73	48.03	47.28	59.26	54.82
Two-stage Model	**57.34**	**54.69**	**81.2**	**80.71**	**82.2**	**82.07**	**50.9**	**50.35**	**61.54**	**57.88**

Table 2: Overall performance on the five datasets. We use the accuracy and the weighted F1 score to evaluate each model. Scores marked by "#" are reported results in authors' paper, while others are based on our re-implementation.

Method	Weighted F1 score	
	CECSP	Clothes
Two-stage model	**54.69**	**80.71**
- Inter-stage	53.68(↓ 1.01)	78.64(↓ 2.07)
- Intra-stage	54.07(↓ 0.62)	78.23(↓ 2.48)
- Intra-stage & Inter-stage	53.53(↓ 1.16)	78.44(↓ 2.27)

Table 3: Results of ablation study on the two representative datasets.

Our proposed two-stage model reaches the new state of the art on all datasets. On the one hand, intra-stage extracts internal correlation features of question sequence and answer sequence in customer part and service part separately. Using attention mechanism with the personalized context of both sequences makes feature extraction sufficient at each time step. On the other hand, we think each customer question is not only associated with the answer behind, but also the answers in other turns, so inter-stage conducts fully interaction between *customer state* and *service state*, which is different from the turn-level approaches in earlier researches. As the result, our model has improved by at least 1%∼3% on F1 score over five datasets, compared with baseline works.

5.2 Ablation Study

To study the impact of the modules in our two-stage model, we evaluate it by removing (1) inter-stage (2) intra-stage (3) intra-stage and inter-stage together. Removing the inter-stage means the we only retain the intra-stage, while removing the intra-stage means only the inter-stage remains. Removing both intra-stage & inter-stage means we no longer separate the dialogue and only retain the customer part of the intra-stage (the whole dialogue as input). We use CECSP and Clothes as the representatives in this study because they are larger datasets with short and long average turns. The results are shown in Table 3.

Here are two sets of comparative experiments. Firstly, let's pay attention to the comparison of the first three rows. Without inter-stage, the weighted F1 score drops by 1.01% on CECSP and 2.07% on Clothes. Without intra-stage, the weighted F1 score drops by 0.62% on CECSP and 2.48% on Clothes. The results imply the importance of both two stages, none of them can be removed.

Secondly, the second and fourth rows of experiments illustrate the advantage of intra attention. Both of them don't have inter-stage, and the only difference between them is whether to split the dialogue into question sequence and answer sequence. As shown in the table, the weighted F1 score drops by 0.15% and 0.20% if we don't apply intra method. Thus we can draw a conclusion, the intra method helps extract the internal correlation of customer context and service context respectively, and indeed improves the performance of our model.

In conclusion, the ablation study proves that both intra-stage and inter-stage play important roles. In particular, the intra method of separating context representations into questions and answers contributes to the improvement of our model.

5.3 Case Analysis

In order to better understand the advantages of our proposed two-stage model, we analyse the case in figure 1. The result shows in figure 4. The heatmap is used to represent the values of attention weights, while darker colors mean larger weights.

Part A illustrates the feature extraction process of a conventional model. As customer's expression contains richer information ("when", "urge",

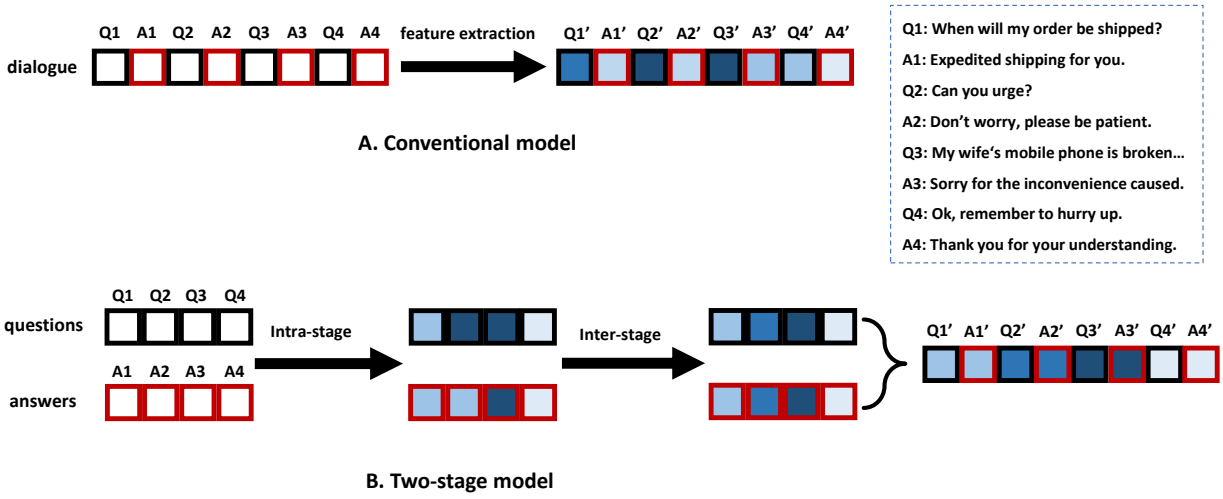

Figure 4: Results of case analysis. Part A represents the feature extraction process of usual model, while Part B represents our two-stage model. The colors of heatmap show the values of attention weights.

"broken") of his problem and emotion, the model will pay more attention to Q1, Q2 and Q3. So, it is likely to ignore the importance of answers, which are critical to deciding whether these questions are solved, thus affecting customer's satisfaction deeply too.

By contrast, Part B illustrates our two-stage model. The dialogue is split into customer questions and chatbot answers, so the model can better learn the inside relations of the two sequences separately in the intra-stage, which ensures the expressions of customers would not attract much more attention than chatbots. In this case, the customer expresses his anxiety and tells the mobile phone is broken in Q2 and Q3, so the weights of those two are larger in the question sequence. Similarly, A3 have larger weights in the answer sequence due to its obvious soothing expression. Then, the inter-stage conducts the interaction to adjust the attention weights of the two parts. In the end, we concatenate two parts and find Q2,A2,Q3,A3 are important utterances of this dialogue. In this dialogue situation, although the customer mainly shows his bad emotion and unsolved problem in Q2 and Q3, the chatbot appeases him in A2 and A3, which leads to a satisfied result. The result of part B appears to be more reasonable.

By comparing the two results, we find the intra-stage of our two-stage model can balance the expression differences of customer questions and chatbot answers, while the conventional model pays more attention to customer questions. What's more, the inter-stage interacts *customer state* with *service state* to adjust the weights of attention, which can help capture the characteristics of dialogue more smoothly.

6 Conclusion

In this paper, we propose a two-stage model for dialogue-level CSP task. We first introduce an intra-stage to discover the relations inside customer part and service part respectively, in which an attention mechanism with masked personalized context is used to fully capture the *customer state* and *service state*. Then, we use an inter attention mechanism to combine those two states in inter-stage and predict the customer satisfaction of the whole dialogue. Experimental results on our in-house dataset and four public datasets indicate our model outperforms all the baseline models on the dialogue-level CSP task.

In the future work, we will further improve our two-stage model by constructing more targeted structures. For example, we can make differentiated design on customer part and service part in intra-stage. Moreover, we will study other interesting tasks in customer service dialogues, such as good case mining.

References

Christophe Cerisara, Somayeh Jafaritazehjani, Adedayo Oluokun, and Hoa T. Le. 2018. Multi-task dialog act and sentiment recognition on mastodon. In *Proceedings of the 27th International Conference on Computational Linguistics, COLING 2018*, pages 745–754.

Aldo Gangemi, Stefano Leonardi, and Alessandro Panconesi, editors. 2015. *Proceedings of the 24th International Conference on World Wide Web, WWW 2015*.

Deepanway Ghosal, Navonil Majumder, Soujanya Poria, Niyati Chhaya, and Alexander F. Gelbukh. 2019. Dialoguegcn: A graph convolutional neural network for emotion recognition in conversation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019*, pages 154–164.

Dilek Hakkani-Tür and Mari Ostendorf, editors. 2010. *2010 IEEE Spoken Language Technology Workshop, SLT 2010*.

Seyyed Hadi Hashemi, Kyle Williams, Ahmed El Kholy, Imed Zitouni, and Paul A. Crook. 2018. Measuring user satisfaction on smart speaker intelligent assistants using intent sensitive query embeddings. In *Proceedings of the 27th ACM International Conference on Information and Knowledge Management, CIKM 2018*, pages 1183–1192.

Devamanyu Hazarika, Soujanya Poria, Amir Zadeh, Erik Cambria, Louis-Philippe Morency, and Roger Zimmermann. 2018. Conversational memory network for emotion recognition in dyadic dialogue videos. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2018*, pages 2122–2132.

Yoon Kim. 2014. Convolutional neural networks for sentence classification. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*, pages 1746–1751.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations, ICLR 2015*.

Dehong Ma, Sujian Li, Xiaodong Zhang, and Houfeng Wang. 2017. Interactive attention networks for aspect-level sentiment classification. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI 2017*, pages 4068–4074.

Navonil Majumder, Soujanya Poria, Devamanyu Hazarika, Rada Mihalcea, Alexander F. Gelbukh, and Erik Cambria. 2019. Dialoguernn: An attentive RNN for emotion detection in conversations. In *The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI 2019, The Thirty-First Innovative Applications of Artificial Intelligence Conference, IAAI 2019, The Ninth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019, Honolulu, Hawaii, USA, January 27 - February 1, 2019*, pages 6818–6825. AAAI Press.

Vinod Nair and Geoffrey E. Hinton. 2010. Rectified linear units improve restricted boltzmann machines. In *Proceedings of the 27th International Conference on Machine Learning (ICML-10)*, pages 807–814.

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP 2014*, pages 1532–1543.

Soujanya Poria, Devamanyu Hazarika, Navonil Majumder, Gautam Naik, Erik Cambria, and Rada Mihalcea. 2019. MELD: A multimodal multi-party dataset for emotion recognition in conversations. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019*, pages 527–536.

Weizhou Shen, Siyue Wu, Yunyi Yang, and Xiaojun Quan. 2021. Directed acyclic graph network for conversational emotion recognition. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP 2021*, pages 1551–1560.

Kaisong Song, Lidong Bing, Wei Gao, Jun Lin, Lujun Zhao, Jiancheng Wang, Changlong Sun, Xiaozhong Liu, and Qi Zhang. 2019a. Using customer service dialogues for satisfaction analysis with context-assisted multiple instance learning. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019*, pages 198–207.

Kaisong Song, Lidong Bing, Wei Gao, Jun Lin, Lujun Zhao, Jiancheng Wang, Changlong Sun, Xiaozhong Liu, and Qiong Zhang. 2019b. Using customer service dialogues for satisfaction analysis with context-assisted multiple instance learning. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 198–207.

Kaisong Song, Shi Feng, Wei Gao, Daling Wang, Ge Yu, and Kam-Fai Wong. 2015. Personalized sentiment classification based on latent individuality of microblog users. In *Proceedings of the Twenty-Fourth International Joint Conference on Artificial Intelligence, IJCAI 2015*, pages 2277–2283.

Kaisong Song, Wei Gao, Ling Chen, Shi Feng, Daling Wang, and Chengqi Zhang. 2016. Build emotion lexicon from the mood of crowd via topic-assisted joint

non-negative matrix factorization. In *Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval, SIGIR 2016*, pages 773–776.

Nitish Srivastava, Geoffrey E. Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. 2014. Dropout: a simple way to prevent neural networks from overfitting. *J. Mach. Learn. Res.*, 15(1):1929–1958.

Kristina Toutanova, Anna Rumshisky, Luke Zettlemoyer, Dilek Hakkani-Tur, Iz Beltagy, Steven Bethard, Ryan Cotterell, Tanmoy Chakraborty, and Yichao Zhou, editors. 2021. *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics, Online.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017*, pages 5998–6008.

Jun Yang, Runqi Yang, Chongjun Wang, and Junyuan Xie. 2018. Multi-entity aspect-based sentiment analysis with context, entity and aspect memory. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18)*, pages 6029–6036.

Riheng Yao, Shuangyong Song, Qiudan Li, Chao Wang, Huan Chen, Haiqing Chen, and Daniel Dajun Zeng. 2020. Session-level user satisfaction prediction for customer service chatbot in e-commerce (student abstract). In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020*, pages 13973–13974.

Sayyed M. Zahiri and Jinho D. Choi. 2018. Emotion detection on TV show transcripts with sequence-based convolutional neural networks. In *The Workshops of the The Thirty-Second AAAI Conference on Artificial Intelligence*, pages 44–52.

Xiangyang Zhou, Lu Li, Daxiang Dong, Yi Liu, Ying Chen, Wayne Xin Zhao, Dianhai Yu, and Hua Wu. 2018. Multi-turn response selection for chatbots with deep attention matching network. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, ACL 2018*, pages 1118–1127.