A Dataset and BenchMark Models for Predicting Human Intentions Expressed in Text-Image Posts in Chinese Online Mental Health Communities

He Zhendong*†
School of Artificial Intelligence
Sun Yat-sen University
Zhuhai, Guangdong, China
hezhd6@mail2.sysu.edu.cn

Huang Zhupeng*†
School of Artificial Intelligence
Sun Yat-sen University
Zhuhai, Guangdong, China
Huangzhp36@mail2.sysu.edu.cn

Xie Yitong*
School of Artificial Intelligence
Sun Yat-sen University
Zhuhai, Guangdong, China
xieyt59@mail2.sysu.edu.cn

Li Feiyu*
School of Artificial Intelligence
Sun Yat-sen University
Zhuhai, Guangdong, China
2416008282@qq.com

He Zongtan*
School of Artificial Intelligence
Sun Yat-sen University
Zhuhai, Guangdong, China
hezt7@mail2.sysu.edu.cn

Peng Zhenhui
School of Artificial Intelligence
Sun Yat-sen University
Zhuhai, Guangdong, China
pengzhh29@mail.sysu.edu.cn

Abstract

Online mental health communities (OMHCs) offer users a public place to exchange social support via posts with text and images. Previous work on modeling posts in OMHCs largely focuses on the sought social support of text-only posts or the features of text-image posts in Western context but overlooks the text-image posts in Chinese context. In this paper, we label 973 text-image posts from three Chinese Baidu Tieba OMHCs with four motivation-based (*e.g.,* sharing, offering help) and nine content-based posting intentions (*e.g.,* personal growth, money) adapted from a human motives taxonomy, and we selected 931 entries as the training data. We then provide benchmark models for predicting each intention given the post's text and image as input. The attentional neural network models have the best performance, with F1 scores above 0.7 in all the 13 binary classification tasks. Our work offers a starting point for understanding and modeling multi-modal content in Chinese OMHCs.

CCS Concepts

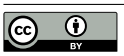
• **Information systems** → **Content analysis and feature selection**; • **Computing methodologies** → **Neural networks**.

Keywords

Intention categories, Mental health, Tieba

ACM Reference Format:
He Zhendong, Huang Zhupeng, Xie Yitong, Li Feiyu, He Zongtan, and Peng Zhenhui. 2024. A Dataset and BenchMark Models for Predicting Human Intentions Expressed in Text-Image Posts in Chinese Online Mental Health Communities. In *Chinese CHI 2024 (CHCHI 2024), November 22–25, 2024, Shenzhen, China.* ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/3758871.3758929

*These authors are from an undergraduate research team
†Both authors contributed equally to this research



CHCHI 2024, Shenzhen, China

ACM ISBN 979-8-4007-1389-7/24/11
https://doi.org/10.1145/3758871.3758929

1 Introduction

Online mental health communities (OMHCs) offer a convenient place for people to share their mentally challenging issues and seek social support via text and images in thread-starting posts. For example, in the Depression Tieba (Yiyuzheng ba in Chinese), an OMHC with 470K members and over 25 million posts as of August 2024, people can pour out their feelings and exchange social support with each other. Existing work on human-computer interaction (HCI) has modelled the community members' various intentions of creating posts in OMHCs [9, 20, 25], which can help to understand communication patterns in OMHCs and design intervention mechanisms to promote members' mental health. For example, Peng et al. [20] built text classifiers to predict the amount (*i.e.,* small, medium, large) of emotional and informational support that a post seeks in Reddit r/depression. Kushner and Sharma [9] compiled user-specified tags of posts from the OMHC Talklife, which include behavioral purposes such as "relationships", "my story", and "self-harm", as well as emotional states like "sad", "frustrated", and "positivity". However, these works largely focus on modeling posters' intentions from posts in the Western context, while overlooking the specific challenges and characteristics of text-image posts in the Chinese context.

In this paper, we seek to understand and model the different intentions of creating text-image posts in Chinese OMHCs. We first sampled 973 text-image posts from the Depression Baidu Tieba and conducted a crowd-sourcing task to label the posters' intentions referring to the human motives taxonomy [22]. The labeled intentions include four categories about motivation (*i.e.,* sharing, venting, seeking help, and offering help) and nine categories about content (*i.e.,* social, financial, social activities, daily life, growth or personal development, culture or communication, psychological state, physical state, and family) With this labeled dataset, we trained multiple classic machine learning models (*e.g.,* KNN, random forest, decision tree) for binary classification tasks using the textual and visual features of text-image posts as input. The random forest models

perform the best in classifying each human intention conveyed through the text-image posts, with F1 scores of 0.80, 0.81, 0.88, and 0.94 for the four motivation-based categories, and F1 scores of 0.75, 0.95, 0.90, 0.70, 0.91, 0.84, 0.62, 0.83, and 0.86 for the nine content-based categories. We also trained neural network models with attention mechanisms for classifying each human intention given the vectorized text and image of the posts as input. These models also perform well, with F1 scores of 0.85, 0.83, 0.87, 0.93 for the four motivation-based categories, and 0.79, 0.92, 0.91, 0.76, 0.94, 0.84, 0.72, 0.84 and 0.92 for the nine content-based categories, respectively.

In summary, we contribute a dataset of human intentions of creating text-image posts in Chinese mental health communities and benchmark computational models to predict the human intentions exhibited in text-image posts.

2 Related Work

Online mental health communities (OMHCs) provide a convenient avenue for users facing health issues to seek and provide social support, which can positively impact their mental health [18]. For example, Rains and Young [21] discovered that long-term participation in these communities can reduce depression, improve quality of life, and enhance self-management of health. However, on the downside, Moorhead et al. [17] pointed out the potential risks associated with the credibility of online health information, which can be incorrect or even harmful. Additionally, comments that lack understanding or respect for others frequently appear, leading to user withdrawal from online forums and further emotional isolation [19]. Previous researchers have explored various factors, *e.g.,* community knowledge [20] and sentiment and topics of the help-seeking posts [12], that may affect the quality of comments that posters received in OMHCs. These prior works highlight the benefits of understanding and predicting the posters' intentions conveyed in their posts for promoting positive community interactions and reduce potential negative impacts.

Existing literature on modeling the posts in OMHCs has primarily focused on the types of social support that the community members seek or provide [11, 12, 20] or the topics and sentiments of the posts [3, 12]. For example, Li et al. [11] analyzed posts in a Chinese depression support community Tulip Sunshine Forum and interviewed forum users [11]. Chen et al. [3] used measurements of eight basic emotions as features of Twitter posts over time to identify users who have depression or are at risk of developing depression. As a closely related work, Li et al. [12] analyzed the text-image posts in Reddit r/GriefSupport. They built computational models to examine the effects of textual, visual, and text-image coherence features of a text-image post on its received social support [12]. We followed this line of literature to label data and build computational models. Different from previous work, we focus on a distinct set of human intentions expressed in the text-image posts and modeling these intentions in Chinese OMHCs.

As the targeted posts contain both text and images, we sought inspirations from related work on multimodal data modeling. For example, multimodal fusion enables machines to extract and integrate information from various sources such as text, images, and audio, thus improving model performance [1]. Zhang et al. [26] introduced both intra-attention and inter-attention mechanisms into a neural network prediction model for the popularity of social media images. Liang et al. [13] proposed a graph model to explore the intra-modal and cross-modal heterogeneity of post images and text, which further improved the accuracy of sarcasm recognition in their dataset [13]. Cai et al. [2] collected Twitter text-image posts with sarcasm emotion labels and extracted image vector representations obtained through the ResNet model, image attribute labels, and text features embedded using GloVe. They then fused these three modalities to predict whether a post expressed user sarcasm [2]. In this work, we also provide benchmark neural network models that fuse textual and visual features for predicting human intentions of creating text-image posts in OMHCs.

3 Data Collection and Labeling

3.1 Research Sites

Our research team conducted an extensive search on the online mental health communities (OMHCs) in which members actively create text-image posts in Baidu Tieba, a popular platform that holds numerous online communities in China. After the search process, we narrowed down our focus on three OMHCs, *i.e.,* Depression Bar, Psychology Bar, and Paradise Chicken Soup Bar, all of which have over 200,000 members as of August, 2024. We collect publicly available posts created between May 2012 to May 2024 from three OMHCs via aiotieba API [16]. First, we filtered out the posts containing images and text, and then randomly selected 1,000 posts. Then we removed the posts that had been deleted, since we cannot track their community knowledge. We also eliminated advertisement posts, because they cannot reflect the psychological status of people in psychological community. After pre-processing, we were left with a total of 973 posts containing images and text: 663 posts from Depression Bar, 183 posts from Psychology Bar, and 127 posts from Paradise Chicken Soup Bar.

3.2 Labels

To compile a nuance set of human intentions of creating text-image posts in OMHCs, we referred to the human motives taxonomy identified in previous studies [22]. This taxonomy compiles 161 human motivations of taking actions and has been applied across various domains of psychological and behavioral studies. Jia et al. [8] adapted this taxonomy to manually annotate photos from the website Unsplash with 28 human intent categories. As suggested by Jia et al. [8], not all 161 human motives in this taxonomy are applicable to the text-image posts in OMHCs. For example, the original set of 161 motives primarily focuses on positive or neutral psychological motives, such as ambition, curiosity, or altruism. However, negative psychological motives — such as sadness, anxiety, frustration, or depression — are frequent and central in spaces like Depression Bar or Psychology Bar.

To start with, three authors of our research team familiarized themselves with the original 161 motives. They then independently annotated 50 randomly sampled text-image posts from the collected dataset, adopting the 161 motives or adding extra labels if applicable. They met and discussed their labels, identifying agreeable labels and resolving the conflicts. Next, they repeated this process twice,

Table 1: Labels used in annotating the human intentions expressed in the text-image posts in Chinese online mental health communities.

Category	Label	Explanation	Num of posts	ICCs
Motivation-based	Sharing	Positive sharing of events or emotions, such as sharing interesting daily life experiences.	337	0.781
	Pouring out	Neutral or negative expression of events or emotions, such as complaints.	302	0.803
	Asking for help	Request for assistance or answers to questions.	191	0.890
	Offering help	Provision of substantial help, whether emotional or rational, such as comforting others or solving problems.	101	0.773
Content-based	Socializing	Interaction with others, including friends, strangers, or romantic partners, such as teasing or sharing relationship dynamics.	199	0.705
	Money	Emphasis on the use, need, or value of money, such as purchasing expensive items.	32	0.740
	Social activities	Engagement in official or social activities, such as work, community involvement, or forum discussions, without focusing on personal relationships.	84	0.663
	Daily life	Reflection on individual daily life events, including diet, sleep, pets, and travel.	297	0.740
	Personal growth	Emphasis on personal progress or stagnation, including educational achievements and determination for improvement.	60	0.607
	Cultural or communicative aspects	Content related to culture, including poetry, technology, festivals, and advertisements.	132	0.731
	Psychological state	Emphasis on mental health and emotional well-being.	500	0.754
	Physical state	Reference to current physical health conditions.	136	0.839
	Family-related themes	Involvement in family relationships and dynamics, including interactions with family members.	94	0.858

first for another 50 randomly sampled posts and then for another 199 posts. After that, they reached an agreement on the appropriate labels for the human intentions expressed in the text-image posts in these Chinese OMHCs. Table 1 summarizes the 13 identified labels. Four labels are motivation-based, *i.e.,* sharing, pouring out, asking for help, and offering help, which encapsulate the core reasons why individuals participate in such spaces. The rest nine labels are content-based, *i.e.,* socializing, money, social activities, daily life, personal growth, cultural or communicative aspects, psychological state, and family-related themes, which reveal the topics of their text-image posts.

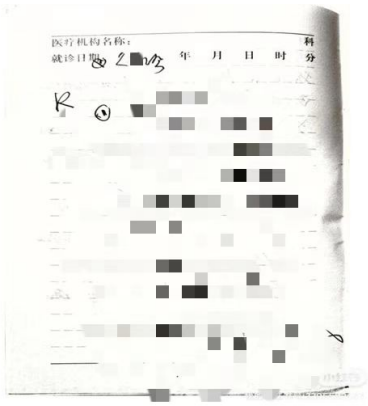
3.3 Annotation Process

After reaching an agreement on the labeling scheme, the three authors independently applied the scheme in a labeling interface to 299 sampled posts used in deriving the labels. We developed the labeling interface customized to our tasks. Each annotator will get the text-image posts one by one in a shuffled order. For each post, the annotator needs to assign only one motivation-based label that the poster mainly conveys and select one, two, or three content-based labels based on the topics represented in the text and images. This multi-layered approach enabled us to capture both the psychological motivations driving each post and the specific life domains to which these motivations were related. For the rest posts in the dataset, we invited 20 (19 males, 1 females; age: mean = 19.95, SD = 0.86) student annotators from a local university. Before commencing the annotation tasks, each annotator received detailed instructions on our labeling schema, including examples illustrating how to apply the motivation-based and content-based labels across different types of posts. During the annotation phase, the platform presented participants with text-image posts one by one from the dataset. Participants were tasked with selecting the labels they deemed most appropriate.

After the labeling processes conducted by our three authors and 20 student annotators, we have a label dataset of 973 text-image posts, each of which was annotated by exactly three annotators. We calculated ICCs values on raw data to verify annotator consistency. For each post, we aggregated label counts from the three annotators: the Motivation-based label with the highest count was assigned, while Content-based labels were included if selected by at least two annotators. Given the diversity and complexity of each post, a strict standardized approach was not applied to ambiguous or missing tags. Instead, annotators made judgments based on guidelines and context, with ICCs ensuring overall consistency. This result in some posts lacking labels. Thus, we ultimately have 931 labeled entries remaining. Table 1 shows the intraclass correlations (ICCs) for each label, all of which are above 0.60, indicating a substantial agreement level.

[redacted]，今年高考结束。现在没有目标，不知道学什[redacted]。有点迷茫有一个嘴硬的妈妈和一个很老实又有点窝囊的爸爸，家庭小康，没什么严重心理问题。但是感觉自己深受原生家庭影响，[redacted]从小到大我想要什么，他们就没有给我顺顺溜溜的买过，但是我又不是完全内向那种，我想改变，我想遇见更好的自己。希望大家给我一些建议谢谢

(a) Asking for help, personal growth and family-related themes



家长求助，可以帮忙看下这个医生写的啥吗？孩子的[redacted]医生写的，谢谢了，我真的很想知道写了啥

(b) Asking for help and cultural or communicative aspects

听轻音乐听哭我承认自己的病情，不再逼自己工作多出色，不再过分伤害自己。 可病症出来，就是坐着都难受不想说话，不想工作。生活所迫，我需要在同事，在客户面前装得像个正常人，[redacted]什么时候是个解脱？

(c) Pouring out and psychological state

Figure 1: Sample posts and their labels

4 Benchmark Models for Classifying Posting Intentions

4.1 Classic Machine Learning Models

For each text-image post in Chinese online mental health communities (OMHCs), we formulate multiple binary classification tasks. Specifically, following [12, 20] we train binary classifiers to predict whether a post expresses intentions that fall into each of the 13 labels (Table 1). The input features for each classic machine learning models are:

Features about the post's text: We employed a sentiment lexicon provided by CNKI, which categorizes emotional words into six main groups: propositions (further divided into "perception" and "regard"), positive emotions, positive evaluations, negative emotions, negative evaluations, and degree adverbs (subdivided into "extremely/most", "very", "more", "-ish", "insufficiently", and "over"). For each post, we designed a 12-dimensional feature vector based on the frequency of these words: [number of proposition "perception" words, number of proposition "regard" words, number of positive emotion words, number of positive evaluation words, number of degree "extremely/most" words, number of degree "very" words, number of degree "more" words, number of degree "-ish" words, number of degree "insufficiently" words, number of degree "over" words, number of negative emotion words, number of negative evaluation words]. In practice, we used the SnowNLP [23] library to convert any traditional Chinese characters to simplified Chinese and utilized the thulac Guo [5] library for Chinese word segmentation, thereby facilitating the counting of words from different categories.

Features about the post's image: We focused on color features, texture features, and subject position features, as inspired by Guo et al. [4].

- Color features: In psychology, previous models have linked colors in images to emotional expressions [6]. For instance, Jafar and Shaikh [7] used common colors to identify the emotions conveyed by images. In our study, we employed basic HSV values, including the proportion of pixels for 10 fundamental colors, brightness levels (very low, low, medium, high, very high), saturation levels (low, medium, high), and hue levels (warm, cool).
- Texture Features: Tamura texture features play a significant role in capturing image emotion [15]. We used the skimage library to extract texture features and analyzed energy, contrast, homogeneity, and correlation for the H, S, and V channels. Energy reflects the uniformity of pixel intensity distribution, with higher values indicating richer texture. Contrast measures the difference in intensity between adjacent pixels, with higher contrast indicating more noticeable texture changes. Homogeneity indicates the similarity between adjacent pixels, with higher values representing smoother pixel variations. Correlation refers to the linear relationship between adjacent pixels of different intensity levels, with higher correlation suggesting stronger linear relationships. Additionally, we incorporated wavelet features by applying a three-level Haar wavelet transform to the HSV channels.
- Subject Position Features: The position of the main subject in an image can reveal a user's emotional state. For instance, a subject positioned slightly to the left might convey unease or introspection, while a rightward position could indicate

positive or proactive emotions [10]. We standardized the image sizes to facilitate comparisons of subject position. For subject recognition, we applied Canny edge detection and computed the x and y coordinates of the image's center of gravity.

Following Guo et al. [4], Peng et al. [20], we trained five classic machine learning models, including K-Nearest Neighbors (KNN), Multilayer Perceptron Classifier (MLP), Support Vector Classifier (SVC), Decision Tree (DT), and Random Forest (RF), to predict whether a text-image post expresses each of the 13 intentions. These models were implemented using the scikit-learn library, taking the extract textual and visual features as input. To evaluate model performance, we adopted Precision, Recall, and F1-score as metrics [24]. The dataset was randomly split into a training set (80%) and a validation set (20%), with 10-fold cross-validation used to determine model hyperparameters. Table 2 shows the results of 13 binary classifiers based on classic machine learning models.

KNN and MLP demonstrate average performance, while SVC exhibits near-perfect recall across multiple labels. Among the classic machine learning models evaluated, the random forest achieves the highest precision and F1 score for the majority of the labels.

4.2 Neural Network Models with Attention Mechanisms

The goal of the attention mechanism is to selectively focus on the more critical information in the task at hand, based on other available information. We designed a neural network model that incorporates an attention mechanism. As shown in Figure 2, we employed an Intra-attention module to capture the relations between images and text.

Text embedding: The text embedding is obtained by a pretrained Chinese-base-BERT model. From this, we derived the original textual vector $\mathbf{x_{original}} \in \mathbb{R}^{L\times 768}$ where L is the length of the text. Then, we processed its shape to obtain a vector $\mathbf{x} \in \mathbb{R}^{50\times 768}$. Specifically, for text representations with fewer than 50 tokens, we applied linear interpolation; for those with more than 50 tokens, we used average pooling.

Image embedding: The image embedding is obtained by a pretrained ResNet50 model. We first resized the images to 256×256 and then removed the final average pooling layer and fully connected layer from ResNet50. We divided each image into 49 regions, thereby obtaining a vector $\mathbf{y_{original}} \in \mathbb{R}^{7\times 7\times 2048}$ and $\mathbf{y} \in \mathbb{R}^{49\times 2048}$ after reshaped to represent the image features.

Compute Attention: Let the text vector representation be $\mathbf{x} = [x_1, \ldots, x_i, \ldots, x_{50}]$, where $\mathbf{x_i} \in \mathbb{R}^{1\times 768}$, and the image vector representation be $\mathbf{y} = [y_1, \ldots, y_i, \ldots, y_{49}]$, where $\mathbf{y_i} \in \mathbb{R}^{1\times 2048}$. Firstly, we applied average pooling to the first dimension of the text vector and image vector to obtain $\bar{\mathbf{x}} \in \mathbb{R}^{1\times 768}$ and $\bar{\mathbf{y}} \in \mathbb{R}^{1\times 2048}$. To compute visual attention, we used the following equations:

$$\mathbf{a_i} = \tanh(\mathbf{y_i} \cdot \mathbf{W_t}) \odot \tanh(\bar{\mathbf{x}} \cdot \mathbf{W_s}) \tag{1}$$

$$\mathbf{a} = \mathrm{ELU}([\mathbf{a_1}, \ldots, \mathbf{a_{49}}] \cdot \mathbf{W_a}) \tag{2}$$

$$\mathbf{b} = \mathrm{ELU}(\mathbf{a} \cdot \mathbf{W_b}) \tag{3}$$

$$\mathbf{h} = \mathrm{ELU}(\mathbf{b} \cdot \mathbf{W_h}) \tag{4}$$

$$\alpha_i = \mathrm{Softmax}(\mathbf{h}) \tag{5}$$

$$\mathbf{y_{attention}} = \sum_{i=1}^{49} \alpha_i \cdot \mathbf{y_i} \tag{6}$$

where $\mathbf{W_t} \in \mathbb{R}^{2048\times 1024}$, $\mathbf{W_s} \in \mathbb{R}^{768\times 1024}$, $\mathbf{W_a} \in \mathbb{R}^{1024\times 512}$, $\mathbf{W_b} \in \mathbb{R}^{512\times 256}$ and $\mathbf{W_h} \in \mathbb{R}^{256\times 1}$. The tanh activation function, which handles negative inputs and bounds the output between [-1, 1], is useful for computing similarity. The final similarity matrix is used to compute the attention score and we can use attention score to get the final attention output vector $\mathbf{y_{attention}}$.

Through a process similar to the one described for image attention computation, we can obtain the final textual attention using the following equations:

$$\mathbf{a_i} = \tanh(\mathbf{x_i} \cdot \mathbf{W_t}) \odot \tanh(\bar{\mathbf{y}} \cdot \mathbf{W_s}) \tag{7}$$

$$\mathbf{a} = \mathrm{ELU}([\mathbf{a_1}, \ldots, \mathbf{a_{50}}] \cdot \mathbf{W_a}) \tag{8}$$

$$\mathbf{b} = \mathrm{ELU}(\mathbf{a} \cdot \mathbf{W_b}) \tag{9}$$

$$\mathbf{h} = \mathrm{ELU}(\mathbf{b} \cdot \mathbf{W_h}) \tag{10}$$

$$\alpha_i = \mathrm{Softmax}(\mathbf{h}) \tag{11}$$

$$\mathbf{x_{attention}} = \sum_{i=1}^{50} \alpha_i \cdot \mathbf{x_i} \tag{12}$$

where $\mathbf{W_t} \in \mathbb{R}^{768\times 1024}$, $\mathbf{W_s} \in \mathbb{R}^{2048\times 1024}$, $\mathbf{W_a} \in \mathbb{R}^{1024\times 512}$, $\mathbf{W_b} \in \mathbb{R}^{512\times 256}$ and $\mathbf{W_h} \in \mathbb{R}^{256\times 1}$.

Intention Predictions: We concatenated the image attention output and the text attention output to form a new vector $\mathbf{v} = [x_{attention_1}, \ldots, x_{attention_{50}}, y_{attention_1}, \ldots, y_{attention_{50}}]$. This vector was then passed through a fully connected layer, batch normalization layer, ELU activation function, and dropout layer to accelerate convergence and reduce over-fitting. The output logits were used for cross-entropy calculation.

Training Settings: We split 931 labeled text-image posts into a training set(80%) and a validation set(20%). The model was trained using the ADAM optimizer, with L2 regularization and early stopping to prevent overfitting. The batch sizes for the training set and validation set were set to 16 and 32, respectively. Additionally, precision, recall, and F1-score were used as evaluation metrics. Table 2 shows the results of 13 binary classifiers based on the neural network models with attention mechanisms.

We can see that such models, though lack of interpretability, outperform those classic machine learning models in terms of precision, and demonstrate superior F1 scores across most labels.

5 Discussion and Conclusion

In this paper, we contribute a dataset of text-image posts with a set of 13 labeled human intentions in Chinese online mental health communities (OMHCs). We designated three annotators per post to achieve a balance between annotation efficiency and a multi-perspective understanding of each data point while ensuring effective resource allocation. To assess the consistency in annotators' understanding of the labels, we decided to calculate the Intraclass Correlations (ICCs) for each label upon completion of the annotation process. We provide benchmark models, including the classic ones like random forests and the neural networks with attention mechanisms, to predict whether a text-image post expresses each of the 13 human intentions. There are several limitations that call for future work. First, the labeled dataset is relatively small, and future work could seek to label more text-image posts to improve

Table 2: The individual performance metrics of each model across distinct categorical labels.

Label	KNN			SVC			MLP			DT			RF			Attention Neural Network		
	Precision	Recall	F1	Precision	Recall	F1	Precision	Recall	F1	Precision	Recall	F1	Precision	Recall	F1	Precision	Recall	F1
Sharing	0.61	0.63	0.62	0.67	**0.99**	0.80	0.71	0.78	0.74	0.69	0.78	0.73	0.69	0.94	0.80	**0.86**	0.84	**0.85**
Pouring out	0.59	0.61	0.60	0.68	**0.99**	0.81	0.68	**0.99**	0.81	0.69	0.89	0.78	0.68	**0.99**	0.81	**0.84**	0.83	**0.83**
Asking for help	0.70	0.77	0.73	0.79	**0.99**	**0.88**	0.81	0.94	0.87	0.81	0.96	**0.88**	0.80	**0.99**	**0.88**	**0.90**	0.85	0.87
Offering help	0.80	0.89	0.84	0.89	**0.99**	**0.94**	0.90	0.95	0.92	0.89	0.96	0.92	0.90	**0.99**	**0.94**	**0.95**	0.91	0.93
Socializing	0.69	0.79	0.74	0.70	0.76	0.73	0.70	0.74	0.72	0.70	0.73	0.71	0.70	0.80	0.75	**0.77**	**0.81**	**0.79**
Money	0.93	**0.97**	**0.95**	0.94	0.96	**0.95**	0.94	0.96	**0.95**	0.93	0.95	0.94	0.94	**0.97**	**0.95**	**0.95**	0.89	0.92
Social activities	0.86	0.90	0.88	0.88	0.91	0.89	0.89	0.91	0.90	0.86	0.91	0.88	0.88	**0.92**	0.90	**0.94**	0.89	**0.91**
Daily life	0.62	0.63	0.62	0.72	0.72	0.72	0.66	0.66	0.66	0.64	0.64	0.64	0.70	0.71	0.70	**0.79**	**0.74**	**0.76**
Personal growth	0.88	**0.94**	0.91	0.88	**0.94**	0.91	0.88	0.92	0.90	0.89	0.92	0.90	0.88	**0.94**	0.91	**0.95**	0.93	**0.94**
Cultural or communicative aspects	0.80	0.85	0.82	0.74	0.86	0.80	0.81	0.82	0.81	0.79	0.82	0.80	0.81	**0.87**	**0.84**	**0.86**	0.83	**0.84**
Psychological state	0.54	0.54	0.54	0.61	0.60	0.60	0.60	0.59	0.59	0.61	0.60	0.60	0.62	0.62	0.62	**0.73**	**0.71**	**0.72**
Physical state	0.75	**0.86**	0.80	0.79	0.82	0.80	0.80	0.84	0.82	0.80	0.83	0.81	0.81	**0.86**	0.83	**0.87**	0.81	**0.84**
Family-related themes	0.85	0.89	0.87	0.82	0.90	0.86	0.85	0.88	0.86	0.84	0.88	0.86	0.82	0.90	0.86	**0.94**	**0.91**	**0.92**

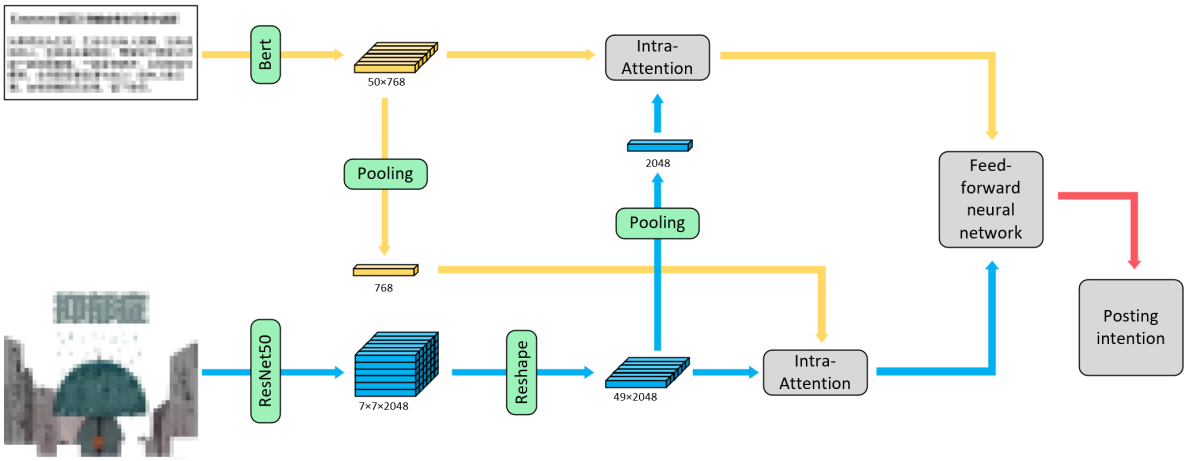


Figure 2: The neural network models with attention mechanisms to predict posting intentions.

the model performances. Second, we only focus on the text-image posts, while the posts in pure text or images in OMHCs may express different types of human intentions [14]. Third, we labeled the human intention from the viewers' perspectives, and a future interview with the posters in Chinese OMHCs would deepen our understandings on their posting intentions. Fourth, gender imbalance among annotators may affect accuracy, as different genders may perceive content differently, introducing bias. Future work will address this by achieving a more balanced gender representation to enhance consistency and objectivity. Our work offers implications for understanding and modeling the multi-modal posts in Chinese OMHCs and urges future researchers to extend its applications to improve people's mental health.

Acknowledgments

This work is supported by the Young Scientists Fund of the National Natural Science Foundation of China with Grant No. 62202509.

References

[1] Tadas Baltrusaitis, Chaitanya Ahuja, and Louis-Philippe Morency. 2019. Multimodal Machine Learning: A Survey and Taxonomy. *IEEE Trans. Pattern Anal. Mach. Intell.* 41, 2 (feb 2019), 423–443. https://doi.org/10.1109/TPAMI.2018.2798607

[2] Yitao Cai, Huiyu Cai, and Xiaojun Wan. 2019. Multi-Modal Sarcasm Detection in Twitter with Hierarchical Fusion Model. 2506–2515. https://doi.org/10.18653/v1/P19-1239

[3] Xuetong Chen, Martin D. Sykora, Thomas W. Jackson, and Suzanne Elayan. 2018. What about Mood Swings: Identifying Depression on Twitter with Temporal Measures of Emotions. In *Companion Proceedings of the The Web Conference 2018* (Lyon, France) *(WWW '18)*. International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, CHE, 1653–1660. https://doi.org/10.1145/3184558.3191624

[4] Qingyu Guo, Siyuan Zhou, Yifeng Wu, Zhenhui Peng, and Xiaojuan Ma. 2022. Understanding and Modeling Viewers' First Impressions with Images in Online

Medical Crowdfunding Campaigns. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems* (New Orleans, LA, USA) *(CHI '22)*. Association for Computing Machinery, New York, NY, USA, Article 361, 20 pages. https://doi.org/10.1145/3491102.3501830

[5] Zhipeng Guo. 2016. *THULAC-Python*. https://github.com/thunlp/THULAC-Python

[6] Simon Haykin. 1998. *Neural Networks: A Comprehensive Foundation* (2nd ed.). Prentice Hall PTR, USA.

[7] Shahra Jafar and Talal Shaikh. 2019. Image Based Emotion Detection. In *2019 International Conference on Computational Intelligence and Knowledge Economy (ICCIKE)*. 323–328. https://doi.org/10.1109/ICCIKE47802.2019.9004259

[8] Menglin Jia, Zuxuan Wu, Austin Reiter, Claire Cardie, Serge Belongie, and Ser-Nam Lim. 2021. Intentonomy: A Dataset and Study Towards Human Intent Understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 12986–12996.

[9] Taisa Kushner and Amit Sharma. 2020. Bursts of Activity: Temporal Patterns of Help-Seeking and Support in Online Mental Health Forums. In *Proceedings of The Web Conference 2020* (Taipei, Taiwan) *(WWW '20)*. Association for Computing Machinery, New York, NY, USA, 2906–2912. https://doi.org/10.1145/3366423.3380056

[10] Namyeon Lee, Kelli Buchanan, and Mansoo Yu. 2020. Each post matters: A content analysis of# mentalhealth images on Instagram. *Journal of Visual Communication in medicine* 43, 3 (2020), 128–138.

[11] Guo Li, Xiaomu Zhou, Tun Lu, Jiang Yang, and Ning Gu. 2016. SunForum: Understanding Depression in a Chinese Online Community. In *Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing* (San Francisco, California, USA) *(CSCW '16)*. Association for Computing Machinery, New York, NY, USA, 515–526. https://doi.org/10.1145/2818048.2819994

[12] Shuailin Li, Shiwei Wu, Tianjian Liu, Han Zhang, Qingyu Guo, and Zhenhui Peng. 2024. Understanding the Features of Text-Image Posts and Their Received Social Support in Online Grief Support Communities. In *International Conference on Web and Social Media*. https://api.semanticscholar.org/CorpusID:237940432

[13] Bin Liang, Chenwei Lou, Xiang Li, Lin Gui, Min Yang, and Ruifeng Xu. 2021. Multi-Modal Sarcasm Detection with Interactive In-Modal and Cross-Modal Graphs. In *Proceedings of the 29th ACM International Conference on Multimedia* (Virtual Event, China) *(MM '21)*. Association for Computing Machinery, New York, NY, USA, 4707–4715. https://doi.org/10.1145/3474085.3475190

[14] Xueer Lin, Jiaxin Cheng, Shun Wang, Kexin Song, Jiao Wang, and Zhenhui Peng. 2024. Exploring Users' Text and Image Posting Behaviors in Weibo Mental Health Communities. In *Proceedings of the Tenth International Symposium of Chinese CHI* (Guangzhou, China and Online, China) *(Chinese CHI '22)*. Association for Computing Machinery, New York, NY, USA, 277–281. https://doi.org/10.1145/3565698.3565797

[15] Jianli Liu, Edwin Lughofer, and Xianyi Zeng. 2015. Aesthetic perception of visual textures: a holistic exploration using texture analysis, psychological experiment, and perception modeling. *Frontiers in computational neuroscience* 9 (2015), 134.

[16] Lumina. 2022. *aiotieba*. https://github.com/lumina37/aiotieba

[17] S Anne Moorhead, Diane E Hazlett, Laura Harrison, Jennifer K Carroll, Anthea Irwin, and Ciska Hoving. 2013. A New Dimension of Health Care: Systematic Review of the Uses, Benefits, and Limitations of Social Media for Health Communication. *J Med Internet Res* 15, 4 (23 Apr 2013), e85. https://doi.org/10.2196/jmir.1933

[18] Mark W. Newman, Debra Lauterbach, Sean A. Munson, Paul Resnick, and Margaret E. Morris. 2011. It's not that i don't have problems, i'm just not putting them on facebook: challenges and opportunities in using online social networks for health. In *Proceedings of the ACM 2011 Conference on Computer Supported Cooperative Work* (Hangzhou, China) *(CSCW '11)*. Association for Computing Machinery, New York, NY, USA, 341–350. https://doi.org/10.1145/1958824.1958876

[19] Kathleen O'Leary, Arpita Bhattacharya, Sean A. Munson, Jacob O. Wobbrock, and Wanda Pratt. 2017. Design Opportunities for Mental Health Peer Support Technologies. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing* (Portland, Oregon, USA) *(CSCW '17)*. Association for Computing Machinery, New York, NY, USA, 1470–1484. https://doi.org/10.1145/2998181.2998349

[20] Zhenhui Peng, Xiaojuan Ma, Diyi Yang, Ka Wing Tsang, and Qingyu Guo. 2021. Effects of Support-Seekers' Community Knowledge on Their Expressed Satisfaction with the Received Comments in Mental Health Communities. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (Yokohama, Japan) *(CHI '21)*. Association for Computing Machinery, New York, NY, USA, Article 536, 12 pages. https://doi.org/10.1145/3411764.3445446

[21] Stephen A Rains and Valerie Young. 2009. A meta-analysis of research on formal computer-mediated support groups: Examining group characteristics and health outcomes. *Human communication research* 35, 3 (2009), 309–336.

[22] Jennifer R. Talevich, Stephen J. Read, David A. Walsh, Ravi Iyer, and Gurveen Chopra. 2017. Toward a comprehensive taxonomy of human motives. *PLOS ONE* 12, 2 (02 2017), 1–32. https://doi.org/10.1371/journal.pone.0172279

[23] Rui Wang. 2013. *snownlp*. https://github.com/isnowfy/snownlp

[24] Reda Yacouby and Dustin Axman. 2020. Probabilistic Extension of Precision, Recall, and F1 Score for More Thorough Evaluation of Classification Models. 79–91. https://doi.org/10.18653/v1/2020.eval4nlp-1.9

[25] Amir Hossein Yazdavar, Mohammad Saeid Mahdavinejad, Goonmeet Bajaj, William Romine, Amit Sheth, Amir Hassan Monadjemi, Krishnaprasad Thirunarayan, John M. Meddar, Annie Myers, Jyotishman Pathak, and Pascal Hitzler. 2020. Multimodal mental health analysis in social media. *PLOS ONE* 15, 4 (04 2020), 1–27. https://doi.org/10.1371/journal.pone.0226248

[26] Wei Zhang, Wen Wang, Jun Wang, and Hongyuan Zha. 2018. User-guided Hierarchical Attention Network for Multi-modal Social Image Popularity Prediction. *WWW '18: Proceedings of the 2018 World Wide Web Conference*, 1277–1286. https://doi.org/10.1145/3178876.3186026