International Journal of Computer Information Systems and Industrial Management Applications
ISSN 2150-7988 Volume 18 (2026) pp. 1-15

https://doi.org/10.70917/ijcisim-2026-1008
Article

An Analysis of the Multimodal Representation and Linguistic Mechanisms of the Principle of (Im) politeness Reciprocity in Social Media Interactions

Renjun Pan [1] and Xiaodong Wang [1,*]

[1] School of Arts and Sciences, Northeast Agricultural University, Harbin, Heilongjiang, 150030, China
* Correspondence author: wangxiaodong@neau.edu.cn

Abstract: Impoliteness is a negative attitude towards a specific behavior in a specific context, and an impolite strategy is a reflection of the type of impolite output. In this paper, we propose a model for impoliteness detection in social media interaction based on sentiment-dependency graph convolutional neural network with modality fusion from the perspective of pragmatics. The model enhances the emotional and syntactic information of text modality through emotion graph and syntactic dependency graph, uses graph convolutional neural network to obtain text information with rich emotional semantics, and then fuses multimodal features by modal fusion, and filters the redundant information by using the self-attention mechanism to detect impolite reciprocity based on the fused information. Based on the real corpus of Chinese and English blog discourse collected in real time, we study the similarities and differences of impolite language in online discourse between the two languages. The experimental results show that the accuracy of the model reaches 84.72, which is 0.64 percentage points higher compared to the better of the comparison models. Under similar online communication contexts, the formation of overall discourse features and local distinctive features of the two corpora stems from the combined effects of dynamic and diverse online communication contexts and communication resources as well as different linguistic and cultural contexts. The study provides valuable new discoveries for the study of social media online discourse features and enlightens people's comprehensive understanding of online language and functions.

Keywords: social media interactions; sentiment-attachment graph; graph convolutional neural network; impoliteness reciprocity

1. Introduction

With the development of society, the increase in the popularity of online social media and the increase in the discourse power of young people, online terms gradually show the characteristics of high abbreviation, youthfulness, and rapid updating and iteration [1]. From popular buzzwords such as “YYDS,” “cheering on,” and “EMO” to exaggerated emotional expressions like “breaking down” and “lying flat,” all of these reflect the new and unique online social discourse system that contemporary young netizens have shaped through online linguistic symbols [2–5]. These terms break the rules of grammar in the traditional sense, and at the same time affect daily life expressions and written formal language [6]. Expressions such as "family members", "baby", "sister" and others that are more intimate have also been gradually used in addressing strangers. More exaggerated expressions like "Help me!" and "Save me!" have gradually replaced the traditional expressions for seeking help. The way of expressing humor and interest has also become more exaggerated, such as "I'm so amused". This new language strategy has triggered new thoughts on the etiquette norms in the new media era [7]. Current research mostly focuses on the semantic evolution or rhetorical features of internet language, but does not delve deeply into the politeness mechanism behind it.

It is well known that the “politeness strategy theory” proposed by Brown, P. & Levinson, S, which has been around for more than 40 years, is characterized as a “universal theory of politeness” [8]. Since its public release in 1978, the theory has triggered heated debates in the academic community, and its influence has gone far beyond the scope of linguistics (pragmatics), spreading to related fields such as sociology, cultural anthropology, psychology, and communication [9]. Since the 1990s, non-Indo-European language researchers, especially Asian language researchers, have criticized that politeness strategy theory is centered on Western languages, and is unsuitable for the study of Asian languages [10-11]. One of the main reasons for the above phenomenon is that descriptive researchers do not distinguish between “descriptive research on politeness” and “theoretical research on politeness” in exploring the phenomenon of politeness [12]. In our view, “politeness description research” refers to research on the system of honorifics, the principles of honorific use, and their cross-cultural comparisons in various languages [13]; "The study of politeness theory" refers to the research on the construction of politeness theories based on the motivation of comprehending various politeness phenomena in different language cultures. Its focus is to explain, illustrate and predict politeness phenomena in different cultures using the same theoretical framework [14-16]. Although we hope that descriptive and theoretical studies can influence each other and develop together, it is regrettable that the two have not played a good role in mutually reinforcing each other in the more than 40 years of politeness phenomenon research [17]. Nonetheless, refining the two different purposes of research has to some extent also avoided the needless controversy brought about by focusing on the different one, and provided the possibility of practically grasping the future direction of politeness research [18].

Impoliteness research, on the other hand, has emerged in response to the current situation of the extreme imbalance between the development of politeness and impoliteness research. While politeness norms are regarded as a conversational norm, impoliteness is viewed as a stumbling block to the realization of polite communication and as a marginal phenomenon [19-20]. In fact, conflict discourse is common in everyday language, such as military training discourse, courtroom discourse, family discourse, and doctor-patient discourse. In recent years, impoliteness research has begun to gain attention. While academics pay more and more attention to impoliteness, Chinese scholars regard linguistic impoliteness as a discordant phenomenon as a result of language intrinsicity [21]. Therefore, on the basis of sorting out the domestic and foreign language impoliteness research in definition and classification, the basic concepts from social cognitive theory and its research are introduced, analyzed and summarized, which can find a theoretical basis for impoliteness research.

In this paper, we propose a multimodal emotion-dependency graph-based convolutional neural network with modal fusion for impoliteness detection model (ADGCN-MFM) to analyze impolite reciprocal behaviors in social media interactions. The method enhances the affective semantic features of textual modalities by constructing sentiment maps and syntactic dependency maps, and effectively fuses multimodal features by modal fusion to further improve the accuracy of multimodal impoliteness detection. The impoliteness of verbal behavior in online feedback discourse triggered by the same hot topic in two social media, Chinese Weibo and English Twitter, is collected to explore the linguistic mechanism of the impoliteness reciprocity principle in social media interactions by combining the impoliteness theory of Culpeper and the impoliteness detection model.

2. Relevant models and techniques

2.1. Graph Neural Networks

Graph convolutional neural networks are computed in a similar way to general convolutional neural networks, but the convolution kernel is different and general graph networks are multilayer structures. Given a relational graph matrix A of n nodes, the computational style process of its convolution is shown in equation (1):

$$h_i^l = \sigma\left(\sum_{j=1}^{n} A_{ij} W^l h^{l-1} + b^l\right) \tag{1}$$

where $h_i^l \in R^{d\times 1}$ is the hidden layer state of the i th node in the l th layer of the graph network, $h^{l-1} \in R^{d\times n}$ is the resultant of the l th layer, and h_i^0 is the initial vector representation of the i th node after the encoding of embedding layers and W^l is the parameter matrix of the linear transformation of the l -layer, b^l is the bias term of the l -layer, σ is the nonlinear activation function, and A_{ij} denotes the existence of an edge between node i and node j . In this way after multiple layers of iterative learning finally get the relevant features of each word and its dependent words.

The graph attention network approach is to change the feature extraction method of convolution to a form of attention to compute, but here the attention is different from the traditional form of attention. The traditional attention is considered for all the words in the sentence, here only each word and the words on which it has a structural relationship are considered for the computation of attention and the computation process is as follows:

Assuming that at a certain stage, the hidden layer states of two nodes in the graph network structure are h_i and h_j, the alignment scores e_{ij} of these two nodes are computed as shown in equation (2):

$$e_{ij} = \mathrm{Re}LU\left(w\left[h_i; h_j\right]\right) \tag{2}$$

where W is the parameter matrix during training and ReLU is the nonlinear activation function. So note that the weights are calculated as shown in equation (3):

$$\alpha_{ij} = \frac{\exp\left(e_{ij}\right)}{\sum_{k\in N(i)} \exp\left(e_{ik}\right)} \tag{3}$$

where $N(i)$ denotes the set of all neighboring points of i in the relation graph, it can be seen from the above equation that the core of the graph attention mechanism is to compute the degree of influence between a word and the words associated with it. So the state update process of graph attention network at l layer is shown in equation (4):

$$h_i^l = \sigma\left(\sum_{j\in N(i)} \alpha_{ij} W^l h_j^{l-1} + b^l\right) \tag{4}$$

The parameters of the above equation are roughly similar to the graph convolutional network, with the difference that the α_{ij} attention weight in the above equation replaces the value of A_{ij} in the matrix. Since each word's own information is also the key information for the classification task, the researcher proposes to set all the values on the main diagonal of the let's-relationship graph matrix to 1 before the graph network feature extraction, i.e., $\tilde{A} = I + A$, and use the new matrix $\tilde{A}$ to go to the feature extraction, which makes the model's effect further improved.

Figure 1 shows the model diagram of graph convolutional network for text categorization task, which has embedding layer, intermediate layer and fully connected layer as the general network model. The process can be described as follows: the input text W is embedded in the word embedding layer to get the initial hidden layer state H_0; then H_0 is input to the graph neural network in the L layer to perform the feature transformation, and each layer of the graph neural network performs the corresponding feature transformation operation and takes the result as the input of the next layer of the graph neural network; after the processing of the L layer, the final hidden layer vector is obtained, and finally the hidden layer vector will be used as the classification vector. After L layers, the final hidden layer vector is obtained, and finally this hidden layer vector will be used as the input of the classification layer to complete the target classification.

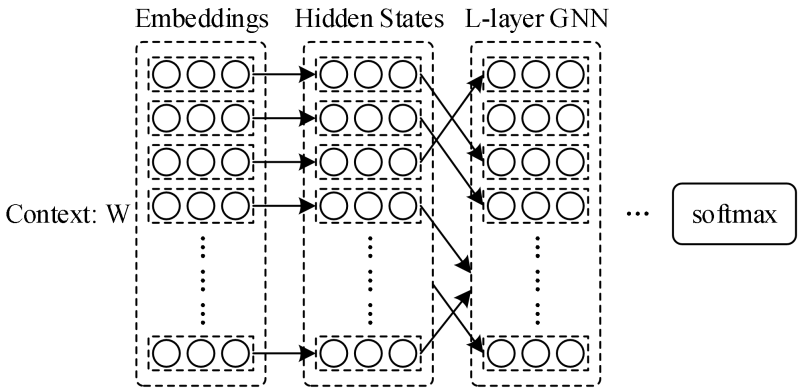

Figure 1. Graph neural network model

In summary graph neural networks differ from previous networks in that graph networks emphasize the use of structured relational information, so it learns richer feature vectors than LSTMs and CNNs,

and achieves good results in the processing of certain tasks.

2.2. Attention mechanisms

Attention mechanism is to judge the size of each feature's contribution to the output by calculating the similarity probability distribution function. Related scholars have proposed a Transformer network structure composed of an encoder-decoder structure, with Self-Attention mechanism as its core structure, which is based on the idea of Attention Mechanism, and assigns the corresponding weight to each input item by learning the interactions between the input contents. Each Transformer's encoder is composed of multiple identical layers, where each layer consists of a Self-Attention layer and two feed-forward layers composed of fully connected networks, and the Self-Attention layer mainly exchanges information about different inputs, and then obtains the weights of the different connections between the inputs and outputs in the same layer of the network, and from this, obtains the the output of each layer of the network model. Since the fully connected network cannot deal with sequences of unknown length, however, the self-attention mechanism can obtain the weights of different connections by establishing long-distance dependencies within the sequences.

Suppose the input sequence is $X=\left[X_1,X_2,X_3,\cdots,X_n\right]$, and meanwhile three matrices are randomly generated as W_Q,W_K,W_V , and the new vectors are obtained by multiplying each word vector by the three matrices, which are the query vectors Q , key vector K , value vector V , and the formula is as follows:

$$Q=W_Q\cdot X \tag{5}$$

$$K=W_K\cdot X \tag{6}$$

$$V=W_V\cdot X \tag{7}$$

$$Attention\left(Q,K,V\right)=V\cdot soft\max\left(\frac{K^T\cdot Q}{\sqrt{d_k}}\right) \tag{8}$$

where d_k is the length of the input vector X .

By combining deep learning with the attention mechanism can make the deep neural network focus on the information of important regions in the image, which also forms the visual attention mechanism. Generally speaking, visual attention mechanism generally includes soft attention mechanism and hard attention mechanism. Among them, soft attention is microscopically and deterministically attentive, which emphasizes more on regions or channels, and can quickly use network learning to generate corresponding weights, so it can learn by forward propagation and backward feedback to the neural network in order to obtain the attention weights. Hard attention, on the other hand, selects information by maximum sampling or random sampling, which focuses more on points, so the functional relationship between the final loss function and the attention distribution is not microscopic. Therefore, soft attention is usually used to replace hard attention.

The soft attention mechanism is generally divided into three major attention domains: spatial domain, channel domain hybrid domain. Among them, the spatial domain is to do spatial mapping of the spatial domain information in the image, to obtain the relevant information and mask scoring based on the space. Channel domain is to indicate the correlation between each channel and the key content by adding weights to each channel, if the weight value is higher, the higher the correlation is indicated, the method is to generate masks for the channels and score them, such as SENet, SKNet and so on. Hybrid domain model is a combination of spatial domain and channel domain attention, since the spatial domain based attention ignores the information in the channel domain, it is necessary to consider the image features in each channel at the same time, and this approach will limit the spatial domain transformation method to the original image feature extraction stage.

2.3. BERT model

In NLP, if the dataset is not large enough and you want to use a very good model, it is common practice to train the model on a large dataset and use that model as an initializer or feature extractor for similar tasks. Good quality models usually have their own training parameters available for fine-tuning by others, which saves time and computational resources and gives better results very quickly.

BERT pre-trains a deep Transformer bi-directional encoder on a large amount of text data to learn contextually relevant word vector representations. These word vector representations can then be used

as input features for downstream natural language processing tasks, such as question and answer, named entity recognition, and sentiment analysis, etc. BERT has achieved good results in a wide range of natural language processing tasks and pushed forward the development of the field of natural language processing, and the structure of its model is shown in Fig. 2.

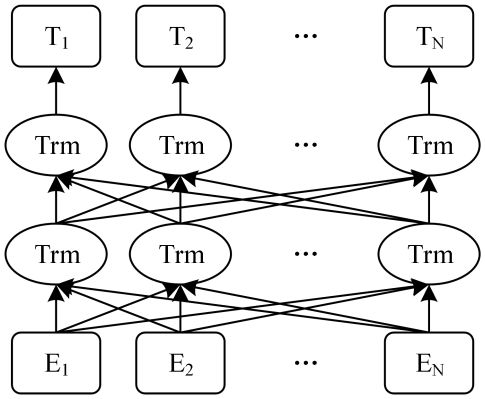

Figure 2. BERT model

The innovation of BERT is the use of Masked Language Modeling (MLM) and Next Sentence Prediction (NSP) for the pre-training task: the MLM allows BERT to simultaneously predict masked words using contextual information, while the NSP is designed to help the BERT model understand the relationship between two sentences.

To train the deep bidirectional representation, the BERT model uses the MLM pre-training task. In this task, the model randomly masks some tokens in the input sequence and then tries to predict these masked tokens. Specifically, 15% of the Word Piece tokens in each sequence are randomly masked during the training process.

However, this training approach has a drawback: it creates a mismatch between the fine-tuning phase and the pre-training. This is because $[MASK]$ tokens will not appear in the fine-tuning phase. Therefore, when generating the training data, for each selected token position, there is an 80% probability that it will be replaced with a $[MASK]$ token, a 10% probability that it will be replaced with a random token, and a 10% probability that it will remain unchanged. Finally the cross-entropy loss is used to predict the original token.

During the training of the BERT model, two pre-training tasks, NSP and MLM, are trained together. These two tasks together form a combined loss function and the goal of the model is to minimize this loss function.

3. multimodal-based impoliteness detection in social media interactions

In this paper, we conduct impoliteness detection research for graphic and textual bimodal data on social media, and propose an impoliteness detection model based on emotion-dependency graph convolutional neural network with modal fusion (ADGCN-MFM), and the structure of the model is shown in Fig. 3.

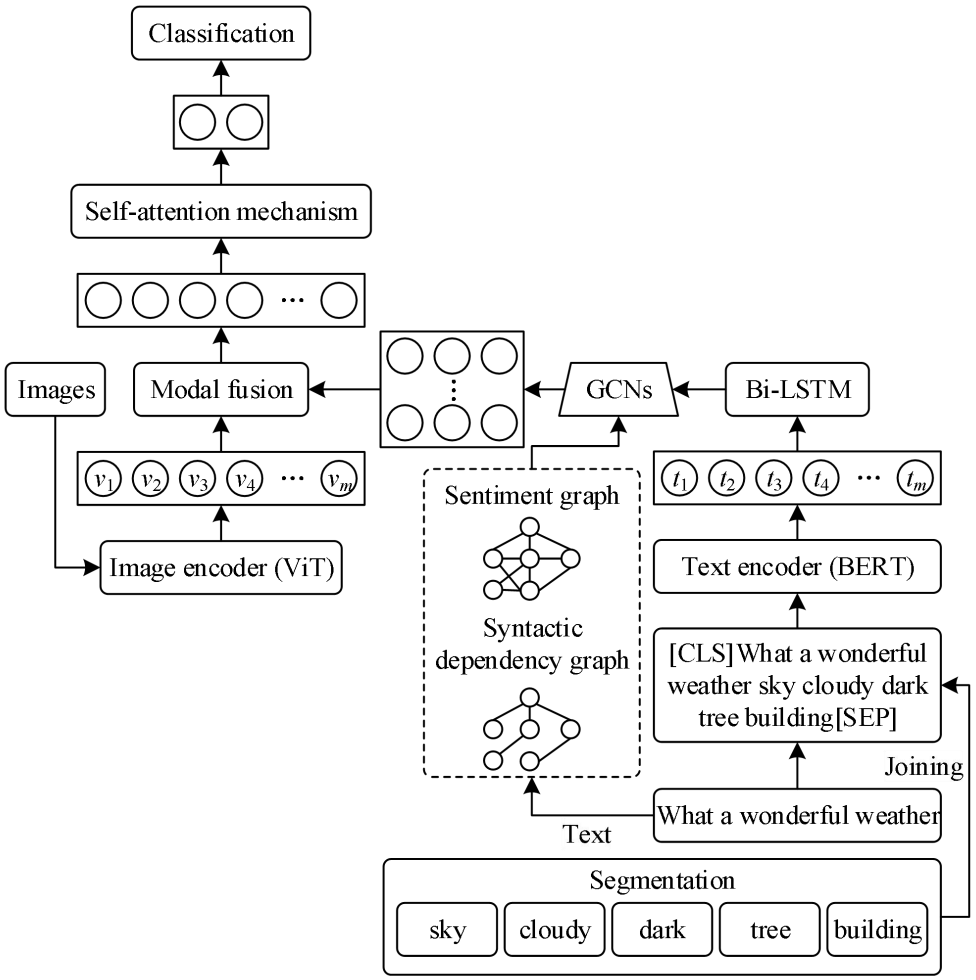

Figure 3. Structure of ADGCN-MFM Model

3.1. Coding layer

(1) Image coding

ViT is a model that does not use convolutional neural network but uses Transformer structure for image classification, in this paper we use ViT for encoding image features.ViT model slices the image into image blocks one by one and each block generates a vector.

The embedding of position information is similar to BERT's [class], where each image block contains the position information and the D -dimensional vector of the image after linear transformation, and the vector Z_i obtained after splicing the position information and the D -dimensional vector is used as the input to the Transformer structure, and the Encoder module of the Transformer consists of the Multihead Attention Mechanism (MSA) and Multilayer Perceptual Machine (MLP) layers. Layer normalization (LN) is applied before each block and residual joining is applied after each block.Transformer Encoder consists of the Encoder module repeated stacked L times, after Transformer Encoder and then layer normalization to get the image feature V_i , this computational process is shown in Equation (9)-Equation (11).

$$Z_i = \left[X_{class}; X_p^1 E; X_p^2 E; \cdots; X_p^N E \right] + E_{pos} \tag{9}$$

$$H_i = TE\left(Z_i\right) \tag{10}$$

$$V_i = LN\left(H_i\right) \tag{11}$$

where X_{class} denotes $[class]token$, $X_p^N E$ denotes N image blocks, E_{pos} is the position encoding, TE is the Transformer encoder, and LN is the layer normalization.

(2) Text encoding

① Expansion of text information

In order to enhance the model impolite reciprocity detection detection performance, the model in this paper considers the image attribute text extracted by ResNet101 as a part of the textual features, and splices the image attribute text after the corresponding text as a supplement and enhancement of the textual information. Through the above operation, the original text s is expanded to s', as shown in Equation (12) and Equation (13).

$$attr = \mathrm{Re}sNet\ 101(image) \tag{12}$$

$$s' = Concat[s; attr] \tag{13}$$

where $attr$ denotes the text of image attributes extracted from the image and the maximum length of the spliced text is 128.

② BERT word embedding

Input the spliced text s' in the form of $[CLS]w_1^c w_2^c, \cdots w_m^c, a_1^c, a_2^c, \cdots, a_s^c$, $[SEP]$ into the BERT pre-training model, the output of the last layer of the model is extracted as the word vector representation of the text, and this process is shown in Equation (14).

$$X_t = BERT([CLS]s'[SEP]) \tag{14}$$

where the text word vector has dimension 768 and $X_t \in \mathbb{R}^{128\times768}$.

3.2. Information Interaction Layer

(1) Information Interaction Based on Convolutional Neural Networks with Emotion-Dependency Maps

① Construction of Sentiment Map

In order to mine inconsistent sentiment expressions in sentences to identify contextual incongruity information and enhance the model's recognition ability for impolite reciprocity detection, this paper utilizes SenticNet, an external sentiment knowledge base, to obtain the sentiment scores of each word in the sentence, and then constructs an adjacency matrix based on the sentiment scores of each word $A^a \in \mathbb{R}^{n\times n}$, and the construction of the adjacency matrix element $A_{i,j}^a$ is shown in Equation (15).

$$A_{i,j}^a = \left| S(w_i^c) - S(w_j^c) \right| \tag{15}$$

where $S(w_i^c) \in [-1,1]$ denotes the corresponding sentiment score of the word, the value of the sentiment score is less than 0 to indicate a negative sentiment, and the value of the sentiment score is greater than 0 to indicate a positive sentiment, and the value of 0 is assigned to the word if it does not have a sentimental meaning in the sentence, i.e., $S(w_i^c) = 0$; and $|\cdot|$ denotes absolute value operation. By the above way, words with opposite sentiment are highly valued.

② Construction of syntactic dependency graph

Syntactic Dependency Tree expresses the whole sentence structure through the dependencies between words, and these dependencies express the semantic dependencies between the components of the sentence. The dependencies between all the words constitute a syntactic tree, and two words with specific syntactic relations can be obtained through the dependencies in the syntactic dependency tree. The construction of the neighbor matrix element $A_{i,j}^d$ in the syntactic dependency graph is shown in Equation (16).

$$A_{i,j}^d = \begin{cases} 1, & i \text{ and } j \text{ are related} \cup i = j \\ 0, & \text{Other} \end{cases} \tag{16}$$

where $A^d \in \mathbb{R}^{n\times n}$.

③ Contextual information representation

In the impolite reciprocity detection detection task, the output of the current moment is not only associated with the forward state of the current moment, but also may be associated with the backward state of the current moment. Therefore, the model in this paper introduces Bi-LSTM structure after BERT word embedding to perform deep feature extraction on the input text in order to more accurately capture the global feature information of the text context and make up for the defect that BERT

pre-training model is easy to forget the contextual information, and the computational process is shown in Eq. (17).

$$H = Bi - LSTM\left(X_t\right) \tag{17}$$

④ Sentiment-Dependency Based Graph Convolutional Neural Networks

In order to capture the parts of the text with contradictory sentiments while preserving the global structure of the sentence for impolite reciprocity detection detection, the model in this paper uses graph convolutional neural network (GCN) to acquire sentiment and syntactic features. The context-hidden representation H, the adjacency matrix A^a of the sentiment map and the adjacency matrix A^d of the syntactic dependency map are taken as inputs to the GCN, and since the construction of the sentiment map relies on the syntactic structure, the model of this paper inputs the syntactic dependency map into the first layer of the convolutional structure in the process of the graph convolution firstly, and the proposed sentiment- based dependent graph convolution network structure, the computational process is shown in Equation (18).

$$\begin{cases} g^l = \mathrm{Re}LU\left(\widetilde{A^a}\,\mathrm{Re}LU\left(\widetilde{A^d}g^{l-1}W_d^l + b_d^l\right)W_a^l + b_a^l\right) \\ \widetilde{A_i} = A_i / \left(E_i + 1\right) \end{cases} \tag{18}$$

where g^{l-1} is the output of the $l-1$ th layer of the node; $\widetilde{A_i}$ is the normalized adjacency matrix; W_d^l and W_a^l are the trainable weights of the l th layer of the GCN; and b_d^l and b_a^l are the biases of the l th layer of the GCN. Finally, the hidden representation of the L -layer GCN is obtained as the final output of the text $V_t = g^L$.

(2) Multimodal feature modal fusion

After obtaining text features and image features, with reference to the fusion method of Cai et al. this paper adopts modal fusion for the interactive fusion of text features and image features. Firstly, the lengths of text features and image features are converted into fixed-length vectors V_m', and then two-layer neural network is used to calculate the attention weights of text features and image features, and this process is shown in Eq. (19)-Eq. (21).

$$V_m' = \mathrm{Re}LU\left(W_{m1} \cdot V_m + b_{m1}\right) \tag{19}$$

$$\alpha = W_{m3} \cdot \mathrm{Re}LU\left(W_{m2} \cdot V_m + b_{m2}\right) + b_{m3} \tag{20}$$

$$\alpha_m = Soft\max\left(\alpha\right) \tag{21}$$

where $V_m \in \left\{V_i, V_t\right\}$; W_{m1}, W_{m2}, and W_{m3} are the trainable weight matrices; b_{m1}, b_{m2}, and b_{m3} are the bias terms; and ReLU activation function is used. Substituting image features V_i and text features V_t into the above formula, V_i' and V_t' as well as the attention weights α_i and α_t are obtained respectively, and weighting operation is performed on the weights and the feature vectors to obtain the final fusion vector V_{fused}, and this computational procedure is shown in equation (22).

$$V_{fused} = \alpha_i \cdot V_i' + \alpha_t \cdot V_t' \tag{22}$$

3.3. Classification prediction layer

In order to reduce the redundant information and noise interference, this paper introduces the self-attention mechanism after the fusion vector V_{fused} for reducing the interference of modal noise.

The self-attention mechanism is shown in Equation (23).

$$SelfAttention\left(Q, K, V\right) = Soft\max\left(\frac{QK^T}{\sqrt{d_K}}\right)V \tag{23}$$

where Q, K, and V are the query vector matrix, key vector matrix, and value vector matrix, respectively, and all three are mapped from the input matrix; $\sqrt{d_K}$ is the scaling factor used for

deflation, and the operation of QK^T results in an attention weight matrix. A self-attention mechanism is used on the fusion vector V_{fused} to obtain the final vector F, a process shown in Equation (24).

$$F = SelfAttention\left(V_{fused}, V_{fused}, V_{fused}\right) \tag{24}$$

The resulting vector F is fed into a fully connected network, and then a Sigmoid layer is used to obtain the probability distribution of the classification and generate the predicted impolite reciprocity detectability labels to accomplish the impolite reciprocity detection detection task.

4. Experimental results and analysis

4.1. Experimental setup

(1) Dataset Construction

Reply chains under controversial hot topics (e.g., political discussions, fan curses) were crawled from Weibo and Twitter between 2021 and 2023. The three sets of linguistics students were manually labeled:

Impoliteness labels: whether a reply contains at least one impoliteness tactic (negative comments, cursing, status devaluation, sarcasm).

Reciprocal labeling: if a reply is impolite and its parent post also contains impoliteness, it is labeled as impolite reciprocal.

Multimodal tagging: record whether text, exaggerated images, etc. were used.

The final result is 12847 conversation posts, of which 2380 are the number of complete conversations (at least 3 rounds of reciprocity), divided into training set, validation set test set according to the ratio of 6:2:2.

(2) Comparing Models

In order to verify that the proposed model in this paper can effectively improve the accuracy of impolite sentiment detection, the following two main types of benchmark models are selected as comparison models: unimodal model approach and multimodal model approach. Among them, unimodal models can be further divided into textual modal methods and image modal methods based on the type of modality used.

Unimodal Model Approach

Text modal methods

1) TextCNN is a CNN based deep learning for handling text categorization task.

2) Bi-LSTM is one of the most popular methods for solving many text classification problems, where a Bi-LSTM network is utilized to learn text features, which are then fed into a classification layer for impoliteness detection.

3) SIARN uses internal attention for text impoliteness sentiment detection.

4) SMSD uses self-matching network and low-rank bilinear pooling for text impoliteness sentiment detection.

5) BERT is a common pre-trained BERT model using “[CLS] text [SEP]” as input.

Image Modal Approach

6) Image uses a pre-trained ResNet-50 network to extract image features from the input image. The image features are then fed into a fully connected layer to compute the probability that the post contains impolite sentiment.

Multimodal Approach

7) Hierarchical Fusion Model (HFM) model is a hierarchical fusion model for multimodal sarcasm detection. Their model takes image features, image attribute features and text features as three modalities, reconstructs and fuses these three modal features and then uses them for prediction.

8) D&R Net model is a model for cross-modal comparison in multimodal sarcasm task. The model uses image features, text features, and adjective-noun pairs extracted from each image as the three modalities, and constructs a decomposition and relation network, the D&R network, to model cross-modal comparison and semantic associations.

(3) Experimental parameter settings

Like for feature extraction, the output dimension of the first linear layer in the image feature extractor is 1024, and the output dimension of the second linear layer is 512. in the text feature extractor, this paper adopts the ‘Bi-LSTM’ provided by the python library 'torch ' model for text feature extraction, the dimension of the hidden layer of the model is 256 and the output dimension is 512.

For the classifier the output dimension of the first layer is set to 512 and the output dimension of the second layer is set to 1. The word vector dimension of the text is set to 512. The maximum length of

the text is set to 75, if the length of the text is more than 75 then the later text is discarded and vice versa, zero is added.

The activation function used in the experiment is sigmoid activation function. During training, the batch_size is 32 and the number of epochs is 15. For the optimizer, the Adam optimizer is chosen in this paper. For the loss function, the binary cross entropy loss function is chosen in this paper.

4.2. Comparison of experimental results

Table 1 shows the experimental results of different models on the dataset, observing the results in the table, it can be found that the accuracy of the proposed model in this paper is better than all the unimodal baseline models.

The results of comparing the model proposed in this paper with textual modal methods in terms of accuracy show that the ADGCN-MFM based model proposed in this paper is better than all textual modal methods, which indicates that adding other modal information can effectively improve the accuracy of impoliteness detection in social media. Comparison of the experimental results shows that the method proposed in this paper gets a significant improvement in Precision, which is 4.93% better than the BERT model, which has the best effect among the textual modal methods, which indicates that the method proposed in this paper can better ensure the correctness of recognizing posts with impolite sentiments compared to the textual modal methods. The reason for analyzing the poor results of text modal may be because the user uses a completely positive or negative language to express the opposite meaning, and the text features alone cannot completely express the user's emotion, and more other information is needed to assist the detection.

Comparing with the image modality approach, the accuracy of Image is 20.00% lower than the ADGCN-MFM model, respectively.The precision, recall, and F1 score of Image are 29.30%, 9.45%, and 20.25% lower than the ADGCN-MFM model, respectively. The results of Image modal are much worse than the model proposed in this paper. The reason for analyzing the poor results of image modality may be due to the fact that separate image representations can generally only express some kind of direct emotional tendency, and it is difficult to embody impolite emotions, which makes it a very challenging task to extract the features related to impolite emotions from images.

Comparing with the multimodal approach, the accuracy of HFM and D&R Net is 1.29% and 0.64% lower than that of the ADGCN-MFM model, the recall of HFM and D&R Net is 84.14% and 83.34%, and the F1 scores of HFM and D&R Net are 1.57% and 1.09% lower than that of MMCAN. Observing the experimental results, it can be found that the overall effect of the ADGCN-MFM model proposed in this paper is better than that of the HFM and D&R Net models, and its precision is much higher than that of the other two multimodal models, which indicates that the ADGCN-MFM model recognizes posts with impolite sentiments with a higher correct rate. The model adopts Bi-LSTM and self-attention mechanism to better extract the impolite features and recognize the inconsistent information in the text.

Table 1. Experimental results of different models in the data set

Modal state	Method	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)
Text	TextCNN	80.07	74.37	76.35	75.34
	Bi-LSTM	81.91	76.71	78.37	77.47
	SIARN	80.46	75.58	75.71	75.62
	SMSD	80.87	76.48	75.14	75.84
	BERT	83.91	78.75	82.31	80.21
Image	Image	64.72	54.38	70.81	61.48
Text+ Image	HFM	83.43	76.49	84.14	80.16
	D&R Net	84.08	77.98	83.34	80.64
	Ours	84.72	83.68	80.26	81.73

4.3. Model Training Analysis

(1) Training cycle analysis

This section will explore the effect of training period on the ADGCN-MFM model, and Figure 4 shows the loss curve when the number of iterations is 200. In general, a shorter training period will make the model unable to comprehensively capture the hidden information in the data, resulting in the model not converging; a longer training period will make the model focus too much on the training data, resulting in the model overfitting the training data. In order to investigate the most suitable training period for the ADGCN-MFM model, the experiments in this section adjust the number of iterations of the model and draw the corresponding Loss curves while keeping the parameters constant.

The performance of the model improves faster between the 1st Epoch and the 10th Epoch, and the

Loss value decreases rapidly, which indicates that the ADGCN-MFM model is able to learn the relevant features of the impolite target from the dataset quickly during this period. The curve flattens out after the 10th Epoch, which indicates that the learning speed of the ADGCN-MFM model slows down, but the ADGCN-MFM still has some room for improvement, and it has not yet fully learned the feature information in the training set. Before the training cycle Epochs reaches the 74th cycle, the Loss value of the validation set tends to stabilize with only fluctuating changes, which indicates that the model is still learning the latent features in the training set during this cycle. After the training cycle Epochs exceeds the 74th cycle, the Loss curve of the validation set is increasing while the Loss of the training set is still decreasing, which indicates that the model has been overfitted at this point and should not be further trained.

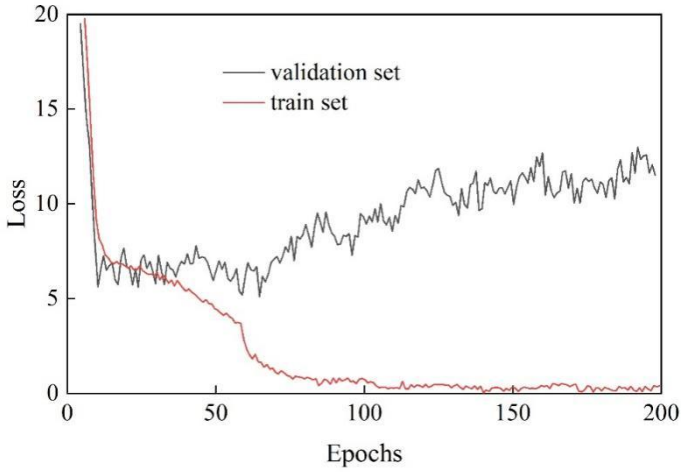

Figure 4. Model performance when epochs=200

(2) Training time analysis

This section analyzes the training time of the comparison model and the ADGCN-MFM model. The results of the training time comparison are shown in Table 2, and the EM accuracy is calculated as the number of samples in which the model successfully recognizes all the impolite targets divided by the total number of samples.

The impolite target recognition method that combines multidimensional features performs better than the sarcastic target recognition method that only focuses on impolite text and images, but also will have a longer training time, this is because it takes a lot of time to extract the multidimensional features of the background information and combine them with the information of impolite comments, due to the fact that the ADGCN-MFM model takes the longest time to train and has the best results. The training time of various pre-trained recognition models based on BERT is lower than that of the models proposed in this paper because these pre-trained models only need to be fine-tuned on the downstream task.The model of TextCNN is the fastest to train because it contains the least number of parameters but also the least effective.

The ADGCN-MFM model also achieves the best performance compared to the other models, although the BERT-based fine-tuning model has a large advantage in training time, but the focus of the satirical target recognition task is on the final F- and EM-values, and the gap in training time will continue to shrink as the hardware performance improves, making the training time for the text model acceptable. In the future, more lightweight and efficient information extraction methods and information perception methods can be adopted to shorten the model training time.

Table 2. The training time comparison

Method	F1-score (%)	EM	Training time (s)
TextCNN	75.34	71.87	111.1
Bi-LSTM	77.47	80.78	131.2
SIARN	75.62	77.98	122.5
SMSD	75.84	76.47	255.7
BERT	80.21	81.69	336.7
Image	61.48	65.12	279.6
HFM	80.16	83.13	2020.5
D&R Net	80.64	84.35	2258.7

Ours	81.73	84.63	2113.3

5. Analysis of the linguistic mechanisms of impolite reciprocity in social media interactions

5.1. Case Selection

In this study, two representative Chinese and English social media, microblogging (Sina Weibo) and Twitter, were selected to tag, classify and analyze the corpus according to the Culpeper impoliteness strategy as well as the newly emerged strategies using the ADGCN-MFM impoliteness detection model of this paper, to conduct a qualitative study, and on the basis of this study, to conduct a quantitative study of statistical comparison. Research.

The researcher chose similar communicative topics in both social media's-On April 9, 2017, several employees on a United Airlines flight violently dragged an Asian passenger off the plane, causing the passenger's face to bruise and bleed; passengers on the flight recorded the process and posted it in a video, which provoked a large amount of user discussion, which expressed indignation about the incident and condemned United Airlines. At the same time, the incident generated a lively discussion on Weibo. The context of similar topics is very suitable for a comparative study of impoliteness in Chinese and English online communication.

Under the same topic of Weibo and Twitter, 200 replies from each of the two types of users were selected for the study. After deleting the replies that were not suitable for the study of impoliteness strategies (e.g., non-Chinese comments in Weibo and non-English replies in Twitter, replies with ambiguous expressions, and replies unrelated to the topic, etc.), the corpus of this paper was determined to be the Weibo corpus (containing 128 replies) and the Twitter corpus (containing 161 replies) and labeled as W1-128 and T1-161 in order, respectively.

5.2. Case Studies

The frequency and proportion of impolite strategies used in both Chinese and English social media were counted, and the results of the frequency and proportion of impolite strategies used in the two corpora are shown in Table 3. The frequency and proportion of impolite strategies used by communicators in the Chinese and English corpora are similar in structure: the most used strategies are positive impolite strategies, followed by negative impolite strategies, and sarcastic, direct impolite, and profile impolite strategies are fewer in that order. The results of related scholars' studies also show that the use of positive impoliteness strategies in online communication is more than the use of negative impoliteness strategies. As a validation, the results of the chi-square test for the comparison of impoliteness strategies in the Chinese and English corpus ($\chi^2 = 0.801$, df = 4, Sig. = 0.935) show that there is no significant difference between the two kinds of online linguistic communication, which suggests that the results of the use of the impoliteness strategies here are similar in the Chinese and English corpus. In other words, communicators using Chinese and communicators using English tend to use similar impoliteness strategies (e.g., high-level strategies versus other levels) to achieve communicative purposes in similar contexts.

Table 3. The frequency of the use of impolite strategies in both kinds of language

Impolite strategy	**Weibo corpus**		**Twitter corpus**	
	Frequency	**Percentage**	**Frequency**	**Percentage**
Direct impolite	7	3.95%	10	4.15%
Positive and impolite	88	49.72%	115	47.72%
Negative behavior	55	31.07%	78	32.37%
Sarcasm	25	14.12%	35	14.52%
Indirect impolite	2	1.13%	3	1.24%
Total	177	100%	241	100%

In the use of impolite strategies, positive impolite strategies and negative impolite strategies accounted for the largest proportion of the use of impolite strategies, and these two high-level strategies contain many specific output strategies, so it is necessary to further compare the Chinese and English corpus impolite strategies at the level of the use of output strategies. A comparison of the frequency of positive impoliteness output strategies in the Weibo and Twitter corpora is shown in Figure 5. Comparing the microblogging corpus with the Twitter corpus, it can be seen that among the seven types of positive impoliteness strategies used in the two, the frequency of the use of each type of output

strategy is more different than the same. Although the proportion of positive impoliteness strategies among all impoliteness strategies is similar in the Chinese and English corpus, the specific output strategies are significantly different in terms of the frequency of the specific strategies actually used in online communication in the two languages. The results of the chi-square test ($\chi^2 = 44.859$, df = 6, Sig. = 0.000) also show that there is a significant difference between the two groups of data, i.e., Chinese-speaking communicators and English-speaking communicators use different positive impoliteness output strategies in similar online contexts.

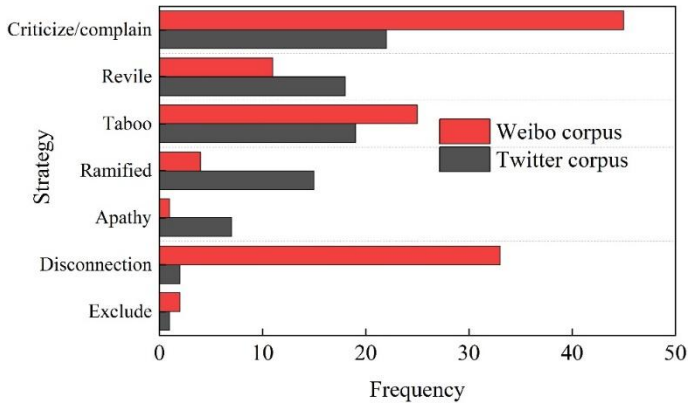

Figure 5. Use frequency comparisons with active impolite output policies

The frequency statistics of negative and impolite output strategies in Chinese and English corpora are shown in Figure 6. There are also differences in the frequency of use of negative and impolite strategies between Chinese and English corpora. For instance, in the Twitter corpus, the strategies of "contempt", "challenge/ question", and "negative emotion expression" are used more frequently than in the Weibo corpus, while in the Weibo corpus, the strategy of "imposing a negative image" is used more frequently than in the Twitter corpus.

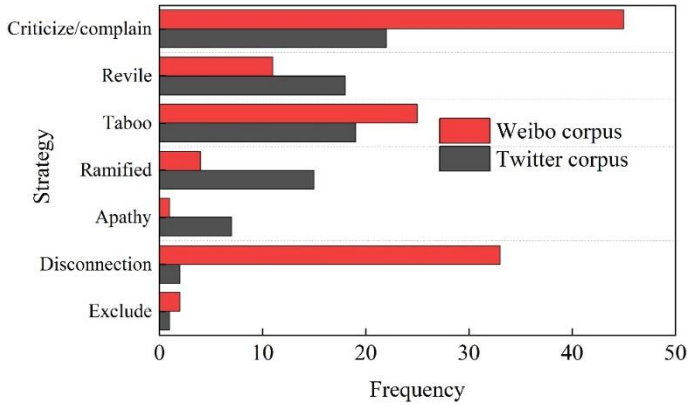

Figure 6. Frequency Statistics of Negative and Impolite Output Strategies in Chinese and English Corpora

5.3. Language Mechanisms

The strategy of "criticism/complaint" is not included in Culpeper's impoliteness model. However, this strategy appears frequently in both Chinese and English corpora. By drawing on the concept of criticism/complaint formulae proposed by Culpeper, this strategy can be defined as: the speaker deliberately criticizes or complains about the addressee or the hearer. The frequent occurrence of the "criticism/complaint" strategy is related to the topic of communication. The frequency of using this strategy in the English corpus is higher than that in the Chinese corpus, and the proportion it occupies in positive impoliteness strategies is also higher. Moreover, the proportion of this strategy being used in combination with other strategies as a usage pattern in the English corpus is also higher. This should also be related to the changes in language and culture among the vast English-speaking population in the contemporary world and in a multi-cultural and cross-cultural context when choosing impolite language to deal with global (racist) issues.

6. Conclusion

In this paper, we combine Culpeper's impoliteness strategy with a sentiment analysis model based on deep multimodal learning to explore the impoliteness reciprocity principle and linguistic mechanisms in social media interactions. Experiments are conducted on a self-constructed multimodal dataset, and compared with the multimodal approach, HFM and D&R Net have 1.29% and 0.64% lower accuracy than the ADGCN-MFM model, and the recalls of HFM and D&R Net are 84.14% and 83.34%, respectively, and the F1 scores of HFM and D&R Net are 1.57% and 1.09% lower than those of MMCAN. Observing the experimental results, it can be found that the overall effect of the ADGCN-MFM model proposed in this paper is better than that of the HFM and D&R Net models, and its precision is much higher than that of the other two multimodal models, which indicates that the ADGCN-MFM model recognizes posts with impolite sentiments with a higher correct rate. The experimental results prove the effectiveness and rationality of the model in this paper. Similar impolite strategies are used in Chinese and English online media postings, and the difference in the use of impolite strategies in Chinese and English corpus is reflected in the use of positive impolite strategies, which are closely related to contextual factors such as topic and object, and are significantly affected by the factors of communicative activity, purpose, tone, and communicative paradigm. The model in this paper has some limitations. On the one hand, due to the limitation of data resources, only one dataset is used to validate the model, and subsequent consideration is given to constructing a Chinese social media graphic dataset on its own, which is used to further validate the model's generalization and robustness. On the other hand, the main research of this paper is to improve the multimodal feature extraction, and in the future, we will focus on the modal information fusion aspect to improve the existing model.

Funding

This work was supported by “The Empowerment Mechanism of Internet Catchwords in Agricultural Product E-commerce Live Streaming in Heilongjiang Province” (Project No. 2026Y040), a General Project of the 2026 Heilongjiang Provincial Language Commission Language Research Project.

About the Author

Renjun Pan was born in Anhui, China, in 2002. He is a master’s student in Foreign Linguistics and Applied Linguistics at Northeast Agriculture University, China, having enrolled in the 2024 academic year. His research interests include pragmatics and discourse analysis. He has published a paper, which has been indexed by A&HCI.

Xiaodong Wang was born in Heilongjiang, China, in 1975. She is a Professor of School of Arts and Sciences at Northeast Agricultural University, China. Her research interests include business English and pragmatics. She has published more than 20 papers, 10 of which has been indexed by CSSCI, EI and A&HCI.

References

1. Iqbal, L., Safi, F., & Ullah, I. (2020). The use of symbols (emoticons) in social media: A shift of language from words to symbols. Global Mass Communication Review, 3, 124-135.
2. Zhang, K., & Yang, L. (2025). Analysis of Code-switching Phenomena in the Evolution of Social Media Language—Take the Mixed Use of Chinese and English by Gen Z as an Example. Journal of Humanities, Arts and Social Science, 9(3).
3. Li, X. (2024). Analysis on the popular phenomenon of network abbreviations. In SHS Web of Conferences (Vol. 183, p. 02024). EDP Sciences.
4. Teng, B. (2025, December). The Inclusive Expression and Root Analysis of Chinese Character Symbols in the Internet Age. In 2025 International Conference on Mental Growth and Human Resilience (MGHR 2025) (pp. 589-597). Atlantis Press.
5. Ren, H., & Cheng, X. (2024). A study on the construction of national image by the variation of network buzzwords from the per-spective of sociolinguistics. J. New Media Econ, 1(2), 133.
6. Bao, P. (2025). Generational linguistic conflict in digital fields: Bourdieu’s cultural capital theory and cross-cultural decoding hegemony of Z-generation internet slang. International Journal of Multilingualism, 1-24.
7. Wang, T., Huang, Y. X., & Xiao, Y. X. (2022). Translation methods of internet Buzzwords and their application. International Journal of Education and Humanities, 5(3), 141-145.

8. Brown, P., & Levinson, S. C. (1987). Politeness: Some universals in language usage (Vol. 4). Cambridge university press.
9. Brown, P., & Levinson, S. (2011). Politeness. The Cambridge encyclopedia of the language sciences, 635-636.
10. Sapitri, P. A., Chasanah, A., Putri, A. A., & Paulima, J. (2019). Exploring Brown and Levinson's politeness strategies: An explanation on the nature of the politeness phenomenon. REiLA: Journal of Research and Innovation in Language, 1(3), 111-117.
11. Jansen, F., & Janssen, D. (2010). Effects of positive politeness strategies in business letters. Journal of pragmatics, 42(9), 2531-2548.
12. Brown, R. (2014). Politeness theory: Exemplar and exemplary. In The legacy of Solomon Asch (pp. 23-38). Psychology Press.
13. Al-Duleimi, H. Y., Rashid, S. M., & Abdullah, A. N. (2016). A Critical Review of Prominent Theories of Politeness. Advances in Language and Literary Studies, 7(6), 262-270.
14. Pishghadam, R., & Navari, S. (2012). A study into politeness strategies and politeness markers in advertisements as persuasive tools. Mediterranean Journal of Social Sciences, 3(2), 161-171.
15. Fathi, S. (2024). Revisiting Brown and Levinson's theory of politeness. European Journal of Language and Culture Studies, 3(5), 1-11.
16. Culpeper, J., Oliver, S. J., & Tantucci, V. (2021). Politeness reciprocity in Shakespeare's dialogue: The case of thanks. Journal of Historical Pragmatics, 22(2), 202-224.
17. Meiratnasari, A., Wijayanto, A., & Suparno, S. (2019). An analysis of politeness strategies in Indonesian English textbooks. ELS Journal on Interdisciplinary Studies in Humanities, 2(4), 529-540.
18. Sorlin, S. (2017). The pragmatics of manipulation: Exploiting im/politeness theories. Journal of Pragmatics, 121, 132-146.
19. Culpeper, J., & Tantucci, V. (2021). The principle of (im) politeness reciprocity. Journal of Pragmatics, 175, 146-164.
20. Zhao, K., Ferguson, E., & Smillie, L. D. (2017). Politeness and compassion differentially predict adherence to fairness norms and interventions to norm violations in economic games. Scientific Reports, 7(1), 3415.
21. Zhao, L. (2025). Revisiting respect and politeness: Insights from metapragmatics of zunzhong in Chinese public spaces. Journal of Pragmatics, 240, 109-121.