EIJL: Popularity Prediction of Social Media Advertisements Based on Multimodal Emotional Interaction and Joint Learning

Jidong Leng, Qiang Yan†

School of Economics and Management, Beijing University of Posts and Telecommunications No. 10 Xitucheng Road, Haidian District, Beijing 100876, China

Keywords: Popularity prediction; Multimodal emotional interaction; Joint learning; Social media; CLIP

Citation: Leng J.D., Yan Q.: EIJL: Popularity Prediction of Social Media Advertisements Based on Multimodal Emotional Interaction and Joint Learning. Data Intelligence, Vol. XX, Art. No.: 2025??XX, pp. 1-18, 2025. DOI: https://doi.org/10.3724/2096-7004.di.2025.0066

ABSTRACT

Predicting the popularity of social media advertisements holds significant value in brand marketing. However, current prediction methods rely heavily on user networks and temporal data, overlooking the role of ad content and user emotions. This study aims to predict ad popularity by integrating multimodal emotional interactions between ad content and user feedback. The research collected image-text advertisement data from social media and employed the CLIP multimodal model to extract multimodal features of the advertisements. By extracting emotional features from advertisement text and images, the study investigates the interaction between text emotion, image emotion, and ad likability on advertisement popularity. An emotional interaction layer was designed, and a multi-task joint learning method that uses classification prediction to assist regression prediction was adopted to predict advertisement popularity. The study found that the interaction between text emotion, image emotion, and ad likability negatively impacts ad sharing. The integration of multimodal information, multi-task joint learning, and the emotional interaction layer effectively enhanced the model's ability to predict popularity, with a significant improvement in the MEA index. This research demonstrates that leveraging multimodal emotional interaction information to enhance neural network predictions of advertisement popularity is effective from the perspective of meme propagation. This approach provides valuable insights and directions for further optimizing popularity prediction.

† Corresponding author: Qiang Yan (E-mail: yan@bupt.edu.cn; ORCID: 0000-0003-0128-8360).

1. INTRODUCTION

In the digital age, advertising has become a core component of business communication. It plays a critical role in shaping brand image, attracting potential customers, and influencing consumer decisions. Current research on social media advertising popularity prediction focuses on three main methods: ensemble methods based on machine learning models, deep learning methods based on feature fusion, and methods based on large models. The most commonly used evaluation metrics are Mean Absolute Error (MAE), Root Mean Square Error (RMSE), Root Mean Square Logarithmic Error (RMSLE), and F1 score. Some studies have extracted textual emotions as features. However, most research still focuses on user network relationships and overlooks the emotional interactions in advertising content. This limits the ability to effectively predict the popularity of user-generated content and affects the accuracy of social media advertising popularity predictions. Additionally, the datasets used, such as those from WeChat, Weibo, and Flickr, reflect the diversity of user-generated content but inadequately cover the dimension of emotional interaction [1].

Emotional interaction, particularly through meme elements in advertisements, promotes rapid content spread and enhances brand-audience interaction and resonance [2]. For example, advertisements on the Little Red Book platform are delivered in a comprehensive format that integrates both text and images, which is highly favored by social media users. An in-depth analysis of the Little Red Book platform facilitates a deeper understanding of advertisement dissemination rules and audience preferences, enabling the development of precise and effective generative advertising strategies for marketers. Additionally, predicting social media advertisement popularity has applications in public opinion monitoring, topic recommendation, and brand public relations [3].

As social media advertising rapidly expands, traditional user relationship models have become increasingly complex and difficult to maintain. At the same time, emotional interaction through meme elements spreads swiftly across social media platforms, enabling advertising information to reach a broader target audience [4]. Despite this, most studies continue to focus on the dissemination paths and timing of advertisements within social networks. In-depth research on the emotional content and its interaction with users remains limited, limiting the effective prediction of the popularity of user-generated advertising content.

In this context, this study analyzes the impact of the emotional characteristics of advertisement content on popularity from the perspective of memetics. This approach aims to move beyond reliance on user relationship networks and complex data, focusing instead on the emotional interaction within advertisement content. Specifically, this paper proposes a social media advertisement popularity prediction method based on multimodal emotional interaction. This method involves mining the emotions in advertisement text and images, designing a multimodal emotional interaction layer, and using a multi-task joint learning approach to predict advertisement popularity. The advantage of this method is that it predicts the popularity of social media advertisements based on content without requiring temporal data and user relationship data, thereby addressing the issue of high data requirements in popularity prediction.

2. RELATED RESEARCH AND THEORETICAL FOUNDATION

2.1 Popularity Prediction Research

Popularity prediction aims to establish a system to forecast the future popularity of information posted by users in social networks. Traditional popularity prediction methods mainly include multi-feature-based prediction and point process modeling. With the development of deep learning, deep learning-based popularity prediction methods have gradually gained researchers' attention.

The core of multi-feature-based prediction is to identify features that contribute to popularity prediction. These features include user characteristics, content features, temporal features, and structural features. Shang used an unsupervised graph neural network to model undirected social homophily, combining attention mechanisms with the graph neural network framework to learn directed and heterogeneous social relationships [5]. Wu developed an activation-decay algorithm, predicting the long-term popularity of online content based on early repost volumes [6]. Yu used dynamic neural networks and partial differential equations to study and extract the dynamic evolution speed of cascade diffusion [7]. Bačić used the XGBoost classifier to analyze 42 collected features, including engagement, positive emotions, and skin conductance peaks [8]. In predicting the popularity of advertisements, Yan used the Factorization Machine (Deep FM) to capture both low-order feature interactions and high-order feature interactions, enhancing the prediction accuracy of advertisement popularity [9]. Arora proposed a novel multimodal online news popularity prediction model based on ensemble learning, using meta-features, text features, and image features to predict news popularity [10]. The key to multi-feature-based prediction is feature extraction, and the effectiveness of prediction depends on the quality of these features.

Point process modeling-based popularity prediction involves modeling the rate function of user repost behavior [11]. Blundell focused on time characteristics and decay effects in their modeling to explain the diffusion of information [12]. Zadeh approaches the development of a model based on multivariate Hawkes processes from temporal and network perspectives, considering follower counts and the activity variations observed in collective resharing behavior in popularity prediction modeling [13]. Tai used graph neural networks to aggregate local neighbor node information in the cascade network, considering the transitivity of information within internal paths during the modeling process [14]. Point process modeling-based prediction methods emphasize early dissemination patterns of data but neglect content analysis, limiting the model's expressive power.

Deep learning-based prediction focuses on the impact of content on dissemination. For short video prediction, Wu proposed the AMPS model, which uses BiLSTM combined with attention mechanisms to learn the relationships between features of different modalities [6]. For news text, Xiong proposed the DNCP model, which uses attention mechanisms to learn news attractiveness and timeliness features [15]. The DFW-PP framework proposed by Viswanatha Reddy focuses on learning the importance of different features in image content over time, predicting based on the dynamic changes in image content [16]. When handling text and image data, Hsu used trained word vectors to encode textual information and semantic image features, then used ensemble regression methods to aggregate these encoded features for

prediction [17]. Word vector technology can abstract multimodal data into higher-dimensional features, helping neural networks learn deeper features. Zhang embedded user relationship vectors into model training, using cross-attention mechanisms to help the model learn the relationship between relational data and textual data [18]. Qian proposed a multimodal joint feature representation model, using attention mechanisms to jointly model textual and image content, enhancing the model's ability to learn multimodal features [19].

Deep learning-based prediction methods improve the semantic representation of vectors. However, Guo [20] pointed out that deep learning methods may introduce model bias and user behavior bias during model training and in applied scenarios. Lee studied the association between social media marketing content and user engagement, finding that messages containing humor and emotion were associated with higher user engagement [21]. Therefore, multimodal emotional interaction can reduce these biases to some extent. However, existing studies have not yet incorporated the relationship between multimodal emotional interaction and advertisement popularity into the model structure. Therefore, this paper will examine the impact of multimodal emotional interaction on advertisement popularity and predict the popularity of image-text advertisements from the perspective of multimodal emotional interaction.

2.2 Multimodal Emotional Interaction

Businesses and enterprises utilize topics, memes, humorous images, and emoticons for social media marketing and brand IP (Intellectual Property) development. They co-create marketing content with users, resulting in a series of classic internet marketing cases [4]. Notable examples include the Chinese cultural icon Li Ziqi, the national brand Hongxing Erke, the Snow King image of Mixue Bingcheng, and KFC's Crazy Thursday. These instances are rich and multidimensional in cultural connotations and emotional information, bearing similarities to "allusions" in Chinese culture. They establish unique emotional connections with users [22], which easily trigger audience participation and secondary dissemination. This closely aligns with the characteristics of meme dissemination [23].

However, there are challenges with this approach. Emotional expression is the core connotation of memes. It is a crucial reason why marketers use memes in advertisement creation. Moreno et al., based on cognitive load theory, proposed the "Modality Principle." This principle reveals the impact of modality forms on individuals' understanding of specific information [24]. Reducing cognitive load helps improve information processing efficiency. Zheng's empirical research has indicated that information overload and social overload contribute to social media fatigue, thereby exacerbating users' SNS anxiety [25]. Anxiety can cause social media users to exhibit avoidance behaviors [26]. These behaviors reduce system usage frequency and advertisement dissemination probability. Meanwhile, Caucci's research indicates that in interactive scenarios, users prefer receiving emotionally consistent combinations of images and texts [27]. The emotional communicability of multimodal meme advertisements can be linked to the social identity of social media users. This link has a certain interactive effect on users' sharing behaviors [28]. Therefore, analyzing the multimodal emotional interaction of social media advertisements from the perspective of meme dissemination can, to some extent, explain the viral spread of social media advertisements.

In summary, popularity prediction has become a research hotspot for many researchers and internet companies. Many methods have been proposed to address various business scenarios or problems. However, these methods have certain shortcomings. Specifically, prediction methods based on social media advertisement content often lack a solid scientific theory as support. This makes it difficult to clearly and comprehensively explain the mechanisms of popularity prediction. In this context, memetics provides an effective solution to overcome the above issues. This paper focuses on analyzing the multimodal emotional interactions in advertisements and predicts the popularity of social media advertisements through the lens of memetics.

3. EXPLORATORY ANALYSIS

Following a series of exploratory experiments, we utilized the Little Red Book app to gather social media advertisements. A total of 78 high-quality marketing cases were identified on the platform. For each case, we collected 100 text-image marketing notes, as illustrated in Figure 1, which include the cover image, title, main text, number of likes, number of favorites, and number of comments. Notably, for the purpose of studying popularity prediction, we also obtained the share counts corresponding to each note as the target variable. The use of historical data is particularly meaningful because, due to the platform's traffic strategy, these notes have a saturated state, and their share counts are unlikely to change significantly.

Figure 1. Example of Little Red Book Advertisements.

Exploratory analysis was conducted to gain deeper insights into the target variable (number of shares) and key features such as image emotion and text emotion. In the collected dataset, the number of shares

is the target variable. The highest number of shares recorded was 24,547, while the lowest was 0. The detailed distribution of the target variable (excluding notes with zero shares) is shown in Figure 2(a). This distribution follows a log-normal pattern, indicating that a small portion of notes have high share counts, whereas the majority have share counts concentrated in the lower range.

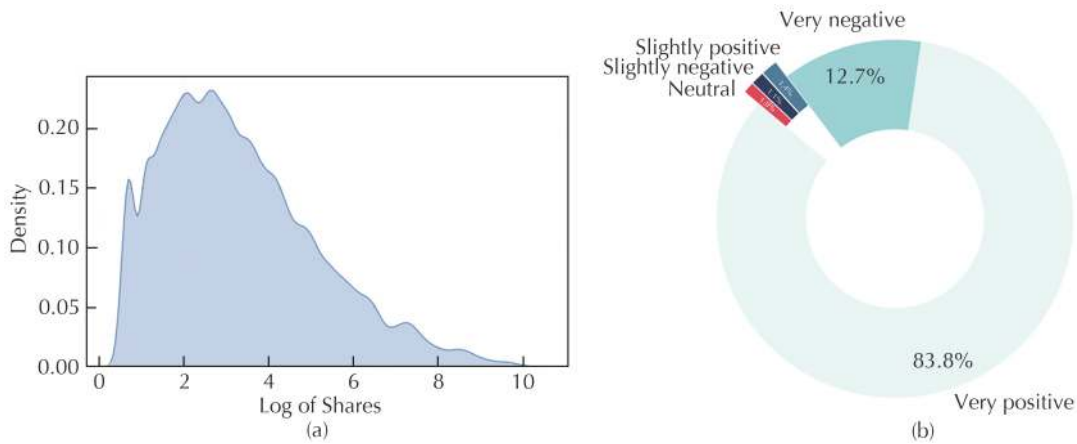

Figure 2. Comparative Analysis of Note Shares and Text Emotion Intensity Distributions.

According to the Selective Attention Theory [29] and the SOR Theory (Stimulus-Organism-Response Theory) [30], the cover of a Little Red Book note largely determines whether users will view the note content. Subsequently, the text content of the note influences user perception. If users like the note content, there is a high probability that they will share the note.

In the exploratory experiments, this study used the DeepFace and SnowNLP libraries to extract facial emotions from the covers and text emotions from the note content. An intriguing finding was observed: the majority of covers did not include faces. Consequently, the emotional intensity of images was classified into two categories based on the presence of a face. Text emotional intensity was classified into five categories based on emotional polarity: very negative, slightly negative, neutral, slightly positive, and very positive. Covers containing facial emotions accounted for 7.45% of the total covers. The distribution of text emotional intensity is shown in Figure 2(b). Similar to the share count, the emotional characteristics exhibit a long-tail distribution. This indicates a "Matthew Effect" in Little Red Book advertising dissemination. This phenomenon suggests that a few attractive or emotionally charged advertisements resonate with users and spread rapidly on social media. In contrast, most advertisements receive less attention and fewer shares due to their lack of uniqueness or standout features.

It is noteworthy that Little Red Book differs significantly from social media platforms like Weibo and Twitter in terms of user interface design. The latter typically allows users to view the number of shares, while the former does not provide this feature. Little Red Book's design aims to guide users to focus on content of interest, stimulating their emotions and cognition, rather than blindly chasing popular content.

Therefore, researching the impact of multimodal emotional interaction on content popularity using Little Red Book data is more meaningful.

By calculating the correlation matrix between likes, favorites, comments, and shares, and plotting the correlation heatmap as shown in Figure 3, it is evident that both comments and shares relate to user engagement, but their correlation is lower than that between likes and shares. This further highlights the importance of emotional interaction in advertisement sharing.

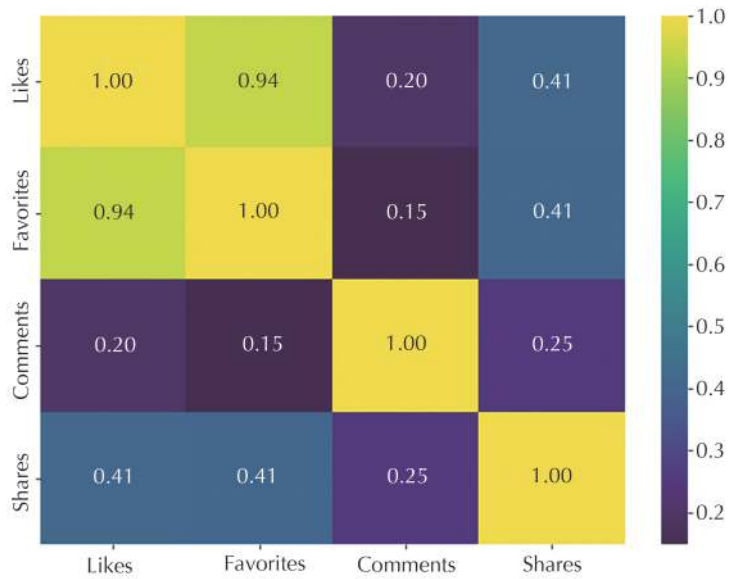

Figure 3. Correlation Heatmap of Likes, Favorites, Comments, and Shares.

Based on the above exploratory analysis, a multi-factor regression model is established as follows:

$$Y = \beta_0 + \beta_1 * \text{txt\_emo} + \beta_2 * \text{img\_emo} + \beta_3 * \text{likability} + \beta_3 * \text{txt\_emo} * \text{img\_emo} + \beta_4 * \text{txt\_emo} * \text{img\_emo} * \text{likability} + \epsilon \quad (1)$$

Where the dependent variable Y represents the advertisement popularity, which is classified into popular and non-popular ads based on the 2:8 ratio of share counts. β_0 is the constant term, and ϵ is the error term. txt_emo represents text emotional intensity, img_emo represents image emotional intensity, and likability represents user likability, , which is categorized into high, medium, and low based on the number of likes. The symbols β_1 to β_5 represent the weights of the explanatory variables and interaction terms.

To thoroughly verify the fit between the model and the actual data, this study used SPSS 27.0 software to analyze the impact of emotional interaction on advertisement popularity. The regression coefficients and significance levels of various influencing factors are shown in Table 1.

Table 1. Regression coefficient and significance.

Variable	Regression Coefficient Estimate	P-value
Text Emotion	–0.030	0.87
Image Emotion	–0.011	0.432
Likability	0.235	$< 0.001^{***}$
Text, Image Emotion Interaction	0.049	0.050^{*}
Text, Image Emotion and Likability Interaction	–0.057	$< 0.001^{***}$

The R-squared value can be used to assess the model's fitting ability and predictive power. The results show that the model's R-squared is 0.234, indicating that the model is statistically significant overall. This means that the emotional interactions explain 23.4% of the variance in advertisement shares. Specifically, likability (regression coefficient = 0.235, $P < 0.001$) has a significant positive impact on advertisement shares, while text emotion (regression coefficient = –0.030, $P = 0.87$) and image emotion (regression coefficient = –0.011, $P = 0.432$) individually do not have a significant effect on shares. The interaction between text and image emotions (regression coefficient = 0.049, $P = 0.050$) has a marginally significant positive effect on advertisement shares, suggesting that a balanced combination of emotions can enhance the likelihood of sharing. However, the three-way interaction between text emotion, image emotion, and likability (regression coefficient = –0.057, $P < 0.001$) shows a significant negative impact, indicating that excessive emotional interaction may increase cognitive load on users, thereby inhibiting the spread of advertisements.

In summary, the interactions between text emotion and image emotion have positive impacts on advertisement shares. However, the three-way interaction among text emotion, image emotion, and advertisement likability has a negative impact on advertisement shares. According to the modality principle of cognitive load theory, this negative impact might be due to the simultaneous increase in text and image emotions, which increases the cognitive load of individuals, thereby reducing the efficiency of information processing and inhibiting advertisement sharing behavior. This further verifies that multimodal emotional interaction can partially explain the popularity of social media advertisements. Therefore, based on memetics, this study designs a multimodal emotional interaction module to improve the effectiveness of advertisement popularity prediction.

4. SOCIAL MEDIA ADVERTISEMENT POPULARITY PREDICTION METHOD

The popularity prediction process is illustrated in Figure 4. First, text information undergoes preprocessing operations such as deduplication and tokenization, and image formats are standardized. Next, the CLIP multimodal model is used to pre-extract vector representations of advertisement text and images. On this basis, the CNN model is employed to learn the extracted vector representations of text and images separately, followed by pooling operations to learn the semantic features in multimodal

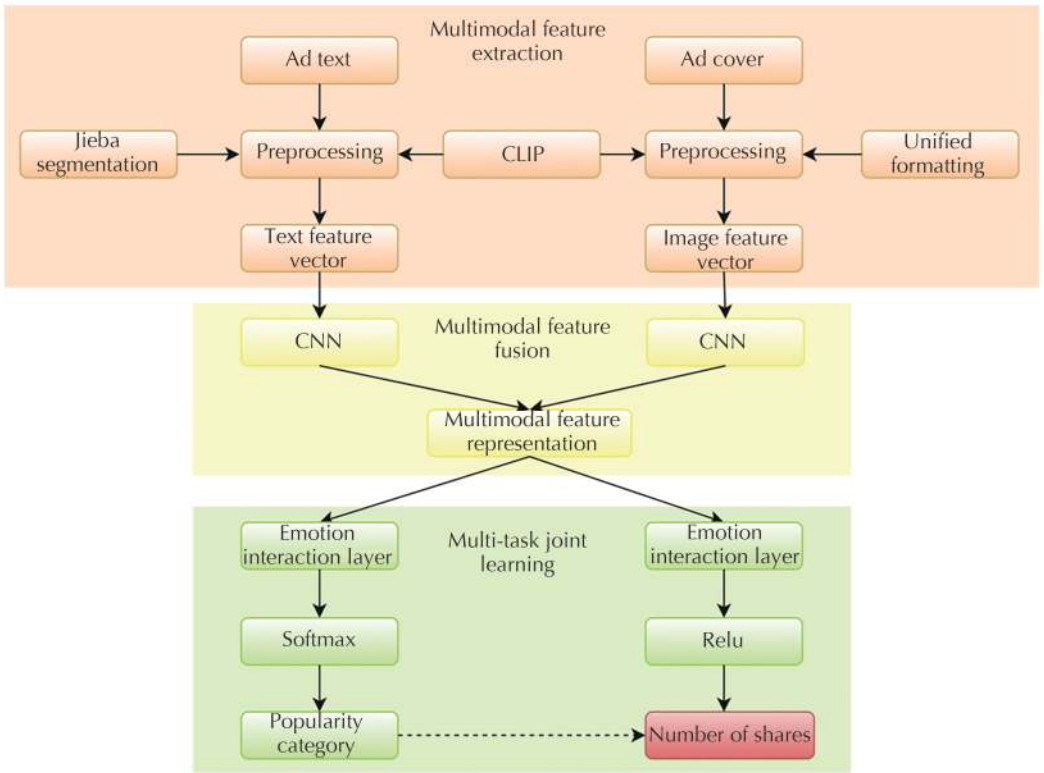

Figure 4. Flow Chart of Popularity Prediction.

information and enhance the robustness of the model's predictions. Subsequently, a joint learning approach is utilized to assist in predicting share counts by predicting popularity categories, thereby achieving popularity prediction. During this process, the features of text and images are concatenated, and attention mechanisms are used to learn the interactions between multimodal features, improving the model's focus on key information within the fused features. Finally, the emotional interaction layer is employed to enhance the model's learning of the multimodal emotional interaction process, thereby achieving the prediction of social media advertisement popularity.

4.1 Emotional Interaction Layer

The multimodal emotional interaction in this study refers to how social media advertisements enhance users' likability of ads through various modes of emotional expression, thereby influencing users' sharing behavior. Image-text advertisements combine features such as images, colors, and text, becoming a new form for social media users to express emotions. Compared to unimodal information, multimodal information offers richer and more complex emotional expressions, and multimodal emotion analysis provides higher stability [31]. Therefore, based on the results of exploratory analysis, this study designed an

emotional interaction layer to improve the effectiveness of popularity prediction. The emotional interaction function is shown in Equation 2.

$$\phi_{emo} = \phi + \omega_1 * \text{txt\_emo} + \omega_2 * \text{img\_emo} + \omega_3 * \text{likability} \quad (2)$$

Where ϕ is the input vector, txt_emo is the emotional intensity of the advertisement text, ω_1 is the weight of the emotional features of the advertisement text, img_emo is the facial emotional intensity on the cover of the advertisement, ω_2 is the weight of the facial emotional intensity on the cover of the advertisement, likability is the user's likability of the meme advertisement, indicated by the number of likes, ω_3 is the weight of the user's likability of the advertisement, and ϕ_{emo} is the output vector.

4.2 Multitask Joint Learning

Multitask joint learning has been proven effective in fields such as natural language processing and computer vision [32]. It typically implements joint learning through parameter sharing, which includes a shared module for handling all tasks and specific modules for handling particular tasks [33]. The shared module helps alleviate the problem of model overfitting during training, while the specific modules can assist each other, learning richer abstract features from the data. Minhwa Cho, in his research on video view prediction, used a joint approach of classification and regression prediction to improve the accuracy of view predictions [34]. However, existing research has not applied joint classification and regression prediction in the context of predicting the popularity of image-text advertisements. Therefore, this study designs a joint learning approach that categorizes the logarithm of advertisement shares into 13 popularity categories. By sharing network parameters, the classification prediction assists in the regression prediction of advertisement shares, thereby enhancing the model's effectiveness in predicting advertisement popularity.

5. EXPERIMENTS

The input for this experiment consists of multimodal social media advertisement information, and the output is the predicted number of advertisement shares. The hardware configuration includes 48 Intel (R) Xeon (R) Silver 4214R CPUs @ 2.40 GHz, an NVIDIA GeForce RTX 3090 GPU, and the operating system is Ubuntu 18.04.5. The software environment for the experiment is Python 3.8.13, with PyCharm used as the development tool. After preprocessing and further filtering the collected data, a final set of 7,758 image-text notes was selected for the experiment. These notes were randomly divided into training and testing sets in an 80% to 20% ratio. For the baseline model, this study compares the proposed method with the multimodal pre-trained model CLIP. Additionally, ablation experiments were conducted to validate the effectiveness of multimodal information, multitask joint learning, and the emotional interaction layer in predicting social media advertisement popularity.

5.1 Evaluation Metrics

The experiment uses the Mean Absolute Error (MAE) metric to evaluate the prediction effect. MAE is the average absolute difference between the predicted and actual values. The smaller the MAE value, the better the model's prediction effect. The formula for MAE is as follows.

$$MAE = \frac{1}{n}\sum_{i-1}^{n}\left|\tilde{y}_i - y_i\right| \tag{3}$$

In this formula, y_i represents the actual value, $\tilde{y}_i$ represents the predicted value, and n represents the number of predictions.

5.2 Experimental Results and Analysis

Given that the dataset includes both text and image modalities, and incorporates an emotional interaction layer and a multitask joint learning method, we have designed corresponding comparative and ablation experiments to comprehensively evaluate the effectiveness and robustness of the proposed method. The comparative experiments aim to compare the performance of the proposed model with the baseline CLIP model [35], as well as to evaluate different multimodal information extraction and fusion methods. The ablation experiments focus on assessing the specific contributions of the joint learning module and the emotional interaction layer. Through these comparative and ablation experiments, we can gain a deeper understanding of the role of each module in predicting advertisement popularity, thereby verifying the overall performance and advantages of the proposed model. The specific experimental setups are detailed below.

Experiment 1: CLIP Model Fine-tuning

Fine-tuning the CLIP multimodal pre-training model using advertisement text and cover data for regression prediction.

Experiment 2: (RoBERTa + ResNet50)_CNN

Concatenating advertisement text vectors extracted by RoBERTa and advertisement cover vectors extracted by ResNet50, learning the concatenated feature vectors using CNN, and performing regression prediction.

Experiment 3: (CLIPTXT + CLIPIMG)_CNN

Concatenating advertisement text vectors and cover vectors extracted by CLIP, learning the concatenated feature vectors using CNN, and performing regression prediction.

Experiment 4: CLIPTXT_CNN + CLIPIMG_CNN

Feeding advertisement text vectors and cover vectors extracted by CLIP into two separate CNN network layers, pooling the vectors after learning, concatenating them, and performing regression prediction.

Experiment 5: (CLIPTXT + CLIPIMG)_CNN_Joint

Concatenating advertisement text vectors and cover vectors extracted by CLIP, learning the concatenated vectors using CNN, pooling the vectors after learning, concatenating them, and using joint learning with popularity label classification prediction to assist share count regression prediction.

Experiment 6: (CLIPTXT_CNN + CLIPIMG_CNN)_Joint

Feeding advertisement text vectors and cover vectors extracted by CLIP into two separate CNN network layers, pooling the vectors after learning, concatenating them, and using joint learning with popularity label classification prediction to assist share count regression prediction.

Experiment 7: (CLIPTXT + CLIPIMG)_CNN_Joint_Emo

Concatenating advertisement text vectors and cover vectors extracted by CLIP, learning the concatenated feature vectors using CNN, processing them through the emotional interaction layer, and using joint learning with popularity label classification prediction to assist share count regression prediction.

Experiment 8: (CLIPTXT_CNN + CLIPIMG_CNN)_Joint_Emo

Feeding advertisement text vectors and cover vectors extracted by CLIP into two separate CNN network layers, pooling the vectors after learning, concatenating them, processing them through the emotional interaction layer, and using joint learning with popularity label classification prediction to assist share count regression prediction.

The experimental results are shown in Table 2, with the overall analysis divided into three experimental groups. Group One is the comparison group for multimodal feature extraction. Group Two is the comparison group for multitask joint learning. Group Three is the comparison group for the emotional interaction layer.Through these three comparison experiments, we can study the role of multimodal features, joint learning, and the emotional interaction layer in the popularity prediction model.

Group One: Multimodal Feature Extraction Comparison

The results of Experiments 1, 2, and 3 show that using neural networks to learn features extracted by CLIP outperforms fine-tuning CLIP. This may be because the dataset is relatively small, and the large number of parameters in the CLIP model can lead to overfitting during direct fine-tuning. In contrast, using CLIP to extract multimodal features and then learning them with a smaller parameter-scale CNN can better avoid overfitting and improve performance. The comparison between Experiments 2 and 3 indicates that

Table 2. Results of Popularity Prediction.

Experiment	Method	MAE
1	CLIP Model Fine-tuning	234
2	(RoBERTa + ResNet50)_CNN	233
3	(CLIPTXT + CLIPIMG)_CNN	223
4	CLIPTXT_CNN + CLIPIMG_CNN	226
5	(CLIPTXT + CLIPIMG)_CNN_Joint	222
6	(CLIPTXT_CNN + CLIPIMG_CNN)_Joint	219
7	(CLIPTXT + CLIPIMG)_CNN_Joint_Emo	199
8	(CLIPTXT_CNN + CLIPIMG_CNN)_Joint_Emo	**187**

training with CNN after feature extraction by CLIP is more effective than feature extraction by RoBERTa and ResNet50. This is because CLIP excels in handling multimodal tasks, learning, and capturing interactions between different modalities, resulting in higher prediction accuracy and generalization capabilities during feature extraction and subsequent neural network training. Therefore, CLIP is selected for feature extraction in subsequent experiments.

Group Two: Multitask Joint Learning Comparison

Analyzing the results of Experiments 3, 4, 5, and 6 reveals that multitask joint learning can enhance model prediction performance. This is because CLIP is designed to handle various downstream tasks, and its vector representations maintain generality and relevance across different tasks. The comparison between Experiments 3 and 4 shows that there is little difference between concatenating CLIP multimodal vectors at the input stage and fusing them after CNN training. However, comparing Experiments 3 and 5, and Experiments 4 and 6 shows that using two separate CNN networks to process text and image vectors respectively, followed by joint learning, results in more significant performance improvements. This may be because more parameters allow the model to learn feature representations more comprehensively, improving prediction performance. In summary, for the task of predicting the popularity of social media text-image advertisements, integrating and utilizing CLIP multimodal feature vectors through joint learning is effective and significantly enhances model performance.

Group Three: Emotional Interaction Layer Comparison

The comparison of Experiments 5, 6, 7, and 8 shows that the emotional interaction module significantly improves the effectiveness of popularity regression prediction. This is because the emotional interaction module effectively captures emotional interactions between users and note content. Introducing the emotional interaction module further optimizes the modeling of user psychology and emotional attitudes,

enabling the model to more accurately assess user interest and emotional inclination towards note content, thereby more accurately predicting popularity.

Analyzing the results of the three groups shows that the use of multimodal features and different methods of feature fusion impact popularity regression prediction to some extent, while multitask joint learning and the introduction of emotional interaction layer further improve the prediction performance. Specifically, CLIP multimodal features provide the model with more comprehensive and rich semantic content, enabling the neural network model to more accurately understand the content and characteristics of advertisements by combining text and image information; multitask joint learning enhances the model's ability to understand CLIP multimodal feature vectors, enabling the neural network to learn the relationship between classification prediction and regression prediction, improving overall model performance; the emotional interaction module, by introducing emotional interaction information, helps the model better analyze the emotional impact of advertisements on users, improving the effectiveness of popularity prediction.

In conclusion, from the perspective of information dissemination and memetics, using multimodal semantic features and multimodal emotional interaction information to enhance neural network prediction of advertisement popularity is effective. This approach provides useful insights for optimizing popularity prediction and helps to explore emotional interactions in social media advertising through the lens of memetic theory.

6. CONCLUSION AND FUTURE WORK

This paper addresses the issue of predicting the popularity of user-generated content by utilizing real social media advertising data. This paper examines the popularity distribution of text-image advertisements on the Little Red Book platform, employing DeepFace and SnowNLP to analyze image and text emotions. The study evaluates how interactions among text and image emotions, along with user preferences, influence sharing behavior. The research addresses the challenge of predicting the popularity of user-generated content using real social media advertising data. From the perspective of meme dissemination theory, it designs a popularity prediction algorithm that focuses on advertisement content and multimodal emotional interaction. Based on traditional neural network prediction methods, the CLIP multimodal model is used to extract features from both advertisement text and images. The study designs a multi-task joint learning network structure for popularity category classification prediction and popularity share regression prediction. By incorporating an emotional interaction layer, the model's understanding and learning of emotional features are enhanced, achieving accurate predictions of social media advertisement popularity. Experimental results show that the inclusion of the emotional interaction module significantly improves the prediction performance, demonstrating the effectiveness of this approach.

This study makes several contributions to the prediction of social media advertisement popularity. Firstly, it verifies the long-tail theory by analyzing Little Red Book note data, finding that the popularity of social media advertisements follows a long-tail distribution, confirming the applicability of the long-tail

theory in social media advertising. The emotional extraction analysis results show that the multimodal emotional intensity of social media advertisements also follows a long-tail distribution, further validating the application of the long-tail theory in the field of emotional dissemination. Secondly, it supplements the modality principle theory by analyzing the influencing factors. It was found that the interaction between text emotion and image emotion had a marginally significant positive effect on advertisement popularity. However, the three-way interaction between text emotion, image emotion, and advertisement likability had a negative impact on advertisement popularity, which supplemented the modality principle in meme dissemination. This suggests that platforms like Little Red Book should recommend advertisements with high text and image emotional intensity but lower user likes in order to enhance advertisement dissemination. Finally, based on these findings, a multimodal emotional interaction layer was designed to enhance the neural network's prediction accuracy for advertisement popularity. Experimental results show that using multimodal semantic features and multimodal emotional interaction information to improve the neural network's prediction accuracy for advertisement popularity is effective.

The findings of this study have significant implications for businesses, enterprises, and marketers. Firstly, they should focus on the multimodal emotional expression of advertisement content, utilizing comprehensive multimodal information to enhance the emotional resonance of advertisements, thereby increasing user engagement and the effectiveness of advertisement dissemination. Secondly, businesses should optimize advertisement placement strategies by combining advertisement content with user feedback to improve the precision and effectiveness of advertisement delivery. Finally, by deeply analyzing users' emotional responses and behavior data, businesses can design advertisements that better meet user needs, further enhancing brand influence and market competitiveness. Future research can be expanded in three main directions: first, by exploring more emotional cues in social media advertisements for multimodal emotional interaction; second, by collecting data from different cultural backgrounds and social media platforms to explore popularity prediction from the perspectives of multicultural and multisource data; thirdly, by studying the impact of emotional interactions in advertisements for different types of products on their popularity. These approaches can further enhance the model's performance and applicability.

AUTHOR CONTRIBUTIONS

Jidong Leng: Proposed the research problems, performed the research, designed the research framework, collected and analyzed the data, and wrote and revised the manuscript.

Qiang Yan: Oversaw the research project and provided strategic direction.

ACKNOWLEDGEMENTS

This work is supported by The First Batch of New Liberal Arts Research and Reform Practice Projects of the Ministry of Education of China in China (2021090003) , the BUPT Innovation and Entrepreneurship

Support Program (2025-YC-T041) and the Supercomputing Platform of Beijing University of Posts and Telecommunications.

REFERENCES

[1] J. O'Callaghan, "How OpenAI's text-to-video tool Sora could change science—and society," *Nature*, vol. 627, no. 8004, pp. 475–476, 2024.

[2] A. Razzaq, W. Shao, and S. Quach, "Towards an understanding of meme marketing: conceptualisation and empirical evidence," *J. Mark. Manag.*, vol. 39, no. 7–8, pp. 670–701, 2023.

[3] Z. L. Li, Y. Ge, and X. Bai, "What will be popular next? Predicting hotspots in two-mode social networks," *MIS Q.*, vol. 45, no. 2, pp. 925–966, 2021.

[4] W. Yao, J. Han, X. Song, and R. Xu, "The application of social media intelligence analysis model based on memetics in the CI," *J. China Soc. Sci. Tech. Inf.*, vol. 35, no. 6, pp. 605–616, 2016.

[5] Y. Shang, X. Wang, Z. Liu, H. Zhang, Q. Chen, and T. Li, "Predicting the popularity of online content by modeling the social influence and homophily features," *Front. Phys.*, vol. 10, Art. no. 915756, 2022.

[6] L. Wu, L. Yi, X.-L. Ren, and L. Lü, "Predicting the popularity of information on social platforms without underlying network structure," *Entropy*, vol. 25, no. 6, Art. no. 916, 2023.

[7] D. Yu, Y. Zhou, S. Zhang, W. Li, M. Small, and K. K. Shang, "Information cascade prediction of complex networks based on physics-informed graph convolutional network," *New J. Phys.*, vol. 26, no. 1, Art. no. 013031, 2024.

[8] D. Bačić and C. Gilstrap, "Predicting video virality and viewer engagement: a biometric data and machine learning approach," *Behav. Inf. Technol.*, vol. 43, no. 12, pp. 2854–2880, 2024.

[9] C. R. Yan, L. J. Zhou, Q. L. Zhang, and X. L. Li, "Research on wide and deep extension of factorization machine," *J. Softw.*, vol. 30, no. 3, pp. 822–844, 2019.

[10] A. Arora, V. Hassija, S. Bansal, S. Yadav, V. Chamola, and A. Hussain, "A novel multimodal online news popularity prediction model based on ensemble learning," *Expert Systems*, vol. 40, no. 8, Art. no. e13336, 2023.

[11] H. Shen, D. Wang, C. Song, and A.-L. Barabási, "Modeling and predicting popularity dynamics via reinforced Poisson processes," in *Proc. AAAI Conf. Artif. Intell.*, vol. 28, no. 1, pp. 291–297, 2014.

[12] C. Blundell, J. Beck, and K. A. Heller, "Modelling reciprocating relationships with Hawkes processes," in *Adv. Neural Inf. Process. Syst.*, vol. 25, pp. 2600–2608, 2012.

[13] A. Zadeh and R. Sharda, "How can our tweets go viral? Point-process modelling of brand content," *Inf. Manag.*, vol. 59, no. 2, Art. no. 103594, 2022.

[14] Y. Tai, H. He, W. Zhang, H. Yang, X. Wu, and Y. Wang, "Predicting information diffusion using the inter- and intra-path of influence transitivity," *Inf. Sci.*, vol. 651, Art. no. 119705, 2023.

[15] J. Xiong, L. Yu, D. Zhang, and Y. Leng, "DNCP: An attention-based deep learning approach enhanced with attractiveness and timeliness of news for online news click prediction," *Inf. Manag.*, vol. 58, no. 2, Art. no. 103428, 2021.

[16] G. V. Reddy, B. S. N. V. Chaitanya, P. Prathyush, M. Sumanth, C. Mrinalini, P. D. Kumar, *et al.*, "DFW-PP: Dynamic feature weighting-based popularity prediction for social media content," *J. Supercomput.*, vol. 80, no. 4, pp. 5708–5730, 2024.

[17] C.-C. Hsu, L.-W. Kang, C.-Y. Lee, J.-Y. Lee, Z.-X. Zhang, and S.-M. Wu, "Popularity prediction of social media based on multi-modal feature mining," in *Proc. 27th ACM Int. Conf. Multimedia*, pp. 2687–2691, 2019.

[18] X. Zhang and W. Gao, "Predicting viral rumors and vulnerable users with graph-based neural multi-task learning for infodemic surveillance," *Inf. Process. Manag.*, vol. 61, no. 1, Art. no. 103520, 2024.
[19] Y. Qian, W. Xu, X. Liu, H. Ling, Y. Jiang, Y. Chai, *et al.*, "Popularity prediction for marketer-generated content: A text-guided attention neural network for multi-modal feature fusion," *Inf. Process. Manag.*, vol. 59, no. 4, Art. no. 102984, 2022.
[20] X. Guo, D. Wu, Q. Wei, and G. Chen, "The problem of biases in machine learning and user behavior: Insights into knowing the deviated and upholding the undeviated," *Manag. World*, vol. 39, no. 5, pp. 145–159, 199, 160–162, 2023.
[21] D. Lee, K. Hosanagar, and H. S. Nair, "Advertising content and consumer engagement on social media: Evidence from Facebook," *Manage. Sci.*, vol. 64, no. 11, pp. 5105–5131, 2018.
[22] W. M. Lim, T. Rasul, S. Kumar, and M. Ala, "Past, present, and future of customer engagement," *J. Bus. Res.*, vol. 140, pp. 439–458, 2022.
[23] R. Dawkins, *The Selfish Gene*, Oxford, U.K.: Oxford University Press, 2016.
[24] J. K. Moreno, "The psychologist as novelist," *Prof. Psychol. Res. Pract.*, vol. 37, no. 2, p. 210, 2006.
[25] H. Zheng, X. Chen, S. Jiang, and L. Sun, "How does health information seeking from different online sources trigger cyberchondria? The roles of online information overload and information trust," *Inf. Process. Manag.*, vol. 60, no. 4, Art. no. 103364, 2023.
[26] N. Bozionelos, "Socio-economic background and computer use: The role of computer anxiety and computer experience in their relationship," *Int. J. Hum.-Comput. Stud.*, vol. 61, no. 5, pp. 725–746, 2004.
[27] G. M. Caucci and R. J. Kreuz, "Social and paralinguistic cues to sarcasm," *Humor*, vol. 25, no. 1, pp. 1–22, 2012.
[28] Y.-C. Shen, C. T. Lee, and W.-Y. Lin, "Meme marketing on social media: The role of informational cues of brand memes in shaping consumers' brand relationship," *J. Res. Interact. Mark.*, vol. 18, no. 4, pp. 588–610, 2024.
[29] D. E. Broadbent, *Perception and Communication*, Amsterdam, Netherlands: Elsevier, 2013.
[30] M. Sharma, D. Kaushal, and S. Joshi, "Adverse effect of social media on Generation Z user's behavior: Government information support as a moderating variable," *J. Retail. Consum. Serv.*, vol. 72, Art. no. 103256, 2023.
[31] L. Zhu, Z. Zhu, C. Zhang, Y. Xu, and X. Kong, "Multimodal sentiment analysis based on fusion methods: A survey," *Inf. Fusion*, vol. 95, pp. 306–325, 2023.
[32] P. Y. W. Myint, S. L. Lo, and Y. Zhang, "Unveiling the dynamics of crisis events: Sentiment and emotion analysis via multi-task learning with attention mechanism and subject-based intent prediction," *Inf. Process. Manag.*, vol. 61, no. 4, Art. no. 103695, 2024.
[33] R. Caruana, "Multitask learning," *Mach. Learn.*, vol. 28, pp. 41–75, 1997.
[34] S. Sarkar, S. Basu, A. Paul, and D. P. Mukherjee, "Vivid: View prediction of online video through deep neural network-based analysis of subjective video attributes," *IEEE Trans. Broadcast.*, vol. 69, no. 1, pp. 191–200, 2023.
[35] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, *et al.*, "Learning transferable visual models from natural language supervision," in *Proc. 38th Int. Conf. Mach. Learn.*, pp. 8748–8763, 2021.

AUTHOR BIOGRAPHY

Jidong Leng is a PhD student at the School of Economics and Management, Beijing University of Posts and Telecommunications (BUPT), and a student research fellow at the AI+ Youth Research Center of BUPT. He previously served as an editorial assistant with the Journal of Beijing University of Posts and Telecommunications. He earned his M.S. in computer science (natural language processing) in 2022 from Beijing Information Science and Technology University, and started his PhD at BUPT in the same year, advised by Professor Qiang Yan. His research interest lies at the intersection of human-computer interaction and natural language processing. His current focus is on sentiment analysis in interactive systems, particularly multimodal sentiment analysis. He previously worked on information extraction, knowledge graphs, automatic text summarization and blockchain-related research.
ORCID: 0000-0001-8921-1784.

Prof. **Qiang Yan** has a PhD in Computer Science and Technology (Peking University, 2003). He is a Professor and the Dean of the School of Economics and Management at Beijing University of Posts and Telecommunications. His research focuses primarily on intelligent human-computer interaction and individual decision-making, AI risk governance, and big data analytics in social media contexts.
ORCID: 0000-0003-0128-8360.