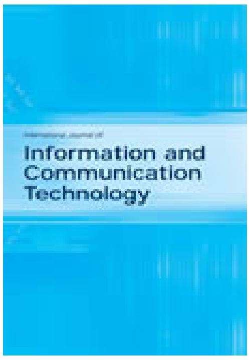
INDERSCIENCE PUBLISHERS
Linking academia, business and industry through research



International Journal of
Information and Communication Technology

International Journal of Information and Communication Technology

ISSN online: 1741-8070 - ISSN print: 1466-6642
https://www.inderscience.com/ijict

Digital media operations prediction based on user sentiment analysis and deep neural networks

Xinyu Chen, Zhenbin Huang

DOI: 10.1504/IJICT.2025.10074866

Article History:

Received:	30 June 2025
Last revised:	23 July 2025
Accepted:	23 July 2025
Published online:	17 December 2025

Digital media operations prediction based on user sentiment analysis and deep neural networks

Xinyu Chen and Zhenbin Huang*

School of Art, Design and Media,
Sanda University,
Shanghai, 201209, China
Email: chenxinyuusm@hotmail.com
Email: leohuang98@126.com
*Corresponding author

Abstract: Against the backdrop of increasingly fierce competition in the digital media industry, how to accurately predict operational effects has become the key to enhancing the competitiveness of media. Aiming at the problem of fusion redundancy caused by the existing research ignoring the mutual influence among cross-modalities, this paper first uses BERT and the improved visual transformer model to extract text and image features respectively. Then, cross-modal shared computing is utilised to enhance the complementarity among the features of each modal. Introduce text gating enhancement and use text information as prior knowledge to guide and improve the representation of image characteristics. Eventually, the fused characteristics are input into the classification layer for prediction. Experimental outcome indicates that the prediction accuracy rate of the suggested approach is 95.3%, which is at least 2.2% higher, significantly improving the accuracy of predicting the operation effect of digital media.

Keywords: digital media; operation effect prediction; sentiment analysis; convolutional neural network; vision transformer.

Reference to this paper should be made as follows: Chen, X. and Huang, Z. (2025) 'Digital media operations prediction based on user sentiment analysis and deep neural networks', *Int. J. Information and Communication Technology*, Vol. 26, No. 46, pp.20–36.

Biographical notes: Xinyu Chen received her Master's degree from the Lancaster University, UK, in 2016. She is currently a Lecturer at the Sanda University, Shanghai. Her research interests include digital media art, AI-assisted design, animation and interaction design.

Zhenbin Huang received his Master's degree from the Donghua University in 2015. He is currently an Associate Professor at the School of Art, Design and Media, Sanda University, Shanghai. His research interests include visual new media and AI-assisted design.

1 Introduction

In the current era of vigorous development of digital technology, digital media has become the core platform for information dissemination and business operations due to its advantages of fast dissemination, strong interactivity, and wide coverage (Jiang et al., 2022). From content recommendations on social media platforms to precise placement of online advertisements, the quality of digital media operations directly affects the breadth of information dissemination and the realisation of commercial value (Ahmed et al., 2019). However, traditional methods for predicting the effectiveness of digital media operations struggle to capture the complex and changing emotional factors of users, leading to deviations between prediction results and actual effects, and failing to fulfil the dynamically shifting operational demands within the digital media landscape (Kennedy et al., 2021). Users, as the core participants in digital media, have a profound impact on operational effectiveness (Friedrich and Hoel, 2023). At the same time, deep neural networks demonstrate unique advantages in processing high-dimensional and complex data due to their strong feature learning and pattern recognition capabilities, enabling the effective mining of hidden patterns in data (Grover et al., 2022). Therefore, combining user sentiment analysis with deep neural networks delivers novel concepts and cutting-edge techniques to anticipate the performance of digital media initiatives.

Digital media operation effectiveness prediction mainly involves identifying users' emotional tendencies to achieve predictions of operational effectiveness. In early research, people mostly used statistical sentiment dictionary methods. Lee et al. (2018) divided sentences based on the number of evaluative objects contained in comments on digital media platforms, and assigned corresponding sentence segments to each evaluative object. Sailunaz and Alhajj (2019) created the SentiCircles dictionary-based platform, specifically for predicting satisfaction on Twitter. The uniqueness of this platform lies in updating the sentiment polarity and rating of words based on their co-occurrence patterns in different contexts. Najafabadi et al. (2024) combined domain knowledge and dimension dictionaries to connect a word with a more general and fine-grained sentiment classification system, and represented it with quantitative calculation intensity to enhance sentiment features. Kuznetsova et al. (2023) used psychological models and sentiment dictionaries to generate discrete Gaussian distributions for the real emotional labels of sentences based on the psychological distance of emotions, thereby improving the prediction effectiveness of digital media operations.

Text sentiment analysis methods in light of conventional machine learning perform sentiment tendency judgments on text data through feature extraction and classifier training. This method first extracts sentiment features from the text, then uses machine learning algorithms to train classifiers, achieving automatic judgment of the sentiment of digital media review texts. Marturana and Tacconi (2013) used supervised learning methods to check aspects and categories in reviews, identifying emotional tendencies related to aspect terms, thereby predicting operational effectiveness. Han et al. (2019) adopted decision tree methods to classify tweet data into multiple categories such as location access, service satisfaction, and the condition and functions of existing facilities, and used the classification results as inputs for promotion strategies applied in web media. Shah et al. (2022) performs a comprehensive sentiment analysis of movie reviews using machine learning techniques like support vector machines (SVM) and Naive Bayes, and provided satisfaction prediction results. Liu et al. (2022) used the TF-IDF model to successfully convert the occurrence frequency of each unit in digital media review texts into corresponding feature values, thereby forming a vector representation of the text.

Conventional machine learning approaches depend significantly on manual feature extraction, demand extensive labelled data, and struggle to grasp nuanced semantic context, limiting their generalisation capabilities. These factors limit the accuracy and efficiency of traditional machine learning in sentiment analysis tasks. Deep learning-based approaches build deep neural network models to mechanically study deep feature representations in text, effectively capturing context information and sentiment tendencies, thus improving prediction accuracy. Zhang (2022) innovatively combined the gate mechanism with CNN for aspect-level sentiment analysis, and this fusion enabled the model to selectively extract sentiment information based on specific aspects. Abid et al. (2024) not only focused on sentiment-related features through sparse attention mechanisms, but also explored broader context semantics and integrated multi-scale features, providing new ideas for processing digital media effectiveness prediction. Long et al. (2019) suggested an LSTM method in light of the attention mechanism, which calculated attention weights by connecting aspect vectors to sentence hidden representations, and obtained users' sentiment tendencies towards digital media through a fully connected layer. The trend of image sharing on digital media is also gradually increasing, and relying solely on text content is no longer sufficient to comprehensively and accurately capture user emotions, prompting more researchers to turn to exploring multimodal digital media operation effectiveness prediction methods. Zhao et al. (2021) used CNN to capture picture characteristics and LSTM to capture text features. These features obtained from text and images were calculated to obtain specific emotional probabilities, and finally, the operational effectiveness was determined through decision-level fusion. Subbaiah et al. (2024) first extracted visual features through recurrent attention. Then, Text-CNN was used to obtain high-level semantic information from user reviews. Finally, the network effectively fused text and visual features through bilinear pooling, improving the efficiency of operational effectiveness prediction.

In summary, the limitations of traditional single-modal digital media operational effectiveness prediction methods in capturing user emotions are becoming increasingly evident. By leveraging deep learning techniques, emotional information can be accurately extracted from massive text and image data, thus achieving complementary and integrated multi-modal information. However, traditional feature fusion methods ignore the mutual influence between modalities, leading to fusion redundancy. To address this, this paper proposes a digital media operational effectiveness prediction method based on user sentiment analysis and deep neural networks. First, dynamic word vectors of review texts were obtained through the BERT pre-trained language model. CNN was used to capture text characteristics, and an attention mechanism was introduced to concentrate on important features. Improving the visual transformer model (KDViT) through knowledge distillation to reduce computational complexity. Extracting feature from review images through the KDViT model. Then, using cross-modal shared computation to guide single-modal feature vector interaction, enhancing modal feature representation through text gate mechanism, and improving the model's focus on key emotional features in multi-modal information. Finally, the cross-modal characteristic integration representation is input into the classification layer to complete digital media operational effectiveness prediction. Experimental outcome implies that the suggested approach improves forecasting accuracy by 2.2%–5.8% compared to the baseline method, and can be well applied to digital media operational effectiveness prediction.

2 Relevant technologies

2.1 Convolutional neural network

CNN represents a canonical deep learning model, with their key strength being the automatic and efficient extraction of spatially organised, hierarchical features directly from data (Zhou, 2020). This characteristic has not only achieved success in the field of image analysis, but also shown great potential in text processing and sentiment analysis tasks. In text sentiment analysis tasks, CNN captures local features in text through convolutional layers, and extracts features of different sizes through convolution operations. These local features enable the network to effectively capture key words and phrases in text, thereby improving the accuracy of sentiment categorisation (Vonder Haar et al., 2023). Unlike the spatial structure of image data, the features of text data are presented in sequence form. Therefore, when designing CNN for text processing, one-dimensional convolution operations are usually adopted to extract features directly from sequences composed of word vectors.

CNN is chiefly made up of convolutional levels, pooling levels, and fully linked levels. The convolutional level applies sliding filter kernels to the input data through convolution computations, enabling local characteristic extraction. Through dimensionality reduction, pooling layers maintain salient characteristics while aggressively compressing characteristic map sizes to optimise computational efficiency. Finally, the fully linked level summarises the learned features and outputs the final judgment or classification result. Applying CNN to sentiment analysis of evaluation texts can automatically learn feature representations that are beneficial for sentiment classification, thus reducing the workload of manually designing features in traditional methods. In addition, the advantage of CNN in capturing local semantic features provides strong technical support for understanding and analyzing users' subtle emotional changes (Ghosh et al., 2020).

2.2 Attention mechanism

After CNN and other deep learning models became a research hotspot, the attention mechanism, which has significant object detection capabilities, has been widely applied to enhance the interpretability of neural network models in deep learning. The attention mechanism dynamically allocates computational resources to relevant input segments by learning context-aware importance weights, thereby enhancing its learning ability and effectively improving the performance of sentiment analysis tasks (Brauwers and Frasincar, 2021). The attention mechanism allows the digital media operation effect prediction model to more accurately adapt to multi-featured and personalised scene characteristics through dynamic weight allocation, while enhancing interpretability and adaptability to sparse data. In the end, it not only improves the prediction index, but also provides a grounded decision-making basis for operation strategies and realises data-driven refined operation.

In its most basic form, at its core, attention implements a learned transformation function that processes query-key-value tuples to produce context-aware representations. For a given Query, by calculating the similarity or match degree between the query and a set of keys, then using these similarity scores to weight the corresponding values, and finally outputting the sum of the weighted values. This process can be summarised as follows.

$$Attention(Q, K, V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{1}$$

where Q stands for the query matrix, K stands for the key matrix, and V stands for the value matrix, d_k is the dimension of the key vector, utilised as a scaling factor for the query-key similarity measures, avoiding excessive dot product magnitudes that would push softmax into its low-gradient saturation regime, thereby affecting the propagation of gradients. The softmax operation is adopted to normalising the dot products through exponentiation and summation to create probabilistic weights, representing the weights of each value.

3 Multimodal digital media review information feature extraction based on deep neural network

3.1 Feature extraction of digital media review text based on convolutional neural network and self-attention mechanism

The effectiveness of digital media operations is often predicted through the sentiment tendency of users' comments. However, users' comments not only contain a large amount of text information but also include some image information. Traditional research has ignored multimodal comment information, leading to insufficient feature extraction and resulting in unsatisfactory prediction results. To solve the above problems, this paper first uses BERT (Bilal and Almazroi, 2023) to generate word embeddings for the review text, and leverages convolutional neural networks and attention to capture hierarchical and context-aware features from reviews. Then, an improved visual transformer (ViT) (Han et al., 2022) model is used to extract image features, obtaining an overall image feature representation, thus laying the foundation for the construction of subsequent digital media operation effect prediction models.

Digital media review text has characteristics such as strong randomness, non-standard language, and heavy word ambiguity, so it has high requirements for text pre-processing. In previous natural language processing tasks, most adopted static word vectors such as Word2Vec or Glove as model inputs. However, static word vectors have defects such as being unable to solve the problem of polysemy. The BERT model, based on a deep bidirectional Transformer encoder architecture, can generate dynamic word vectors according to different contexts, which is more in line with the original meaning of the text.

Therefore, in the study of digital media operation effectiveness, this paper adopts the BERT pre-training model as the word embedding layer of the model to fully learn the contextual semantic information of the text. The pre-processed text is input into the BERT pre-training model, and the text is segmented at the character level. The maximum sentence processing length of the BERT model is set to n, and the excess parts are truncated, while the insufficient parts are zero-padded. By learning the context information, each character is converted into a 768-dimensional dynamic word vector.

Then, the entire text T is transformed into a two-dimensional matrix $x \in R^{n \times 768}$, where $x_i \in R^{768}$ represents the vector representation of the i^{th} character in the text.

$$X = BERT(T) = \{x_1, x_2, \ldots, x_n\} \tag{2}$$

The dynamic word vectors generated by BERT are pre-trained on large-scale general corpora, but still need to be combined with other models for further feature extraction to improve the semantic accuracy of text vector representation. This paper uses TextCNN for text feature extraction. TextCNN is a text classification model based on CNN. Compared with CNN in traditional image processing tasks, the network structure of TextCNN has no changes. It extracts local information of different scales in text through convolutional kernels of different sizes. Usually, after convolutional calculations, a pooling layer is used for dimensionality reduction. However, the pooling operation may cause problems such as loss of feature information. Consequently, this article adopts the attention mechanism to replace the pooling operation, and assigns weights to the feature matrix after convolutional operations to highlight important local feature information.

The textual feature matrix X derived from BERT's pre-trained representations serves as the input to the TextCNN architecture. Three convolutional kernels of various sizes are used to extract local characteristics of the text. The expression of the feature c_{ri} extracted by the i^{th} convolution operation with a convolutional kernel of window size r is as follows.

$$c_{ri} = f\left(W \cdot x_{i:i+r-1} + b\right) \tag{3}$$

where $f(.)$ is the activation operation, W is the parameter of the convolutional kernel; $x_{i:i+r-1}$ is the word vector from the i^{th} row to the $i + r - 1^{\text{th}}$ row; b is the bias term; r is the size of the convolutional kernel. In the text vector matrix, a total of $n - r + 1$ convolution operations can be performed, and the extracted local characteristic vector c_r is expressed as bellow.

$$c_r = \left[c_1, c_2, \ldots, c_{n-r+1}\right] \tag{4}$$

Usually, after CNN extracts features through convolution, a max pooling layer is used for feature dimensionality reduction, but this method often causes problems such as feature loss. In sentiment analysis tasks, keywords with strong emotional connotations in the text often have a significant impact on the overall sentiment tendency of the text. Assigning larger weights to these keywords using the attention mechanism is of great significance for elevating the sentiment polarity recognition capability. Therefore, the suggested methodology employs attention weighting as an alternative to conventional max pooling operations, and assigns attention weights to the feature vectors obtained through convolution operations, mitigating the feature degradation inherent in max-pooling operations. The calculation equation is as follows.

$$h_r = \tanh\left(W \cdot c_r + b\right) \tag{5}$$

After normalising the weight vector h_r, the attention score α_t can be obtained. Then, the attention score is calculated with the sub-vector of the feature matrix to obtain the text feature vector s_c. The computation equation is as bellow.

$$\alpha_r = \frac{\exp(h_r)}{\sum_{i=1}^{n} \exp(h_r)} \tag{6}$$

$$s_c = \sum_{i=1}^{n} \alpha_r h_r \tag{7}$$

Then, the feature vectors s_c obtained from three different convolutional kernels are concatenated to obtain the final digital media review text feature vector output s.

3.2 Digital media review image feature extraction based on improved vision transformer

When processing images, alleviating the representational impoverishment characteristic of max-pooling transformations, which leads to high computational complexity. To tackle the previously outlined problems, this article introduces the idea of knowledge distillation to optimise the ViT model. The KDViT model introduces a distillation token into the transformer structure, collaboratively interacting with both classification and patch tokens during attention-based feature learning. The primary objective of the class token is to align with ground truth labels, whereas the distillation token aims to match the teacher model's predicted labels. In this way, the KDViT learns not only from the real labels but also from the predictions of the teacher model during training, thereby effectively extracting and integrating knowledge. During training, the KDViT adopts the knowledge distillation method (Song et al., 2022). If it is soft distillation, the model calculates a distillation loss with the output of the teacher model and adds this loss to the model's classification loss for backpropagation. If it is hard distillation, the model calculates the cross-entropy among the output results and the output of the teacher model, and adds this cross-entropy loss to the model's classification loss for backpropagation. In this way, the KDViT model can continuously optimise its performance and enhance its generalisation ability on complex tasks.

For the input image data $I = \{i_1, i_2, i_3, \ldots, i_i, \ldots, i_n\}$, in which i_i stands for the embedded representation of the i^{th} input image. Using the self-attention (SA) scheme in the Transformer, the KDViT model performs global attention focusing on the embedded sequence I to extract the relationships among elements in the sequence. Then, through attention focusing, the model generates a weighted sum of each embedded block, forming a sequential representation. Finally, the obtained sequential representation is sent to the global pooling level to achieve the overall image characteristic representation, and the image feature representation F_v is as follows.

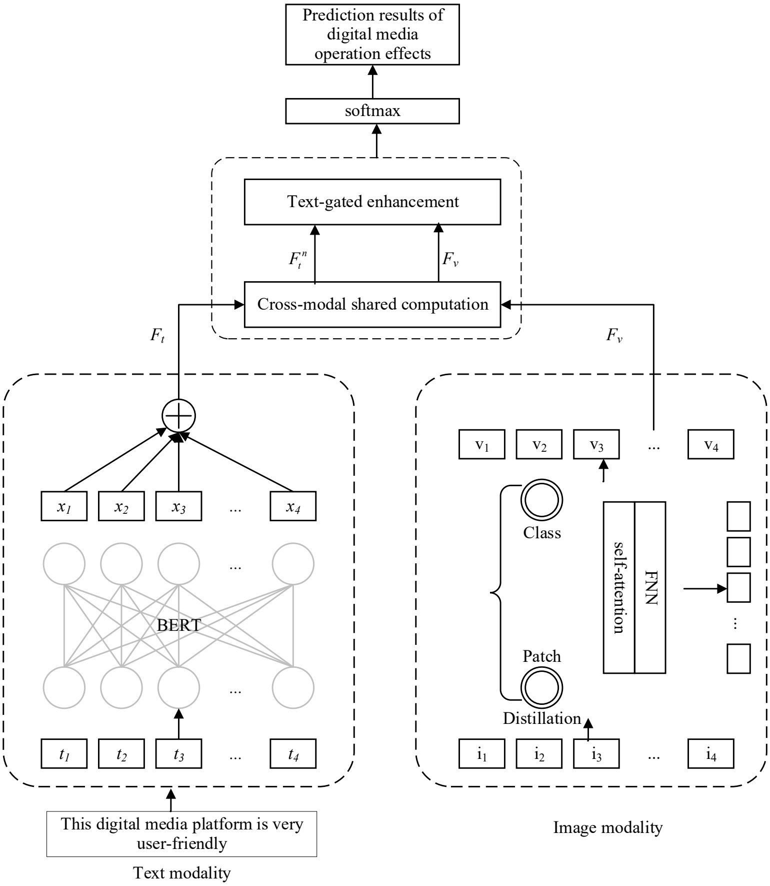
$$F_v = KDViT(I) = \{v_1, v_2, v_3, \ldots, v_n\} \tag{8}$$

4 Digital media operation effect prediction based on user multimodal sentiment analysis

4.1 The structure of the digital media operation effect prediction model

As the digital media platforms quickly growing, the comment data generated by users on digital media have increasingly shown a multimodal trend, mainly in the form of combinations of images and text. The emergence of this multimodal data has made traditional unimodal sentiment analysis methods insufficient in capturing user emotional expressions. Directly applying multimodal sentiment analysis methods to data on multimodal digital media platforms faces a series of challenges. The single text data on digital media platforms are shorter compared to the text traditionally used for sentiment analysis, and they contain relatively less information. Existing text sentiment analysis methods find it difficult to efficiently and accurately extract modal features and sentiment tendencies. Secondly, digital media platform data have a strong emotional characteristic. Different features may describe the same emotion or attitude, which means that information redundancy may occur during feature fusion. Therefore, during the fusion process, more attention should be paid to more important and valuable features.

Figure 1 The structure of the digital media operation effect prediction model



To cope with the above issues, this chapter puts forward a digital media operation effect prediction model based on multimodal affective analysis, as shown in Figure 1. The model is classified into four parts: text feature extraction, image feature extraction, cross-modal shared enhancement, and user emotional output of operation effect. In the text characteristic extraction phase, the BERT model is used to obtain text embedding representation, and CNN is introduced to capture the global characteristics of the text. In the image characteristic extraction phase, KDViT is adopted for picture characteristic vector representation. In the characteristic fusion phase, a cross-modal shared enhancement computing mechanism is used, which enhances the robustness of different modal affective feature expressions by establishing close interactions between different modal data, thus reducing fusion redundancy caused by multimodal data fusion. This mechanism can effectively integrate the correlation between text and images, improving the expressive ability of fused features. Finally, in the digital media operation effect prediction layer, a softmax classifier is used to complete the digital media operation effect prediction task. This layer performs the final emotional classification of digital media operation effect by integrating the bimodal features. The entire model structure covers multiple key layers, and can combine text and image information at the same time, achieving more accurate operation effect prediction.

4.2 Multimodal feature fusion based on text reinforcement

In user emotion analysis, the text modality occupies a dominant position. Compared to removing other modality data, removing the text modality data results in a significant decline in model accuracy and performance. Text and image data on digital media have deep semantic relevance. Directly concatenating single-modality feature vectors easily causes emotion loss and fusion redundancy. Considering that different modality data provide different semantic richness in emotion analysis, a modality-sharing enhancement module is designed in the multimodal feature fusion stage. This module consists of two key parts: cross-modal shared computing and text gate enhancement.

1 *Cross-modal shared computing:* This process is built based on the text modality. Taking the text modality feature representation as the core, first, the text modality improves its characteristic representation via SA, enabling deep exploration of the rich semantic features within the text. At the same time, the text modality guides the interaction with the image modality, completing the complementary fusion between modalities using the attention mechanism. The characteristic representation of the text modality can guide the parsing direction of the image modality, providing necessary semantic information to help the image modality better understand the scene and objects. The model achieves feature complementarity between modalities through this process.

   This part consists of a cross-attention (CA) part and a SA part. In CA, s_c is used as the input for the text modality feature, and F_v is used as the input for the image modality feature. First, to achieve normalisation and collaborative training between different modalities, the input feature vectors s_c and F_v are normalised using layer normalisation. The calculation process is shown in equations (9) and (10).

$$s_C = LN(s_c) \tag{9}$$

$$F_V = LN(F_v) \tag{10}$$

The processed feature representations are used with the CA mechanism to establish associations between different modality representations, and the image-related feature representation Fs_{v-c} is obtained by balancing the contributions of different modalities, as shown below.

$$Fs_{v-c} = CA_v(F_V, s_c) = softmax\left(\frac{s_c W_{Q_c} W_{K_V}^T F_V^T}{\sqrt{d}}\right) F_V W \tag{11}$$

where W_{Q_c}, W_{K_V}, and W_{V_V} represent the weight matrices for the query vector, key vector, and value vector, respectively. Then, the SA is applied to the text modality feature s_C to dynamically modify the attention weights in the text modality feature, obtaining the enhanced feature representation s_c^n, as shown below.

$$s_c^n = SA_c(s_C) = softmax\left(\frac{F_C W_{QC} W_{KC}^T s_C^T}{\sqrt{d}}\right) s_C W_{VC} \tag{12}$$

The adaptive fusion network is used to process the text-enhanced feature representation s_c^n and the image-related feature representation Fs_{v-c}, as shown in equations (13) and (14).

$$G = \sigma\left(s_c^n * W_c + F_{v-c}^* W_{v-c} + b\right) \tag{13}$$

$$F_c = G \odot s_c^n + (1 - G) \odot s_{v-c} \tag{14}$$

where σ stands for the sigmoid nonlinear operation, ⊙ stands for element-wise multiplication. By learning the parameters W_t, W_V, and b, the proportions of F_c^n and F_v are determined to filter out incorrect information generated by cross-modal interactions, and to measure the fusion ratio of the two modalities. Then, F_c^n and F_v are processed through a feed-forward layer (PPF), as shown in equations (15) and (16).

$$F_{vm} = PFF\left(LN(F_v)\right) + F_v \tag{15}$$

$$F_{cm}^n = PFF\left(LN(F_c^n)\right) + F_c^n \tag{16}$$

2 *Text gate reinforcement:* After cross-modal shared computation, a unified dimension feature of multi-modal is obtained. This process not only ensures the consistency of different modal information in dimension, but also fully integrates the unique information of each modality. Considering the complex relationships within each modality and the complementarity between cross-modal, a text gate reinforcement mechanism is further introduced. This mechanism uses the rich semantic information and context clues in the text modality as prior knowledge to guide the self-enhancement of the image, balance the contribution of different modalities, make the model more accurately identify and parse key features in the image, and effectively improve the quality and robustness of multi-modal feature representation. The framework of this process is indicated in Figure 2.

Considering the complementarity between the internal emotional characteristics of each modality and the cross-modal emotional features, while performing attention processing on F_{cm}^{n} and F_{vm}, F_{cm}^{n} is used as a gate (g), and learning parameters θ_g to activate or deactivate the corresponding vector channels as needed, as shown in equations (17) and (18).

$$g = \sigma\left(Linear\left(F_{cm}^{n}; \theta_g\right)\right) \tag{17}$$

$$gF_v = (1+g) \odot F_{cn}^{n} \tag{18}$$

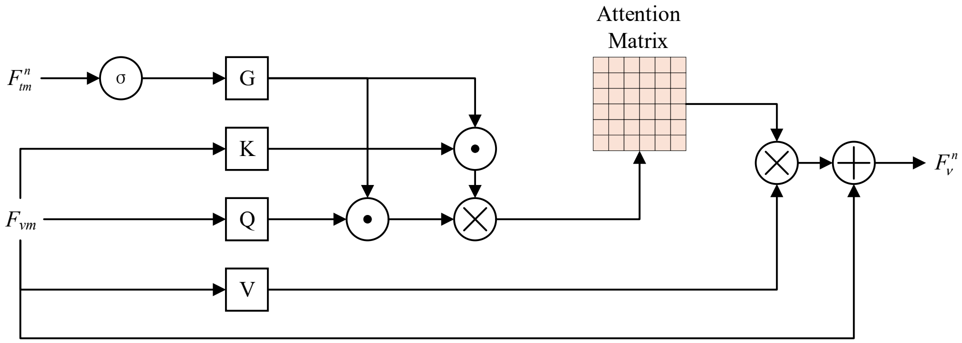
Using the text modality as a gate to control and adjust the key vectors and query vectors from the image modality to obtain the image-enhanced feature representation F_v^n, which dynamically adjusts the relationship between different modalities to better capture cross-modal information interaction, as shown in equation (19).

$$F_v^n = softmax\left(\frac{gF_{vm}W_{Q_v}W_{K_v}^{T}gF_{vm}^{T}}{\sqrt{d}}\right)F_{vm}W_{Vv} + F_{vm} \tag{19}$$

where W_{Q_v}, W_{K_v}, and W_{V_v} represent the weight matrices of the query vector, key vector, and value vector, respectively. Finally, the text-enhanced feature representation F_c^n and image-enhanced feature representation F_v^n obtained through modal shared enhancement are concatenated to obtain the final multimodal feature vector F, as shown in equation (20).

$$F = F_{cm}^{n} \oplus F_v^n \tag{20}$$

Figure 2 Text gating enhanced architecture diagram (see online version for colours)



4.3 Output of digital media operation effect prediction

Using the learned multimodal feature fusion vector F as input, the softmax activation operation is adopted as the sentiment polarity classifier to complete the sentiment classification task. The softmax activation function converts the input real values into the probability distribution of digital media operation effect prediction, guaranteeing that every output element lies in the interval [0, 1], with their summation constrained to 1. This makes the model's output interpretable as the probability of each category, which is convenient for intuitive understanding and interpretation of the model's output. The explicit calculation process is implied in equation (21).

$$\rho = softmax(F) \tag{21}$$

The model uses the cross-entropy loss function (Shim, 2024) to calculate the loss between the model output ρ and the true label y. The calculation of the loss function is shown in formula (22).

$$Loss = -\sum_{i=1}^{N} y_i \cdot \log\left(\rho_i\right) \tag{22}$$

where y_i is the i^{th} element of the true label, and ρ_i is the i^{th} element of the probability distribution of the model output.

5 Experimental results and analyses

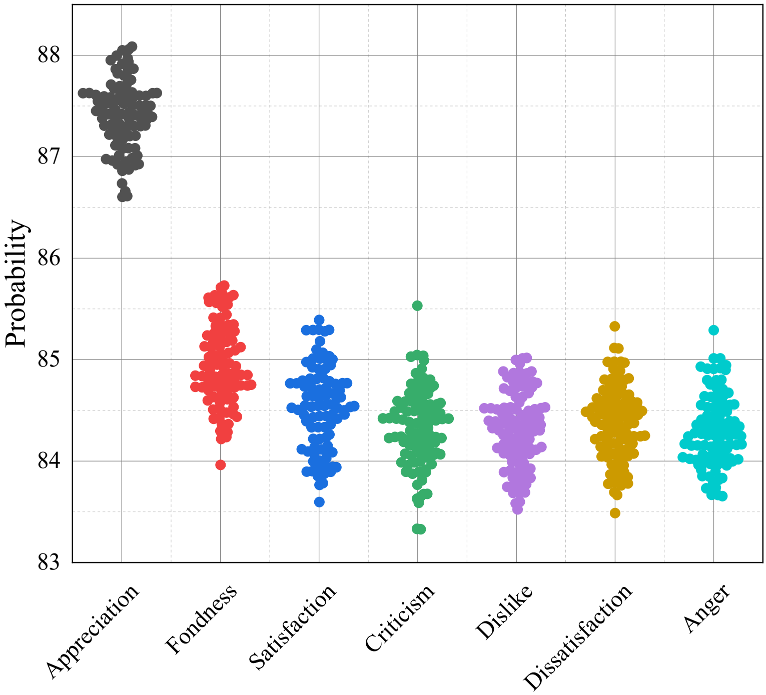
This paper uses the multimedia platform comment text information collected in the literature (Liu et al., 2024) as the dataset. After data processing, the dataset comprises 21,532 valid samples, split into training, validation, and test sets at an 8:1:1 ratio to ensure model generalisation capability. The experiment adopts ten-fold cross-validation to enhance the reliability of the results, and final outcome are based on the average of these ten experiments. The GPU type used in the experiment is GeForce RTX 3090, with a memory capacity of 24 G. The model is executed with the PyTorch framework within the Windows operating system, and the Python version is 3.7. During the experiment, considering the parameters and computational load of the model, the word embedding dimension of the comment text features is set to 768, the maximum length of the text is set to 128, and the studying rate is set to 0.0001, the batch size of the MVSA dataset is set to 64, the batch size is set to 32, and the dropout technique is used to reduce the overfitting of the model, with the parameter set to 0.1. Adam is used as the optimiser of the model to enhance the convergence speed and generalisation ability of the model.

The bee swarm graph of different sentiment tendencies for digital media operation effect prediction in the suggested approach USA-DNN is implied in Figure 3. The x-axis represents the number of days of digital media platform usage, and the y-axis stands for the probability of digital media operation effect prediction. The distribution of points for samples of digital media operation effect prediction in the first to seventh days of various users fully reflects the distribution of sentiment tendency data. As can be seen from Figure 3, as the amount of days increases, the prediction probabilities of both positive and negative sentiments tend to stabilise on the fourth day, with the maximum fluctuation of the negative sentiment probability reaching 18%, while the maximum fluctuation of the positive sentiment transition probability is only 3%.

To further verify the prediction effect of the suggested approach USA-DNN, this paper selects SVM-NB (Shah et al., 2022), CNN-GRU (Zhang, 2022), LSTM-AM (Long et al., 2019), CNN-LSTM (Zhao et al., 2021) and TCNN-RM (Subbaiah et al., 2024) as baseline methods. The performance evaluation indicators are accuracy, recall, precision, specificity, and F1. The prediction accuracy of different methods is shown in Figure 4. SVM-NB needs to iterate 140 times for its prediction accuracy to converge to 89.5%.

CNN-GRU, LSTM-AM, and CNN-LSTM need to iterate 130 times, and their prediction accuracy can converge to 90.8%, 91.2%, and 92.6%, respectively. TCNN-RM needs to iterate 100 times for its prediction accuracy to converge to 93.1%. However, USA-DNN only needs to iterate 70 times for its prediction accuracy to converge to 95.3%. USA-DNN integrates soft attention mechanism and image global sentiment perception technology to obtain accurate single-modal feature representation. During the cross-modal characteristic integration process, the text modality is used as the core carrier of information expression, and interacts and enhances with the image modality by utilising its rich semantic information and context clues, thereby improving the accuracy of digital media operation effect prediction.

Figure 3 Prediction probability of emotional tendency for digital media operation effect (see online version for colours)



The prediction performance indicators of different methods are compared in Table 1. The harmonic mean F1 of USA-DNN's recall and precision is 95.04%, which is improved by 6.51%, 5.62%, 3.9%, 2.2%, and 1.61% respectively compared to SVM-NB, CNN-GRU, LSTM-AM, CNN-LSTM, and TCNN-RM. The specificity of USA-DNN is 0.9982, which is improved by 1.64%–9.92% compared to the baseline methods. To address the interaction issues between different modalities, USA-DNN constructs a cross-modal shared computing module, which realises effective interaction between text and picture information in the shared space. Through the text-dominated cross-modal attention module, text information can receive more attention during the fusion process. To more accurately measure the weights of different modalities during the fusion process, an adaptive fusion network is introduced, which dynamically determines the fusion ratio according to the feature representation of different modalities, and controls the contribution of different modalities during the fusion process. To further improve the fusion efficiency, the model performs a jump processing on the feature representation to eliminate the differences and conflicts between different modalities. In terms of cross-modal reinforcement, a text-gated reinforcement module is used, which uses the information of the text modality as a gating signal to screen and reinforce the features of the image modality, deeply exploring the semantic information on the time series, while reducing the redundant information in the multi-modal fusion process and ignoring the noise interference in the image modality. Through the combined action of the above modules, the model demonstrates stronger modal learning ability in the digital media operation effect prediction task.

Figure 4 The prediction accuracy of different methods (see online version for colours)

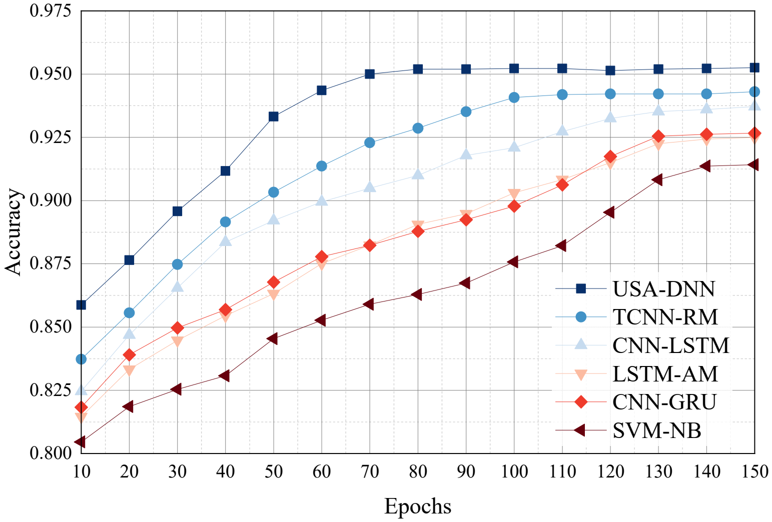


Table 1 Performance comparison of digital media operation effect prediction

Month	*July*	*August*	*September*	*October*	*November*	*December*
Number of contracts	1,117	1,345	1,724	1,325	7,545	5,163
Transportation costs before optimisation	834.6	787.3	840.1	810.8	842.4	682.4
Optimised transportation costs	745.1	603.6	751.3	722.9	767.5	432.6
Proportion	89.26%	76.67%	89.44%	89.16%	91.11%	63.40%

6 Conclusions

Multimodal data on digital media platforms often exhibit the characteristics of short text information and casually styled images. This feature makes traditional feature fusion methods difficult to accurately capture the mutual influence between modalities, leading to the problem of unsatisfactory operational effect prediction. To address this, this paper proposes an operational effect prediction method for digital media based on user sentiment analysis and deep neural networks. First, dynamic word vectors of comment texts are obtained through the BER model, CNN is adopted to capture text characteristics, and employed to augment critical feature representations. A visual transformer model improved by knowledge distillation is adopted to capture characteristics from comment pictures. Then, cross-modal shared computing is used to guide the interaction of single-modal feature vectors. A text gating mechanism is used to optimise the modal characteristic representation and improve the model's focus on key emotional features in multimodal information. Finally, the cross-modal characteristic integration representation is input into softmax to accomplish the operational effect prediction of digital media.

This study only considers the text and image data in digital media platforms. Future research can be carried out from the following perspectives.

1 Video data, as a composite carrier integrating multiple modal information such as text, speech, and images, is increasingly showing analytical value. Therefore, future research can explore digital media operational effect analysis that includes videos.

2 In the actual data collection and processing process, due to various reasons, it is difficult to ensure that each sample contains complete modal information. When some samples in the data set are missing one or more modalities, the phenomenon of modal missing occurs, which makes the model unable to fully utilise all available data, thus reducing its overall performance. How to enhance the generalisation capability of the model under the condition of modal missing, so that it can better handle the problem of modal loss, will be a vital direction for future research.

Declarations

All authors declare that they have no conflicts of interest.

References

Abid, F., Rasheed, J., Hamdi, M., Alshahrani, H., Al Reshan, M.S. and Shaikh, A. (2024) 'Sentiment analysis in social internet of things using contextual representations and dilated convolution neural network', *Neural Computing and Applications*, Vol. 36, No. 20, pp.12357–12370.

Ahmed, R.R., Streimikiene, D., Berchtold, G., Vveinhardt, J., Channar, Z.A. and Soomro, R.H. (2019) 'Effectiveness of online digital media advertising as a strategic tool for building brand sustainability: evidence from FMCGs and services sectors of Pakistan', *Sustainability*, Vol. 11, No. 12, pp.34–46.

Bilal, M. and Almazroi, A.A. (2023) 'Effectiveness of fine-tuned BERT model in classification of helpful and unhelpful online customer reviews', *Electronic Commerce Research*, Vol. 23, No. 4, pp.2737–2757.

Brauwers, G. and Frasincar, F. (2021) 'A general survey on attention mechanisms in deep learning', *IEEE Transactions on Knowledge and Data Engineering*, Vol. 35, No. 4, pp.3279–3298.

Friedrich, K. and Hoel, A.A. (2023) 'Operational analysis: a method for observing and analyzing digital media operations', *New Media & Society*, Vol. 25, No. 1, pp.50–71.

Ghosh, A., Sufian, A., Sultana, F., Chakrabarti, A. and De, D. (2020) 'Fundamental concepts of convolutional neural network', *Recent Trends and Advances in Artificial Intelligence and Internet of Things*, Vol. 4, pp.519–567.

Grover, P., Kar, A.K. and Dwivedi, Y.K. (2022) 'Understanding artificial intelligence adoption in operations management: insights from the review of academic literature and social media discussions', *Annals of Operations Research*, Vol. 30, No. 2, pp.177–213.

Han, J., Fang, M., Ye, S., Chen, C., Wan, Q. and Qian, X. (2019) 'Using decision tree to predict response rates of consumer satisfaction, attitude, and loyalty surveys', *Sustainability*, Vol. 11, No. 8, pp.23–36.

Han, K., Wang, Y., Chen, H., Chen, X., Guo, J., Liu, Z., Tang, Y., Xiao, A., Xu, C. and Xu, Y. (2022) 'A survey on vision transformer', *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 45, No. 1, pp.87–110.

Jiang, X., Mao, T. and Tian, J. (2022) 'The application of digital technology in the complex situation of news dissemination from the perspective of new media art', *Computational Intelligence and Neuroscience*, Vol. 20, No. 1, pp.16–24.

Kennedy, H., Kunkel, T. and Funk, D.C. (2021) 'Using predictive analytics to measure effectiveness of social media engagement: a digital measurement perspective', *Sport Marketing Quarterly*, Vol. 30, No. 4, pp.265–277.

Kuznetsova, Y.M., Chuganskaya, A.A. and Chudova, N.V. (2023) 'Organization of emotional reactions monitoring of social networks users by means of automatic text analysis', *Artificial Intelligence and Decision Making*, No. 2, pp.64–75.

Lee, S.Y., Qiu, L. and Whinston, A. (2018) 'Sentiment manipulation in online platforms: an analysis of movie tweets', *Production and Operations Management*, Vol. 27, No. 3, pp.393–416.

Liu, H., Chen, X. and Liu, X. (2022) 'A study of the application of weight distributing method combining sentiment dictionary and TF-IDF for text sentiment analysis', *IEEE Access*, Vol. 10, pp.32280–32289.

Liu, Z., Yang, T., Chen, W., Chen, J., Li, Q. and Zhang, J. (2024) 'Sentiment analysis of social media comments based on multimodal attention fusion network', *Applied Soft Computing*, Vol. 164, pp.11–20.

Long, F., Zhou, K. and Ou, W. (2019) 'Sentiment analysis of text based on bidirectional LSTM with multi-head attention', *IEEE Access*, Vol. 7, pp.141960–141969.

Marturana, F. and Tacconi, S. (2013) 'A machine learning-based triage methodology for automated categorization of digital media', *Digital Investigation*, Vol. 10, No. 2, pp.193–204.

Najafabadi, A.J., Skryzhadlovska, A. and Valilai, O.F. (2024) 'Agile product development by prediction of consumers' behaviour; using neurobehavioral and social media sentiment analysis approaches', *Procedia Computer Science*, Vol. 232, pp.1683–1693.

Sailunaz, K. and Alhajj, R. (2019) 'Emotion and sentiment analysis from Twitter text', *Journal of Computational Science*, Vol. 36, pp.10–23.

Shah, P., Swaminarayan, P. and Patel, M. (2022) 'Sentiment analysis on film review in Gujarati language using machine learning', *International Journal of Electrical and Computer Engineering*, Vol. 12, No. 1, pp.1030–1039.

Shim, J.W. (2024) 'Enhancing cross entropy with a linearly adaptive loss function for optimized classification performance', *Scientific Reports*, Vol. 14, No. 1, pp.27–35.

Song, J., Chen, Y., Ye, J. and Song, M. (2022) 'Spot-adaptive knowledge distillation', *IEEE Transactions on Image Processing*, Vol. 31, pp.3359–3370.

Subbaiah, B., Murugesan, K., Saravanan, P. and Marudhamuthu, K. (2024) 'An efficient multimodal sentiment analysis in social media using hybrid optimal multi-scale residual attention network', *Artificial Intelligence Review*, Vol. 57, No. 2, pp.34–45.

Vonder Haar, L., Elvira, T. and Ochoa, O. (2023) 'An analysis of explainability methods for convolutional neural networks', *Engineering Applications of Artificial Intelligence*, Vol. 117, pp.10–17.

Zhang, R. (2022) 'Digital media teaching and effectiveness evaluation integrating big data and artificial intelligence', *Computational Intelligence and Neuroscience*, Vol. 20, No. 1, pp.46–52.

Zhao, J., Lin, J., Liang, S. and Wang, M. (2021) 'Sentimental prediction model of personality based on CNN-LSTM in a social media environment', *Journal of Intelligent & Fuzzy Systems*, Vol. 40, No. 2, pp.3097–3106.

Zhou, D-X. (2020) 'Universality of deep convolutional neural networks', *Applied and Computational Harmonic Analysis*, Vol. 48, No. 2, pp.787–794.