International Journal for Housing Science and Its Applications
Publish August 3, 2025. Volume 46, Issue 3 Pages 1020-1030

https://doi.org/10.70517/ijhsa46378

# Sentiment analysis and propagation trend modeling of social media video content based on frequency domain feature extraction and computational methods

**Jingyu Zhang[1,*] and Kang An[2]**
[1] Shanghai Documentary Academy, Shanghai University of Political Science and Law, Shanghai, 201701, China
Corresponding authors: (e-mail: zhangjingyu0129@126.com).

**Abstract** In this paper, a multimodal joint analysis method based on frequency domain feature extraction and deep learning is proposed. Firstly, frequency domain decomposition and threshold denoising of video frames are performed using wavelet transform to improve the image characterization ability by retaining low-frequency key information and high-frequency detailed features. Second, the RNN model is improved by combining the multiple-pate notice machine-made to realize the emotion-semanteme fusion across visual-text modalities. Finally, the probabilistic clustering of communication sequences based on the GMM pattern is engaged in analysis the spatio-temporal evolution pattern of opinion diffusion. The test consequences indicate that the proposed method realize an image extraction precision of 92.59% on the VOT-2024 dataset and an F1 value of 84.29% for RNN-AM in the CMU-MOSI sentiment analysis task, which outperforms existing mainstream models. When applied to the COVID-HATE dataset, it successfully discovers the sentiment relevance of video content and captures the 48-hour intervention window and 7-day decay cycle of video dissemination, indicating that online public opinion intervention behaviors need to be carried out within 48 hours in order to achieve better results.

**Index Terms** content sentiment analysis, propagation trend analysis, frequency domain feature extraction, RNN-AM, GMM

## I. Introduction

In the rapid progress of Internet technology, diversified information carriers have gradually enriched our information consumption mode. From early text blogs and news reports, to social media posts incorporating images and emoticons, and even the current popularity of short videos, each carrier has demonstrated its unique information characteristics [1], [2]. Public attitudes toward these diverse media vary, and these attitudes not only affect individuals, but also have far-reaching implications for businesses and even governments [3], [4]. Video platforms have been transformed into a new generation of social media, with a more significant impact than traditional social platforms, covering a richer and more diverse range of social information content, and showing a significant increase in netizens' discussion and participation in their content [5]-[7]. By virtue of its unique form of expression, i.e., the content is more specific, more coherent, more impactful, and easier to stimulate emotions, video provides netizens with a kind of immersive experience and real feelings as if they were “in the scene” or “experiencing the scene” [8]-[10]. Therefore, the use of video sentiment analysis techniques to mine the emotional tendency and emotional state of video content can be helpful for the improvement of products and services, the monitoring of online public opinion, and the research and development of human-computer interaction systems, such as emotional dialog robots [11], [12].

Sentiment analysis is used to reason about the viewpoints embedded in the target object, which narrowly includes only textual information [13]. With the popularity of multimedia social platforms, more and more users post videos to express their subjective opinions and share their views on things in the form of single monologue or multi-person conversations, and the target object of sentiment analysis in a broad sense includes multimodal information such as text, audio, and visual information that is separated from the video [14], [15]. The core difficulty of video sentiment analysis lies in reasoning from multimodal features to the most relevant ones for speaker sentiment discrimination [16]. Specifically, the video sentiment analysis model should be able to effectively fuse multimodal heterogeneous features and eliminate redundant features on the one hand, and enrich the sentiment semantics of video clips on the other hand [17], [18]. For this reason, video content sentiment analysis methods based on frequency domain feature extraction and computation methods show great potential for application when facing the above problems.

In this paper, we utilize wavelet transform for frequency domain feature extraction of social media video images, and improve the quality of image features through denoising process. A sentiment analysis model based on RNN and notice machine-made is engaged in deeply mine the sentiment data in the video. Combined with GMM model, the propagation trend of social media videos is modeled and analyzed. The PF and MS algorithms are introduced to compare with the suggested arithmetic to assessment the perform of the wavelet frequency domain based image feature extraction model. The effectiveness of the RNN-AM model is explored through experimental validation on CMU-MOSI and CMU-MOSEI ensemble of communication. The COVID-HATE ensemble of communication is selected to carry out the study to reveal the propagation law of social media videos based on multi-label sentiment and propagation trend analysis.

## II. Social media video content sentiment and communication trend analysis model construction

With the quickly development of society medium systems, video content has become one of the core carriers for the public to express their views and convey their emotions. However, the unstructured nature of massive video data and the complex communication environment pose a serious challenge to sentiment analysis and trend prediction: on the one hand, video images are often interfered by equipment noise and transmission distortion, which leads to difficulties in extracting key features. On the other hand, sentiment expression is characterized by multimodal spatial and temporal coupling, and traditional unimodal analysis methods are difficult to capture semantic associations across visual and textual domains. In addition, public opinion communication is jointly driven by individual behavior and group effect, and dynamic modeling methods are urgently needed to reveal its spatio-temporal evolution law. Aiming at the hereinbefore issues, this article suggest a joint emotion-communication analysis framework for social media videos that integrates frequency domain characteristic extraction and deep learning, which systematically improves the accuracy of sentiment recognition and the reliability of communication prediction.

### *II. A. Wavelet frequency domain based image feature extraction*

#### II. A. 1) Wavelet frequency domain based image preprocessing

Images in social media videos are susceptible to a variety of noise interferences and influences, which may originate from defects in the image acquisition equipment, errors in the transmission process, or uncertainties in the processing. The presence of noise will have an adverse effect on image feature extraction, so it is necessary to remove noise from social media video images, which helps to improve the accuracy of image feature extraction. Therefore, firstly the social media video images are preprocessed using wavelet frequency domain i.e. image noise removal process.

The wavelet decomposition process of social media video images is shown in Fig. 1. In the figure, $f(x,y)$ is the original social media video image; $A(x)$, $A(y)$, $B(x)$, $B(y)$ are the auxiliary functions applied in the wavelet decomposition process; $D^0(x,y)$ is the first-stage smoothing approximation of $f(x,y)$; $D^1(x,y)$, $D^2(x,y)$, $D^3(x,y)$ are the details of $f(x,y)$.

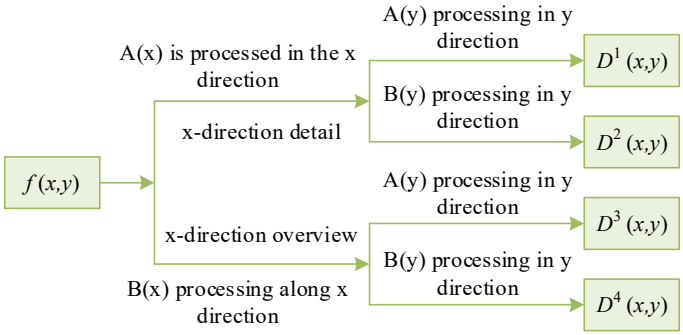



Figure 1: Wavelet decomposition of social media video images

As can be seen in Fig. 1, the social media video image is decomposed into four sub-frequency images, $D^0(x,y)$, $D^1(x,y)$, $D^2(x,y)$ and $D^3(x,y)$, by using the wavelet transform frequency division method. Among them, the low-frequency image $D^0(x,y)$ is mainly the image key information, the high-frequency images $D^1(x,y)$, $D^2(x,y)$ and $D^3(x,y)$ are mainly the image edge information and detail information, and the noise information mainly exists in the high-frequency images. In the standard case, the wavelet coefficient value of the high-frequency component is

small, while the wavelet coefficient value of the noise information is large, so the wavelet threshold denoising way can be engaged in delete the noise in the high-frequency image, and the denoised social media video image can be obtained by wavelet reconstruction.

Wavelet frequency domain algorithm is mainly based on the difference between the wavelet coefficients of the key information of the image (larger) and the wavelet coefficients of the noise information (smaller), set a fixed threshold, retain the small wave modulus bigger than the value range, and set the small wave modulus smaller than the threshold to 0, therefore to realize the objective of noise data removal. Formulate the social media video graphic delete the noise process according to small wave frequency domain arithmetic, the specific steps are shown in Fig. 2.

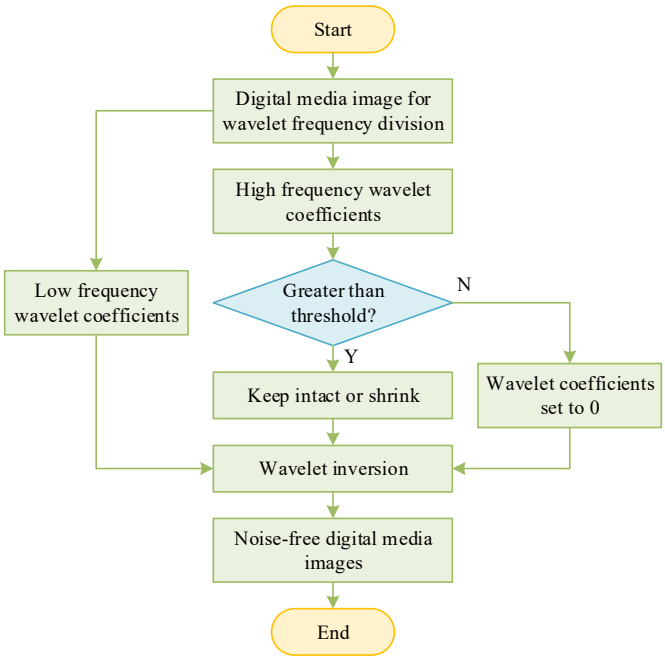



Figure 2: Denoising process based on wavelet frequency domain

Executing the process shown in Fig. 2 can complete the preprocessing of social media video images, that is, social media video image noise removal processing, to obtain noise-free social media video images, denoted as $g(x,y)$.

#### II. A. 2) Image Feature Extraction

Based on $g(x,y)=\{\Omega_1,\Omega_2,\cdots,\Omega_m\}$, social media video image features are extracted based on Tiansi operator.

The social media video image feature extraction based on Tiansi operator is:

$$\zeta^{\gamma}(\Omega_i)=\lim\frac{1}{h^{\gamma}}\times\sum_{k=1}^{n}(-1)^k\times\frac{\psi(\gamma+1)}{\psi(\gamma-k+1)\cdot k!} \tag{1}$$

where: $\zeta^{\gamma}(\Omega_i)$ is the feature information of the social media video image segmentation region $\Omega_i$; $\gamma$ is the auxiliary constant of the Tiansi operator, and the performance of the Tiansi operator plays stable only in the range of $[0.4,0.6]$; $\psi(\cdot)$ is the image feature information finding function.

The extraction of social media video image feature information is accomplished according to equation (1), which is written as $\zeta=\left\{\zeta^{\gamma}(\Omega_1),\zeta^{\gamma}(\Omega_2),\cdots,\zeta^{gamma}(\Omega_m)\right\}$.

### *II. B.RNN-AM based content sentiment analysis*

#### II. B. 1) Recurrently Neurally Networks

Recurrently Neurally Networks (RNN) can use the outputs of some nodes as inputs to the same nodes, and are therefore suitable for portraying dynamic sequential scenes, and are often used in the field of NLP to capture long-

distance dependencies of sentences.RNNs can be categorized into LSTM networks as well as GRU networks according to the network structure.

LSTM network composition of three parts: import port, oblivion port, and efference port, and the construction is indicate in Fig. 3(a), which is suitable for tasks with long processing intervals and delays. When the value of the sigmoid function at the input gate is close to 0, the data stream is not passed backward; when the value of the sigmoid function at the forgetting gate is close to 0, the stored data is forgotten; and the value of the sigmoid function at the output gate will determine whether the current data can be output or not. Eqs. (2) to (6) give the specific calculations:

$$f_t = \delta_g(W_f x_t + U_f h_{t-1} + b_f) \tag{2}$$

$$i_t = \delta_g(W_i x_t + U_i h_{t-1} + b_i) \tag{3}$$

$$o_t = \delta_g(W_o x_t + U_o h_{t-1} + b_o) \tag{4}$$

$$c_t = f_t \square\ c_{t-1} + i_t \square\ \delta_c(W_c x_t + U_c h_{t-1} + b_c) \tag{5}$$

$$h_t = o_t \square\ \delta_h(c_t) \tag{6}$$

where $x_t \in \square^d$, is the import feature of the network; $f_t \in R^h$ expression the forgetting port; $i_t \in \square^h$ expression the import port; $o_t \in \square^h$ expression the import port; $h_t \in \square^h$ expression the efference port; $c_t \in \square^h$ expression the conceal condition of the network; $W \in \square^{h \times d}$, $U \in \square^{h \times h}$, $b \in \square^h$ denotes the matrix during training and $\square$ represents the Hadamard product.

LSTM networks have more parameters and are more complicated to train. Therefore GRU network is proposed to simplify the LSTM network, the concrete construction is indicate in Fig. 3(b). The main innovation of this network is to unify the forgetting port and import port into a simplex “update gate”. The specific calculations are shown in Eq. (7) to Eq. (10).

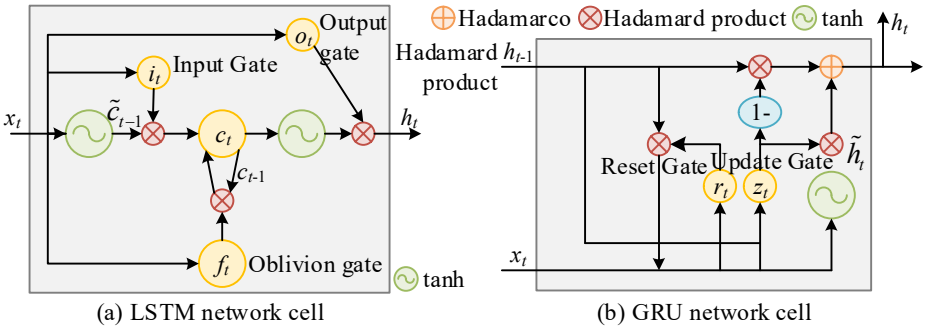
$$z_t = \sigma_g(W_z x_t + U_z h_{t-1} + b_z) \tag{7}$$

$$r_t = \sigma_g(W_r x_t + U_r h_{t-1} + b_r) \tag{8}$$

$$\hat{h}_t = \phi_h(W_h x_t + U_h(r_t \square\ h_{t-1}) + b_h) \tag{9}$$

$$h_t = z_t \square\ h_{t-1} + (1 - z_t) \square\ \hat{h}_t \tag{10}$$

where $x_t \in \square^d$ is the import feature complexor, $h_t \in \square^h$ is the output feature vector, $\hat{h}_t \in \square^h$ is the efference feature complexor, $z_t \in \square^h$ is the update gate vector, $r_t \in \square^h$ is the reset gate vector, $W \in \square^{h \times d}$, $U \in \square^{h \times h}$, $b \in \square^h$ are the parameter matrices to be trained with bias terms.




(a) LSTM network cell (b) GRU network cell

Figure 3: Cell structure of LSTM and GRU networks

**II. B. 2) Attention mechanisms**

Humans have an attentional bias when observing images and reading passages, consciously focusing on the main part of the picture and key phrases, and weakening or even ignoring unimportant components. In order for computers to simulate human attentional behavior, the Attention Mechanism (AM) was created. This mechanism

helps to quickly and accurately filter out useful information from a large number of information sources and aggregate them to get new global information. Attention mechanisms in deep learning are usually categorized into three types: global supple notice, local hard notice, and self-notice. In this article, we concentrate on supple notice, whose input contains two objects: query sequences and key-value pair sequences. Denote the query matrix as $Q$, $Q=[q_1,q_2,\cdots,q_N]\in\mathbb{R}^{d\times N}$, the key matrix as $K$, $K=[k_1,k_2,\cdots,k_M]\in\mathbb{R}^{d\times M}$ and the value matrix as $V$, $V=[v_1,v_2,\cdots,v_M]\in\mathbb{R}^{d\times M}$. The specific implementation process of the attention mechanism is as follows:

(1) Match each query vector $q_i$ with the column vectors in the key matrix to compute the value of the attention score. Generally, $q_i$ is calculated by dot-multiplying it with the vector $k_j$ of the corresponding key, so it is also called dot-propagation attention (DPA), as shown in Eq. (11).

$$e_{q_i,k_j}=q_i\cdot k_j \tag{11}$$

(2) In order to obtain the attention distribution of the query vector $q_i$ to all key vectors, the obtained attention score value is passed through the Softmax normalization function to acquire the attention weight distribution, which indicates the degree of attention of the query vector to the different key vectors, and the computation process is shown in Eq. (12).

$$\alpha_{q_i,k_j}=Soft\max(e_{q_i,k_j})=\frac{\exp(q_i\cdot k_j)}{\sum_{j'=1}^{M}\exp(q_i\cdot k_{j'})} \tag{12}$$

(3) Weighted summation of the value complexor by $\alpha_{q_i,k_j}$ to get the final attention result, obtaining the representation of the context vector at the $i$ th position after the attention calculation, where each value vector corresponds to the corresponding key, the specific calculation process is shown in Eq. (13).

$$Attention(q_i,K,V)=\sum_j\alpha_{q_i,k_j}v_{k_j} \tag{13}$$

With the above steps, each word in the input sentence can be associated with its corresponding query, key and value vectors. The above process can be abbreviated into matrix form as shown in equation (14).

$$Attention(Q,K,V)=Vsoft\max\left(\frac{K^TQ}{\sqrt{d}}\right) \tag{14}$$

where, the division by the normalization factor $\sqrt{d}$ is to ensure that the variance of the dot product result does not vary with the dimension $d$ to alleviate the gradient vanishing situation when computing Softmax.

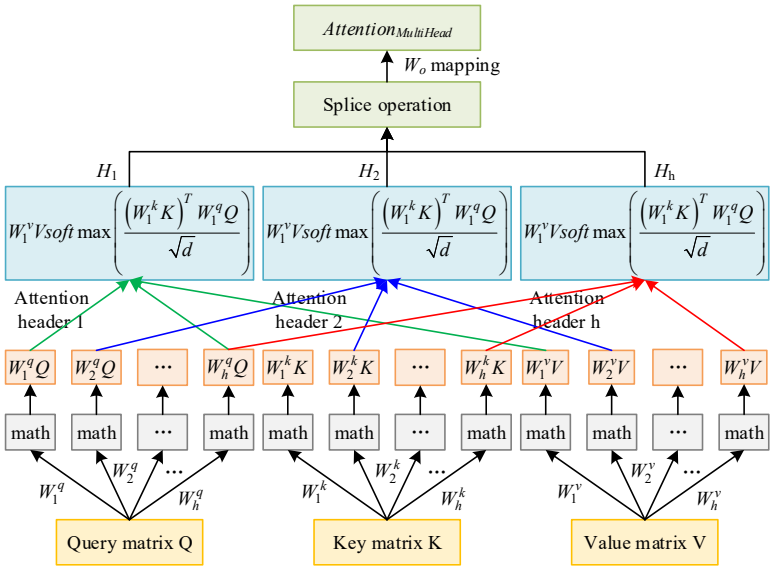



Figure 4: Computational process of multi-head attention mechanism

In order to realize the execution of dot product operations in different subspaces in parallel to improve the computational efficiency, the multi-head attention mechanism is designed, and its computational process is shown in Fig. 4. Assuming that there are $h$ subspaces for performing the multi-head attention operation, firstly, through the $h$ group mapping matrices $W_i^q$, $W_i^k$, $W_i^v \in \mathbb{R}^{d/h\times d}\left(i=1,2,\cdots,h\right)$ will $Q$, $K$, and $V$ are mapped to the corresponding subspaces, respectively, to obtain the result of the attention computation $H_i \in \mathbb{R}^{d/h\times N}$ for the $i$ th attention head, as shown in Eq. (15). Finally, the computed results of all attention heads are spliced along the hidden layer dimension and the spliced results are mapped back to the original space using the output mapping matrix $W_o$, as shown in Equation (16).

$$H_i = Attention\left(W_i^q Q, W_i^k K, W_i^v V\right) \tag{15}$$

$$Attention_{MultiHead} = W_o\left(H_1 \oplus H_2 \oplus \cdots \oplus H_h\right) \tag{16}$$

### *II. C.GMM-based communication trend analysis*

Due to the complexity and diversity of objectives that characterize spatio-temporal big data, a number of analysis ways have arisen, including but not limited to clustering, prediction, and change detection. As one of the most important ways, clustering has been widely engaged in many applications such as image segmentation problems in medicine and transportation.

Spatio-temporal data clustering is usually based on spatial and temporal similarity to divide datasets with similar behaviors into spatio-temporal objects, and in doing so, it should be maintained that the differences between groups of divided datasets should be as large as possible for the datasets within the same group should be as small as possible.

In this article, a clustering arithmetic according to Gaussian Mixture Model (GMM) is selected, because its mathematical form is simple and the expression of its parameters is closed, which can achieve better clustering performance in complex multimodal data, and effectively solve the problem of poor clustering performance of multimodal data.

The core idea of the algorithm is that the GMM composition of several Gaussian distributions, the original data are generated from these distributions, and data obeying the same independent Gaussian distribution are considered to belong to the same cluster. Its advantages are that it gives a more realistic probability of belonging, is relatively scalable by changing the distribution, the number of clusters, etc., and is supported by well-developed statistics. The disadvantages, on the other hand, are that it involves many parameters that have a great impact on the clustering results and has a relatively high time complexity.

The GMM-based clustering arithmetic composition of two subissues. First, we must estimate the pattern parameters. Second, we need to determine the number of components in the GMM.

#### II. C. 1) Setting the GMM parameters

First, we address the problem of pattern parameter estimation by assuming that the training ensemble of communication $D_j$ is generated by a finite Gaussian mixture pattern consisting of $M$ components. If the labels of these components are known, then the issue reduces to the usual parameter estimation problem and we can use maximum likelihood estimation (MLE).

GMM-based clustering methods use MLE to find the maximum logarithmic similarity probability of each data point, a value that represents the maximum probability that this data point will be classified into that cluster and the minimum probability that it will be classified into other clusters. In this method, each component of the data element is associated with some probability capability so that their sum will be equal to one.

Suppose that each sample $x_j$ comes from a superpopulation $D$, which is a mixture of a finite number $M$ of clusters $D_1,\cdots,D_M$ in a certain ratio $\alpha_1,\cdots,\alpha_m$ respectively, where $\sum_{j=1}^{M}\alpha_i = 1$, $\alpha_i \geq 0\left(i=1,\cdots,M\right)$. We can now pattern the data $D=\{x_i\}_{n=1}^{n}$ as a mixture of densities arising independently from the following:

$$p(x_i \mid \Theta) = \sum_{j=1}^{M}\alpha_j p_j(x_i \mid \theta_j) \tag{17}$$

$$L(\Theta) = \sum_{i=1}^{n}\ln\left[\sum_{j=1}^{M}\alpha_j p_j(x_i \mid \theta_j)\right] \tag{18}$$

Here $p_i(x_i \mid \theta_i)$ corresponds to mixture $j$ and is parameterized by $\theta_j$, $\theta=(\alpha_1,\cdots,\alpha_m,\theta_1,\cdots,\theta_m)$ denotes all unknown parameters related to the density of the $M$ -component mixture. In general, Eq. (18) is difficult to optimize because it contains the logarithmic function $\ln$. However, this equation is greatly simplified when there are unobserved (or incomplete) samples.

We now briefly describe maximum likelihood estimation, where the first step of the arithmetic is to maximize the log-likelihood function for expectation using the current parameters and conditioning on the observed samples. In the second step of the arithmetic, the parameter values are recalculated.The EM arithmetic iterates through these two steps until convergence is reached. For a multiple normal distribution, the expectation value $E[\cdot]$, expression by $p_{ij}$, is the probability that a Gaussian mixture $j$ produces data point $i$, which is given by:

$$p_{ij}=\frac{\left|\hat{\Sigma}_j\right|^{-1/2} e^{\left\{-\frac{1}{2}(x_i-\hat{\mu}_j)^t\hat{\Sigma}_j^{-1}(x_i-\hat{\mu}_j)\right\}}}{\sum_{l=1}^{M}\left|\hat{\Sigma}_l\right|^{-1/2} e^{\left\{-\frac{1}{2}(x_i-\hat{\mu}_l)^t\hat{\Sigma}_l^{-1}(x_i-\hat{\mu}_l)\right\}}} \tag{19}$$

$$\hat{\alpha}_j^k=\frac{1}{n}\sum_{i=1}^{n}p_{ij} \tag{20}$$

$$\hat{\mu}_j^k=\frac{\sum_{i=1}^{n}x_i p_{ij}}{\sum_{i=1}^{n}p_{ij}} \tag{21}$$

$$\hat{\Sigma}_j^k=\frac{\sum_{i=1}^{n}p_{ij}(x_i-\hat{\mu}_j^k)(x_i-\hat{\mu}_j^k)^t}{\sum_{i=1}^{n}p_{ij}} \tag{22}$$

#### II. C. 2) Clustering

Once the GMM is fitted to the training data, we can use the pattern to predict the labels for each cluster. The assignment of labels is done using the maximum likelihood (MLE) procedure. The discriminant function $g(\cdot)$ given by the MLE principle is shown below:

$$g_i(x)=-\ln\left|\Sigma_i\right|-(x-\mu_i)^t\left|\Sigma_i\right|^{-1}(x-\mu_i) \tag{23}$$

For each characteristic complexor, we assign a cluster label $i$ if $g_i(x)$ is the largest among all cluster labels.

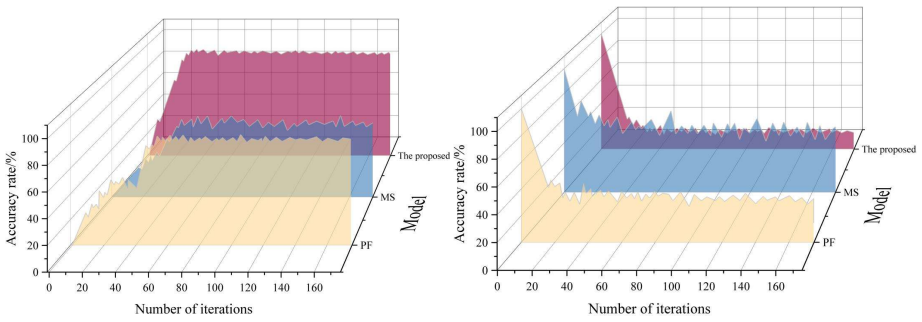
## III. Empirical Analysis of Sentiment and Communication Trends of Social Media Video Content Based on Frequency Domain Feature Extraction

### *III. A. Performance level validation*

#### III. A. 1) Image Feature Extraction

In order to assessment the perform of the wavelet frequency domain based image feature extraction model, this paper performs simulation experiments and introduces the Particle Filtering (PF) algorithm and Mean Shift (MS) algorithm to compare with the proposed algorithm. In this paper, we choose to use the Visual Object Tracking 2024 (VOT-2024) dataset for the simulation experiments, and randomly 80% of it is used for training and 20% is used for testing.Each video sequence in the VOT dataset contains a number of frames, each of which contains one or more annotated objects. The annotations are usually bounding boxes of the objects or finer markers such as pixel-level segmentation of the objects. These annotations enable training as well as testing of video image feature extraction algorithms. Moreover, the dataset contains evaluation frameworks that allow the calculation of various performance metrics.

The accuracy as well as robustness of the three image feature extraction models are tested and the test consequences are indicate in Fig. 5(a~b), respectively. From Fig. 5, it can be seen that the model suggested in this article has better perform in terms of accuracy and robustness, and its training effect also performs well and can reach the optimal state with fewer training times. The algorithm in this paper can reach the optimal state in about 27 iterations, and its accuracy is 92.59%, which is 14.91% and 27.12% ahead of the PF algorithm and the MS algorithm, respectively; and its optimal robustness is 12.64%, which is 20.45% and 37.89% less than the PF arithmetic and the MS arithmetic, respectively. The higher accuracy and lower robustness ensure that the arithmetic in this article can maintain high accuracy and have a low error rate in video image feature extraction.




(a)Accuracy rate (b)Robustness

Figure 5: Accuracy and robustness test of the three algorithms

**III. A. 2) Content Sentiment Analysis**

Because CMU-MOSI and CMU-MOSEI ensemble of communication are widely engaged in the field of emotion analysis, they are chosen in this article to validate the feasibility of the suggested model.CMU-MOSI, as a widely engaged in dataset in the field of emotion analysis, is edited from 93 videos posted by 89 different narrators on YouTube, which contain a total of 2,199 discourse videos, with video characteristics extracted at a sampling rate of 15Hz sampling rate is extracted, while the text modality is segmented by words and represented as discrete word embeddings.CMU-MOSEI is much larger than CMU-MOSI in terms of data volume.Clipped from 3837 videos posted by 1000 different narrators on YouTube, a total of 22,856 discourse videos are included, and the video characteristics are extracted at a sampling rate of 15Hz.

The proposed pattern is compared with some of the existing condition-of-the-art baseline patterns according to unaligned data sequences, including Early Fusion LSTM (EF-LSTM) based on word-forced alignment, Late Fusion (LF-LSTM) Multimodal Factor Decomposition Model (MFM), Recursive Participant Variable Embedding Network (RAVEN), Multi-Layer Feature Fusion Network HFFN, Multimodal Cyclic Translation Network (MCTN), Modal Invariant and Specific Representation (MISA), MSAF, and Word Unaligned Multiple modal Transformer based MulT, Low-rank Fusion Transformer (LMT-MULT), learn modality-fused representations with CB- Transformer (LMR-CBT), Self-Supervised Multiple-Task Learning (Self-MM), Unimodal Reinforced Transformer (UR-Transformer), Progressive Modal Reinforcement (PMR), Weightedcross-modal attention mechanism).

The word-aligned setting requires additional steps to manually align the visual streams in word resolution and perform text and video cross-modal fusion at word-aligned time steps. The non-aligned setup does not require explicit alignment of different modal data sequences. The test consequences of each method under both MOSI and MOSEI datasets are shown in Tables 1 and 2. Compared with other baselines, the precision Acc2 and F1 values of the pattern in this paper are only lower than those of the MSAF pattern in the word-aligned setup. Compared with other baseline models in the non-aligned setting, the pattern in this paper performs optimally overall under four different evaluation criteria.

First, compared to the MSAF pattern that performs optimally in the word-based alignment setting on the CMU-MOSI dataset, this paper's model improves 4.11% in accuracy and 4.16% in F1 value. While on the CMU-MOSEI ensemble of communication, according to the MSAF pattern that performs optimally in the word alignment setting, this paper's model decreases by 0.55% in accuracy and 0.48% in F1 value, and it performs optimally in two evaluation metrics, MAE and Corr. The test consequences indicate that the sentiment analysis pattern incorporating the attention mechanism suggested in this article is able to better focus on the interactions between different modalities on important time steps.

Second, based on the unaligned setting, the proposed model maintains the same accuracy on the CMU-MOSEI ensemble of communication with a 0.13% improvement in the F1 value compared to the overall optimal performance of the weighted cross-modal attention mechanism model, as well as a 1.42% improvement in the Corr assessment metric. On the CMU-MOSEI dataset, this paper's model increased accuracy by 0.83% and F1 value by 0.78% compared to the overall best performing weighted cross-modal attention mechanism model. In the experiments, comparing this paper's model with other word-aligned and non-aligned baseline patterns, we were able to demonstrate that this paper's model has a smaller model performance gain compared to the baseline pattern in the non-aligned setting, and a significant pattern perform gain when comparing it to the baseline pattern in the word-aligned setting. This is due to the large number of parameters and high computational complexity of the direct cross-modal emotion analysis pattern in the non-aligned setting compared to the word-aligned setting.

Table 1: Results of emotion analysis in aligned/unaligned MOSI dataset

| Pattern | Acc-2/% | F1-Score/% | MAE | Corr |
|---|---|---|---|---|
| CMU-MOSI(Word Alignment) | | | | |
| EF-LSTM | 76.35 | 76.12 | 1.045 | 0.611 |
| LF-LSTM | 77.12 | 77.03 | 1.022 | 0.623 |
| MFM | 78.44 | 78.32 | 1.019 | 0.628 |
| RAVEN | 78.62 | 78.17 | 0.974 | 0.642 |
| HFFN | 79.25 | 79.13 | 0.937 | 0.653 |
| MCTN | 79.51 | 79.45 | 0.883 | 0.668 |
| MISA | 79.74 | 79.62 | 0.851 | 0.674 |
| MSAF | 80.27 | 80.13 | 0.819 | 0.703 |
| The proposed | 84.38 | 84.29 | 0.716 | 0.779 |
| CMU-MOSI(Word misaligned) | | | | |
| MulT | 80.46 | 80.18 | 0.863 | 0.634 |
| LMT-MULT | 82.53 | 82.37 | 0.996 | 0.668 |
| LMR-CBT | 81.94 | 80.88 | 0.827 | 0.641 |
| Self-MM | 79.38 | 79.11 | 0.794 | 0.712 |
| UR-Transformer | 80.92 | 80.04 | 0.735 | 0.701 |
| PMR | 81.48 | 81.22 | 0.758 | 0.608 |
| weightedcross-modal attention mechanism | 84.38 | 84.16 | 0.734 | 0.637 |
| The proposed | 84.38 | 84.29 | 0.716 | 0.779 |

Table 2: Results of emotion analysis in aligned/unaligned MOSEI dataset

| Pattern | Acc-2/% | F1-Score/% | MAE | Corr |
|---|---|---|---|---|
| CMU-MOSEI(Word Alignment) | | | | |
| EF-LSTM | 78.47 | 77.92 | 0.683 | 0.629 |
| LF-LSTM | 80.31 | 80.29 | 0.645 | 0.648 |
| MFM | 81.48 | 81.25 | 0.636 | 0.659 |
| RAVEN | 81.77 | 81.62 | 0.507 | 0.681 |
| HFFN | 82.69 | 82.51 | 0.628 | 0.693 |
| MCTN | 83.46 | 83.33 | 0.614 | 0.698 |
| MISA | 83.77 | 83.62 | 0.603 | 0.711 |
| MSAF | 84.83 | 84.67 | 0.574 | 0.729 |
| The proposed | 84.28 | 84.19 | 0.437 | 0.801 |
| CMU-MOSEI(Word misaligned) | | | | |
| MulT | 82.43 | 82.26 | 0.539 | 0.668 |
| LMT-MULT | 82.77 | 82.64 | 0.633 | 0.693 |
| LMR-CBT | 82.93 | 82.71 | 0.602 | 0.701 |
| Self-MM | 83.25 | 83.03 | 0.594 | 0.749 |
| UR-Transformer | 83.36 | 83.35 | 0.587 | 0.694 |
| PMR | 83.41 | 83.39 | 0.561 | 0.736 |
| weightedcross-modal attention mechanism | 83.45 | 83.41 | 0.508 | 0.728 |
| The proposed | 84.28 | 84.19 | 0.437 | 0.801 |

### *III. B. Analysis of application effects*

#### III. B. 1) Background of the study

DuringDuring COVID-19, society medium provided a way for the public to express their emotions and exchange opinions, as well as a large amount of data for researchers to conduct public sentiment analysis. Among them, Twitter, with more than 330 million monthly active users, is often used as a source of data for social sentiment studies such as hate speech. During the COVID period, social media recorded online events and public hate speech during the global pandemic, as well as the traces of anti-hate campaigns in real life, and social media data-based analysis of hate speech and anti-hate speech is of great practical significance and theoretical value.

Therefore, this paper conducts a study using the COVID-HATE ensemble of communication, an open-access ensemble of communication containing 206 million tweet IDs and corresponding discourse labels, including hate, neutral, and anti-hate. In this paper, we randomly sampled 6836 Twitter sample IDs, further collected social media

video information posted by relevant users, and analyzed the sentiment and dissemination trends of video content using the pattern suggested in this article.

**III. B. 2) Multi-label Sentiment Analysis**

The RNN-AM pattern suggested in this article is engaged in multi-label emotion recognition, and the video emotions include six kinds, which are: happiness, sadness, anger, fear, disgust, and surprise, and the correlation analysis results are shown in Fig. 6.

The depth of colors in the heatmap indicates the strength of the correlation between different categories of emotions, with darker colors representing higher correlations and lighter colors representing lower correlations. The correlation coefficient between the emotional category "sadness" and "disgust" was 0.098, indicating that there was a positive correlation between the two, that is, the probability of their co-occurrence was higher, and the correlation coefficient between the emotional category "happiness" and "disgust" was -0.15, indicating that there was a negative correlation between the two, that is, they were usually mutually exclusive and less likely to appear at the same time.

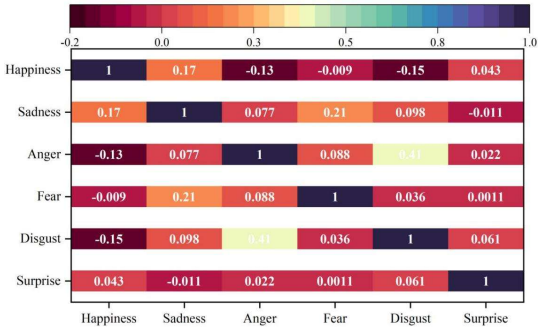
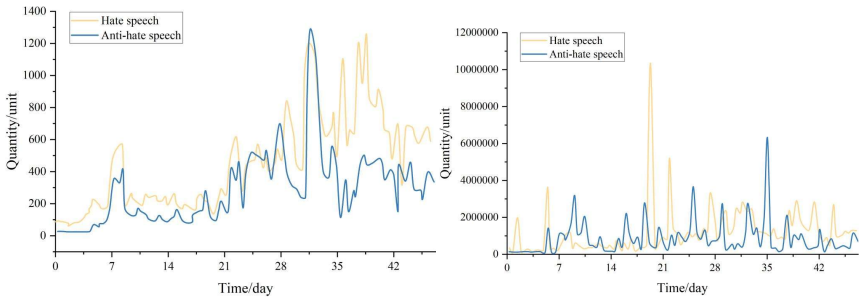



Figure 6: Results of emotional correlation analysis

**III. B. 3) Analysis of communication trends**

Using the GMM model to analyze the spatio-temporal trends of hate speech and anti-hate speech, the number of videos and the number of followers change values against as shown in Fig. 7 (a~b), from Fig. 7, some time series oriented patterns can be found.




(a)Number of videos (b)Number of fans

Figure 7: Comparison of the change value of the number of videos and followers

As shown in Figure 7(a), the average time interval from the video posting to the arrival of the peak, there are six significant peaks. When the interval is 48 hours, most of the peaks can be localized. Meanwhile, almost every high exposure value is followed by a video peak. The average time interval in between is about 7 days on the dataset.

Figure 7(b) shows the trend of video exposure. The degree of information exposure is mainly measured by the number of nodes that can receive this data. For example, in time, when a user with 1,000 followers posts a video, this video has an exposure of 1,000. And when 10 users with 5,000 followers post some videos, the exposure of all the videos equals 50,000 over time. Comparing the number of videos and the amount of exposure at the same time, a relationship was observed between peak video and peak exposure. The peaks caused by high exposure also vary at different stages, with steeper exposures implying a step change - in the dataset, there are two very high peaks of exposure, followed by the hate opinion moving to the next numerical stage.

## IV. Conclusion

In this paper, we design a social media video content sentiment and dissemination trend analysis model based on frequency domain feature extraction, and examine its effectiveness and application effect through experiments.

The algorithm in this paper can reach the optimal state in about 27 iterations, and its accuracy is 92.59%, which is 14.91% and 27.12% ahead of the PF algorithm and the MS algorithm, respectively; its optimal robustness is 12.64%, which is 20.45% and 37.89% less than the PF arithmetic and the MS arithmetic, respectively. Compared to the MSAF model, which performs optimally in the word-based alignment setting on the CMU-MOSI ensemble of communication, the RNN-AM pattern enhances 4.11% in accuracy and 4.16% in F1 value. Based on the non-aligned setting, the RNN-AM model maintained the same accuracy on the CMU-MOSEI ensemble of communication compared to the overall best-performing weighted cross-modal attentional machine-made pattern, with a 0.13% improvement in the F1 value and a 1.42% improvement in the Corr assessment metric. On the CMU-MOSEI dataset, this paper's model performed optimally overall on four different evaluation criteria.

On the COVID-HATE dataset, the correlation coefficient between the emotional category "sadness" and "disgust" is 0.098, indicating that there is a positive correlation between the two, that is, the probability of their simultaneous occurrence is higher, and the correlation coefficient between the emotional category "happiness" and "disgust" is -0.15, indicating that there is a negative correlation between the two, that is, they are usually mutually exclusive and less likely to appear at the same time. The average time interval from the release of the video to the peak is 48 hours, the average duration is 7 days, and the peaks caused by high impressions are also different at different stages, and steep exposures mean a stepwise change.

## Acknowledgements

This work was supported by “2024 Shanghai Municipal Key Course Construction Project for Universities and Colleges: AI+ Data Journalism”.

## References

[1] Zhu, H., Wei, H., & Wei, J. (2024). Understanding users’ information dissemination behaviors on Douyin, a short video mobile application in China. Multimedia Tools and Applications, 83(20), 58225-58243.

[2] Wu, H., & Cheng, M. (2022). Trust of information during the dissemination of popular science web videos in the new media era. Computational Intelligence and Neuroscience, 2022(1), 1746472.

[3] Oyighan, D., & Okwu, E. (2024). Social media for information dissemination in the digital era. RAY: International Journal of Multidisciplinary Studies, 10(1), 1-21.

[4] Zhu, X., & Luo, M. (2022). A Study on Short Video Marketing Dissemination for Rural Tourism Based on SIPS Model. Forest Chemicals Review, 1803-1810.

[5] Lan, Z. H. A. N. G. (2024). Study on factors influencing dissemination effect of WeChat short videos in popular science journals. Chinese Journal of Scientific and Technical Periodicals, 35(4), 466.

[6] Zhao, Z., & Ge, C. (2024). Live Streaming Interaction and Content Information Dissemination on TikTok Short Video Platform. IEIE Transactions on Smart Processing & Computing, 13(2), 113-119.

[7] Yang, B., Zhang, R., Cheng, X., & Zhao, C. (2023). Exploring information dissemination effect on social media: an empirical investigation. Personal and Ubiquitous Computing, 27(4), 1469-1482.

[8] Yin, H. (2024). From Virality to Engagement: Examining the Transformative Impact of Social Media, Short Video Platforms, and Live Streaming on Information Dissemination and Audience Behavior in the Digital Age. Advances in Social Behavior Research, 14, 10-14.

[9] Xia, Z. (2025). Research on short videos of secondary creation dissemination based on TPB theory. In Connecting Ideas, Cultures, and Communities (pp. 52-56). Routledge.

[10] Shonhe, L. (2017). A literature review of information dissemination techniques in the 21st century era. Library Philosophy and Practice (e-journal), 1731.

[11] Stappen, L., Baird, A., Cambria, E., & Schuller, B. W. (2021). Sentiment analysis and topic recognition in video transcriptions. IEEE Intelligent Systems, 36(2), 88-95.

[12] Nawaz, S., Rizwan, M., & Rafiq, M. (2019). Recommendation of effectiveness of YouTube video contents by qualitative sentiment analysis of its comments and replies. Pakistan Journal of Science, 71(4), 91.

[13] Deori, M., Kumar, V., & Verma, M. K. (2023). Analysis of YouTube video contents on Koha and DSpace, and sentiment analysis of viewers' comments. Library Hi Tech, 41(3), 711-728.

[14] Bozkurt, A. P., & Aras, I. (2021). Cleft lip and palate YouTube videos: content usefulness and sentiment analysis. The Cleft Palate-Craniofacial Journal, 58(3), 362-368.

[15] Li, Z., Li, R., & Jin, G. (2020). Sentiment analysis of danmaku videos based on naïve bayes and sentiment dictionary. IEEE Access, 8, 75073-75084.

[16] Al-Azani, S., & El-Alfy, E. S. M. (2020). Enhanced video analytics for sentiment analysis based on fusing textual, auditory and visual information. IEEE Access, 8, 136843-136857.

[17] Rao, A., Ahuja, A., Kansara, S., & Patel, V. (2021, February). Sentiment analysis on user-generated video, audio and text. In 2021 International Conference on Computing, Communication, and Intelligent Systems (ICCCIS) (pp. 24-28). IEEE.

[18] Xu, C., Liu, L., Jin, L., Du, G., Guo, Z., Zhao, Y., ... & Li, R. (2024). Infer Induced Sentiment of Comment Response to Video: A New Task, Dataset and Baseline. Advances in Neural Information Processing Systems, 37, 103737-103750.