IIETA International Information and Engineering Technology Association
Advancing the World of Information and Engineering

Traitement du Signal
Vol. 41, No. 6, December, 2024, pp. 3039-3047
Journal homepage: http://iieta.org/journals/ts

Image Content Analysis for Social Media Public Opinion Monitoring and Response Strategies

Lina Lin, Dezhi Wei*

Information Engineering School, Jimei University Chengyi College, Xiamen 361021, China

Corresponding Author Email: aide@jmu.edu.cn

https://doi.org/10.18280/ts.410621

Received: 29 June 2024
Revised: 16 October 2024
Accepted: 10 November 2024
Available online: 31 December 2024

Keywords:
social media, public opinion monitoring, image content analysis, deep learning, sentiment analysis, public opinion response strategies

ABSTRACT

With the widespread use of social media, the formation and dissemination speed of online public opinion has accelerated, and the influence of public opinion events has become increasingly significant. Traditional public opinion monitoring methods mainly rely on text analysis. However, in the context of social media, multimedia content such as images and videos has become an important carrier of public opinion dissemination. Images not only convey emotional information in a direct manner but also play a key role in public opinion events. Therefore, image-based public opinion monitoring has become a research hotspot and a challenge. Existing studies mainly focus on text analysis, with insufficient in-depth analysis of image content, and there are certain limitations in areas such as semantic understanding and sentiment orientation judgment. This paper aims to explore how to enhance the accuracy of social media public opinion monitoring and response strategies through image content analysis. Firstly, the paper analyzes the shortcomings of traditional public opinion monitoring methods in terms of semantic usage and proposes improvement ideas. Secondly, an image content analysis model for social media public opinion monitoring is constructed, using deep learning and other technologies to extract emotional and social inclination information from images. Finally, based on the results of image content analysis, response strategies for social media public opinion are proposed, providing theoretical support and practical guidance for public opinion management and crisis response. This study not only addresses the shortcomings of existing methods and improves the accuracy of public opinion monitoring but also provides feasible suggestions for responding to social media public opinion, offering significant application value.

1. INTRODUCTION

With the widespread application of social media, the formation and dissemination speed of online public opinion have greatly accelerated, and the influence of public opinion events has also increased significantly [1, 2]. In this context, how to effectively monitor and manage public opinion information on social media platforms has become a focal point of concern for society [3-6]. Social media not only carries a large amount of text information but also includes a large number of images, videos, and other multimedia content, which play an increasingly important role in the dissemination of public opinion events [7, 8]. Therefore, how to use image content analysis techniques to deeply mine public opinion information on social media has become an urgent problem in the field of public opinion monitoring.

The significance of research on public opinion monitoring and response lies in its ability to help governments, enterprises, and other relevant departments understand the emotional changes and public opinion trends of society in a timely manner, and also provide a scientific basis for crisis management. Traditional public opinion monitoring methods mostly focus on the analysis of text information, neglecting the potential value of image content [9-13]. However, in the current social media environment, images and videos often convey emotions and information that are more intuitive and richer than text. Especially in the dissemination of emergencies or sensitive topics, image content can quickly attract widespread attention [14-18]. Therefore, research on image content-based public opinion monitoring methods can not only improve the accuracy of public opinion monitoring but also provide more comprehensive decision-making support for public opinion management and response.

Although some progress has been made in the field of public opinion monitoring, most studies are still limited to the processing of text information, lacking in-depth analysis of image content. In addition, existing image content analysis methods often focus on basic feature extraction of images, without fully mining the potential emotional tendencies and social influence within the images [19-24]. Moreover, traditional public opinion monitoring models often have biases in their use of semantics, leading to insufficient accuracy and effectiveness of monitoring results. Therefore, existing research methods are difficult to meet the practical application needs in the face of the complex and dynamic social media environment.

This paper aims to explore image content-based social media public opinion monitoring and response strategies. The

research mainly includes three parts: First, analyzing the issues related to the improper use of semantics in traditional public opinion monitoring methods and proposing improvement strategies; second, constructing an image content analysis model for social media public opinion monitoring to improve the precision of image information interpretation and emotional judgment capability; third, based on the results of image content analysis, proposing corresponding public opinion response strategies for social media, to help relevant institutions respond to public opinion crises in a timely and effective manner. This study will not only help improve the accuracy of public opinion monitoring technology but also provide governments and enterprises with more scientific public opinion response solutions, with important theoretical value and practical significance.

2. ANALYSIS OF IMPROPER SEMANTIC USE IN SOCIAL MEDIA PUBLIC OPINION MONITORING

Social media platforms such as Weibo, WeChat, Instagram, etc., often feature user-generated content that includes images and videos. These images carry rich emotional expressions and social viewpoints during dissemination, and are often more intuitive and attention-grabbing than pure text. For example, in the case of emergencies, social movements, or political controversies, images may be more likely to trigger emotional responses from the public, quickly sparking widespread discussions. An image content analysis model for social media public opinion monitoring can be applied to extract valuable information from the vast image data on social media platforms to assist in real-time monitoring and sentiment judgment. The model can use techniques such as image recognition, sentiment analysis, and emotion recognition to automatically identify people, scenes, expressions, and the emotional tendencies behind them in the images, thus determining the direction and potential risks of public opinion.

In image content analysis for social media public opinion monitoring, a serial decoder structure based on the attention mechanism can couple visual features with semantic features to achieve image-to-text conversion. However, this method of visual-semantic feature coupling may lead to issues of improper semantic use, primarily reflected in the model's dependence on vocabulary. Specifically, the strong coupling of visual and semantic features means that semantic information is often learned from a specific word bank in the image training set, which typically has limited corpus size and high noise content. In social media public opinion monitoring, the semantic information carried by images may have emotional overtones or implied intentions, and different users express themselves in very different ways. This makes semantic models based on a fixed word bank susceptible to noise and limitations, leading to inaccurate or inappropriate semantic inference. Especially when there are few training samples and the vocabulary of the training set is small, the coupled semantic information may be biased towards specific expressions, affecting the comprehensive and precise understanding of public opinion. Figure 1 illustrates the framework structure of the attention mechanism decoder with visual-semantic feature coupling.

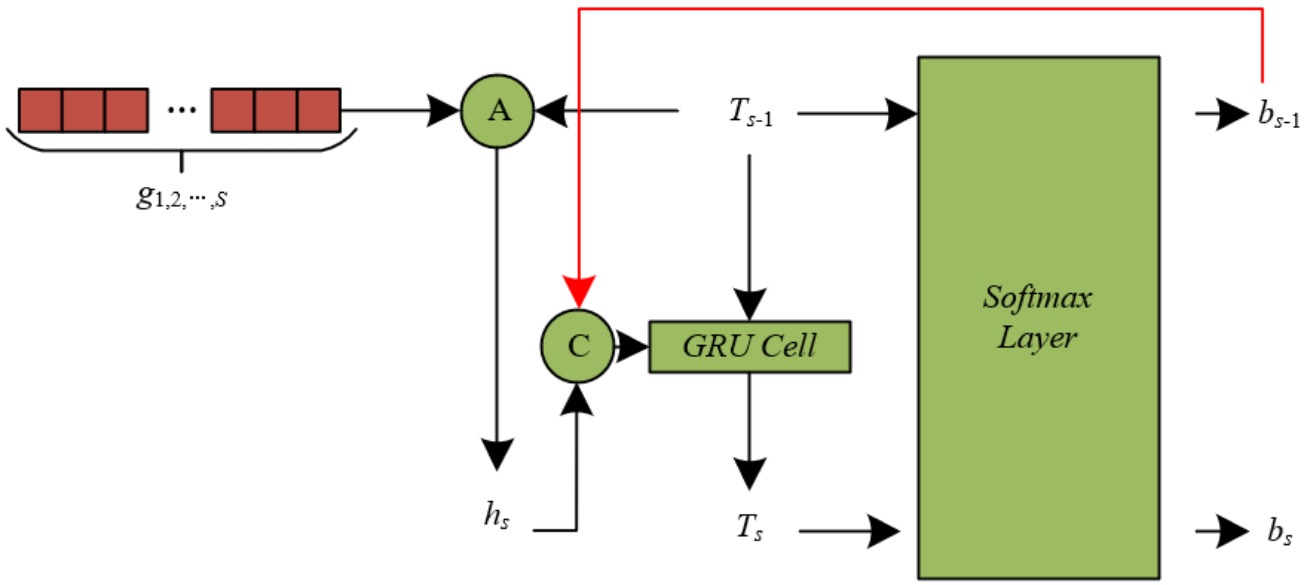

Figure 1. Framework structure of the attention mechanism decoder with visual-semantic feature coupling

The decoder based on visual-semantic feature coupling combines visual features and semantic features through serial decoding, gradually generating text predictions. This design essentially builds a character-level language model implicitly through the attention mechanism, relying on the predicted character from the previous step and the visual features of the current step for character-level language modeling. However, it is this coupling method that leads to potential improper use of semantic information, which is particularly evident in social media public opinion monitoring tasks. Specifically, suppose the one-dimensional visual feature sequence obtained after the input image passes through Convolutional Neural Network (CNN) and bidirectional Long Short-Term Memory (LSTM) is represented as $g_{1,2,\ldots,s}$, the hidden state from the previous time step is represented as T_{s-1}, and the predicted character from the previous time step is represented as B_{s-1}. We first treat ts-1 as a query, and $g=(g_{1,2,\ldots,s})$ as both the key and value. After passing through an attention mechanism module, the character visual feature h_s at step s can be obtained:

$$h_s = X\left(t_{s-1}, g, g\right) \tag{1}$$

Let the character embedding mapping function be denoted as $d(*)$. Further, the hidden state at step s can be expressed as:

$$t_s = GRUCell\left(t_{s-1}, Z\left(h_s, d\left(b_{s-1}\right)\right)\right) \quad (2)$$

Similarly, the predicted character b_s at step s is obtained from the hidden state t_s through a Softmax layer.

From the above process, it can be seen that due to the strong coupling of visual and semantic features in the decoder, the model's semantic modeling ability is constrained by the vocabulary in the training set. In the task of social media public opinion monitoring, social media images often contain rich emotions, implicit social intentions, and even specific cultural backgrounds and contextual information. However, traditional visual-semantic coupling decoders can only rely on the vocabulary included in the image training set for semantic learning. Since these word banks are often limited and may contain much noise, such as random symbols or non-standard terms, this noise can adversely affect the model's semantic learning, making it difficult for the decoder to effectively capture the true semantics behind the images. For example, some images may contain text with sarcastic or ambiguous meanings, and due to the limitations of the word bank, these complex semantics are hard to recognize and understand in traditional decoders.

In addition, the language model in the model can only predict subsequent characters based on known character sequences, and cannot consider the global semantic information comprehensively. In the context of social media public opinion monitoring, the dissemination of public opinion sentiment is often highly time-sensitive and contextual. The semantics of information may not only be determined by a single character sequence but is also influenced by multiple factors, such as context, image content, and social culture. Therefore, when the decoder relies only on the predicted character from the previous step and the limited vocabulary in the training set for character prediction, it lacks an accurate grasp of global semantics, which may lead to improper semantic use or an over-reliance on local information, thus affecting the comprehensive recognition and precise understanding of potential public opinion on social media.

More seriously, due to the possible presence of significant noise and non-standard expressions in social media images, the visual-semantic feature coupling decoder is easily influenced by this chaotic information during the learning process. Especially in certain special scenarios, such as meme images, malicious comments, or intentionally distorted content, the model may misinterpret these non-standard symbols or expressions as valid semantic information, thereby reducing the model's ability to identify public opinion. The openness and diversity of social media inevitably introduce unrelated or erroneous semantic data into the training set vocabulary, and traditional coupling decoders struggle to effectively distinguish and eliminate this noise, leading to incorrect public opinion analysis conclusions by the model.

In summary, the primary cause of improper semantic use in social media public opinion monitoring arising from the visual-semantic feature coupling decoder is its over-reliance on a limited and potentially noise-affected vocabulary, lacking accurate modeling of global semantics. This results in the model being unable to effectively handle the complexity and diversity of social media content. To address this issue, this paper proposes constructing a visual-semantic feature decoupled image content analysis model for social media public opinion monitoring, which can effectively reduce this dependence, enhance the model's semantic modeling ability, and thus more accurately analyze and identify public opinion information on social media.

3. VISUAL-SEMANTIC FEATURE DECOUPLING IMAGE CONTENT ANALYSIS MODEL CONSTRUCTION

3.1 Model architecture overview

The model proposed in this paper is designed with a decoupled structure, which is its key advantage. By separating the visual decoder and the semantic decoder, the model is able to independently extract pure visual features and semantic features. This decoupling design effectively reduces the dependency on the vocabulary and avoids the negative impacts caused by limited corpus and noise interference. The visual decoder focuses on extracting visual information from the image, while the semantic decoder captures global semantic information through bidirectional encoding, ensuring the accuracy and comprehensiveness of the information. Additionally, self-supervised pretraining tasks, such as word spelling correction, are introduced to further enhance the semantic modeling capabilities of the semantic module. By utilizing a large corpus of freely available vocabulary, the model continuously optimizes its semantic understanding during training. This innovative training method not only reduces reliance on high-quality annotated data but also improves the model's performance in the complex social media environment. Figure 2 illustrates the architecture of the proposed visual-semantic decoupled image content analysis model.

Specifically, the model is based on the Aster framework with improvements to meet the needs of social media image content analysis. Targeted optimizations have been made to the architecture. First, the model uses a Spatial Transformer Network (STN) module based on Thin Plate Spline (TPS) transformations to geometrically correct the input images, addressing the issue of arbitrary-shaped text found in social media images and ensuring effective recognition of text lines in various formats. Next, the visual feature extractor, which includes a 45-layer ResNet and two layers of bidirectional LSTM, extracts features from the corrected image, producing an original visual feature sequence $g^{(n)}$ with a length of S and dimensionality of F, providing efficient visual information for subsequent semantic modeling.

In the decoding phase, the decoupling of the visual and semantic information decoders enables the model to more efficiently handle the complex emotions and potential public opinion information embedded in social media images. In the visual decoding part, the model first processes the raw visual feature sequence $g^{(n)}$ through a visual encoder, then passes these features to the visual decoder for stepwise decoding. To ensure that the visual features of each character are accurately extracted and aligned with the corresponding semantic information, the model introduces an innovative mechanism: when calculating the alignment of character features, the visual decoder relies not only on the visual features of the current step but also incorporates the hidden state from the semantic module. The core of this design is to leverage the semantic module's assistance to help the visual decoder locate and decode accurate characters in the highly variable social

media images, while ensuring that the visual features $t^{(n)}$ for each character are derived from the weighted average of the original visual feature sequence, maintaining the purity of the visual information and avoiding interference from image noise or irregular expressions. Figure 3 illustrates the decoding process of the visual feature decoder.

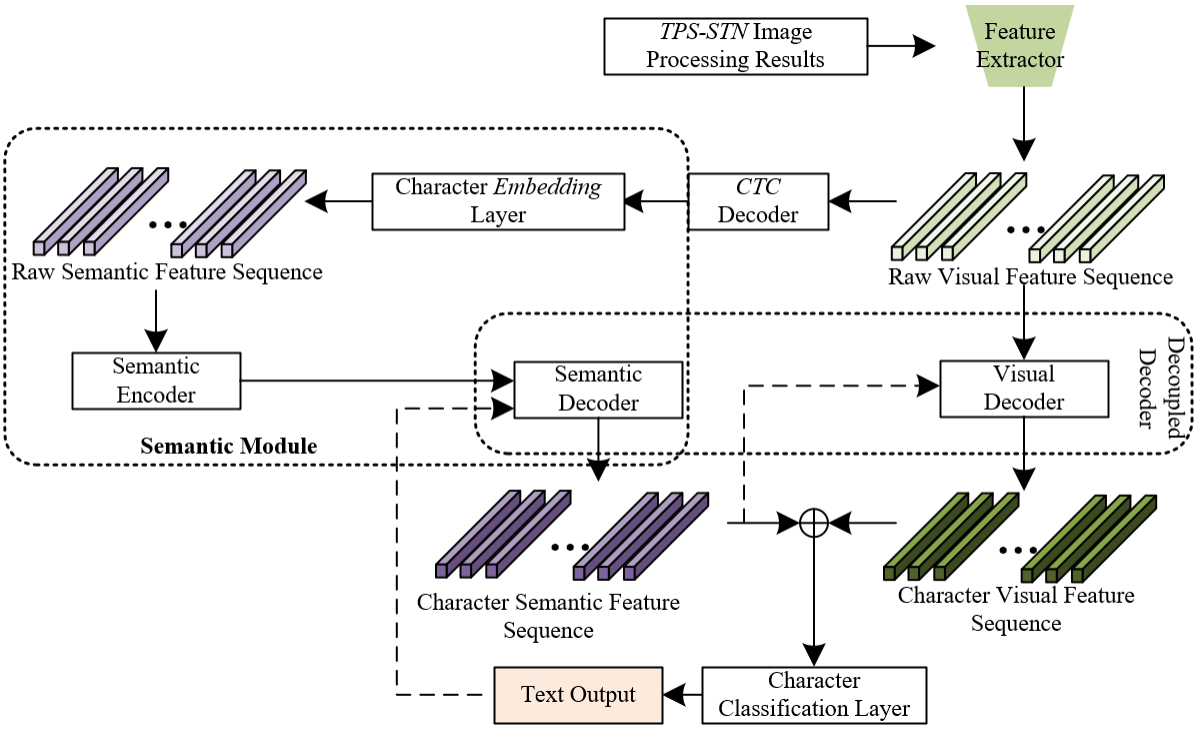

Figure 2. Architecture of the visual-semantic decoupled image content analysis model

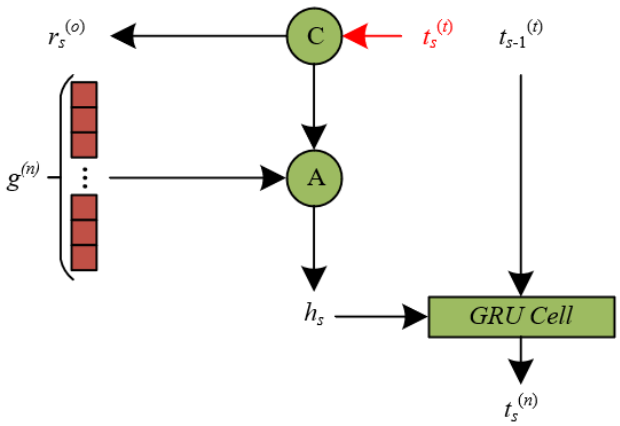

Figure 3. Decoding process of the visual feature decoder

In the semantic modeling part, model V decouples the processing of semantic features from the visual features. To capture the emotions, implicit social intentions, and potential cultural background behind the image text, the model first maps the raw text sequence S into an initial semantic feature sequence $g^{(t)}$ through a character embedding layer, and then uses a semantic encoder to extract contextual features from $g^{(t)}$, thus modeling global semantic information. On this basis, the semantic decoder, as a serial character-level language model, further processes each step's prediction. In the decoding process, the semantic decoder not only utilizes global semantic information but also combines the embedding of the previously predicted character to form a more accurate text output. Through this design, the model avoids the strong coupling of semantic and visual information found in traditional models, independently modeling semantic features. This overcomes the limitations of the training corpus vocabulary and enables the model to flexibly respond to the rich, diverse, and non-standardized images and text content on social media, accurately identifying complex public opinion information.

In social media public opinion monitoring, images usually contain multi-layered, diversified content, involving emotions, implicit social viewpoints, or cultural backgrounds. These types of information may not be fully expressed through visual or semantic features alone. Therefore, during the recognition phase, the model concatenates the $t^{(n)}$ and $t^{(t)}$ features extracted by the visual and semantic decoders, respectively, to form a richer feature representation, which is then passed to the character classification layer for processing. The concatenated feature combination captures not only the visual information in the image but also effectively integrates the contextual knowledge at the semantic level, ensuring accurate text recognition when facing potential emotional cues or social backgrounds in social media images. In the character classification phase, the concatenated visual and semantic features are used as input to a character classification layer. At this point, the decoding process is serial, where each step is based on the previous character prediction and the current visual-semantic features for decision-making. In each decoding step s, the model first produces a probability distribution O_s, where Z represents the number of categories, indicating the probability of the current character belonging to each category. Then, the model selects the category index corresponding to the highest probability to obtain the character prediction result b_s.

3.2 Visual feature decoder

Social media images often contain a mixture of visual elements, such as text, emojis, image backgrounds, and

specific cultural symbols. The visual feature decoder in this model must be capable of handling these complex and diverse inputs. In the decoding process, the visual feature decoder first concatenates the already obtained semantic features $t^{(t)}_s$ and the embedding of the current position $r^{(o)}_s$, generating the query vector for the current step. Then, by combining these query vectors with the original visual feature sequence $g^{(n)}$, the attention mechanism calculates the attention weights β_s, which determine the visual regions that need to be focused on for the current character.

$$z_{s,u} = Q^{S\tanh}\left(I\left[t_s^{(t)}; r_s^{(o)}\right] + Ng_u^{(n)} + y\right) \tag{3}$$

$$\beta_{s,u} = \frac{\exp\left(z_{s,u}\right)}{\sum_{u'=1}^{S}\exp\left(z_{s,u'}\right)} \tag{4}$$

The above process is essentially a character alignment operation, where the original visual feature sequence is weighted using a soft mask to highlight the visual features related to the current decoded character, ensuring accurate extraction of the complex visual information in social media images. With this attention mechanism, the model is able to focus on the most relevant visual content at each decoding step, whether handling clear text or more complex or blurry scenes in the image. Clearly, β_s can also be seen as a soft mask of the specific location of the character to be decoded in $g^{(n)}$. The expressions for the original and new visual features of the character are as follows:

$$h_s = \sum_{u=1}^{S}\beta_{s,u}t_u \tag{5}$$

$$t_s^{(t)} = GRUCell\left(t_{s-1}^{(t)}, h_s\right) \tag{6}$$

3.3 Semantic module

In traditional visual-semantic fusion models, semantic information is often directly extracted from the original visual features. However, the semantic module in this model starts with the text itself for encoding, avoiding interference from visual noise and improving the accuracy of global semantic modeling. This approach is especially beneficial in social media content, where the emotional tone, context, and implied meaning of the text are often very complex. By leveraging a language model-based structure, the semantic module is better equipped to handle these challenges. Figure 4 illustrates the decoding process of the semantic feature decoder. In the proposed model, the semantic module processes the text features independently using a character-level self-encoding language model, providing precise semantic understanding of the text content in social media images. The module's Connectionist Temporal Classification (CTC) decoder first makes an initial prediction of the text in the social media image, obtaining a predicted text sequence. This sequence is then mapped to the raw semantic feature sequence $g^{(t)}$ through a character embedding layer. The character embedding not only maps each character directly but also helps the subsequent semantic encoder better understand the text's language structure and semantic content. The semantic encoder processes $g^{(t)}$ by inputting it into a two-layer bidirectional Gated Recurrent Unit (GRU), capturing the contextual information in the text and establishing a comprehensive understanding of the input text at the global semantic level. Further, the semantic module integrates the sequence features extracted from the bidirectional GRU through a fully connected layer to produce a feature vector t_h, which encapsulates the global semantic information of the input string. The obtained feature vector t_h not only provides a holistic understanding of the text but also helps the model capture underlying emotional information, public opinion trends, and social intentions. Finally, the semantic decoder further maps the global semantic information t_h to the initialized semantic features $t^{(t)}_0$ and the semantic encouragement features e at each step. These features guide the prediction of each character during the decoding process, improving the accuracy of text recognition in the public opinion monitoring task, especially when analyzing image content with specific emotions, viewpoints, or cultural backgrounds. Assuming the embedding of the character predicted at step $s-1$, is denoted as b_{s-1}, the semantic features $t^{(t)}_s$ for step s are computed as follows:

$$t_s^{(t)} = GRUCell\left(t_{s-1}^{(t)}, \left[d\left(b_{s-1}\right); e\right]\right) \tag{7}$$

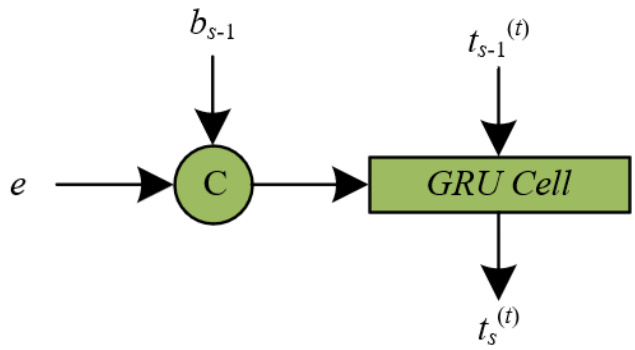

Figure 4. Decoding process of the semantic feature decoder

3.4 Loss function

The optimization objective function during the training phase of the proposed model consists of four components, each corresponding to the CTC decoder, visual decoder, semantic decoder, and the final classification layer. Each decoder independently outputs string predictions through its own character classification layer, helping the model to perform supervised learning at multiple levels. Let the CTC loss function be represented as $LOSS_{CTC}$, and the cross-entropy loss functions for the visual decoder, semantic decoder, and the classification layer be represented as $LOSS_{SJ}$, $LOSS_{YY}$, and $LOSS_{FL}$, respectively. The hyperparameter for these losses is denoted by η, and the overall loss function can be expressed as:

$$\begin{aligned} LOSS = {} & \eta_{CTC}LOSS_{CTC} + \eta_{SJ}LOSS_{SJ} \\ & + \eta_{YY}LOSS_{YY} + \eta_{FL}LOSS_{FL} \end{aligned} \tag{8}$$

3.5 Comparison with other semantic-enhanced models

The visual-semantic feature decoupling model proposed in this paper offers unique advantages in the image content analysis of social media for public opinion monitoring, especially in character localization and text semantic modeling.

This model is particularly effective in addressing common challenges in social media environments, such as blurred images, text overlap, and dense characters. Compared to other semantic-enhanced models, the model presented in this paper explicitly decouples visual and semantic features, processing and optimizing these two modalities independently through separate visual and semantic decoders. This decoupling design is especially important in social media public opinion monitoring tasks, as the text in social media images often contains multi-layered emotional expressions and complex social intentions. Traditional visual-semantic coupled models, such as DAN and SRN, typically rely on visual features for character alignment or localization. While these models may perform well in simple scenarios, their single visual feature-based localization approach is insufficient for handling complex situations in social media images, such as dense text, overlapping characters, or blurry scenes. In particular, during public opinion monitoring, images often carry complex information related to emotions and social intentions. The proposed model improves character alignment using more accurate semantic features, ensuring precise location prediction and avoiding misjudgments and information loss. This enhancement boosts the model's ability to understand complex image content effectively.

In social media public opinion monitoring applications, especially for tasks like sentiment analysis and public opinion trend identification, semantic information modeling is crucial. The SRN model, which relies solely on visual features for character alignment, can handle simple text recognition tasks but struggles in social media images, especially when text spacing is small or image quality is poor. For example, social media images often contain user-generated content with strong emotional tones, influenced by individual emotions or social events. The SRN's reliance on single visual features leads to poor performance in these complex contexts. By introducing the semantic decoder and adopting a decoupled bimodal framework, the proposed model integrates semantic information into the character alignment stage, enabling the model to combine visual data and semantic context. This integration not only improves prediction accuracy but also enhances the model's ability to capture the underlying emotions and social intentions in the text. Thus, in social media public opinion monitoring, the model can more accurately analyze the emotional tendencies, social attitudes, and potential public opinion shifts in image text, providing strong support for subsequent public opinion response strategies.

4. SOCIAL MEDIA PUBLIC OPINION RESPONSE STRATEGY BASED ON IMAGE CONTENT ANALYSIS RESULTS

Based on the analysis results, developing a targeted communication plan is key to effectively responding to public opinion. Firstly, for the large amount of negative emotional images spread on social media or other platforms, a prompt response is necessary. These images often have a strong emotional impact, and if not dealt with in time, they may cause long-lasting negative effects on the brand, enterprise, or individual's image. The response strategy should first include a public statement to clarify the true situation, express attention to public concerns, and commit to taking appropriate actions. Crisis public relations measures are also indispensable, with a professional crisis handling team analyzing the severity of the situation, timely releasing clarification information, or issuing public apologies to defuse public dissatisfaction. It is worth noting that the tone of the response should be sincere and transparent, avoiding minimizing key issues or overly aggressive language, as this could escalate the spread of negative emotions.

In contrast, when the analysis results show a large amount of positive emotional images, this public opinion advantage should be actively utilized for positive publicity or image maintenance. By sharing these positive images, a more positive public impression can be established, enhancing public trust and goodwill. Positive information can be disseminated more intensively through various means, such as issuing thank you statements, praising supporters, and highlighting positive stories, thus further consolidating and expanding this positive influence.

Moreover, the path and speed of image dissemination are also critical factors when developing a communication plan. By analyzing image sharing, likes, comments, and other interaction data, the main channels and key nodes of image dissemination can be identified, allowing the tracking of information diffusion trends in real time. Opinion leaders and influential figures on social media platforms are often key guides of public opinion, and their attitudes and statements have a significant impact on public emotions. Leveraging these opinion leaders' resources can help guide public opinion in a positive direction at critical moments. For example, by contacting relevant bloggers or public figures and inviting them to participate in positive topic discussions or share positive images, a positive feedback loop can be formed to reduce the spread of negative emotions.

Finally, the effectiveness of public opinion response needs to be regularly summarized and evaluated. By tracking changes in public opinion and analyzing the actual effect of response measures, strategies can be adjusted in a timely manner to ensure the long-term effectiveness of public opinion management. At the same time, image content analysis technology should be continuously optimized to improve the ability to identify emotions, themes, and potential risks in images, thus enhancing the accuracy of public opinion monitoring. Furthermore, the public opinion management strategy should keep pace with the times, adjusting response methods and processes according to changes in the social media environment and public psychology.

5. EXPERIMENTAL RESULTS AND ANALYSIS

From the data in Table 1, it can be observed that the accuracy of word auto-encoding using the FastText model significantly varies with different acquisition methods. On the training set, the combination of initialization (init) and step-by-step training (step) (init & step) performs the best, with an accuracy of 88.9%, far surpassing the other two methods (init and step, which are 52.3% and 42.5%, respectively). For the test set, Init & step also shows a significantly higher accuracy compared to the other two methods, achieving 88.5%, while init and step only reached 7.3% and 6.2%, respectively. This indicates that the combination of initialization and step-by-step training can significantly improve the model's generalization ability on the test set, and that the handling of semantic features for social media public opinion monitoring is particularly crucial in image content analysis tasks.

Table 1. The impact of utilization and acquisition methods of global semantic features in social media public opinion on word auto-encoding accuracy

Acquisition Method	FastText	FastText	FastText	Encoder
Utilization Method	Init	Step	Init & Step	Init & Step
Training Set	52.3	42.5	88.9	98.5
Test Set	7.3	6.2	12.5	88.5

Table 2. The impact of human involvement frequency in word error correction on underfitting in word error correction learning

Involvement Frequency	100%	5%
Accuracy	58.8	75.4

From the data in Table 2, it can be seen that the frequency of human involvement in word error correction significantly affects the accuracy of word error correction. When the frequency of human correction is 100%, the accuracy is 58.8%. However, when the human involvement frequency is reduced to 5%, the accuracy significantly increases to 75.4%. This result suggests that, although human correction improves accuracy to some extent, as the frequency of human intervention decreases, the model's learning ability and automatic error correction ability are better utilized, leading to an improvement in accuracy. This could be because human intervention, in some cases, limits the model's space for autonomous learning, which affects its overall understanding and adaptability to public opinion content.

Table 3. Comparison of social media public opinion recognition accuracy across different models

Method	Flickr30k	Twitter Image Dataset	Social Media Image Dataset	ImageNet	Memes Dataset	Open Images Dataset
MNB	86.2	82.3	81.2	66.5	72.3	81.2
BNB	83.2	81.5	88.9	67.9	71.2	82.1
Linear SVM	91.2	85.6	92.6	68.4	72.5	81.9
Text-CNN	93.2	87.9	93.5	77.8	81.2	82.3
Deep RNN	92.3	85.6	92.4	68.5	75.6	83.4
BERT	92.4	91.2	92.5	77.5	78.9	82.5
RoBERTa	92.8	92.3	91.5	78.6	83.2	82.4
ExtraTrees	93.6	88.9	92.6	73.5	81.5	83.6
K-Means++	94.5	88.5	94.5	78.5	82.6	88.9
CTM	92.6	89.5	92.6	76.2	82.3	82.4
sLDA	93.6	92.5	93.6	81.5	84.5	86.9
GraphSAGE	92.6	88.7	92.5	75.6	77.9	78.9
Proposed Model	93.8	92.9	94.3	84.2	85.5	87.2

From the comparison of public opinion recognition accuracy across different models in Table 3, it can be seen that the proposed model outperforms other traditional models on multiple datasets. Specifically, the proposed model shows high accuracy across several datasets, including Flickr30k (93.8%), Twitter Image Dataset (92.9%), Social Media Image Dataset (94.3%), ImageNet (84.2%), Memes Dataset (85.5%), and Open Images Dataset (87.2%). Notably, on the Social Media Image Dataset and Flickr30k datasets, the accuracy is 94.3% and 93.8%, respectively, both leading among all models. This demonstrates that the proposed model exhibits excellent semantic understanding and sentiment judgment capabilities across various social media image datasets, enabling effective recognition and classification of public opinion information in social media images.

Figure 5 presents recognition samples of the proposed model, demonstrating its outstanding performance in image content analysis, especially in the context of social media public opinion monitoring. The model efficiently identifies and classifies complex image information. Testing on different datasets shows that the model excels at recognizing difficult characters in key social media public opinion words, whether they appear at the beginning or in the middle of the word. Additionally, while identifying social media images' public opinion information, the model also exhibits strong strategy generation capabilities, effectively leveraging global semantic information, not just unidirectional semantics. This makes it easier for decision-makers to respond more scientifically and promptly during public opinion crises.

According to the data in Table 4, the loss functions in the visual and semantic decoders significantly impact model training. The proposed model achieves high accuracy when using the $LOSS_{SJ}$ loss function, with a training set accuracy of 92.3% and a test set accuracy of 92.6%. Using the $LOSS_{YY}$ loss function results in training and test set accuracies of 92.5% and 92.5%, respectively, slightly outperforming the $LOSS_{SJ}$ function. Combining both loss functions yields the best performance, with accuracies of 92.1% for the training set and 92.8% for the test set. This demonstrates that combining visual and semantic loss functions effectively enhances model training and achieves high accuracy on the test set.

Figure 5. Recognition samples of the proposed model

Table 4. The role of loss functions in the visual and semantic decoders on model training

Method	$LOSS_{SJ}$	$LOSS_{YY}$	Training Set	Test Set
Proposed Model			91.2	91.5
	√		92.3	92.6
		√	92.5	92.5
	√	√	92.1	92.8

These experimental results further confirm the critical role of loss functions in image content analysis models for social media public opinion monitoring. Specifically, the combination of $LOSS_{SJ}$ and $LOSS_{YY}$ optimizes the model's performance in visual and semantic feature decoding, enabling it to better interpret image content and assess sentiment, thus improving public opinion monitoring accuracy. The proposed model, by optimizing the relationship between image and semantic features, avoids the common issues of improper semantic usage found in traditional public opinion monitoring systems.

Table 5. The impact of different query requests on social media public opinion recognition accuracy

Query	Training Set	Test Set
Previous Visual Hidden Layer State	91.2	82.5
Current Semantic Features	92.6	81.9

The data in Table 5 shows the impact of different query requests on social media public opinion recognition accuracy. Specifically, it compares the effects of "Previous Visual Hidden Layer State" and "Current Semantic Features" on model accuracy in the training and test sets. In the training set, using "Current Semantic Features" improves accuracy to 92.6%, compared to 91.2% when using "Previous Visual Hidden Layer State." However, in the test set, the "Previous Visual Hidden Layer State" yields better accuracy (82.5%) than "Current Semantic Features" (81.9%), indicating that the model relies more on visual information when processing the test set, while the influence of semantic features is slightly weaker.

These results suggest that visual and semantic features do not play equal roles in social media public opinion monitoring, with each affecting the model's training and testing accuracy differently. Semantic features contribute more significantly in the training set, likely because they help the model better understand the emotions and sentiments behind the image content, thus improving its ability to recognize complex social media public opinion. In contrast, in the test set, the performance of visual hidden layer states is more pronounced, likely due to the complexity of the image content in the test set, where the model relies more on direct visual features for judgment. In practical applications, the combination of visual and semantic information should be adjusted based on the specific task and data characteristics.

6. CONCLUSION

This study explored the use of image content analysis for social media public opinion monitoring and response strategies, proposing an innovative model that demonstrates its potential and advantages through experiments. The first part of the research analyzed the issues of improper semantic use in traditional public opinion monitoring methods, highlighting the limitations of text-based approaches in dealing with image content, especially the differences in expressing image sentiment and semantics. To address these issues, we introduced an improved strategy by constructing a model that combines visual and semantic analysis. This model successfully enhances the accuracy of image interpretation and sentiment judgment. The experimental results on various datasets show that the proposed model outperforms traditional methods in public opinion recognition accuracy, especially on image-heavy datasets. It is effective in capturing the sentiment and public opinion dynamics behind images. Additionally, by introducing different loss functions and query request methods during training, the model can flexibly optimize the fusion of visual and semantic features, improving its adaptability to complex social media public opinion.

However, there are still limitations to this research. While the image content analysis model performs well on training and test sets, it may struggle with recognizing extreme or latent emotions in certain situations. Furthermore, the experiments primarily relied on publicly available datasets, and the complexity and diversity of real-world social media data may lead to some variation in the model's performance in practical applications. Lastly, although the fusion of image and semantic information has improved public opinion recognition accuracy, further optimization of this fusion and the rapid generation of effective response strategies in different types of public opinion crises remain key areas for future research. Future studies could extend the application of this model by incorporating multimodal data, such as video and audio analysis, to improve the comprehensive understanding of complex social media information. As social media platforms evolve, new image content and emotional expression methods constantly emerge, making it crucial to develop more dynamic and flexible models that can adapt to these changes. Researchers could explore more refined feature extraction techniques and deep learning methods to further enhance the model's ability to recognize subtle public opinion signals. Additionally, integrating the model's analysis results with response strategies to provide more actionable decision support will be an important direction for future research.

ACKNOWLEDGEMENTS

This paper was supported by Natural Science Foundation of Xiamen, China (Grant No.: 3502Z202474006); The program of cultivating outstanding young scientific research talents in Universities of Fujian Province (Grant No.: ZX17033); the doctoral research initiation Fund Program (Grant No.: CK18013), and the program of Fujian Provincial Department of Education (Grant No.: JAT201035).

REFERENCES

[1] Zhang, Y., Chen, F., Rohe, K. (2022). Social media public opinion as flocks in a murmuration: Conceptualizing and measuring opinion expression on social media. Journal of Computer-Mediated Communication, 27(1): 21. https://doi.org/10.1093/jcmc/zmab021

[2] Anstead, N., O'Loughlin, B. (2015). Social media analysis and public opinion: The 2010 UK general election. Journal of Computer-Mediated Communication, 20(2): 204-220. https://doi.org/10.1111/jcc4.12102

[3] Zhou, H.Z., Li, X.W. (2021). Quantitative research on the evolution stages of we-media network public opinion based on a logistic equation. Tehnicki Vjesnik-Technical Gazette, 28(3): 983-993. https://doi.org/10.17559/TV-20210316155352
[4] McGregor, S.C. (2019). Social media as public opinion: How journalists use social media to represent public opinion. Journalism, 20(8): 1070-1086. https://doi.org/10.1177/1464884919845458
[5] McGregor, S.C. (2020). "Taking the temperature of the room" how political campaigns use social media to understand and represent public opinion. Public Opinion Quarterly, 84(S1): 236-256. https://doi.org/10.1093/poq/nfaa012
[6] Dubois, E., Gruzd, A., Jacobson, J. (2020). Journalists' use of social media to infer public opinion: The citizens' perspective. Social Science Computer Review, 38(1): 57-74. https://doi.org/10.1177/0894439318791527
[7] Yang, S. (2022). Analysis of network public opinion in new media based on BP neural network algorithm. Mobile Information Systems, 2022(1): 3202099. https://doi.org/10.1155/2022/3202099
[8] An, L., Hu, J., Xu, M., Li, G., Yu, C. (2021). Profiling the users of high influence on social media in the context of public events. Journal of Database Management (JDM), 32(2): 36-49. https://doi.org/10.4018/JDM.2021040103
[9] Liu, P. (2021). Information dissemination mechanism based on cloud computing cross-media public opinion network environment. International Journal of Information Technologies and Systems Approach (IJITSA), 14(2): 70-83. https://doi.org/10.4018/IJITSA.2021070105
[10] Lin, L., Jiang, A., Zheng, Y., Wang, J., Wang, M. (2021). New media platform's understanding of Chinese social workers' anti-epidemic actions: an analysis of network public opinion based on COVID-19. Social Work in Public Health, 36(7-8): 770-785. https://doi.org/10.1080/19371918.2021.1954127
[11] Zareer, M.N., Selmic, R.R. (2024). Modeling interactions in social media networks using an asynchronous and synchronous opinion dynamics. Social Network Analysis and Mining, 14(1): 1-28. https://doi.org/10.1007/s13278-024-01402-x
[12] Cheng, Y., Huang, Y.H.C., Chan, C.M. (2017). Public relations, media coverage, and public opinion in contemporary China: Testing agenda building theory in a social mediated crisis. Telematics and Informatics, 34(3): 765-773. https://doi.org/10.1016/j.tele.2016.05.012
[13] Lee, F.L. (2016). Opinion polling and construction of public opinion in newspaper discourses during the Umbrella Movement. Journal of Language and Politics, 15(5): 592-611. https://doi.org/10.1075/jlp.15.5.05lee
[14] Herbst, S. (2004). Illustrator, American icon, and public opinion theorist: Norman Rockwell in democracy. Political Communication, 21(1): 1-25. https://doi.org/10.1080/10584600490273245-1910
[15] Xie, C., Huang, Q., Lin, Z., Chen, Y. (2020). Destination risk perception, image and satisfaction: The moderating effects of public opinion climate of risk. Journal of Hospitality and Tourism Management, 44: 122-130. https://doi.org/10.1016/j.jhtm.2020.03.007
[16] Maracchione, F., Sciorati, G., Combei, C.R. (2024). Changing images? Italian Twitter discourse on China and the United States during the first wave of COVID-19. The International Spectator, 59(2): 77-94. https://doi.org/10.1080/03932729.2023.2299452
[17] Schnoll, R.A., Wileyto, E.P., Hornik, R., Schiller, J., Lerman, C. (2007). Spiral computed tomography and lung cancer: Science, the media, and public opinion. Journal of Clinical Oncology, 25(36): 5695-5697. https://doi.org/10.1200/JCO.2007.13.5228
[18] Tang, H.L., Hanka, R., Ip, H.H.S. (2003). Histological image retrieval based on semantic content analysis. IEEE Transactions on Information Technology in Biomedicine, 7(1): 26-36. https://doi.org/10.1109/TITB.2003.808500
[19] Wei, P., He, F., Zou, Y. (2020). Content semantic image analysis and storage method based on intelligent computing of machine learning annotation. Neural Computing and Applications, 32(7): 1813-1822. https://doi.org/10.1007/s00521-020-04739-4
[20] Ruan, S.L., Zhang, K., Chen, E.H. (2024). Color enhanced cross correlation net for image sentiment analysis. IEEE Transactions on Multimedia, 26: 4097-4109. https://doi.org/10.1109/TMM.2021.3118208
[21] Evans, W. (2000). Teaching computers to watch television: Content-based image retrieval for content analysis. Social Science Computer Review, 18(3): 246-257. https://doi.org/10.1177/089443930001800302
[22] Chang, R.I., Lin, S.Y., Ho, J.M., Fann, C.W., Wang, Y.C. (2012). A novel content based image retrieval system using k-means/KNN with feature extraction. Computer Science and Information Systems, 9(4): 1645-1661. https://doi.org/10.2298/CSIS120122047C
[23] Huo, C., Bhattacharya, P. (2001). Content-based indexing of volumetric images using principal component analysis. International Journal of Pattern Recognition and Artificial Intelligence, 15(8): 1299-1309. https://doi.org/10.1142/S0218001401001441
[24] Che, X., Kong, J., Dai, J., Gao, Z., Qi, M. (2011). Content-based image hiding method for secure network biometric verification. International Journal of Computational Intelligence Systems, 4(4): 596-605. https://doi.org/10.2991/ijcis.2011.4.4.16