Journal of Combinatorial Mathematics
and Combinatorial Computing
www.combinatorialpress.com/jcmcc

Technical research on real-time multimodal interaction architecture based on AI big language model for dynamic adjustment of semantic features in intelligent customer service robots

Haizhi Wei[1,✉]

[1] *Kexun Jialian Information Technology Co., Ltd, Hefei, Anhui, 230000, China*

ABSTRACT

The service efficiency of intelligent customer service robots affects the service operation efficiency of enterprises and plays an important role in maintaining customer resources. This paper applies multimodal interaction technology to intelligent customer service system, takes multimodal big language model Qwen-VL as the core, proposes a two-stage relationship multimodal relationship extraction framework based on big language model, realizes multimodal relationship extraction with the help of high-quality auxiliary knowledge, integrates dynamic semantic features and static structural features to complete the multimodal emotion polarity prediction, and constructs multimodal retrieval Q&A system to improve the performance of smart robot performance. Applying the intelligent customer service system in this paper for service practice, the conversation between the intelligent customer service robot and the customer usually ends in about 50 rounds, and the service efficiency is relatively efficient. In the face of customer emotional sentences labeled as happy, complaining and angry, the recognition accuracy under multimodal sentiment analysis is greater than 99%, and the behavior of "notification" and "confirmation" service behavior accounts for the largest proportion of behaviors, and the number of behaviors reaches 560,365 times, 365976 times, which is in line with the expected service behavior of intelligent customer service robots.

Keywords: multimodal interaction technology, multimodal language model, large language model, auxiliary knowledge, intelligent customer service

1. Introduction

With the rise of the digital economy, the digital and intelligent transformation in the field of customer service has become one of the important entry points for Chinese industries to carry out the transformation of the landing work. Intelligent customer service is an application technology in the

✉ Corresponding author.
E-mail address: hzweixf@163.com (H. Wei).
Received 01 February 2024; Revised 20 April 2024; Accepted 10 November 2024; Published Online 15 April 2025.
DOI: 10.61091/jcmcc127a-219

field of natural language processing, human-computer interaction under the rapid development of computers and the Internet, which is realized on the basis of large-scale knowledge processing [1-3]. Compared with traditional customer service, intelligent customer service uses natural language processing, artificial intelligence, speech recognition and other technologies to improve the processing capacity of text and language, and has outstanding advantages in access channels, response efficiency, data management and analysis, which greatly improves the efficiency of customer service [4-7].

Semantic matching is a natural language processing technology, which can analyze and judge the content of text or statements by calculating the similarity of text or statement vector features [8-9]. Due to the Chinese language diversity of expression and the rich expression of linguistic meanings, it leads to a more difficult word separation between words and the latter utterances, with an extremely large number of similar semantic meanings, and the difficulty of distinguishing the boundaries between features [10-13]. In addition to this, the capture of semantic information at this stage only considers the linear extraction of text sequences, and lacks the deep interactive fusion of dynamic semantics and static structure, resulting in the inability to accurately model the relationship between aspectual words and viewpoint words [14-16]. Therefore, on the basis of having deep semantic matching and analysis, research on semantic feature dynamic adjustment technology, so that intelligent customer service has the ability to analyze fine-grained utterances and provide users with better intelligent answering services.

Literature [17] shows that the electric power intelligent customer service system is able to deeply understand the user's intention and semantics through natural language processing and artificial intelligence technology in order to accurately identify the user's problems and needs, and provide relevant feedback through a preset knowledge base to realize intelligent service. Literature [18] proposes a multi-semantic relationship knowledge graph construction model TransCD, which divides relationships into multiple semantics and calculates their deterministic weights, establishes a spatial projection model of entity relationships through dynamic mapping matrices, realizes the construction of multi-semantic knowledge graph, and improves the performance of intelligent customer service in dealing with complex relationships. Literature [19] proposed a short text-specific semantic mining method by combining fuzzy c-mean clustering algorithm and immune monophilic genetic algorithm for enhancing the understanding of contextual information by intelligent customer service robots, which maintains a reasonable semantic segmentation performance regardless of the high or low level of short text-specific semantic transfer. Literature [20] utilizes icsBERTs algorithm to achieve automatic word extraction and multi-task training for intelligent customer service business, and optimizes the text processing capability of intelligent customer service language model by improving the performance of the pre-trained language model. Literature [21] constructs an intelligent customer service system supported by a fuzzy control system, enriches the expertise of the intelligent customer service robot by establishing a domain ontology library, and introduces an interrogative semantic vector model on the basis of its dialog structure, which significantly improves the accuracy of the customer service robot's answer feedback. Literature [22] optimizes the round-by-round interaction scheme of an intelligent chatbot and proposes a multi-round response triggering model (MRTM), which relies on semantic matching relations with contextual information for language model training and combines the self-attention mechanism to judge whether an utterance is responsive or not, so that the chatbot is able to make a well-timed response after aggregating multiple discourse information.

On the dynamic adjustment of semantic features of the intelligent customer service robot, this study chooses to take the multimodal interaction technology as the research breakthrough point, and builds an intelligent customer service service system by combining the multimodal interaction technology to improve the service performance of the intelligent customer service robot from the aspects of multimodal language, multimodal relationship extraction, and multimodal sentiment analysis, and builds the multimodal retrieval Q&A system. Taking the multimodal language model Qwen-VL as the basis, a two-stage relationship extraction framework PLFM based on the multimodal large language

model is proposed to guide the large language model to generate high-quality auxiliary knowledge, and the obtained auxiliary knowledge is inputted into the downstream text model together with the original text to realize the final multimodal relationship extraction. A multimodal sentiment analysis model that integrates dynamic semantic and static structural features is proposed. Dynamic semantic features are obtained by dynamically adjusting the weight adapter, while static structural features are obtained by using the constituent graph attention module and the dependency graph attention module, and finally, the dynamic semantic and static structural features are fused to obtain the deep-level text features. Combined with the multimodal retrieval enhanced Q&A dataset and the FEC-MRQA model, the multimodal retrieval Q&A system is built to provide technical support for intelligent customer service robots and even intelligent customer service systems. Simulation experiments are conducted to test the performance of each multimodal module in the intelligent customer service system in this paper. At the same time, using the TCSIC dataset, the potential and value of this system in practical application scenarios are discussed in depth.

2. Intelligent customer service system based on multimodal interaction

Intelligent customer service robots are more and more widely used in enterprises to meet the function of automatic response to customer messages. Customer resources is an important resource for enterprises to maintain sustainable development, the operation strategy to fully consider the customer, can enhance the management efficiency of enterprises, retaining enterprise resources. From the customer's point of view, differentiated services can be realized to enhance the competitiveness of the enterprise. In this paper, we will design an intelligent customer service system based on multi-modal interaction architecture to improve the defects and problems such as the cumbersome operation of customer service system and the low efficiency of intelligent customer service robot's service ability in answering questions, which are common nowadays.

2.1. Overall system architecture

Multi-modal interaction technology is applied to the intelligent customer service system, and the overall structure of the designed intelligent customer service system is shown in Figure 1. From Figure 1, the designed intelligent customer service system includes network hardware layer, foundation platform layer, data layer, intelligent service layer and function layer, the system is supported by the Web server using HTTP protocol and MSN protocol to provide network support, and the foundation platform layer utilizes agent's multi-modal interaction middleware to provide the system with operation support mechanism. System functions include statistical analysis and role management. The system utilizes configuration components, logging components, security components, etc. to provide intelligent guidance for the system, and the SQL database is selected to provide data support for the system.

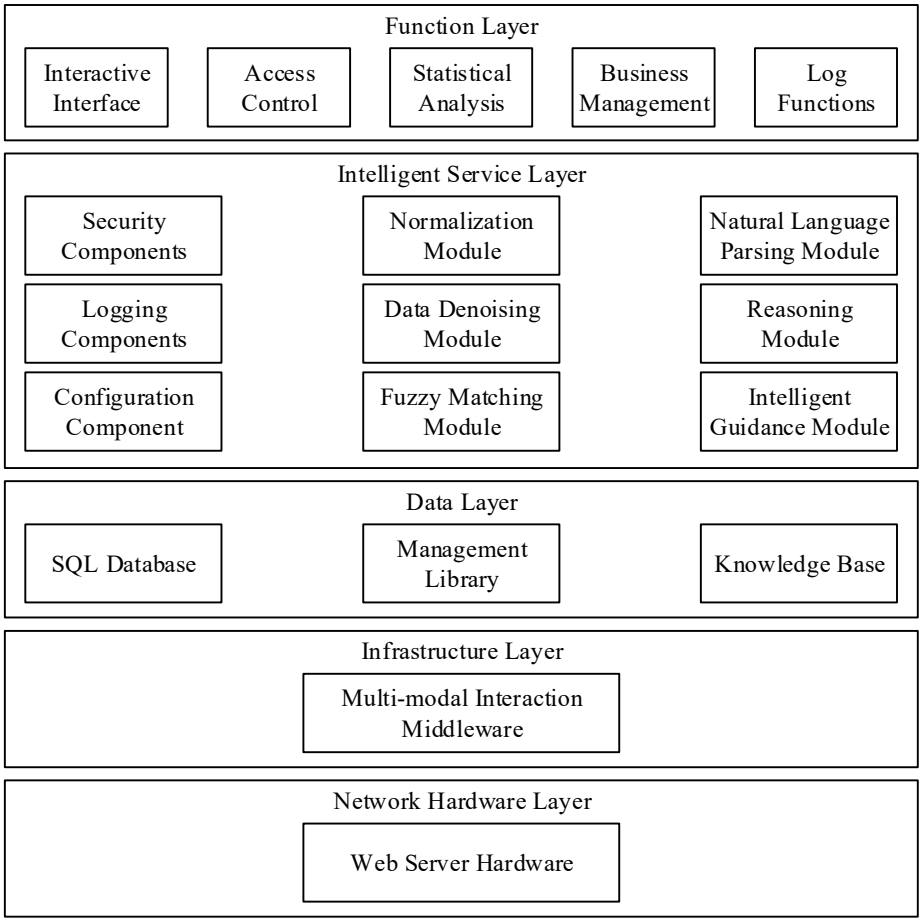


Fig. 1. System structure

2.2. Multimodal AI Large Language Modeling

The multimodal language model Qwen-VL is an advanced multimodal big language model carefully developed by Aliyun, which integrates multiple input modes such as text, image and detection frames, and is capable of generating text and locating detection frames as outputs. The core construction of the model is based on the strong foundation of Qwen-7B pre-trained language model and Openclip ViT-bigG visual coder, which is co-trained to improve cross-modal comprehension by incorporating a single-layer random initialization cross-attention mechanism between the two and utilizing graphic and textual data with a size of about 1.5B. The following are its key features and optimization innovations.

1) Superior performance: Qwen-VL achieves the best performance with the same number of parameters on four mainstream English multimodal tasks, which demonstrates its powerful multimodal comprehension and generation capabilities.

2) Multi-image interactive dialog. The model is capable of adapting and processing successive inputs from multiple images, performing comparative analysis, executing specified picture quizzes, and supporting complex scenarios such as literary creation based on multiple images.

3) The first generic model for open domain localization in Chinese. Qwen-VL is the first general-purpose model that supports annotating the location of a target region in an image via a Chinese open-domain language representation.

In a series of multimodal downstream tasks, Qwen-VL has demonstrated exceptional application effectiveness and strong comprehension and generation potential. Especially in complex situations where image and text information are highly complementary, Qwen-VL is able to efficiently fuse the information content of the two modalities to provide more accurate and in-depth analysis conclusions.

As the technology iterates and continues to be optimized, Qwen-VL will undoubtedly play an even more critical role in future multimodal AI research and applications, and continue to push the frontiers of the field.

2.3. Multimodal Relational Extraction Models

In this paper, we propose a two-stage framework that aims to enhance the performance of multimodal relation extraction by prompting a large language model to generate high-quality auxiliary knowledge, called PLFM.The PLFM model is mainly divided into two key phases, which are the phase of high-quality auxiliary knowledge generation and the phase of relation prediction based on high-quality auxiliary knowledge. In the first phase, a multimodal similarity detection module is first employed to select relevant examples from a set of manually predefined samples for each sample in the dataset. These selected examples are then integrated into format-specific multimodal cue templates that are used to inspire the Big Language Model to generate high-quality auxiliary knowledge. In the second stage, the raw text is spliced together with the auxiliary knowledge generated by the grand language model. This spliced input is then fed into the Transformer encoder for generating fusion features. Finally, these fused features are fed into the Entity Relationship Prediction layer to predict the relationship between two entities in the original text.

2.3.1. Context learning in multimodal relation extraction. Big language models are pre-trained autoregressive language models based on massive datasets. In the inference process, new downstream tasks are accomplished on the pre-trained large language model by context less sample learning in the form of text sequence generation tasks. Specifically, given a test input x with n context examples in context $C=\{c_1,\cdots,c_n\}$, output y is predicted based on formatting cues $p(C,x)$ as conditions. Where C and x are both image text sequences and output $y=\{y_1,\cdots,y_I\}$ is a text sequence of length L. For the l th character in the sequence, the decoding process can be represented as:

$$y_l = \arg\max P_{LLM}\left(y_l \mid p, y_d\right) \tag{1}$$

LLM represents the weights of the pre-trained large language model that are frozen in the downstream task. Each context example $c_i=\left(x_i,y_i\right)$ is an input-output pair for the task in question, which were manually constructed or sampled from the training set.

2.3.2. High-quality supporting knowledge generation phase. In this paper, two different schemes are tried for what kind of auxiliary knowledge can improve the model performance [23]. The first scheme adopts a similar approach to most existing methods by directly allowing the large language model to generate predictions of entity relationships. Since there are as many as 23 categories of entity relations in the MNRE dataset, the results of entity relation prediction through context less sample learning are not satisfactory. In fact, the relationship extraction task essentially consists of two key subtasks. The first is to discriminate the categories of two entities in a text, and the second is to determine the relationship between these two entities. Taking this into account, this paper tries the second scheme, i.e., letting the big language model generate the auxiliary knowledge related to the entity categories first, and then using this auxiliary knowledge for the prediction of entity relationships in the next stage. By manually analyzing the auxiliary knowledge generated by the two schemes, this paper concludes that the auxiliary knowledge generated by the second scheme is more accurate and reliable.

A general multimodal pre-training model usually consists of a multimodal encoder F and decoders suitable for different downstream tasks, where the multimodal encoder F includes a visual encoder, a text encoder and a multimodal fusion module. By means of the multimodal encoder F, image-a-text input pairs are encoded into multimodal fusion features H:

$$H = F(t,v) \tag{2}$$

Previous studies have projected input text t and image v into a high-dimensional hermitian space via a multimodal fusion module to obtain fusion features H before feeding H directly into a decoder for result prediction. Since in high-dimensional hermitian space, approximate samples are more likely to have the same mapping method and entity class, the fusion feature H can be used as a basis for judging similar samples.

Specifically, the cosine similarity between the fusion features between each artificial predefined sample and the input samples is first calculated, and then the N artificial predefined samples with the highest cosine similarity are selected as context examples. Context examples C are defined as follows:

$$I = \underset{j\in\{1,2,\ldots N\}}{\arg TopN} \frac{H^T H_j}{\|H\|_2 \|H_j\|_2} \tag{3}$$

$$C = \left\{ \left(t_j, v_j, y_j \right) \mid j \in I \right\} \tag{4}$$

where I is the set of the first N similar samples in the artificial predefined sample, and t, v and y refer to the text, images and original labels. After obtaining contextual examples C, a complete heuristic augmentation cue can be constructed in combination with the test inputs to utilize the few-sample learning capability of the large language model for entity category determination.

2.3.3. Relationship prediction phase based on high-quality supporting knowledge. The auxiliary knowledge generated by the big language model through contextual learning is $Z = \{z_1, \cdots, z_m\}$, and the auxiliary knowledge is connected to the original text $T = \{t_1, \cdots, t_n\}$ to obtain $[T; Z]$. Referring to the previous method, the two entities to be evaluated for the relationship are first wrapped using the special markers, and at the same time the categorization markers [CLS] and separator markers [SEP] are inserted at both ends of the text and between the two text segments, which are then fed into the Transformer Encoder E:

$$\left\{ h_{[CLS]}, h_1, \cdots, h_n, h_{[SEP]}, h_{n+1} \cdots, h_{n+m}, h_{[SEP]} \right\} = E\left(addToken\left([T; Z] \right) \right) \tag{5}$$

Since the textual description of the corresponding image is already included in the auxiliary knowledge, and the feature vector $h_{[CLS]}$ of the classification markers covers the relevant cues of the auxiliary knowledge Z through the attention mechanism of the Transformer, there is no need to fuse it directly with the original image features, avoiding the introduction of redundant information. The input $h_{[CLS]}$ is fed into a decoder consisting of a fully connected layer and a Softmax classification layer, which defines the probability of the existence of a relationship y between two entities in a sentence given the input sentence T:

$$P(y \mid T, Z) = Soft\max\left(W h_{[CLS]}^L \right) \tag{6}$$

Finally, the parameters of the Transformer encoder and the fully connected layer decoder are fine-tuned by minimizing the negative log-likelihood (NLL) loss:

$$L_{NLL}(\theta) = -\sum_i^M \log\left(P_\theta\left(y^i \mid T^i, Z^i \right) \right) \tag{7}$$

where M denotes the number of samples in the training set.

2.4. Multimodal Sentiment Analysis Models

As two important linguistic features of textual information, semantic features and structural features serve a crucial role in aspect-level sentiment analysis tasks. In this section, we will discuss both

dynamic semantic feature extraction and static structural feature extraction, and then the two semantic features will be fused, so as to effectively model the relationship between aspectual words and opinion words, and ultimately achieve a more accurate prediction of sentiment polarity [24].The CDSSSF model consists of four parts: the text encoding layer, the dynamic semantic feature acquisition module, the static structural feature acquisition module, and the fusion and classification layer.

2.4.1. Text encoding layer. For a sentence consisting of n word, given a length of m aspect words $A=\{\omega_{r+1},\omega_{r+2},\ldots,\omega_{r+m}\}$, each word in the sentence is embedded into a low-dimensional vector using the word embedding matrix $E\in\square^{|V|\times d_e}$ of the BERT pre-trained model, where $|V|$ is the lexicon size and d_e is the word embedding dimension, and the sentence embedding result is $s=\{s_1,s_2,\ldots,s_n\}$ and the aspect word embedding result is $a^s=\{a_{r+1},a_{r+2},\ldots,a_{r+m}\}$ [25].

For the final embedding result of aspect words, this paper discusses two cases. If the aspect word sequence is a single word, such as the embedding of "food", then the embedding result of the aspect word is the same as a^s. If the aspect sequence contains multiple words, such as the case of "systemmemory" which contains two words, the final embedding result of the aspect word is expressed as the average value of the embedding of each word in the aspect word. The aspect embedding process can be expressed as shown in equation (8):

$$a=\begin{cases}a_r & m=1\\ \left(\sum_{j=r+1}^{r+m}a_j\right)\Big/m & m>1\end{cases} \tag{8}$$

Since BiLSTM has strong ability to process sequence data, better ability to understand textual contextual information, and better ability to capture textual sequence long-term dependencies [26]. Therefore, for the initialized sentence embedding result S, in this paper, sentence encoding is performed by BiLSTM to generate the hidden state vector as follows

$$\begin{cases}\overrightarrow{h_i}=LSTM\left(\overrightarrow{h_{i-1}},s_i\right)\\ \overleftarrow{h_i}=LSTM\left(\overleftarrow{h_{i-1}},s_i\right)\end{cases} \tag{9}$$

$$H^c=\left[\vec{H}^c;\overleftarrow{H}^c\right] \tag{10}$$

In Eq. (2), $\overrightarrow{h_i}$ and $\overrightarrow{h_{i-1}}$ are the hidden states of the word in the $i,i-1$rd state from forward to backward, and $\overleftarrow{h_i}$ and $\overleftarrow{h_{i-1}}$ are the hidden states of the word in the ith and $i-1$th states from backward to forward, respectively, and Eq. (3) splices the results of the forward hidden state $\overrightarrow{H^c}=\{\vec{h}_1,\overrightarrow{h_2},\ldots,\overrightarrow{h_n}\}$ and the backward hidden state $\overleftarrow{H^c}=\{\overleftarrow{h_1},\overleftarrow{h_2},\ldots,\overleftarrow{h_n}\}$, and then the final output can be obtained as H^c. H^c is the final representation of the BiLSTM encoding, which contains the context between aspectual word and viewpoint word information.

2.4.2. Dynamic semantic feature acquisition module. The inputs to the dynamically adjusted weight adapter are the result after sentence embedding s and the result after aspect word embedding a. In each step, in this paper, the dynamically adjusted weight function F is first utilized to select the most relevant words to the aspect words from the input sequence $s=\{s_1,s_2,\ldots,s_n\}$ as shown in Eqn. (11), and a_l is the output of the dynamically adjusted weight function F, as well as the information about the viewpoint words that are most relevant to the particular aspect, as selected in step l:

$$a_l=F\left(\left[s_1,s_2,\ldots,s_n\right],h_{l-1},a\right) \tag{11}$$

Then, this paper utilizes the gated recurrent unit (GRU) to further simulate the semantic

comprehension process of human beings in specific contexts, which is consistent with the process of people comprehending texts based on semantics [27]. The redundant and irrelevant information is effectively eliminated from the semantic representation through the gating mechanism and the semantic representation of the sentence is updated to obtain the aspect-aware dynamic semantic representation as shown in Equation (12):

$$h_l = GRU\left(a_l, h_{l-1}\right), t \in [1, L] \tag{12}$$

In Equation (12) L is the dynamic reweighting length of the sentence, which represents the human cognitive threshold. $h_0 = H^c$ is the initial state of the dynamic adjustment weight function F, h_l is the hidden state of the output of the dynamic adjustment weight function F in step l, and h_L is the final dynamic semantic feature obtained after L times of reweighting. The following discussion focuses on the dynamic adjustment weight function F.

In this paper, the attention mechanism is utilized to realize the dynamic adjustment weight function F, whose purpose is to select the most relevant view word information with specific aspects in each step, and the calculation formula is:

$$M = W_s s + \left(W_d h_{l-1} + W_a\right) \otimes E \tag{13}$$

$$\begin{aligned} m &= \left\{m_i \mid i = 1, 2, \ldots, n\right\} \\ &= \omega^T \tanh\left(M\right) \end{aligned} \tag{14}$$

In Eq. (13) and Eq. (14) s is the original sentence embedding sequence $s = \left\{s_1, s_2, \ldots, s_n\right\}$, W_s, W_d, W_a and $\boldsymbol{\omega}$ are trainable parameters, $E \in \mathbb{R}^n$ is a vector whose row vectors are all 1, $\otimes$ denotes the outer product, and M is the result of fusion of the aspect word with the sentence. In order to better encode aspect-aware semantics, this paper selects a most important viewpoint word for the aspect word in each step as shown in the following equation:

$$\alpha_i = \frac{\exp\left(m_i\right)}{\sum_{k=1}^{n} \exp\left(m_k\right)} \tag{15}$$

$$a_l = s_j, \left(j = Index\left(\max\left(\alpha_i\right)\right)\right) \tag{16}$$

where m_i is the hidden state of the ind word, α_i is the attention score of the ith word in the sentence, and a_l is the information about the view word that is most relevant to the particular aspect at step l. Since the $Index\left(\max(\cdot)\right)$ operation has no derivative, its gradient cannot be calculated. Inspired by the softmax function, Eq. (16) is modified in this paper as follows:

$$a_l = \sum_{i=1}^{n} \frac{\exp\left(\lambda m_i\right)}{\sum_{k=1}^{n} \exp\left(\lambda m_k\right)} s_i \tag{17}$$

2.4.3. Static structural characterization module. In order to obtain the static structural features of the text, this paper discusses two aspects of the Dependency Graph Attention Module and the Component Graph Attention Module to model the structural information of the text, so as to obtain the static structural features using more comprehensive syntactic information.

1) Graph Attention Network. Obtaining node features.The input to GAT is an undirected graph $G = (v, e)$, where v denotes the set of nodes and e denotes the set of edges. A node adjacency matrix $A \in \mathbb{R}^{n \times n}$ can be obtained from graph G, where n represents the size of node v. The adjacency matrix A_{ij} reflects the connectivity between node i and node j. $A_{ij} = 1$ when node i is connected to node j. $A_{ij} = 0$ when node i is not connected to node j. For node i at layer l,

its hidden state g_i^l is updated by Eq. (18):

$$g_i^l = \Big\|_{z=1}^{Z} \sigma\left(\sum_{j\in N^l(i)} \alpha_{ij}^{lz} W_g^{lz} \hat{g}_j^{l-1}\right) \tag{18}$$

where $\|$ denotes the splicing of vectors, σ denotes the activation function, $\hat{g}_j^{l-1}$ is the hidden representation of node j in layer $l-1$, generated by layer $l-1$ of the GAT. Z denotes the number of heads in the multi-head attention, W_g^{lz} is the trainable weight vector in layer l, and α_{ij}^{lz} is the normalized attention weight coefficient of node i and its neighbor j, computed as follows:

$$\alpha_{ij}^{lz} = \frac{\exp\left(f\left(g_i^{l-1}, g_j^{l-1}\right)\right.}{\sum_{j'\in N^l(i)} \exp\left(f\left(g_i^{l-1}, g_j^{l-1}\right)\right)} \tag{19}$$

In Eq. (19), $f(\cdot)$ denotes the score function that measures the relevance of two words, usually the nonlinear activation function LeakyReLU function.

2) Attention module of dependency graph. In this paper, Eq. (20) is utilized to transform the dependency tree T into the dependency probability matrix $DA \in \mathbb{R}^{n\times n}$, where n denotes the number of words in a sentence:

$$DA_{i,j} = \begin{cases} 1 & When\ \omega_i\ and\ \omega_j\ are\ directly \\ & connected\ at\ the\ dependency\ tree \\ 0 & Other \end{cases} \tag{20}$$

By inputting the hidden state vector H^c and the syntactic dependency probability matrix DA into the N-layer graph attention network, the final output of the dependency graph attention module is obtained as $\hat{g}_{dep}$ by using Equation (17).

3) Component graph attention module. Component trees, as widely used static graphs nowadays, are capable of capturing phrase-based syntactic relations in one or more sentences.

Since there are multiple neighbor matrices obtained from the component analysis tree, when using the graph attention network for the analysis, this paper adopts the bottom-up order to enter the graph attention network, firstly from the Layer-1 layer as the neighbor matrix, meanwhile, $\hat{g}^0 = H^c$ is used as the input of the first GAT, and the hidden state of Layer i is obtained by calculating the hidden state of Layer g_i^l through Eqn. (17), meanwhile, according to Eqn. (21), the final feature of Layer i is updated with $\hat{g}_i^l$ representation:

$$\hat{g}_i^l = FC\left(g_i^l + \hat{g}_i^{l-1}\right) \tag{21}$$

4) Static structure feature fusion. In order to obtain comprehensive static structural feature information, in this paper, the feature $\hat{g}_{dep}$ extracted by the dependency graph attention module and the feature $\hat{g}_{con}$ extracted by the component graph attention module are spliced, which can lead to the final static structural feature g_{str}, as shown in equation (22):

$$g_{str} = \left[\hat{g}_{dep}; \hat{g}_{con}\right] \tag{22}$$

2.4.4. Fusion and Classification Layer. In this paper, the acquired dynamic semantic feature h_L and static structural feature g_{str} are feature fused and the result is obtained as:

$$O = h_L + g_{str} \tag{23}$$

Where O is the final vector representation obtained by fusing dynamic semantic features and

static structural features, and 0 is used as the input to the fully connected layer to classify the final sentiment polarity by softmax function, i.e., [28]:

$$y = soft\max\left(W_i O + b_i\right) \tag{24}$$

Eq. (18) in which W_i and b_i are the weights and bias terms of the fully connected layer, respectively. In this paper, the model is trained by gradient descent algorithm to accomplish the classification task:

$$loss = -\sum_{i=1}^{S}\sum_{j=1}^{C} y_i^j \cdot \ln\left(\hat{y}_i^j\right) + \lambda \left\|\theta\right\|_2 \tag{25}$$

where S is the number of training samples, C is the number of categories, y_i^j and $\hat{y}_i^j$ are the true and predicted labels of the training set, respectively, θ is all trainable parameters, and λ is the L_2 regularization term coefficient.

2.5. Multimodal Search and Quiz System

In this section, we will combine the multimodal relationship extraction and multimodal sentiment analysis methods proposed in the above paper to construct a multimodal retrieval Q&A system for intelligent customer service robots to improve the service performance of intelligent robots.

2.5.1. Multimodal fact retrieval algorithms. Multimodal fact retrieval algorithms are used to retrieve models containing multiple data types (e.g., current news text, images, etc.). In order to quickly obtain the candidate facts related to the question, first, the Chinese-Clip model is selected, which is based on Openai's Clip model obtained by secondary pre-training using large-scale Chinese data, and outputs the candidate multimodal facts after combining with the KFE-KRR for retrieval recall screening. First, the multimodal facts related to the question need to be quickly recalled, and the multimodal question Q generates the relevant question keyword set K_Q, and the candidate current affairs knowledge is recalled using BM25 for K_Q. Then, Ernie-Dual-Model is used to calculate the relevance score $Score_{Text}$. In addition, Chinese-Clip is used to output the image relevance scores from 0 to 1 for the issue image and the fact image, and the weighted merger of the two is used as the retrieval score sorting and filtering, which provides the basis for the subsequent steps.

The specific process regarding the Chinese-Clip model is as follows. Given the loser problem image Q, the factual image set D is $\{D_1, D_2, \ldots, D_k\}$, and k represents the number of images. At the same time, the Encoder of Chinese-Clip model is utilized to encode them to obtain the problem image encoding representation V_Q and the fact image set encoding representation V_{D_i}:

$$V_Q = f_{Encoder}(Q) \tag{26}$$

$$V_{D_i} = f_{Encoder}\left(D_i\right) \tag{27}$$

Next, the parallel computation of the similarity between the problem image Q and each factual image D_k is continued $sim\left(Q, D_k\right)$. Then, the maximum of all computed similarity scores is selected as the similarity score $Score_{\mathrm{Image}}$ between the problem image Q and the entire factual image D:

$$\mathrm{Image\_Score} = \max_{i\in\{1,2,\ldots,k\}} \cos\left(V_Q, V_{D_i}\right) = \max_{i\in\{1,2,\ldots,k\}} \frac{V_Q \cdot V_{D_i}}{\left|V_Q\right|_2 \cdot \left|V_{D_i}\right|_2} \tag{28}$$

$$Score_{Final} = w_t \cdot Score_{Text} + w_i \cdot Score_{\mathrm{Image}} \tag{29}$$

Where ScoreFinal is the composite retrieval score, w_t and w_i are the weights of the text score and the image relevance score, respectively, which are adjusted according to specific application scenarios and data characteristics. Finally, all candidate facts are sorted and filtered using the comprehensive

retrieval score to provide the most relevant multimodal facts to the subsequent algorithms.

2.5.2. Multimodal Enhanced Qualified Q&A Algorithm. In multimodal augmented-qualified Q&A algorithms, the model generates precise answers to user questions based on relevant multimodal facts (including text, images) obtained from multimodal fact retrieval algorithms. First, key fact extraction is required, and then a rigorous and precise answer is generated by combining the key facts through questions. To achieve this goal, the Qwen-VL model is used here to process text and image content, and Qwen-VL is trained on a multimodal retrieval-enhanced Q&A dataset with Doka LoRA to enhance the model's multimodal language comprehension and generation capabilities in high-stakes domains.

In multimodal augmented-qualified question-answering task training and inference, given question Q, multimodal facts Fs (obtained by extracting fact D_i), and responses A directly related to question Q, these are infused into a LoRA-trained Qwen-VL model. In this process, the model needs to learn how to synthesize the multimodal facts to generate accurate responses A. This process can be represented as the following task:

$$A = MM - GenerateEnhancedAnswer(Q, Fs; \Theta) \tag{30}$$

where MM-Generate Enhanced Answer is the multimodal augmented qualified question-and-answer algorithm and Θ denotes the fine-tuned model parameters. In order for the model to better understand and process multimodal data and generate accurate answers, the model is trained here using supervised learning, where the training set consists of the question-multimodal fact-answer ternary, and the model is optimized by minimizing the following objective function:

$$\mathrm{L}(\Theta) = \sum_{(Q,Fs,A)\in \mathrm{T}} -\log P(A \mid Q, Fs; \Theta) \tag{31}$$

In addition, T denotes the multimodal augmented-qualified Q&A training set, and $P(A \mid Q, Fs; \Theta)$ is the conditional probability of the answer A given by the model under parameter Θ. After the training is completed, the Qwen-VL model can be used for actual multimodal augmented-qualified Q&A tasks.

After completing the training of the model, this study will use the inference multimodal augmented qualification Q&A task along with the implementation of the fact-qualification-correction chain of thought (FCM-CoT). This process aims to ensure the accuracy and reliability of responses by effectively qualifying factual information and accurately correcting the model output. Combining the autoregressive properties of the model, this study designed a mathematical qualification correction process.

Under the FCM-CoT algorithm, the model performs keyword extraction while generating each response and evaluates the consistency of these keywords with the core factual elements. If an inconsistency is detected between the keywords and the factual elements, the model will regenerate the response. After reaching the preset maximum number of attempts $\max_a ttempts$, if the model still fails to generate a compliant answer, the model outputs an alert message and suggests the user to refer to relevant news sources, which is expressed as "Sorry, I cannot answer your question at the moment, please consider browsing relevant news" and D. This strategy ensures high accuracy and factual reliability of the model output, which enhances the user's confidence in the model. This strategy ensures high accuracy and factual reliability of the model output, thus increasing the user's trust in the model.

In a practical application scenario, when a user asks Question Q, the algorithm inputs the question with the associated multimodal factual data into the trained Qwen-VL model. Generate an accurate response A that is highly relevant to Question Q, guided by the FCM-CoT. Such answers are not only based on textual information, but also synthesize visual information to provide users with comprehensive and in-depth answers, thus demonstrating excellent performance in multimodal augmented-qualified question-and-answer tasks in the domain of current affairs news.

3. Multimodal Interaction Simulation Experiment of Intelligent Customer Service System

In the above paper, in order to improve the intelligent customer service robot answering questions service ability of low efficiency defects such as low defects, the introduction of multimodal interaction architecture, from the multimodal language extraction, aspect set sentiment analysis of multiple aspects, constructed a multimodal retrieval of intelligent customer service question system, greatly improving the intelligent customer service robot answering service performance. In this chapter, multimodal interaction simulation experiments will be carried out to test the effectiveness of the functions of the modules in the intelligent customer service system based on multimodal interaction constructed in this paper.

3.1. Multi-modal relation extraction experiments

In this paper, a two-stage framework is adopted in multimodal relationship extraction in intelligent customer service system, and a PLFM model is constructed to improve the performance of multimodal relationship extraction. In this section, the PLFM model of this paper is compared with some classical unimodal relationship extraction models and mainstream multimodal relationship extraction models, and the experimental comparison results are shown in Table 1. From the experimental results, it can be seen that the F1 value of the proposed model is increased by 6.49 percentage points compared with the single-modal relation extraction model MTB, and the F1 value of the best-performing multimodal relation extraction model MEGA is increased by 0.78 percentage points compared with other models.

Table 1. Comparison result

Model	Accuracy rate(%)	Recall rate(%)	F1(%)
PCNN	62.9	49.7	55.44
MTB	64.53	57.79	60.9
UMT	62.87	61.22	62.6
UMGF	64.41	66.28	65.28
AdapCoAtt-BERT	64.7	57.97	61.08
VisualBERT	56.3	58.38	57.38
ViLBERT	64.43	61.96	63.24
MEGA	64.61	68.4	66.51
Model of this article	65.6	69.24	67.39

3.2. Multimodal Sentiment Analysis Experiment

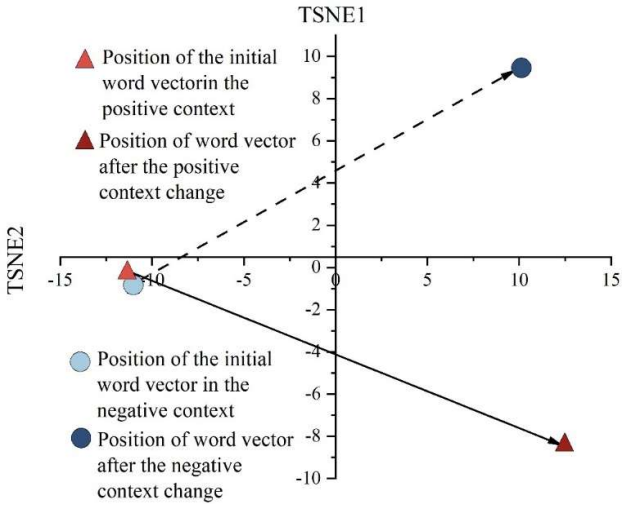
1) Comparison experiment. The CDSSSF model applied in the intelligent customer service customer care system of this paper is compared with the more advanced models in multimodal sentiment analysis task such as TFN, LMF, MFM, etc. The experimental results of different multimodal sentiment analysis task models including this paper's model on CMU-MOSI and CMU-MOSEI datasets are given, as shown in Table 2. Compared with the baseline model, this paper's model shows significant

improvement in both classification accuracy and E_{MA} task loss. The model in this paper achieves the optimal results in E_{MA}, r, A_2 and F_1 metrics on the two public datasets, which is mainly due to the fact that the model in this paper effectively improves its performance by adding the dynamic semantic feature module.

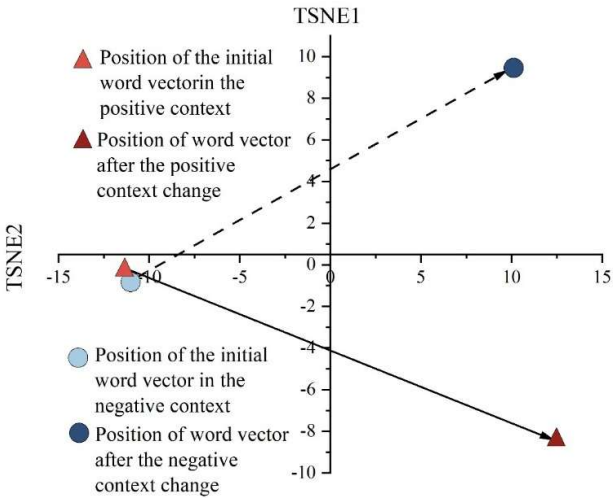
Table 2. Emotional analysis

Model	CMU-MOSI				CMU-MOSEI			
	$E_{...}\downarrow$	$r\uparrow$	$A_2\uparrow$	$F_1\uparrow$	$E_{MA}\downarrow$	$r\uparrow$	$A_2\uparrow$	$F_1\uparrow$
ATAE-LSTM	0.895	0.708	—/80.7	—/80.72	0.592	0.692	—/82.51	—/82.11
MGAN	0.909	0.685	—/82.6	—/82.41	0.622	0.669	—/82	—/82.15
IAN	0.886	0.704	—/81.8	—/81.62	0.564	0.725	—/84.4	—/84.34
DR-BERT	0.861	0.719	—/83.11	—/83.18	0.573	0.72	—/84.24	—/84.2
R-GAT	0.863	0.715	81.51/84.12	80.62/83.91	0.576	0.707	—/82.56	—/82.3
BiSyn-GAT	0.796	0.767	80.78/82.11	80.78/82.04	0.558	0.729	82.58/82.42	82.66/83.98
KSD-GCN	0.728	0.799	82.48/84.31	82.62/84.31	0.541	0.752	83.81/85.21	83.71/85.12
CDSSSF	0.601	0.854	86.43/89.01	86.38/89.02	0.516	0.801	84.74/86.95	84.64/86.82

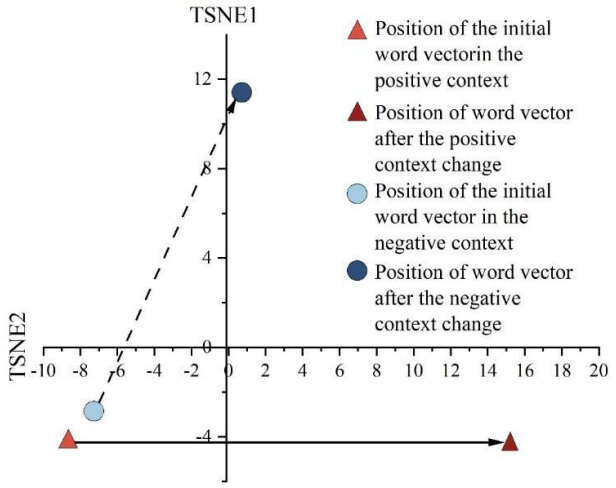
2) Visualization experiments. The word vector representations of the three words "its", "character" and "interesting" were randomly selected from a large number of examples, and the initial word vector representations and the word vector representations after the dynamically adjusted semantic module were projected into the two-dimensional space by using t-random proximity embedding (TSNE), and their position changes in the positive and negative contexts were observed, as shown in Figure 2, (a)~(c) were "character", "character", " interesting" and "its".The initial positions of the three words in positive and negative contexts are very close to each other, but after the interactive fusion of dynamic semantic information and static structural information in this paper's model, their positions in space change drastically, and this change is not only reflected in the size but also in the direction.



(a)Character



(b)Interesting



(c)Its

Fig. 2. The position change of word vectors before and after dynamic adjustment

The heat map is used to visualize the word vector weights of some instances in the CUM-MOSI dataset. The visualization results of randomly selected instances “c7UH_rxdZv4_28” and “tIrG4oNLFzE_1” are shown in Figure 3. The Dependency Graph Attention Module and the Component Graph Attention Module contribute equally to the model of this paper and help the model of this paper to predict the sentiment values more accurately. In the example "Now i am a huge huge fan of nicho-las sparks", a larger weight is given to "huge", and the two "huge" enhance the emotion, which makes the model predict a relatively strong positive sentiment value in the end. Similarly, for the words in the instance "He was the only character that was slightly interesting", a larger weight is assigned to "only" and "interesting", which makes the proposed model finally predict the instance as a negative sentiment value.

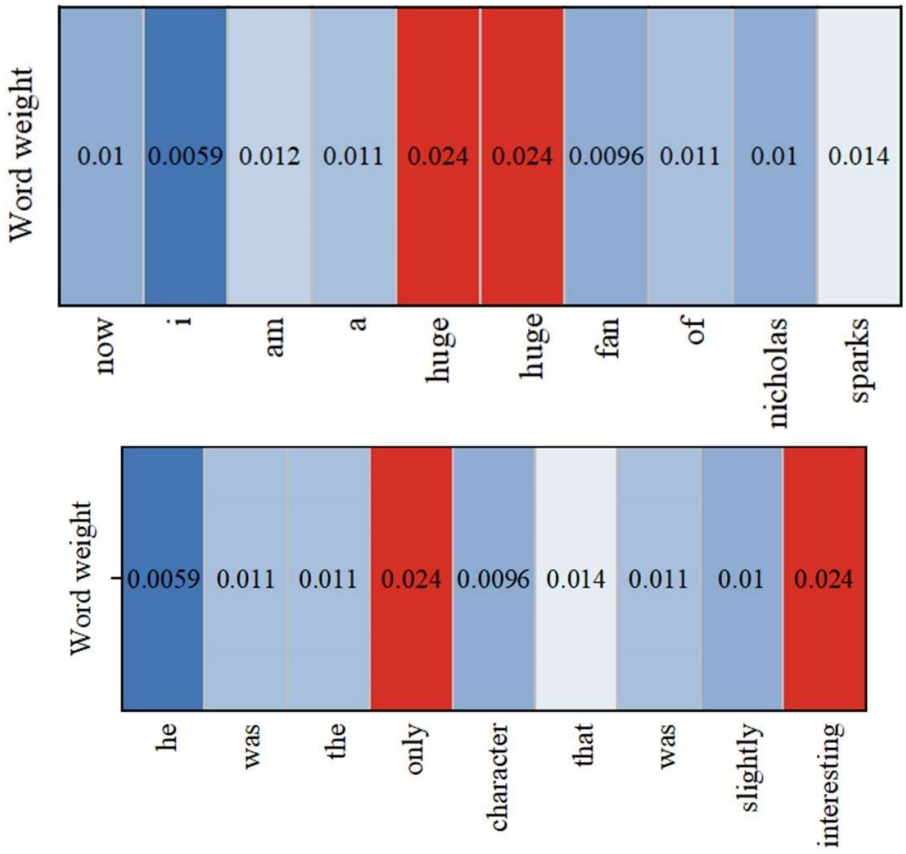


Fig. 3. Visual results

4. Intelligent customer service system application practice

The TCSIC dataset is collected from user-human conversations with customer service in real customer service scenarios of operators. These conversations are enriched with relevant information in the interaction process between users and customer service, including but not limited to conversational intentions, service behaviors, emotional actions, system feedbacks, and so on. In this chapter, the TCSIC dataset will be used as a research sample to apply the intelligent customer service service system based on multi-modal interaction constructed in this paper to carry out the practice of intelligent customer service robot service application.

4.1. Statistical analysis of service call rounds

In this application practice, the system in this paper has carried out a total of 20,000 call services, and the distribution of the number of conversation rounds for each service call is specifically shown in Figure 4. It can be observed that most of the conversation rounds in the dataset are concentrated in the range of 10 to 50, which means that the conversation between the user and the intelligent customer service robot usually ends within about 50 rounds, and the service efficiency is more efficient.

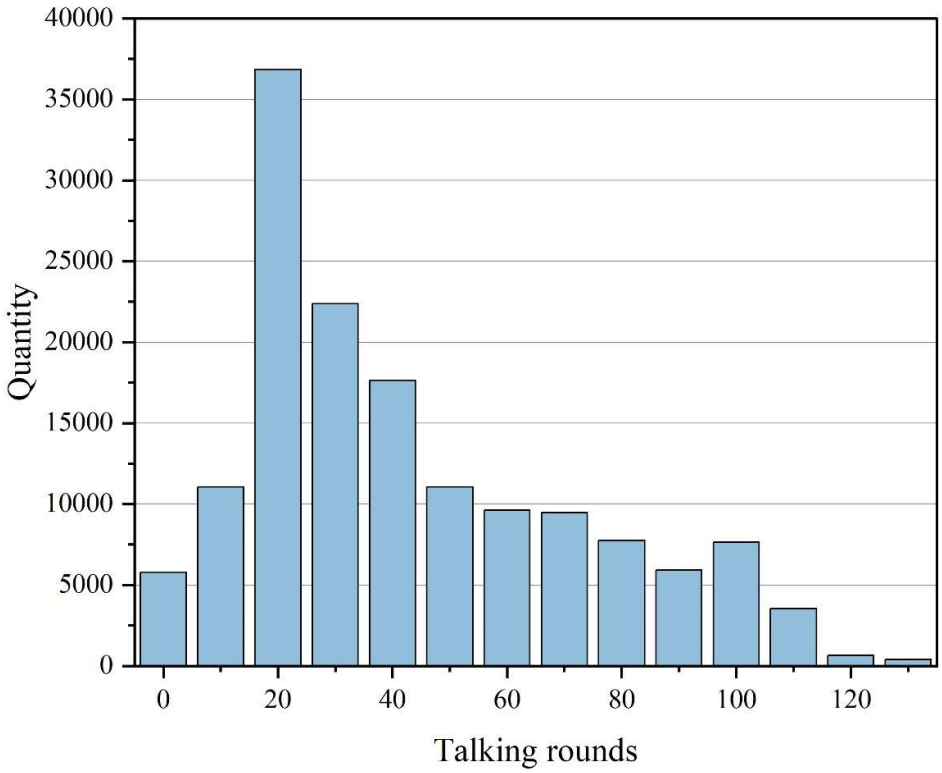


Fig. 4. Talking rounds

4.2. Statistical analysis of user sentiment recognition

Counting the emotion labels of all the users in 20,000 call services, the number of emotionless sentences is 1147459 and the number of emotional sentences is 321993. And among the emotional sentences, 112684 of the actual emotion labels are happy, and 104,811 and 105,389 are complaints and anger, respectively. The user sentiment analysis results derived from the multimodal sentiment analysis of this paper's system are specifically shown in Table 3. It can be seen that the system in this paper achieves 100% accuracy in recognizing the sentiment of emotionless sentences, and is able to completely and accurately recognize the sentiment of emotionless sentences. Among the sentences with emotions, the number of errors made by the system in this paper for the sentences with the actual emotion labels of happy, complaining and angry are 167, 999 and 948, respectively, and the accuracy rate of emotion recognition is greater than the 99% level in all of them.

Table 3. Emotional recognition statistics

Sentence	Emotion tag	Actual quantity	number of correctly identified	Number of error identification	Accuracy rate
Ruthless sentence	-	1158568	1158568	0	100%
Sentence with emotion	Pleasure	112684	112517	167	0.998518
	Whine	104811	103812	999	0.990469
	Anger	105389	104441	948	0.991005

4.3. Statistical Analysis of Intelligent Customer Service Robot Service Behavior

Statistics on the service behavior of the intelligent customer service robot in the application practice, as shown in Figure 5. As you can see, the largest proportion of tags are "Notification" and "Confirmation", and the number of behaviors reaches 560365 and 365976. "Notification" means that the agent needs to inform the user of relevant information, such as the result of an information query, feedback on the execution of an action, etc. "Confirmation" means a second confirmation based on the user's inquiry or the information provided above. Since the common content of the dialogue in the customer service field is to solve the user's problem, the labels of "notification" and "confirm" account for a high proportion in each conversation round, and the service behavior of the intelligent customer service robot in this system is in line with expectations.

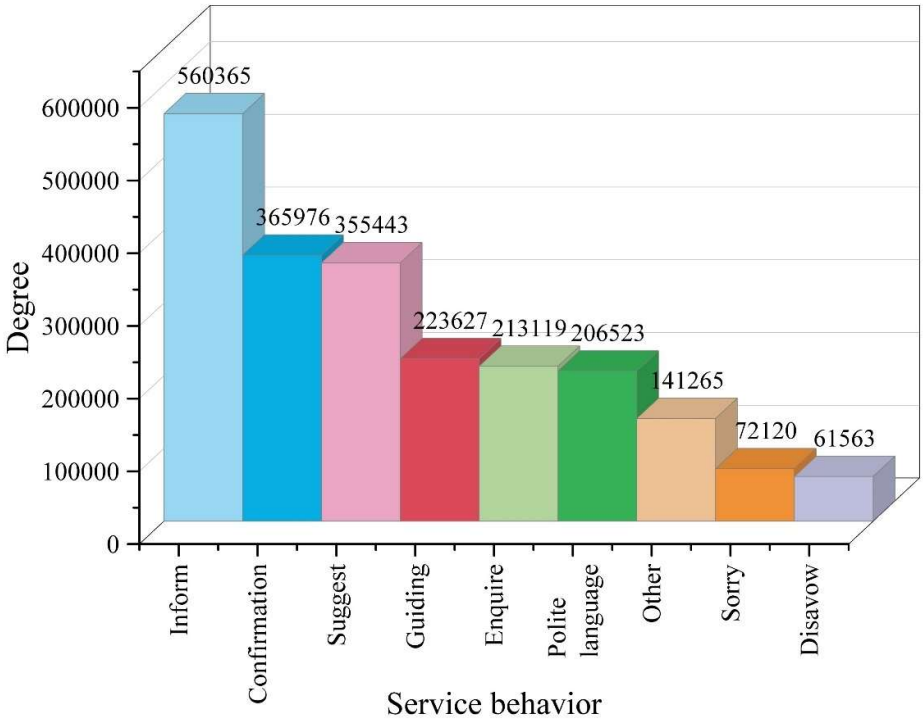


Fig. 5. Service behavior of intelligent customer service robot is

Overall, the intelligent customer service system based on multi-modal interaction constructed in this paper shows the potential and value in practical application scenarios, and is able to provide reference significance for future large-scale promotion.

5. Conclusion

In order to realize the dynamic adjustment of semantic features of intelligent customer service robots and improve the service performance of intelligent customer service robots, this paper integrates the multimodal large language interaction technology to build an intelligent customer service system based on multimodal interaction. Through the multimodal interaction simulation experiment to test the effectiveness of the multimodal module function of this paper's system. In the multimodal relationship extraction experiment, the F1 value of PLFM model in this paper is improved by 6.49% and 0.78% respectively compared with the unimodal relationship extraction model MTB and the best-performing multimodal relationship extraction model MEGA. In the multimodal sentiment analysis task, optimal results are obtained on metrics E_{MA}, r, A_2 and F_1 on both CMU-MOSI and CMU-MOSEI datasets. The multimodal sentiment analysis model in this paper can also predict the sentiment values more accurately in the face of the two sentiment analysis examples "c7UH_rxdZv4_28" and "tIrG4oNLFzE_1", predicting that the two examples of the instances of the sentiment are, respectively, more strongly positive sentiment values, and negative sentiment values. TCSIC data set as an example, the intelligent customer service system constructed in this paper to carry out application practice analysis. In the application practice, most of the conversation rounds between the intelligent customer service robot and the customer are concentrated in the range of 10 to 50, and basically can end the service conversation within 50 rounds, with high service efficiency. The system in this paper also performs well in the multimodal emotion method analysis of customer call statements, the emotion recognition accuracy of emotionless sentences reaches 100%, and in the face of emotion labeled as happy, complaining and angry emotion sentences, the accuracy of emotion recognition is also higher than 99%. The service behavior of the intelligent customer service robot is also mainly based on "notification" and "confirmation", with the number of behaviors reaching 560,365 and 365,976 respectively, which is in line with service expectations. Overall, the intelligent customer service system constructed in this paper can effectively improve the service performance and efficiency of the intelligent customer service robot, showing the application potential and value in the actual scene.

References

[1] Liu, X. (2024, June). Intelligent Customer Service System of Subway Based on Semantic Decision-Making Judgment. In 2024 4th International Conference on Machine Learning and Intelligent Systems Engineering (MLISE) (pp. 374-378). IEEE.

[2] Bassano, C., Piciocchi, P., & Pietronudo, M. C. (2018). Managing value co-creation in consumer service systems within smart retail settings. Journal of Retailing and Consumer Services, 45, 190-197.

[3] Gergely, T., Halmay, E., Szőts, M., Suciu, G., & Cheveresan, R. (2017, July). Semantics driven intelligent front-end. In 2017 International Conference on Speech Technology and Human-Computer Dialogue (SpeD) (pp. 1-8). IEEE.

[4] Zou, Y., Liu, X., Xu, H., Hou, Y., & Qi, J. (2021). Design of intelligent customer service report system based on automatic speech recognition and text classification. In E3S Web of Conferences (Vol. 295, p. 01064). EDP Sciences.

[5] Ji, M., & Zhang, X. (2022). Research on semantic similarity calculation methods in Chinese financial intelligent customer service. International Journal of Computer Applications in Technology, 68(2), 143-149.

[6] Yijing, W. (2018). Intelligent customer service system design based on natural language processing. In Proceedings of 2018 5th International Conference on Electrical & Electronics Engineering and Computer Science, ICEEECS.

[7] Song, S., Wang, C., Chen, H., & Chen, H. (2021, June). An emotional comfort framework for improving user satisfaction in E-commerce customer service chatbots. In Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Industry Papers (pp. 130-137).

[8] Zhang, B., Lin, H., Zuo, S., Liu, H., Chen, Y., Li, L., ... & Yuan, B. (2020, May). Research on Intelligent Robot Engine of Electric Power Online Customer Services Based on Knowledge Graph. In Proceedings of the 2020 the 4th International Conference on Innovation in Artificial Intelligence (pp. 216-221).

[9] Tan, Y., Xu, H., Wu, Y., Zhang, Z., An, Y., Xiong, Y., & Wang, F. (2021). Research on knowledge driven intelligent question answering system for electric power customer service. Procedia Computer Science, 187, 347-352.

[10] Cao, X., Liu, X., Zhu, B., Miao, Q., Hu, C., & Xu, F. (2018). Semantic Matching Using Deep Multi-Perception Semantic Matching Model with Stacking. In CCKS Tasks (pp. 77-88).

[11] Fu, R., & Zhou, X. (2024). Intelligent Recommendation Algorithm for Product Information on E-commerce Platforms Based on Robot Customer Service Automatic Q&A. Scalable Computing: Practice and Experience, 25(6), 5070-5078.

[12] Hu, Y., & Sun, Y. (2023). Understanding the joint effects of internal and external anthropomorphic cues of intelligent customer service bot on user satisfaction. Data and Information Management, 7(3), 100047.

[13] Wu, Y., Mao, W., & Feng, J. (2022). AI for online customer service: Intent recognition and slot filling based on deep learning technology. Mobile Networks and Applications, 27(6), 2305-2317.

[14] Yang, M., Cao, S., Hu, B., Chen, X., Cui, H., Zhang, Z., ... & Li, X. (2021, April). Intellitag: An intelligent cloud customer service system based on tag recommendation. In 2021 IEEE 37th International Conference on Data Engineering (ICDE) (pp. 2559-2570). IEEE.

[15] Afandi, Y., Maskur, M., Fiernaningsih, N., & Fauzi, A. (2023). SEMANTIC AND NATURAL LANGUAGE PROCESSING DEVELOPMENT APPLICATIONS FOR CHATBOTS TO ENHANCE ONLINE STORE CUSTOMER SERVICE. INTERNATIONAL JOURNAL ON ADVANCED TECHNOLOGY, ENGINEERING, AND INFORMATION SYSTEM, 2(4), 242-255.

[16] Dai, Y., Huang, Z., He, W., Khan, N., & Yang, Y. (2024). Emotional dialogue generation model of electronic commerce intelligent customer service based on topic expansion. Engineering Applications of Artificial Intelligence, 138, 109429.

[17] Zhong, L., Yu, S., Ningning, K., & Weifang, Z. (2023, October). Research on Intelligent Customer Service System for Power Industry Based on Semantic Understanding. In 2023 IEEE 3rd International Conference on Data Science and Computer Application (ICDSCA) (pp. 324-329). IEEE.

[18] Zhang, D., Cheng, S., & Yin, D. (2021, May). A Multi-semantic Knowledge Graph Construction Scheme Suitable for Intelligent Power Customer Service. In International Conference on Frontiers of Electronics, Information and Computation Technologies (pp. 1-7).

[19] Zhang, Y., Zhang, D., Tang, W., Zhang, J., & Zhao, Z. (2024, October). Power intelligent question answering customer service robot short text specific semantic mining. In Ninth International Symposium on Advances in Electrical, Electronics, and Computer Engineering (ISAEECE 2024) (Vol. 13291, pp. 1453-1459). SPIE.

[20] Liu, S., Peng, C., Wang, C., Chen, X., & Song, S. (2023). icsBERTs: Optimizing pre-trained language models in intelligent customer service. Procedia Computer Science, 222, 127-136.

[21] Wei, D. (2021). e-Commerce Online Intelligent Customer Service System Based on Fuzzy Control. Journal of Sensors, 2021(1), 4867222.

[22] Liu, C., Jiang, J., Xiong, C., Yang, Y., & Ye, J. (2020, August). Towards building an intelligent chatbot for customer service: Learning to respond at the appropriate time. In Proceedings of the 26th ACM SIGKDD international conference on Knowledge Discovery & Data Mining (pp. 3377-3385).

[23] Dan Song,Shuqi Dai,Wenhui Li,Tongwei Ren,Zhiqiang Wei & An An Liu. (2024). STVformer: A spatial-temporal-variable transformer with auxiliary knowledge for sea surface temperature prediction. Applied Ocean Research104218-104218.

[24] Li Zhou,Li Jiajia,Xie Gengming,Arya Varsha & Li Hao. (2024). Enhanced Safety in Multi-Lane Automated Driving Through Semantic Features. International Journal on Semantic Web and Information Systems (IJSWIS)(1),1-13.

[25] Carlos Abel Córdova Sáenz& Karin Becker. (2024). Understanding stance classification of BERT models: an attention-based framework. Knowledge and Information Systems(1),419-451.

[26] Yuhai Li,Youchen Fan,Shunhu Hou,Yufei Niu,You Fu & Hanzhe Li. (2024). Reconstruction of OFDM Signals Using a Dual Discriminator CGAN with BiLSTM and Transformer. Sensors(14),4562-4562.

[27] Hou Zhi Wen,Di Qian Bin & Chen Xiao Long. (2024). Forecasting carbon emissions in Chinese coastal cities based on a gated recurrent unit model. Energy Reports5747-5756.

[28] Jawa Taghreed M.,Fatima Nahid,SayedAhmed Neveen,Aldallal Ramy & Mohamed Mohamed Said. (2022). Residual and Past Discrete Tsallis and Renyi Extropy with an Application to Softmax Function. Entropy(12),1732-1732.