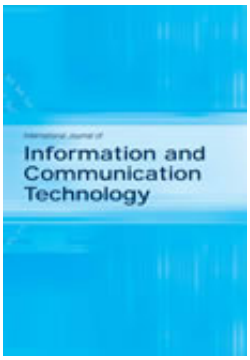


International Journal of Information and Communication Technology

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

Tracking the evolution of youth ideological public opinion based on multimodal transformer and SHAP attribution

Dan Yang

DOI: 10.1504/IJICT.2025.10075266

Article History:

Received:	25 September 2025
Last revised:	22 October 2025
Accepted:	25 October 2025
Published online:	12 January 2026

Tracking the evolution of youth ideological public opinion based on multimodal transformer and SHAP attribution

Dan Yang

School of Marxism,
Suzhou Polytechnic University,
Suzhou, 215104, China
Email: 19907114909@163.com

Abstract: Existing methods for tracking the evolution of ideological and political public opinion struggle to fully uncover intermodal correlations and exhibit low tracking accuracy. To address this, this paper first employs the Shapley additive explanations algorithm optimised by random forests to screen key influencing indicators. These selected indicators undergo a Gramian angular field transformation to generate a two-dimensional image. Subsequently, the Shapley additive explanations optimises the self-attention mechanism of the Transformer model while enhancing locally significant features that substantially influence tracking outcomes. The improved transformer model and bidirectional encoder representations from transformers model are employed to extract image and text features, respectively. Contrastive learning is introduced to align features across modalities. Multimodal fusion features undergo classification via the softmax function, enabling the tracking of public opinion evolution. Experimental results demonstrate that the proposed model achieves a tracking accuracy of 92.1%, exhibiting outstanding tracking efficiency.

Keywords: ideological and political public opinion tracking; transformer model; SHAP algorithm; multimodal feature fusion; contrastive learning.

Reference to this paper should be made as follows: Yang, D. (2025) 'Tracking the evolution of youth ideological public opinion based on multimodal transformer and SHAP attribution', *Int. J. Information and Communication Technology*, Vol. 26, No. 50, pp.17–34.

Biographical notes: Dan Yang is a Lecturer in the School of Marxism at Suzhou Polytechnic University, Suzhou, China. She received her PhD from Wuhan University, China, in 2009. Her research interests include philosophy, ideological education, contrastive learning, and psychology.

1 Introduction

As the new media technology deeply growing, ideological and political public opinion carriers have expanded from traditional text to multimodal forms integrating text, images, etc. (Polli and Santonocito, 2024). This type of public opinion information features rapid dissemination speed, complex emotional orientation, and concealed evolution paths. It is prone to cause cognitive deviations due to misinformation spread and negative emotion

contagion, posing potential risks to ideological security and social stability. Therefore, achieving full-cycle, high-accuracy evolution tracking of ideological and political public opinion and precisely identifying key influencing factors and transmission nodes has become a core demand in current online ideological governance (Su, 2024). Traditional public opinion analysis methods focus mainly on single-text data and struggle to handle the semantic correlation and dynamic evolution patterns of multimodal information (Chen and Wei, 2023). Recently, transformer models have demonstrated significant advantages in multimodal information fusion through self-attention mechanisms, providing technical support for capturing multi-source data correlations in ideological and political public opinion (Tan et al., 2021). However, existing transformer-based public opinion analysis models still suffer from black-box problems. Although they can output public opinion evolution results, they cannot explain the decision-making basis or identify key influencing factors. This hampers the ability to establish the traceable and manageable governance of ideological and political discourse that is actually required (Du et al., 2024). Therefore, the development of an efficient and interpretable model for tracking the evolution of ideological and political public opinion constitutes a significantly complex challenge.

Considerable scholarly work, both domestic and international, has contributed to advances in tracking the evolution of ideological and political discourse. Conventional approaches for monitoring the evolution of ideological and political discourse often rely on manual monitoring, questionnaires, or simple statistical analysis, which suffer from inherent limitations such as strong subjectivity, low efficiency, and obvious lag, making it difficult to cope with massive, high-speed, and rapidly changing online public opinion data. Machine learning models can continuously self-learn and optimise based on new data to adapt to the dynamic changes of ideological and political public opinions. As time goes by and the public opinion environment changes, the model can automatically adjust parameters and algorithms to better capture the new features and changing trends of public opinion, and maintain effective tracking of the evolution of public opinion. Through real-time monitoring and analysis of ideological and political public opinion data by machine learning models, abnormal changes in public opinion data can be detected in a timely manner, potential negative public opinions, false information or events that may trigger public opinion crises can be quickly identified, providing early warnings for relevant departments so that they can take timely measures to deal with the situation and defuse public opinion risks in their infancy. Machine learning can automatically learn patterns and rules from massive complex data, and possesses powerful predictive and generalisation capabilities. Applying it to ideological and political public opinion research holds promise for achieving a paradigm shift from descriptive analysis to predictive judgement. Kirkizh et al. (2024) designed an ideological and political public opinion detection tracking method based on the life cycle and evolution mechanism of online public opinion, using entropy weight method and decision tree. However, the tracking results were not satisfactory. Liang et al. (2020) established a topic tracking classifier system that uses probability models based on simple Bayesian classification algorithms. For different reports, this system adopts different evaluation formulas to calculate the similarity between the current report and a specific topic. Sun et al. (2025) used Shapley additive explanations (SHAP) to analyse influencing factors of ideological and political public opinion and designed a K-nearest neighbours (KNN) classifier for topic tracking with an accuracy rate of only 72.6%. Wei-Dong et al. (2018) proposed a hierarchical topic identification and tracking method, where the multi-level

clustering algorithm achieved good results. Hayadi and Maulita (2025) proposed an opinion tracking strategy that combines KNN with support vector machine (SVM), applying the SVM algorithm to the field of public opinion tracking. Samih et al. (2023) proposed an ideological and political public discourse evolution tracking approach based on extreme gradient boosting (XGBoost) and SHAP, verifying the importance of indicators through feature screening, and combining it with SHAP attribution to further analyse key driving factors of public opinion risks. Al-Laith and Shahbaz (2021) introduced location and time information into topic tracking, and proposed news tracking based on thematic elements, improving tracking efficiency.

Compared to conventional machine learning techniques and qualitative analysis methodologies, deep learning-based ideological and political public opinion evolution tracking methods are better suited for the core characteristics of ideological and political public opinion being multimodal, dynamic, and highly interconnected due to their strong feature representation capability, temporal modelling capability, and multimodal fusion capability. These can significantly enhance the effectiveness of public opinion tracking. Moreover, traditional machine learning methods usually require manual feature engineering, relying on domain expert knowledge to select and construct features. This is not only time-consuming and labour-intensive, but also may fail to fully capture the key features in complex public opinion data. Qualitative analysis methods mainly rely on manual description and interpretation of text and other data, making it difficult to conduct systematic feature extraction and quantitative analysis on large-scale data. Deep learning models possess strong dynamic learning capabilities, capable of continuously adjusting and optimising model parameters based on new data to adapt to the dynamic changes of public opinion data, thereby enhancing the efficiency of public opinion evolution tracking. Chen and Du (2023) used convolutional neural network (CNN) and Word2Vec to represent images and texts, concatenated multi-modal features for representing multi-modal public opinions, and obtained final tracking results through a fully connected network. Lin and Bu (2022) obtained GloVe word vector representations of the text, then used long short-term memory (LSTM) to further capture text characteristics, CNN model was continued for image feature extraction, concatenating both types of features. Finally, SHAP was used for interpretability analysis of the tracking results. Yan et al. (2022) innovatively constructed an evolution tracking approach based on CNN and bidirectional gated recurrent unit (BiGRU), which can capture and utilise deep dependencies existing between different modalities. With the increasing popularity of transformer, scholars have begun to explore its attention mechanism for multi-modal feature fusion representation (Bashiri and Naderi, 2024). Sun et al. (2024) combined multi-modal fusion with feature extraction from social network data, proposing an ideological and political public opinion tracking model based on transformer and LSTM, which improves the effectiveness of online public opinion tracking. Du et al. (2025) used a 12-layer visual transformer for image feature extraction, a 6-layer transformer encoder for text feature extraction of public opinion, and another 6-layer transformer encoder for multi-modal feature fusion, achieving an accurate tracking rate of 86.9%.

From the analysis of existing ideological and political public opinion evolution tracking research, it can be seen that current studies often fail to fully mine correlation information between different modalities when processing multi-modal data for ideological and political public discourse, are insufficient in analysing the causes of public opinion evolution, resulting in poor tracking effects. To overcome these

limitations, this study introduces a novel framework for tracking ideological and political opinion evolution that integrates a multi-modal transformer architecture with SHAP attribution analysis. The principal contributions of this model are delineated in the following four aspects.

1 Adopt the SHAP algorithm optimised by random forest to evaluate the importance of ideological and political public opinion influence indicators under strong coupling conditions, and perform Gramian angular field transformation on indicators with Shapley values greater than 0 to generate two-dimensional images. The SHAP algorithm integrates splitting path information from all trees in the forest to calculate the true contribution value of each indicator to ideological and political public opinion, thereby eliminating redundant indicators.

2 Optimise transformer's self-attention mechanism using the SHAP algorithm, utilising Shapley values to quantify the contributions of input features to prediction results, helping identify and enhance local features with significant impact on tracking results. Use the improved transformer to capture characteristics from ideological and political public opinion images, and introduce the bidirectional encoder representations from transformers (BERT) model to extract text features for ideological and political public opinion.

3 Use contrastive learning methods to align features between different modalities. At the same time, employ momentum distillation to make the image-text features generated by the momentum model serve as negative samples for contrastive learning; fuse the aligned image-text features at the token level and input them into a gated recurrent unit (GRU) model for feature fusion, obtaining fused features. Classify the fused features through a softmax function to obtain ideological and political public opinion categories, thus achieving tracking of ideological and political public opinion evolution.

4 Establish a corresponding momentum model and update it using the original model at a certain ratio; simultaneously input image-text pairs into the momentum model to obtain predicted classification categories from the momentum model, thereby achieving tracking of ideological and political public discourse evolution. Experimental findings demonstrate that the proposed model attains tracking accuracy and F1 scores of 92.1% and 93.9%, respectively, outperforming comparison models, providing new ideas and technical support for ideological and political public discourse monitoring and management, helping relevant departments promptly grasp public opinion dynamics and scientifically formulate response strategies.

2 Relevant theory

2.1 Transformer model

The transformer represents a deep learning architecture fundamentally built upon the self-attention mechanism (Nassiri and Akhloufi, 2023). It can process all positions in an input sequence simultaneously, effectively capturing global relationships and long-range dependencies without being constrained by sequence length. Compared with traditional deep learning models for example CNNs and RNNs, the transformer utilises the parallel

computing capabilities of the self-attention mechanism to achieve high parallelism, allowing the model to handle large-scale and long-sequence data more efficiently.

Each word in the input sequence is first embedded into a vector. Since the transformer focuses on global information and does not consider the order of words, positional information needs to be encoded into its vector representation to introduce sequence information (Valeriani et al., 2023). The dimension of position encoding (PE) is the same as that of the embedding vector, and the computation equation is as below, where pos represents the word's position in the input sequence, d_{model} indicates the embedding dimension, and i is the index for the encoding dimension with a range of [0, d/2 – 1]. The above formula ensures that encodings for adjacent positions are different. By adding the word embedding vector and PE together, the final representation X at each position in the input sequence can be obtained.

$$PE(pos, 2i) = \sin\left(\frac{pos}{10{,}000^{2i/d_{\text{model}}}}\right) \tag{1}$$

$$PE(pos, 2i+1) = \cos\left(\frac{pos}{10{,}000^{2i/d_{\text{model}}}}\right) \tag{2}$$

The encoder module consists of six encoders, and each encoder layer consists of two primary components: a multi-head attention mechanism and a feed-forward neural network (FFN), and normalisation layers. The decoder component similarly comprises six identical layers, each containing an additional masked multi-head attention sublayer compared to the encoder architecture, while maintaining structural parity in all other aspects. To ensure that each position only pays attention to its own current position and previous ones, a mask matrix is added. The output of each decoder becomes the input for the next one, forming a multi-level hierarchical architecture. This design enables the model to progressively adjust its attention on the input while decoding, allowing it to efficiently handle tasks that require sequential generation. After achieving the ultimate output from the decoder module, it undergoes linear transformation and softmax activation to generate the final probability distribution of outputs.

2.2 SHAP attribution algorithm

SHAP calculates the average contribution (i.e., the Shapley value) of a feature across all possible subsets of features to obtain the SHAP value for that feature (Chen et al., 2023). This approach ensures fairness and consistency in evaluating the contribution of each feature. Regardless of how the features are distributed or associated with other features in the dataset, it yields objective marginal contribution quantification results, avoiding the shortcomings of traditional feature importance rankings, which often overemphasise prominent features while neglecting critical features from small samples (Wang et al., 2024).

Assume that the i^{th} sample under study is x_i, the j^{th} characteristic of the i^{th} sample is x_{ij}, the model's prediction for this sample is y_i, and the base value of the whole model is y_{base}. Then, the SHAP value follows equation (3).

$$y_i = y_{base} + f(x_{i1}) + f(x_{i2}) + \ldots + f(x_{ik}) \tag{3}$$

where $f(x_{ij})$ represents the SHAP value of x_{ij}; intuitively, $f(x_{i1})$ indicates the contribution of the first characteristic in the i^{th} sample to y_i. When $f(x_{i1})$, it suggests that the characteristic increases the forecasting value, thus playing a positive role; conversely, if not, it implies that the characteristic reduces the forecasting value and has an adverse effect.

SHAP is a sample-based local explanation method, with its greatest advantage being the ability to reflect the impact of each feature within every sample through SHAP values. Calculating SHAP values involves determining the feature's contribution to the prediction; simply put, it measures the difference between the contribution before and after introducing a specific feature. The calculation method is as follows.

$$\varphi_i = \sum_{S \subseteq N \setminus \langle i \rangle} \frac{|S|!(M-|S|-1)!}{M!} \left[f_x(S \cup \{i\}) - f_x(S) \right] \tag{4}$$

where $N = \{1, 2, \ldots, M\}$ stands for the index of the characteristic variables in the dataset; M is the total amount of characteristic variables; S is a subset of $\{1, 2, \ldots, i-1, i+1, \ldots, M\}$; $|S|$ indicates the total amount of elements in S; $f_x(S \cup \{i\})$ denotes the model's prediction value when only characteristics in $S \cup \{i\}$ are present, and $f_x(S)$ represents the model's prediction value when only characteristics in S are present. The discrepancy between these two values constitutes the marginal contribution of the i^{th} feature variable within the subset S.

3 Selection of ideological and political public opinion influence indicators based on improved SHAP attribution

Ideological and political public discourse is influenced by multiple factors, and precisely identifying key indicators is a prerequisite for effective tracking and governance. Traditional methods have limitations when dealing with complex nonlinear relationships, whereas combining random forests with SHAP attribution offers new insights into indicator selection for ideological and political public opinion. Random forests can efficiently process high-dimensional data and evaluate feature importance (Chen et al., 2016). SHAP quantifies the contribution of features to model outputs based on game theory; integrating these two approaches helps improve the accuracy and explainability of indicator selection (Jahin et al., 2024).

Currently, most use one-to-one methods such as Pearson's correlation coefficient to analyse correlations and determine parameters based on the calculated coefficients (Mu et al., 2018). However, pairwise correlation calculation methods neglect synergistic relationships among indicators; the correlation of a single indicator cannot reflect its importance after coupling with other indicators. In contrast, the SHAP attribution algorithm based on Shapley values considers the coupling between data elements. The SHAP algorithm generates global feature importance assessments by integrating combinations of all features and their marginal contributions. Additionally, random forest algorithms are used to optimise the SHAP algorithm for processing nonlinear relationships among influencing factors.

In random forests, each tree produces a set of local feature importance evaluation values; the SHAP algorithm calculates the true contribution of indicators to ideological and political public opinion by aggregating split-path information from all trees in the

forest. After aligning sample data with lag duration, we calculate an optimised Shapley value ϕ using random forests and rank correlations among influencing indicators based on the SHAP algorithm. Indicators with a Shapley value greater than 0 undergo Gramian angular field transformation, and each selected indicator is converted into a GAF image separately. Due to different pattern characteristics in each phase, such as steep increases during outbreak phases where light areas are concentrated in images; decrease sequences during decline phases create distinctive textures in the images; deep learning models can learn these unique visual patterns and classify new sequence images.

4 Multimodal ideological and political public opinion data feature extraction

4.1 Text feature extraction for ideological and political public opinion based on the BERT model

This paper introduces the BERT model to extract text feature vectors and uses an improved transformer model to extract image feature vectors. As a classic pre-trained language model, BERT demonstrates multiple advantages when dealing with ideological and political public opinion texts. Its greatest advantage lies in its ability to accurately capture the implicit political stance, emotional tendency and complex semantic relationships in the text through bidirectional context modelling and deep semantic understanding, especially when dealing with the polysemy, metaphorical and context-dependent content unique to the ideological and political field. The text and image feature vectors are fused at the token level, and vector updates are performed using GRU. The procedure for multimodal feature extraction from ideological and political public discourse data operates as below.

The first six layers of the BERT model are used for feature extraction. Unlike conventional unidirectional language models, BERT constitutes a pre-trained deep bidirectional transformer architecture capable of generating contextualised language representations through simultaneous forward and backward context analysis. Since BERT is pre-trained using masked language model (MLM), after obtaining the pre-trained model, only an additional output layer needs to be added for fine-tuning, which enables its application in various downstream tasks and demonstrates strong transferability.

The BERT model mainly consists of two parts. The first part is the embedding layer, including position embeddings, segment embeddings, and token embeddings. The second part is the encoder layer composed of multiple transformer encoders stacked together. This paper adopts the BERTBase configuration for text feature extraction, where the encoder contains 12 transformer encoders. Each transformer encoder uses 12 attention heads, and the fully linked levels in the transformer encoder consist of 768 hidden units; therefore, the output vector size from this model is also 768 dimensions.

To preserve the sequential information of words in the sequence, BERT introduces position embeddings. For an input sequence of length n, position embeddings add a unique vector to each position i, which is summed with the word embedding vector to form the final embedding vector, where W_{tok} and W_{pos} are the weight matrices for word embeddings and position embeddings, respectively.

$$E_i = W_{tok}\left[Tokenid_i\right] + W_{pos}[i] \tag{5}$$

When the input contains multiple sentences, segment embeddings are used to distinguish different sentences. This is achieved by assigning different embedding vectors to each sentence and adding them to the corresponding word embeddings, where W_{seg} is the weight matrix of segment embeddings.

$$E_i' = E_i + W_{seg}[Segment\ ID] \tag{6}$$

For each token i, the final embedding vector E_i consists of the following elements.

$$E_i = W_{tok}\left[Tokenid_i\right] + W_{pos}[i] + W_{seg}[SegmentID] \tag{7}$$

The input embeddings are processed through a series of transformer layers. Each layer contains two main parts; the first part is the self-attention level. The self-attention layer allows the model to consider information from the entire sequence when processing each word. The new representation of i at layer l is computed as below:

$$h_i^l = Attention\left(Q_i^{l-1}, K_i^{l-1}, V_i^{l-1}\right) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right) \tag{8}$$

where Q, K and V are the query, key and value matrices respectively, d_k is the dimension of key vectors. This mechanism helps the model to capture local features while maintaining global dependencies.

The second layer is the feedforward network. After self-attention processing, each token's output enters the feedforward network, which is a fully linked level with ReLU activation. For token i, its output at each layer is further updated through an FFN, in which W_1, W_2, b_1 and b are network parameters. ReLU is a nonlinear activation function. The final text feature x_w is obtained.

$$x_w = h_i^l = FFN\left(h_i^l\right) = ReLU\left(h_i^l W_1 + b_1\right)W_2 + b_2 \tag{9}$$

4.2 Image feature extraction for ideological and political public opinion based on the improved transformer model

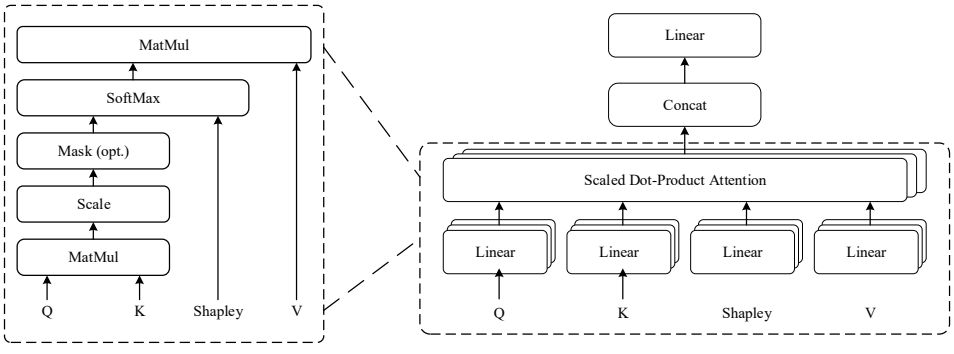
Although the transformer model performs well in modelling global dependencies, it often pays insufficient attention to local features. This is because the multi-head self-attention mechanism primarily allocates attention weights based on similarities between positions without adjusting according to the actual contribution of features to the prediction task, potentially leading to overlooked local information fluctuations and failure to fully utilise all data details for predictive analysis.

To cope with this issue, this paper adopts the SHAP algorithm to enhance the self-attention mechanism of the transformer (STransformer), leveraging Shapley values to quantify the contribution of each input characteristic to the prediction result, helping identify and enhance local features that significantly influence the prediction outcome. Specifically, first, contributions of different parameters at various time steps are obtained based on Shapley values and represented as a weight matrix S; then, this weight matrix is used in attention calculations to adjust the attention scores, as shown in Figure 1. The optimised attention weights $Attention_{SHAP}(Q, K, V, S)$, with calculation formula as

follows, where ∗ denotes element-wise multiplication, S represents the weighted matrix formed by Shapley values.

$$Attention_{SHAP}(Q, K, V, S) = softmax\left(\frac{QK^T}{\sqrt{d_k}} * S\right)V \tag{10}$$

Figure 1 Optimising the self-attention mechanism of transformers using the SHAP algorithm



After introducing Shapley values, the model can assign higher weights to local key parameters during attention calculation, thereby enhancing transformer's ability to capture local features without altering the original model structure. For input ideological and political public opinion images, using the PatchEmbed class in *timm* library, the input image is classified into small image patches. Each segmented image patch is converted from a two-dimensional matrix into a one-dimensional sequence through convolution and flatten operations to obtain a one-dimensional sequence that can be input into STransformer. A learnable one-dimensional sequence with dimensions matching those of the image patch sequence is generated using the torch.nn.Parameter () function, serving as a [CLS] token concatenated to the image patch sequence. Its role is to extract global features from the images in transformer modules for subsequent feature fusion. Adding the [CLS] token can be represented as $X = [x_{cls}; x_p^1, x_p^2, \ldots, x_p^n]$.

Similarly, a learnable one-dimensional sequence (with elements incrementing sequentially starting from 0 and length matching that of the image patch sequence) is generated using the torch.nn.Parameter () function, serving as position embedding added to the corresponding image patch sequences. Its role is to represent positional information for each image patch sequence within STransformer modules. Adding positional embeddings can be represented as follows.

$$X = \left[x_{cls}; x_p^1, x_p^2, \ldots, x_p^n\right] + x_{pos} \tag{11}$$

Define an encoder and stack it sequentially to form the main part of feature extraction from images. The segmented image patch sequences are input into the defined encoder section. During propagation through the encoder, normalisation is first performed on each pixel across all channels, with calculation formula given below:

$$y = \frac{x - E(x)}{\sqrt{Var(x) + \varepsilon}} * \gamma + \beta \tag{12}$$

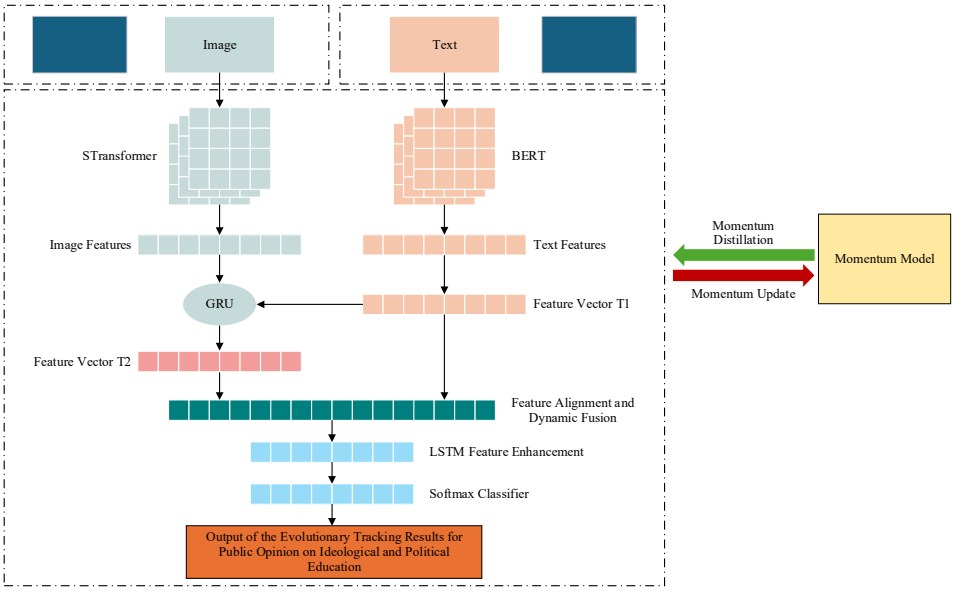
Then, it passes through a multi-head self-attention mechanism layer improved with SHAP attribution. For the self-attention mechanism, there are single-head and multi-head variants; the latter can extract multiple features and has better robustness. Finally, an image feature x_{fet} is obtained through a fully linked level.

5 Evolutionary tracking of ideological and political public sentiment based on SHAP attribution and transformer model

5.1 Structure of the ideological and political public opinion evolution tracking model

In the era of omnimedia, thought-politics public opinion exhibits multimodal and dynamic evolution characteristics. Traditional analytical methods struggle to deeply interpret their complex internal mechanisms. To address the issue that current thought-politics public opinion tracking methods have low efficiency in fusing multimodal information, leading to poor tracking performance, this paper proposes a thought-politics public opinion evolution tracking model based on multimodal feature fusion and an improved transformer, as shown in Figure 2.

Figure 2 Structure of the designed evolution tracking model (see online version for colours)



Use the STransformer model to extract image features and use the first six layers of the BERT model to extract text features. Use contrastive learning methods for feature alignment across different modalities. At the same time, employ momentum distillation (Lin and Hu, 2023) so that image-text features generated by the momentum model act as negative samples in the contrastive learning process; align and fuse the image-text features at the word token level, input them into a GRU model for further feature fusion to obtain fused features. Apply the softmax function to classify these fused features to

determine the thought-politics public opinion categories, thereby achieving tracking of their evolution. Finally, establish the corresponding momentum model and update it using the original model at a certain ratio. At the same time, input image-text pairs into the momentum model to obtain the predicted categories from the momentum model, and incorporate these predictions into the loss function for backpropagation.

5.2 Feature alignment of multimodal ideological and political public opinion data

For the extracted image-text features of thought-politics public opinion, they are often misaligned in most cases; that is to say, image-text pairs with identical semantics may appear unrelated in the feature space or their distributions may be irregular. Even worse, an irrelevant pair of images and texts might have close vector distances in the feature space while a semantically consistent image-text pair could have distant vector distances.

In summary, before the feature fusion of multimodal thought-politics public opinion data, it is necessary to align the extracted image-text features of thought-politics public opinion data so that the features of image-text pairs with identical semantic information are as close as possible in the feature space while those of unrelated image-text pairs are as far apart as possible. This paper uses contrastive learning to align the features of image-text pairs, thereby providing better image-text pair features for multimodal thought-politics public opinion data feature fusion.

Input each batch of data into the momentum model and use the global feature [CLS] token generated by the momentum model from the image-text features as negative samples added to the negative sample queue. Use these as negative samples for contrastive learning in order to achieve alignment between image-text pairs. Input each batch of data into the model and extract the global feature [CLS] tokens from the image-text pair features produced by the model. Treat the same set of image-text features generated by the model as positive sample pairs, and treat the global features extracted by the model and those in the negative sample queue obtained from the momentum model as negative sample pairs. Perform alignment using contrastive learning methods. The loss function for this contrastive learning is as follows:

$$L_{ITC} = \frac{1}{2}\left[H\left(y^{i2t}(I),\, p^{i2t}(I)\right) + H\left(y^{t2i}(T),\, p^{t2i}(T)\right)\right] \tag{13}$$

During the forward propagation process of the model, input the image-text features extracted by the model and the momentum model into the loss function to calculate the loss. At the same time, before feature fusion, perform the first backpropagation based on the contrastive learning loss function to update the image and text feature encoders, thereby aligning as much as possible the image and text features extracted by the image and text encoders, providing aligned image-text features for the feature fusion encoder so that the model has better transferability for downstream tasks.

5.3 Multimodal integration of ideological and political public opinion data features

Input the aligned text-image features into a GRU network to perform multimodal feature fusion. The model uses an attention mechanism at the word-segmentation level to

integrate the text and image features of thought-politics public opinion data. The text feature vector x_w and image feature x_{fet} of thought-politics public opinion are simultaneously input into a gated recurrent unit (GRU). The updated hidden state is as follows:

$$e_i^W = f_{GRUCELL}\left(x_{fet}, x_w\right) \tag{14}$$

In the process of multimodal information fusion, the result can be represented as e_i^W. The initial state h_0 is typically set to zero vector. At the first time step, input h_0 and the representation of multimodal information E^w into the GRU, which produces a new target representation h after updating. In each subsequent time step t, feed the target representation from the previous time step h_{t-1} together with E^w into the GRU again. The GRU module processes this information and continuously updates the target representations. At the last time step of processing, the obtained target representation contains the fused multimodal information. The final result of information processing is as follows:

$$x_{fu} = h_t = f_{GRUCEL}\left(E^W, h_{t-1}\right), \qquad t \geq 1 \tag{15}$$

This method enables the model to consider relevant visual context when processing each word, thereby more accurately understanding and representing text information that contains visual elements. Through the GRU's update mechanism, thus enhancing the final ideological and political public opinion classification or recognition performance.

5.4 Identification of ideological and political public opinion information

Use the fused features for ideological and political public discourse information recognition by extracting the [CLS] token of the features extracted from the feature fusion encoder as the global fusion feature of the image-text pair, which is then input into the classifier. In the classifier, first pass the input image-text global features through a linear projection layer to map the feature dimensions into a higher-dimensional feature space. Next, apply a ReLU() activation function for nonlinear mapping, and finally another linear projection layer maps the features to category dimensions, with the final output passing through a Softmax() function to produce probability distribution of class predictions.

During the model training process, the predicted categories of the momentum model are considered in the loss function calculation. Each batch of data is input into the momentum model to extract image-text fused features for mapping classification and obtain a class probability distribution. This combined with the original model's output classification probability distribution is used to calculate the loss function, then perform a second backpropagation of gradients and update the model parameters. The loss function calculation is as follows, where y represents the true category probability distribution, p represents the model's output category probability distribution, α indicates the weights considering the original model and momentum model. In the process of predicting categories by the model, only the original model is used without consideration of the momentum model. Use the MAX() function to obtain the index with the maximum probability from the output probability distribution as the class for that image-text pair, which is then outputted.

$$L_{mod} = (1-\alpha)H(y,\ p) - \alpha\left(logsoftmax(p) \times softmax(p) \times softmax\left(p_{mod}\right)\right) \tag{16}$$

Sometimes the texts generated by the momentum model corresponding to images can describe the images better than the original texts; more matched image-text pairs enable models to learn better features, which improve the representational ability of the model and allow it to be more effectively applied to downstream tasks.

5.5 Tracking the evolution of ideological and political public opinion

In sudden ideological and political public discourse events in society, the immense pressure generated by online public opinion could potentially escalate the situation further, posing an even greater threat to national security and social stability. Therefore, for current social hotspots and sensitive issues, governments and relevant departments should guide and control them within a certain scope; that is, strengthen supervision of public opinion entities to prevent the situation from worsening. Based on this, tracking the evolution of ideological and political public discourse is essential. Public opinion tracking refers to the objective fact of interactions with specific people or objects under particular spatiotemporal conditions. Therefore, the essence of tracking is a guided learning process, which classifies multimodal ideological and political public opinion information. From a mathematical perspective, public opinion tracking is a mapping process. It maps recent public opinions to topics pre-defined by the system, which can be many-to-one or one-to-many. Therefore, this public opinion can either be positive or negative in nature, with its expression equation as follows:

$$\varpi sim\left(x_i, t_i\right) : A \to B \tag{17}$$

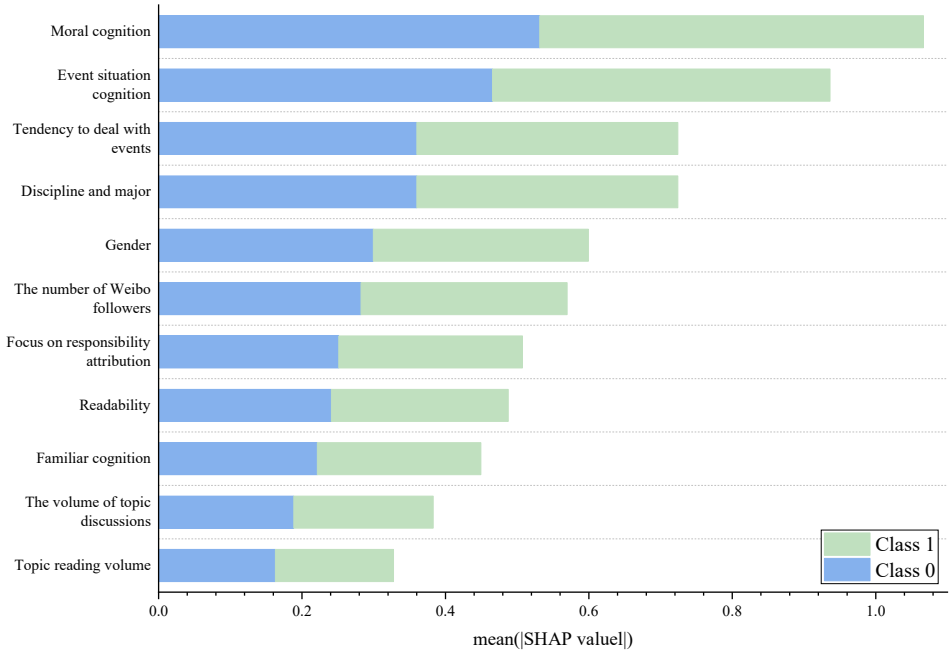
where B represents the current set of hot topics, A represents public opinion, ϖ represents the tracking model. This rule is derived by the system based on sample sets of known topics to formulate discriminative rules for determining internal topic relationships. When new information is input, the system assigns it into a positive or negative public opinion category according to ϖ. By designing a public opinion classification recognition model, the tracking model for ideological and political public opinion can be obtained; this constitutes the training phase. When new public opinion event information appears, based on the classification algorithms discussed in previous sections, the model undergoes comparative analysis against established classification benchmarks to detect and monitor both supportive and critical ideological-political public sentiments.

6 Experimental results and performance analysis

This paper uses web crawlers to collect 203,526 microblog public opinion messages from 24 July 2024 to 31 December 2024, among which 13,698 pieces of university ideological and political public opinion information are selected as data. This dataset contains rich image and text data. The dataset is randomly allocated into the training set, validation set, and test set for the model in a ratio of 8:1:1. The experiment environment uses the VS Code development tool, Python-3.7 programming language, pytorch-1.9.1 deep learning framework; the operating system is Windows64, with an Intel(R) Xeon(R) Gold 6230R CPU @ 2.10 GHz processor used for execution. In the experiments, the number of

network layers in transformer was set to 12, hidden layer size to 768, batch size to 32, learning rate to 5e-5, and weight decay rate to 0.04.

Figure 3 Analysis of the importance of influence indicators based on SHAP values (see online version for colours)



Focusing on investigating the important factors influencing the generation of ideological and political public opinion under negative events, this study ranks the importance of influence indicators based on SHAP values. The importance of influence indicators is shown in Figure 3. The x-axis represents the average sum of |SHAP value|, the y-axis represents feature variables, color represents the predicted value of samples, and length represents the magnitude of the average sum of |SHAP value|. A longer line indicates a greater contribution from the influence indicator variable to the predicted variable. From Figure 3, it can be seen that the most important influence indicator is moral cognition, while the least important is topic reading volume.

To further study the effectiveness of the SHAP-MTRANS model in tracking ideological and political public opinion evolution compared to CNN-BiGRU and MSTRANS models, this paper statistically analyses the time series tracking results distribution for the three models, as shown in Figure 4. From the statistical chart showing the tracking results over time sequences, it can be observed that tracking effectiveness decreases over time. The fastest decline is seen with CNN-BiGRU. This occurs because the multi-modal characteristics of ideological and political public opinion change over time, new features are continuously introduced during classification, while original features are not attenuated, leading to less distinct new features and hence poor classification performance. The tracking results of MSTRANS at later stages are also unsatisfactory, because when new features are introduced, only certain weight values are added without considering the differences in importance among new features, leading to

poor tracking performance. Compared with CNN-BiGRU and MSTRANS, the SHAP-MTRANS model exhibits a slower decline in its tracking results over time; as time progresses further, the tracking results of SHAP-MTRANS outperform the other two methods increasingly, fully demonstrating the superiority of SHAP-MTRANS.

Figure 4 Distribution of tracking results across time series for various models (see online version for colours)

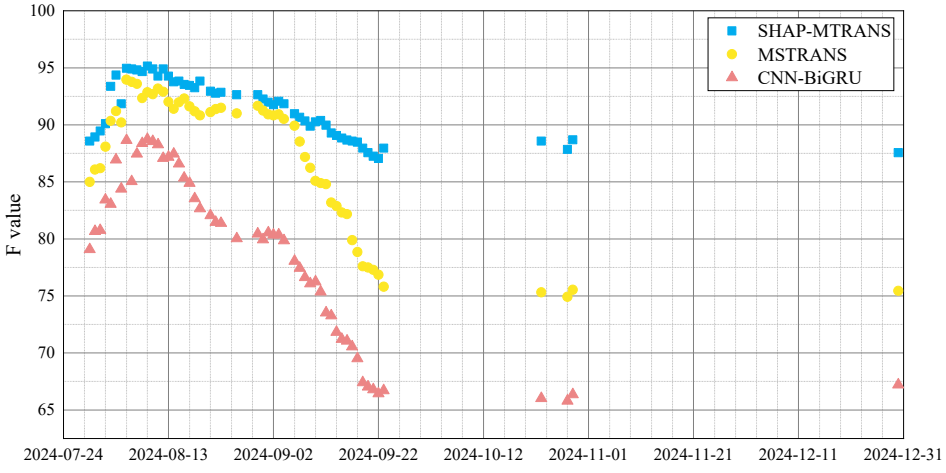
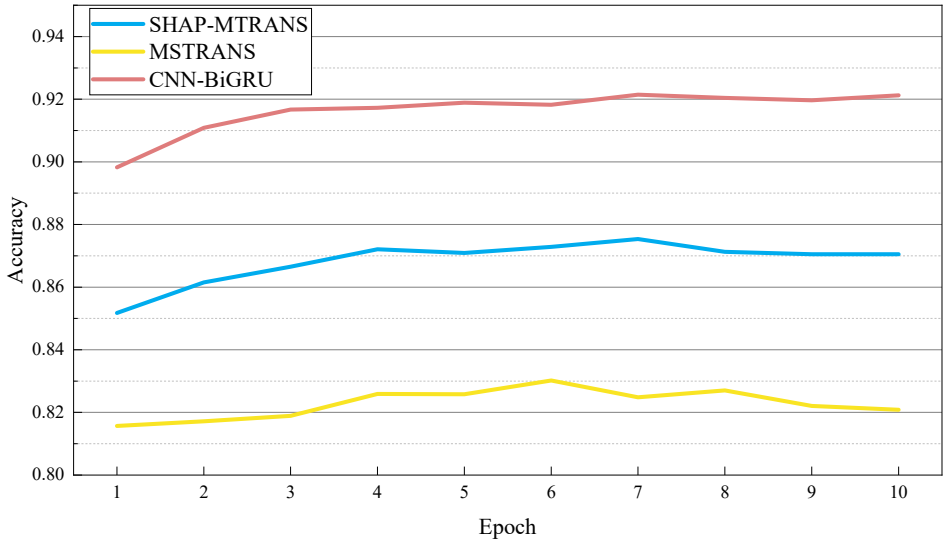


Figure 5 Tracking accuracy trend chart by epoch (see online version for colours)



For the goal of ensuring reliability of the experimental results, the three models are trained using identical datasets under consistent configuration environments. The tracking accuracy trends on test sets for the three comparison methods over epochs are shown in Figure 5. The tracking accuracies of CNN-BiGRU, MSTRANS, and SHAP-MTRANS

are 82.2%, 87.4%, and 92.1% respectively. SHAP-MTRANS achieves better tracking results than other models. The CNN-BiGRU model realises ideological and political public opinion evolution tracking through a hybrid model combining CNN and BiGRU. The bi-directional structure of the BiGRU implies that it must handle both forward and backward states in sequences, resulting in parameter counts and computational costs approximately twice those of ordinary GRUs. When combined with the CNN feature extractor, the overall depth and complexity of the entire model are significantly increased. In addition, features automatically extracted by CNNs are difficult to express in ways understandable to humans, and the state evolution of BiGRU is even more complex; this method struggles to make positive or negative judgements about public discourse. The MSTRANS model achieves ideological and political public opinion evolution analysis through a transformer, but the transformer model lacks explicit input-output mappings, its internal working mechanism is relatively complex, and attention weights are difficult to interpret intuitively. Compared with some other machine learning models, it is even more challenging to explain how this model derives corresponding analytical results and predictive conclusions based on the input ideological and political public opinion data, which poses difficulties for ideological workers in understanding and trusting the model's output. The SHAP-MTRANS model improves the interpretability of the transformer model through SHAP attribution, achieving better tracking performance compared to the other two methods.

This paper also analyses the tracking effects of different models by selecting commonly used metrics for evaluating ideological and political public opinion evolution tracking: precision, recall, F1 and AUC. As shown in Table 1, the precision and recall of SHAP-MTRANS are 94.2% and 93.6%, respectively, which represent improvements of 12.8% and 7.9% over CNN-BiGRU and 8.3% and 5.3% over MSTRANS. The F1 scores for CNN-BiGRU, MSTRANS, and SHAP-MTRANS are 83.5%, 87.1%, and 93.9%, respectively; SHAP-MTRANS achieves improvements of 10.4% and 6.8% over CNN-BiGRU and MSTRANS, respectively. Furthermore, comparing the tracking accuracy metric AUC, SHAP-MTRANS outperforms CNN-BiGRU and MSTRANS by 9.47% and 6.03%, respectively. This is because SHAP-MTRANS not only removes non-critical influence indicators through SHAP, but also improves the multi-head attention mechanism of the transformer via SHAP, achieving efficient multimodal feature extraction and fusion, thereby achieving better tracking performance.

Table 1 Effectiveness analysis of ideological and political public opinion monitoring

Model	*Precision*	*Recall*	*F1*	*AUC*
CNN-BiGRU	81.4%	85.7%	83.5%	0.8966
MSTRANS	85.9%	88.3%	87.1%	0.9257
SHAP-MTRANS	94.2%	93.6%	93.9%	0.9815

7 Conclusions

Ideological and political public opinion is of great significance for social stability, and accurately tracking its evolution process is key to effectively addressing public opinion challenges. Existing methods often struggle to sufficiently mine the correlations between different modalities when handling multimodal data in ideological and political public

opinion, and are also lacking in terms of analysing the causes behind public opinion evolution. To address this issue, this paper proposes an evolution tracking model based on a multimodal transformer and SHAP attribution. First, we use a random forest-optimised SHAP algorithm to assess the importance of influence indicators for ideological and political public opinion under strongly coupled conditions, and apply Gramian angular field transformation to generate 2D images from those indicators with Shapley values greater than zero. Then, this paper uses the SHAP algorithm to improve the self-attention mechanism of the transformer and utilise Shapley values to quantify the contribution of each input characteristic to the forecasting results, helping identify and enhance local features that have a significant impact on tracking outcomes. We employ an improved transformer model to extract image features and introduce the BERT model for extracting text features. Contrastive learning methods are used for feature alignment between different modalities, followed by multimodal fusion via GRU. The fused features are classified using a softmax function to obtain ideological and political public opinion categories, thereby achieving tracking of the evolution of ideological and political public discourse. Experimental outcome show that the proposed model's tracking accuracy improved by 9.47% and 6.03%, respectively, compared to baseline models, fully verifying its effectiveness.

Declarations

This work is supported by t School Teaching Reform Project of Suzhou Polytechnic University (Key Projects) (No. SZDJG-23005).

The author declares that she has no conflicts of interest.

References

Al-Laith, A. and Shahbaz, M. (2021) 'Tracking sentiment towards news entities from Arabic news on social media', *Future Generation Computer Systems*, Vol. 118, pp.467–484.

Bashiri, H. and Naderi, H. (2024) 'Comprehensive review and comparative analysis of transformer models in sentiment analysis', *Knowledge and Information Systems*, Vol. 66, No. 12, pp.7305–7361.

Chen, H., Covert, I.C., Lundberg, S.M. and Lee, S-I. (2023) 'Algorithms to estimate Shapley value feature attributions', *Nature Machine Intelligence*, Vol. 5, No. 6, pp.590–601.

Chen, J., Li, K., Tang, Z., Bilal, K., Yu, S., Weng, C. and Li, K. (2016) 'A parallel random forest algorithm for big data in a spark cloud computing environment', *IEEE Transactions on Parallel and Distributed Systems*, Vol. 28, No. 4, pp.919–933.

Chen, K. and Wei, G. (2023) 'Public sentiment analysis on urban regeneration: a massive data study based on sentiment knowledge enhanced pre-training and latent Dirichlet allocation', *PloS One*, Vol. 18, No. 4, pp.85–101.

Chen, M. and Du, W. (2023) 'The predicting public sentiment evolution on public emergencies under deep learning and internet of things', *The Journal of Supercomputing*, Vol. 79, No. 6, pp.6452–6470.

Du, H., Yu, Q. and Chen, J. (2025) 'Research on sentiment analysis of online public opinion based on multimodal big language modeling', *Journal of Computational Methods in Sciences and Engineering*, Vol. 4, No. 1, pp.25–37.

Du, J., Jiang, Y. and Liang, Y. (2024) 'Transformers in opinion mining: addressing semantic complexity and model challenges in NLP', *Transactions on Computational and Scientific Methods*, Vol. 4, No. 10, pp.43–52.

Hayadi, B.H. and Maulita, I. (2025) 'Sentiment analysis of public discourse on education in indonesia using support vector machine (SVM) and natural language processing', *Journal of Digital Society*, Vol. 1, No. 1, pp.68–90.

Jahin, M.A., Shovon, M.S.H., Mridha, M., Islam, M.R. and Watanobe, Y. (2024) 'A hybrid transformer and attention based recurrent neural network for robust and interpretable sentiment analysis of tweets', *Scientific Reports*, Vol. 14, No. 1, pp.82–96.

Kirkizh, N., Ulloa, R., Stier, S. and Pfeffer, J. (2024) 'Predicting political attitudes from web tracking data: a machine learning approach', *Journal of Information Technology & Politics*, Vol. 21, No. 4, pp.564–577.

Liang, H., Ganeshbabu, U. and Thorne, T. (2020) 'A dynamic Bayesian network approach for analysing topic-sentiment evolution', *IEEE Access*, Vol. 8, pp.54164–54174.

Lin, H. and Bu, N. (2022) 'A CNN-based framework for predicting public emotion and multi-level behaviors based on network public opinion', *Frontiers in Psychology*, Vol. 13, pp.39–51.

Lin, R. and Hu, H. (2023) 'Multi-task momentum distillation for multimodal sentiment analysis', *IEEE Transactions on Affective Computing*, Vol. 15, No. 2, pp.549–565.

Mu, Y., Liu, X. and Wang, L. (2018) 'A Pearson's correlation coefficient based decision tree and its parallel implementation', *Information Sciences*, Vol. 435, pp.40–58.

Nassiri, K. and Akhloufi, M. (2023) 'Transformer models used for text-based question answering systems', *Applied Intelligence*, Vol. 53, No. 9, pp.10602–10635.

Polli, C. and Santonocito, C.S. (2024) 'Reputation at risk: sentiment analysis and social media listening tools under the lens of critical multimodal discourse studies', *Hermes*, No. 64, pp.331–352.

Samih, A., Ghadi, A. and Fennan, A. (2023) 'Enhanced sentiment analysis based on improved word embeddings and XGboost', *International Journal of Electrical and Computer Engineering*, Vol. 13, No. 2, pp.1827–1836.

Su, J. (2024) 'Innovation of guiding mechanism of ideological and political education in colleges and universities under the background of network public opinion', *Transactions on Comparative Education*, Vol. 6, No. 3, pp.50–55.

Sun, M., Wei, Y., Jiang, S. and Jia, G. (2024) 'A comprehensive framework for predicting public opinion by tracking multi-informational dynamics', *Frontiers of Computer Science*, Vol. 18, No. 4, pp.18–34.

Sun, R., An, L., Li, G. and Yu, C. (2025) 'Predicting social media rumours in the context of public health emergencies', *Journal of Information Science*, Vol. 51, No. 2, pp.338–353.

Tan, X., Zhuang, M., Lu, X. and Mao, T. (2021) 'An analysis of the emotional evolution of large-scale Internet public opinion events based on the BERT-LDA hybrid model', *IEEE Access*, Vol. 9, pp.15860–15871.

Valeriani, L., Doimo, D., Cuturello, F., Laio, A., Ansuini, A. and Cazzaniga, A. (2023) 'The geometry of hidden representations of large transformer models', *Advances in Neural Information Processing Systems*, Vol. 36, pp.51234–51252.

Wang, H., Liang, Q., Hancock, J.T. and Khoshgoftaar, T.M. (2024) 'Feature selection strategies: a comparative analysis of SHAP-value and importance-based methods', *Journal of Big Data*, Vol. 11, No. 1, pp.44–56.

Wei-Dong, H., Qian, W. and Jie, C. (2018) 'Tracing public opinion propagation and emotional evolution based on public emergencies in social networks', *International Journal of Computers Communications & Control*, Vol. 13, No. 1, pp.129–142.

Yan, C., Liu, J., Liu, W. and Liu, X. (2022) 'Research on public opinion sentiment classification based on attention parallel dual-channel deep learning hybrid model', *Engineering Applications of Artificial Intelligence*, Vol. 116, pp.105–118.