# Deep Learning-Based Prediction of Polarization in Social Media Among Youth

**Xieer Jiang**
School of Journalism and Communication, Communication University of Zhejiang, Hangzhou 310018, Zhejiang, China
Corresponding author: Xieer Jiang (e-mail: jiangxieer@163.com)

**ABSTRACT** The stochastic nature of youth social media polarization poses challenges for traditional early-stage prediction. This study constructs a heterogeneous temporal network to represent multifaceted interactions among users, text, and hashtags, treating the evolution of online opinion as a spatiotemporal signal-propagation process in networked communication environments. Based on the DyGCN-GRU framework, the model captures the spatiotemporal propagation patterns of public-opinion evolution, integrates sentiment kurtosis, semantic variance, and community opposition density to generate polarization-intensity sequences, and finally employs a Transformer decoder to predict polarization risk. The network incorporates platform-specific embeddings, youth-oriented semantic features, and behavioral attributes to improve the detection of weak early signals. Experimental results on Weibo and Zhihu data show that the proposed method achieves an F1-score of 0.915, a root mean square error of 0.068, and an early-warning lead time of 11.3 minutes. The results verify the model's ability to quantitatively perceive and dynamically predict early risks of youth public-opinion polarization, and provide a technical basis for online risk monitoring and intelligent information-governance systems.

**INDEX TERMS** public opinion polarization,dynamic graph convolutional network,heterogeneous temporal network,signal propagation,transformer

## I. INTRODUCTION

SOCIAL media has become the core digital arena for young people to express social concerns and participate in public discussions. Platforms such as Weibo and Zhihu, with their instant interaction and algorithm recommendation mechanisms, have greatly lowered the threshold for information dissemination and opinion aggregation, but at the same time, they have also given rise to negative phenomena such as "information cocoons" and "echo chamber effects," causing the public opinion ecology to rapidly slide towards emotional extremes and opinion opposition under specific topics [1-2]. This polarization of public opinion not only distorts the rational basis of public discussion, but also may trigger online violence, group conflicts, and even offline social risks [3-4].

This article proposes a spatiotemporal joint prediction framework that integrates dynamic graph convolutional networks and gated recurrent units. The main contributions are as follows:

(1) A heterogeneous temporal network representation method for youth social media is constructed, which maps the three entities of users, text posts, and topic tags into dynamic heterogeneous composition snapshot sequences, preserving the semantic and structural information of diverse interactions.

(2) A multidimensional feature fusion strategy was designed, incorporating BERT semantic embedding and aggressive semantic intensity quantification, combined with user behavior attributes, to enhance the model's sensitivity to early polarized weak signals.

(3) A dual-mode architecture of DyGCN-GRU was proposed, which differentiated the modeling of neighborhood interaction weights through a multi-head attention mechanism, and utilized gated recurrent units to transfer polarization evolution memory across windows, achieving joint modeling of spatiotemporal propagation patterns.

(4) The polarization intensity sequence can be generated by integrating three interpretable indicators, namely emotional kurtosis, semantic variance, and community opposition density, and using Transformer decoder to predict polarization risk, which has been effectively validated on real datasets.

The organizational structure of this article is as follows: it reviews the relevant work on polarization analysis of social media public opinion and deep learning prediction methods; This article elaborates on the complete method flow of constructing heterogeneous temporal networks, extracting multidimensional features, dynamically modeling DyGCN-

GRU, and predicting Transformer decoders; This article validates the performance of the model through comparative experiments, ablation experiments, and sensitivity analysis, summarizes the entire text, and looks forward to future research directions.

## II. RELATED WORK

As social media gradually becomes the core field for emotional expression and viewpoint collision among young people, the problem of information fragmentation and algorithmic amplification driven polarization of public opinion is becoming increasingly severe [5]. Juditha [6] classified user generated content into emotional categories and analyzed the interaction patterns between different opinion groups. Javed and Javed [7] integrated relevant theories and empirical research from communication, political science, and computer science to construct a mechanism model of how algorithms influence public opinion. Prabayanti and Nabilah [8] explored how digital cultural practices reflect the specific emotions and demands of young Indonesians, and analyzed the phenomenon of public opinion polarization it has triggered. Ashraf et al. [9] explored how the design and dissemination mechanisms of social media platforms not only shape public opinion, but also promote the formation of political polarization. Siddique and Tariq [10] conducted a comparative analysis of the differences between social media and traditional media in shaping public opinion and their impact on the democratic process.

The massive unstructured data generated by social media contains rich information on public sentiment and market trends. Choudhary [11] used sentiment analysis algorithms based on natural language processing and machine learning to process, classify, and mine trends in text data from social media. Swastiningsih et al. [12] pointed out that digital media such as social media far exceed traditional media in terms of information dissemination speed, interactivity, and personalization, but they are also more prone to information fragmentation and emotionalization as a result. Petrova and Tapsoba [13] proposed combining media data analysis and public opinion monitoring for early warning and dynamic prediction of conflicts. The popularity of public opinion generated by social media platforms such as Weibo is highly dynamic and complex. Jiang et al. [14] proposed a Weibo public opinion popularity analysis and prediction model that integrates BERT model and X-means algorithm. Peng and Wang [15] fine tuned the GPT-3 model to generate text content that matches social platform features, and combined BERT and LSTM models to perform sentiment analysis and temporal modeling prediction on the generated text and real data. The existing work focuses on the identification or post attribution of polarization phenomena, lacking operable prediction mechanisms for future short-term windows. This article constructs a heterogeneous temporal network to dynamically represent the diverse interactions between users, text, and topic tags. The DyGCN-GRU dual-mode architecture is introduced to collaboratively capture spatiotemporal propagation patterns, and multidimensional polarization tendency indicators are integrated to achieve risk probability output for future windows.

## III. METHOD

This study collected 14 hot topics that sparked widespread discussions on Weibo and Zhihu platforms from January 2024 to June 2025, covering multiple fields such as social livelihood, cultural and entertainment hotspots, and technological controversies. The raw data was obtained through the platform's public API, and after deduplication, desensitization, and filtering of invalid accounts, a total of 286000 active users and 3.3 million valid interaction records (including original posts, comments, shares, and references) were retained. Each event is divided into continuous windows in chronological order, with the first 80% of the windows used for training, 10% for validation, and 10% for testing.

### A. CONSTRUCTION OF HETEROGENEOUS TEMPORAL NETWORKS AND WINDOW SEGMENTATION

In response to the instantaneous and high-frequency characteristics of youth social media interaction, the system performs graph mapping based on Weibo and Zhihu data. To mitigate cross-platform common-sense bias, we introduce a platform-specific embedding layer $E_p$ to account for the inherent differences in discourse styles (e.g., the emotional volatility of Weibo vs. the structured argumentation of Zhihu). This ensures that the heterogeneous temporal network preserves platform-specific interaction logic while modeling global polarization trends [16-17]. The network node set is defined as three entities: users, text posts, and topic tags, while the edge set covers forwarding, commenting, mentioning, and semantic similarity associations between texts. The definition of heterogeneous timing diagram structure is:

$$G_t = (V, \mathcal{E}_t, \phi, \psi). \tag{1}$$

$G_t$ is the heterogeneous timing diagram within the time window $t$; $V$ is a set of nodes; $\mathcal{E}_t$ is the set of edges within the time window $t$; $\phi : V \rightarrow \{\text{user}, \text{post}, \text{hashtag}\}$ is the node type mapping function; $\psi : \mathcal{E}_t \rightarrow \{\text{retweet}, \text{comment}, \text{mention}, \text{similarity}\}$ is an edge type mapping function. To capture weak signals during the budding stage of public opinion, a fixed-step sliding window with a granularity of $\Delta t = 5$ minutes is used to discretize the interaction flow. This granularity ensures that the "burstiness" of youth interactions is captured without over-smoothing, thereby providing a logical basis for the 11.3-minute lead time by allowing the model to detect shifts across at least two prior windows before a peak occurs. The time interval covered by the $k$th time window is:

$$W_k = [t_0 + (k-1)\Delta t,\ t_0 + k\Delta t), \quad \Delta t = 5 \text{ min}. \tag{2}$$

$t_0$ is the starting time; $\Delta t$ is the window step size. Within each time window, the establishment of interactive relationships follows strict temporal logic, which only preserves the evolutionary contributions of the current window and historical windows to the current state. For cross window long-range dependencies, maintaining a global node index table ensures consistency in the identification of the same user across different time windows. In the preprocessing stage, a deep first search algorithm is used to identify and eliminate meaningless zombie account nodes, retaining only active entities that have generated at least one effective interaction during the observation period [18].

Through the above method, the original social communication is transformed into a series of heterogeneous graph snapshots with spatiotemporal topological properties. Table 1 shows the statistical characteristics of 14 hot topic event data after window segmentation.

Figure 1 shows the evolution heatmap of the emotional polarity distribution of a hot topic event within a continuous time window. From the graph, it can be observed that in the early stages of the event outbreak (windows 1-6), emotional polarity exhibits a relatively dispersed distribution pattern; As the discussion deepens (window 7-15), positive and negative emotions gradually gather towards both ends, forming a bimodal pattern; From window 16-20, the expression of emotions in the middle position significantly decreased, and a polarization trend was initially formed. This evolutionary pattern provides a visual basis for subsequent quantification of polarization intensity [19].

### *B. MULTIDIMENSIONAL FEATURE EXTRACTION AND STATE MAPPING*

The feature extraction module is divided into two parallel branches: semantic feature vectorization and behavioral attribute discretization. To tailor the model to the youth demographic, the semantic branch integrates a custom lexicon of youth-specific internet slang and meme-based sentiment weights, which are often precursors to polarization in this group. Additionally, we incorporate “reply-to-post” ratios as a behavioral attribute to reflect the high-frequency interaction habits characteristic of younger users on Weibo and Zhihu. At the semantic level, the pre-trained BERT base model is called to embed 768 dimensional features into each interactive text, followed by the insertion of a Softmax regression layer with 128 hidden neurons to calculate the probability distribution of sentiment polarity [20-21]. The distribution is:

$$q = \text{Softmax}(W_s \cdot \text{BERT}(x) + b_s). \tag{3}$$

$q$ is the probability distribution vector of emotional polarity; $W_s$ is the fully connected layer weight matrix; $x$ is the input text; $b_s$ is the bias term. The intensity of aggressive semantics is quantified by trained irony and hate speech detectors, which output scalar values between 0 and 1, denoted as $s_{tox}$. At the level of behavioral characteristics, the system automatically calculates the user’s posting frequency $f$, average reply response delay $d$, and the structural density of the attention list $\kappa$ within the current window. After layer normalization, all features are fused into node initial state vectors through concatenation operators:

$$h_v^{(0)} = \text{Concat}\left(q, s_{tox}, \text{LayerNorm}([f, d, \kappa])\right). \tag{4}$$

$h_v^{(0)}$ is the initial state vector of node $v$; $f$ is the frequency of user posts; $d$ is the average reply response delay; $\kappa$ is the density of the attention list structure. This vector not only contains the instantaneous emotional intensity expressed by the youth group, but also implicitly expresses the level of user activity in the social relationship chain. This high-dimensional feature representation method ensures that the model can capture subtle psychological changes and social preferences of young users in specific topics, thereby enhancing the model’s sensitivity to early polarization risk perception [22-23].

### *C. DYGCN-GRU DYNAMIC INTERACTION PROPAGATION MODELING*

The core of dynamic evolutionary modeling lies in alternating the execution of spatial convolution and temporal cyclic updates. On the spatial dimension, the Dynamic Graph Convolutional Network (DyGCN) adopts a three-layer stacked structure, with each layer embedded with a multi-head attention mechanism. The number of attention heads is set to 8, and differential modeling of polarization propagation paths is achieved by calculating the attention weights of different interaction types (such as support, opposition, and reference) between neighboring nodes [24]. The attention coefficient of node $v$ to neighbor $u$ in layer $l$ is:

$$\alpha_{v,u}^{(l)} = \frac{\exp\left(\text{LeakyReLU}\left(\mathbf{a}^T\left[W^{(l)}h_v^{(l-1)} \| W^{(l)}h_u^{(l-1)}\right]\right)\right)}{\sum_{k\in N(v)}\exp\left(\text{LeakyReLU}\left(\mathbf{a}^T\left[W^{(l)}h_v^{(l-1)} \| W^{(l)}h_k^{(l-1)}\right]\right)\right)}. \tag{5}$$

$\alpha_{v,u}^{(l)}$ is the attention coefficient of node $v$ to neighboring node $u$ in the $l$-th layer; $\mathbf{a}$ is the learnable parameter vector of the attention mechanism; The superscript $T$ is transposed; $W^{(l)}$ is the weight matrix of the $l$-th layer; $h_v^{(l-1)}$ and $h_u^{(l-1)}$ are the eigenvectors of nodes $v$ and $u$ in the $l-1$st layer, respectively; $N(v)$ is the set of neighboring nodes of node $v$. The spatially aggregated features are fed into a dual layer gated recurrent unit (GRU), whose hidden layer dimension is consistent with the input vector, for transmitting and updating the temporal memory of nodes. The update mechanism of GRU is:

$$z_t = \sigma(W_z u_t + U_z s_{t-1} + b_z). \tag{6}$$

$z_t$ is the update gate vector; $\sigma$ is the sigmoid function; $W_z$ and $U_z$ are weight matrices; $u_t$ is the input of the current time step (i.e. the node features after spatial aggregation);

**TABLE 1.** Statistical Characteristics of Heterogeneous Time Series Diagrams after Sliding Window Segmentation

| Indicator | Mean | Std Dev | Min | Max |
|---|---|---|---|---|
| Number of nodes per window | 1,247 | 386 | 412 | 2,893 |
| Number of edges per window | 5,831 | 1,724 | 1,856 | 13,427 |
| Proportion of user nodes | 68.3% | 5.2% | 57.1% | 76.8% |
| Proportion of post nodes | 26.5% | 4.8% | 18.3% | 35.2% |
| Proportion of hashtag nodes | 5.2% | 1.3% | 2.6% | 7.9% |

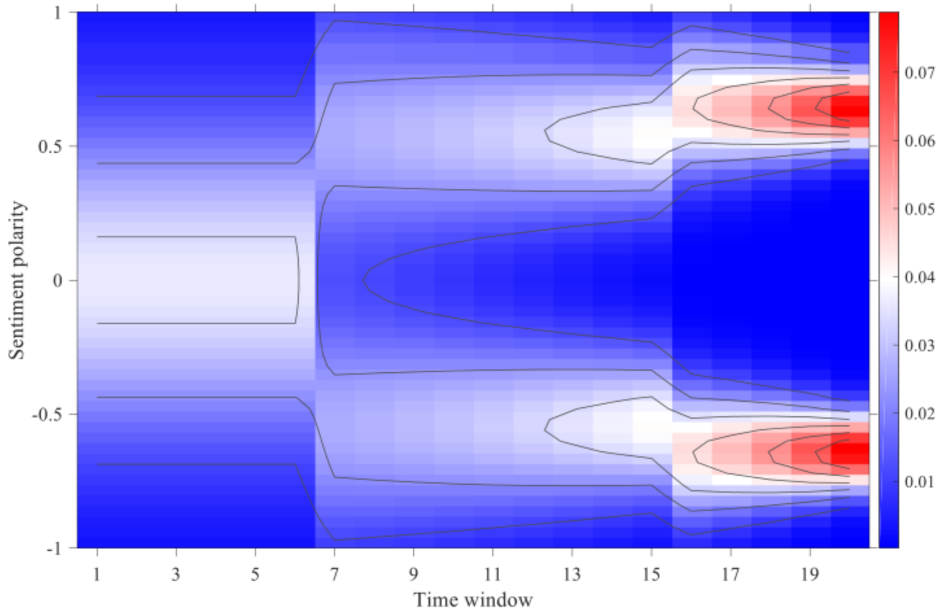


**FIGURE 1.** Evolution of emotional polarity distribution

$s_{t-1}$ is the hidden state of the previous time step; $b_z$ is the bias term.

$$r_t = \sigma(W_r u_t + U_r s_{t-1} + b_r). \tag{7}$$

$r_t$ is the reset gate vector; $W_r$ and $U_r$ are weight matrices; $b_r$ is the bias term.

$$\tilde{s}_t = \tanh(W_h u_t + U_h(r_t \odot s_{t-1}) + b_h). \tag{8}$$

$\tilde{s}_t$ is a candidate hidden state; $W_h$ and $U_h$ are weight matrices; $b_h$ is the bias term.

$$s_t = (1 - z_t) \odot s_{t-1} + z_t \odot \tilde{s}_t. \tag{9}$$

$s_t$ is the final hidden state of the current time step. The introduction of GRU enables the model to remember the polarization trend in the preceding window and smooth out the burst interactions in the current window. In each layer of computation, the dropout rate is set to 0.3 to prevent overfitting, and the learning rate is dynamically adjusted using the cosine annealing algorithm. This spatiotemporal collaborative architecture can accurately depict how young people form opinion clusters through interaction in a short period of time, and achieve exponential diffusion of polarized emotions in complex network structures [25-26]. Figure 2 shows the variation curves of training loss and validation loss with epochs.

The training loss and validation loss decrease rapidly in the first 15 epochs, followed by a slower and more stable decline rate; The validation accuracy steadily increased from the initial 51.2% to 91.5%, indicating that DyGCN-GRU can effectively learn polarization propagation patterns from time series diagrams and has good generalization ability.

### *D. POLARIZATION TENDENCY INDEX FUSION AND SEQUENCE GENERATION*

After obtaining the spatiotemporal feature representation of the nodes, the system quantifies the polarization state of the subgraph through three physical indicators. Firstly, calculate the kurtosis of the emotional distribution to measure whether opinions are highly concentrated towards extreme poles:

$$\text{Kurt} = \frac{\frac{1}{N}\sum_{i=1}^{N}(p_i - \bar{p})^4}{\left(\frac{1}{N}\sum_{i=1}^{N}(p_i - \bar{p})^2\right)^2} - 3. \tag{10}$$

$p_i$ is the sentiment polarity value of the $i$-th sample; $\bar{p}$ is

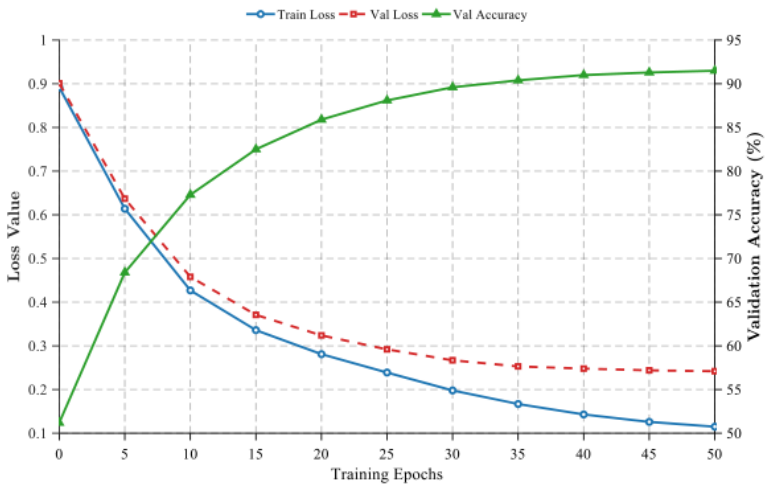


**FIGURE 2.** Training Process Performance

the mean of emotional polarity. Next, calculate the semantic variance ($Var_{sem}$) to reflect the degree of loss of viewpoint diversity. Within the context of youth discourse, $Var_{sem}$ is defined as the statistical dispersion of BERT-based semantic vectors representing youth subculture keywords and slang. A decreasing $Var_{sem}$ indicates "semantic homogenization," where diverse expressions collapse into repetitive polarized slogans, mathematically represented as the average squared Euclidean distance between node embeddings:

$$Var_{sem} = \frac{1}{M}\sum_{i<j}\left(\|h_i - h_j\|_2 - \bar{d}\right)^2. \tag{11}$$

$M$ is the number of node pairs; $h_i$ and $h_j$ are the eigenvectors of nodes $i$ and $j$, respectively; $\bar{d}$ is the average semantic distance between all nodes. Finally, potential opposing clusters are identified through community discovery algorithms, and the ratio of inter cluster edge density to intra cluster edge density is calculated:

$$R = \frac{\rho_{inter}}{\rho_{intra}} = \frac{|\mathcal{E}_{inter}|/(|C_{inter}|)}{|\mathcal{E}_{intra}|/(|C_{intra}|)}. \tag{12}$$

$R$ is the ratio of inter cluster edge density to intra cluster edge density; $\rho_{inter}$ and $\rho_{intra}$ are the inter cluster and intra cluster edge densities, respectively; $|\mathcal{E}_{inter}|$ is the number of edges between clusters; $|C_{inter}|$ is the number of possible connections between clusters; $|\mathcal{E}_{intra}|$ is the number of edges within the cluster; $|C_{intra}|$ is the number of possible connections within the cluster. These three indicators form a three-dimensional evaluation vector, which is then connected to a fully connected layer containing 256 neurons. Nonlinear fusion is performed using the ReLU activation function, and finally mapped to a continuous polarization intensity scalar $P_t$. As the time window slides, the system automatically generates a polarization intensity sequence $\{P_1, P_2, \ldots, P_T\}$. As a macroscopic representation of polarization evolution, this sequence effectively shields redundant noise from underlying interactions and compresses complex topological changes into dynamic curves with temporal correlations. This logical closed loop enhances the interpretability of polarization evaluation. Specifically, to distinguish structural polarization from transient intense debates, we define polarization as the convergence of high sentiment kurtosis (ideological extremity) and high community opposition density (structural fragmentation). Unlike general debates where viewpoints may overlap, our model triggers a polarization warning only when the semantic variance exceeds a sociologically-grounded threshold of 0.65, indicating a breakdown in cross-cluster communication. This significantly enhances the quantification accuracy of the state evaluation [27-28].

### E. TRANSFORMER PREDICTION DECODER AND WARNING OPTIMIZATION

The sequence prediction stage adopts a decoder model based on Transformer architecture. The model receives the polarization intensity sequence of the first 6 time windows (30 minutes in total) as encoding input, and uses self attention mechanism to capture the nonlinear fluctuation patterns inside the sequence [29-30]. The decoder directly outputs the polarization probability values $\hat{y}_{T+1:T+2}$ for the next 2 windows (i.e. the next 10 minutes) through causal mask control. To ensure objective supervision, we define "Polarization Risk" as a tri-level gradient rather than a binary state: Level I (Low, $P_t < 0.4$), representing normal debate; Level II (Moderate, $0.4 \leq P_t \leq 0.7$), signaling emerging echo chambers; and Level III (High, $P_t \geq 0.7$), defined as a "risk event" where community opposition density exceeds structural stability, necessitating algorithmic intervention. Samples were labeled based on these historical intensity thresholds to train the risk classifier. To solve the problem of uneven distribution of polarized samples, the loss function adopts weighted cross entropy, setting the penalty weight of moderately polarized samples to 1.5 times that of low polarized samples. Weighted cross entropy loss is defined as:

$$\mathscr{L}=-\frac{1}{N}\sum_{i=1}^{N} w_{y_i}\cdot\log\hat{y}_i \tag{13}$$

$\mathscr{L}$ is the value of the weighted cross entropy loss function; $w_{y_i}$ is the weight corresponding to the real category $y_i$; $\hat{y}_i$ is the probability that the $i$-th sample predicted by the model belongs to that category. The number of layers in the Transformer is set to 4, and the dimension of the forward feedback network is 512. During the training process, the Adam optimizer is used for parameter updates, with an initial learning rate set at 1e-4. This decoder no longer relies on complex regression assumptions, but learns the premonitory features of polarization bursts through the evolutionary trajectories of massive historical events, thereby providing high confidence warning prediction values during the golden window period before the polarization phenomenon is officially formed [31]. Table 2 presents the polarization intensity grading criteria and corresponding sample distribution and weight settings.

## IV. RESULTS AND DISCUSSION

### A. PREDICTIVE PERFORMANCE ANALYSIS

To verify the effectiveness of the DyGCN-GRU combined with Transformer decoder (hereinafter referred to as "this model") proposed in this article in predicting the polarization of youth social media public opinion, four baseline models were selected for comparative experiments: ARIMA (differential order=1, autoregressive order=2, moving average order=2), LSTM (hidden layer 128 dimensions, dropout rate 0.2, learning rate 1e-3), ST-GCN (spatial convolution layer 3, temporal convolution kernel size 3), and linear regression. All models were run under the same training/validation/test set partitioning, and the average of 5 rounds of cross validation was taken as the final result. Table 3 summarizes the quantitative performance of each model on the above indicators.

Compared to the suboptimal ST-GCN, the F1 score of this model improved by 16.6% (from 0.785 to 0.915), the warning advance increased by 2.8 minutes, and the MAE and RMSE decreased by 41.6% and 39.3%, respectively. Traditional time series models ARIMA and linear regression lack the ability to model the topology and nonlinear interactions of social networks, resulting in large prediction errors in the polarization intensity transition window and warning lead times generally less than 5 minutes. LSTM can capture long-term dependencies, but is not sensitive to graph structure information. Its F1 score and warning lead are between traditional models and graph neural network models. ST-GCN introduces spatial convolution to further improve performance, but has not yet been specifically designed for attention differentiation and emotional kurtosis in polarized propagation. This model effectively identifies the formation of early polarization clusters through the multi-head attention mechanism of DyGCN and dynamic updates of GRU, combined with polarization tendency fusion indicators, thereby achieving earlier and more accurate early warning.

This article extracted F1 score fine-grained data from different polarization levels of each model on the test set, as shown in Figure 3.

Figure 3 reveals a common pattern: the F1 score of all models in high polarization scenarios is lower than that in low polarization scenarios, indicating that the deeper the polarization degree, the greater the difficulty of prediction. However, this model achieved 0.851 under difficult high polarization conditions, while ST-GCN was only 0.672, indicating that the proposed method has stronger robustness to extreme polarization states. In addition, the linear regression model's F1 score dropped to 0.412 under high polarization, almost losing its predictive ability, once again confirming that relying solely on linear assumptions is not sufficient for the nonlinear evolution prediction task of public opinion.

### B. ABLATION EXPERIMENT

On the test set, DyGCN was removed sequentially (replaced with a regular GCN), BERT features (replaced with TF IDF), multi-head attention (replaced with average pooling), and GRU (only static GCN was retained). The training was conducted for 50 epochs with a fixed window step size of 5 minutes. The performance of each variant was recorded as shown in Table 4 below.

The complete model significantly outperforms all ablation variants in all indicators. After removing DyGCN, F1 score decreased by 10.3% and warning advance decreased by 2.6 minutes, indicating that dynamic graph convolution is crucial for capturing polarization propagation paths. The removal of BERT features resulted in the largest performance degradation (a 13.2% decrease in F1 score), indicating that semantic embedding is essential for early recognition of sentiment polarity. After removing the multi-head attention mechanism, the model still retains some spatial aggregation ability, but the F1 score drops to 0.858, confirming that differentiated neighborhood weight allocation can effectively suppress noise interference. After retaining only static GCN and removing GRU, the model lost cross window temporal memory, and the warning advance was reduced to 7.2 minutes, further demonstrating the core role of temporal cycle updates in predicting polarization evolution. Each module collaborates with each other to support the superior performance of the model in high accuracy and early warning.

### C. SENSITIVITY ANALYSIS

This article sets three different sliding window granularities: 1 min, 5 min (default), and 15 min. Other hyperparameters (DyGCN GRU structure, BERT features, multi-head attention, etc.) can be fixed and subjected to 5 rounds of cross validation on the test set. Table 5 summarizes the key performance indicators (mean ± standard deviation) of the model on the test set at different time granularities.

**TABLE 2.** Polarization Intensity Grading and Loss Weight

| Polarization Level | Intensity Range | Sample Proportion | Loss Weight |
|---|---|---|---|
| Low Polarization | $[0, 0.4)$ | 58.3% | 1.0 |
| Moderate Polarization | $[0.4, 0.7)$ | 34.1% | 1.5 |
| High Polarization | $[0.7, 1.0]$ | 7.6% | 2.0 |

**TABLE 3.** Performance Comparison of Different Models

| Model | F1-score | MAE | RMSE | Early Warning Lead Time (min) |
|---|---|---|---|---|
| ARIMA | 0.624 | 0.157 | 0.203 | 4.2 |
| LSTM | 0.713 | 0.112 | 0.146 | 6.8 |
| ST-GCN | 0.785 | 0.089 | 0.112 | 8.5 |
| Linear Regression | 0.587 | 0.178 | 0.231 | 3.1 |
| Proposed | 0.915 | 0.052 | 0.068 | 11.3 |

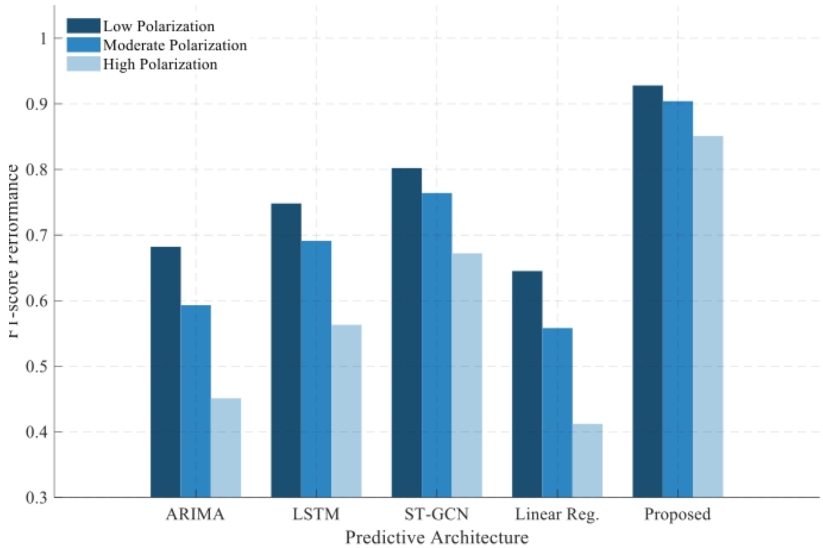


**FIGURE 3.** F1 score at different polarization levels

**TABLE 4.** Ablation Experiment Comparison Results

| Model Variant | F1-score | MAE | RMSE | Early Warning Lead Time (min) |
|---|---|---|---|---|
| Proposed (full) | 0.915 | 0.052 | 0.068 | 11.3 |
| w/o DyGCN (replace with GCN) | 0.821 | 0.097 | 0.126 | 8.7 |
| w/o BERT features (use TF-IDF) | 0.794 | 0.112 | 0.143 | 7.9 |
| w/o Multi-head Attention (avg pooling) | 0.858 | 0.078 | 0.101 | 9.5 |
| w/o GRU (static GCN only) | 0.802 | 0.105 | 0.134 | 7.2 |

**TABLE 5.** Model performance at different time granularities

| Time Granularity | F1-score | MAE | RMSE | Early Warning Lead Time (min) |
|---|---|---|---|---|
| 1 min | $0.928 \pm 0.012$ | $0.047 \pm 0.005$ | $0.061 \pm 0.004$ | $8.6 \pm 0.8$ |
| 5 min | $0.915 \pm 0.009$ | $0.052 \pm 0.004$ | $0.068 \pm 0.005$ | $11.3 \pm 0.6$ |
| 15 min | $0.874 \pm 0.015$ | $0.073 \pm 0.007$ | $0.094 \pm 0.008$ | $10.5 \pm 0.9$ |

From Table 5, it can be seen that the 1-minute granularity achieved the highest F1 score (0.928) and the lowest

MAE/RMSE, but its warning advance was significantly lower than that of the 5-minute granularity (only 8.6 minutes). This is because the overly narrow time window captures a large amount of instantaneous interactive noise. Although the model has higher fitting accuracy for polarization states, it needs to accumulate enough continuous windows to form a stable polarization trend judgment, resulting in a relatively late warning triggering time. The F1 score of 5-minute granularity is slightly lower than that of 1-minute granularity (0.915 vs 0.928), but the warning advance reaches 11.3 minutes, about 2.7 minutes earlier than that of 1-minute granularity, while maintaining a low level of error (MAE=0.052). Although the 15 minute granularity reduces computational overhead, due to the coarse window, the key transient features of polarization germination are lost, resulting in a decrease in F1 score to 0.874, a significant increase in MAE/RMSE, and a reduction in early warning. Taking into account both accuracy and timeliness, 5 minutes is the optimal time granularity.

## V. CONCLUSION

This article proposes an end-to-end prediction framework that integrates heterogeneous temporal networks, DyGCN-GRU dynamic propagation modeling, and Transformer decoders to address the early prediction challenges of polarization in youth social media public opinion. This framework achieves high-precision early warning of polarization risks. However, this study still has certain limitations: The data sources are limited to Weibo and Zhihu platforms, and may overlook polarization patterns in other social media platforms; The computational complexity of the model is high, and real-time deployment still needs to be optimized. Future work can explore cross platform heterogeneous data fusion and lightweight graph neural network architecture, and introduce causal inference methods to reveal the deep driving mechanisms of polarization bursts, further enhancing the interpretability and generalization ability of early warning.

## AUTHOR CONTRIBUTIONS

Xieer Jiang designed, collected and analyzed the data, and drafted the manuscript. Xieer Jiang conducted the study, critically revised the manuscript for important intellectual content, and gave final approval of the version to be published. Xieer Jiang participated fully in the work, take public responsibility for appropriate portions of the content, and agreed to be accountable for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved.

## CONFLICTS OF INTEREST

The authors declare no conflict of interest.

## FUNDING

This research received no external funding.

## ACKNOWLEDGEMENTS

Not applicable.

## REFERENCES

[1] M. S. Alam, M. S. H. Mrida, and M. A. Rahman, “Sentiment analysis in social media: How data science impacts public opinion knowledge integrates natural language processing (NLP) with artificial intelligence (AI),” American Journal of Scholarly Research and Innovation, vol. 4, no. 01, pp. 63-100, 2025, doi: 10.63125/r3sq6p80.

[2] M. Mudhofi, I. Supena, A. Karim, et al., “Public opinion analysis for moderate religious: Social media data mining approach,” Jurnal Ilmu Dakwah, vol. 43, no. 1, pp. 1-27, 2023, doi: 10.21580/jid.v43.1.16101.

[3] P. L. Kharvi, “Understanding the impact of AI-generated deepfakes on public opinion, political discourse, and personal security in social media,” IEEE Security & Privacy, vol. 22, no. 4, pp. 115-122, 2024, doi: 10.1109/MSEC.2024.3405963.

[4] S. Arifin and J. A. Rojak, “Social Media and Public Opinion Formation: Information Dissemination and Polarization,” Journal of Social Science Studies, vol. 2, no. 1, pp. 139-142, 2022.

[5] W. Mengke and L. Gang, “Research on the construction of quantitative index of microblog group polarization in university public opinion events,” The Journal of Engineering, vol. 2022, no. 3, pp. 285-294, 2022, doi: 10.1049/tje2.12113.

[6] C. Juditha, “The Phenomenon of ‘Indonesia Gelap’ on Social Media: Sentiment Analysis and Public Opinion Polarization,” Jurnal Komunikasi, vol. 17, no. 1, pp. 157-170, 2025, doi: 10.24912/jk.v17i1.33846.

[7] U. Javed and U. Javed, “The influence of social media algorithms on political polarization and public opinion,” Online Media and Society, vol. 4, no. 2, pp. 44-52, 2023, doi: 10.71016/oms/2ffw9391.

[8] H. R. Prabayanti and P. S. Nabilah, “Digital Cultural Practices and Public Opinion Polarization in the #KaburAjaDulu Narrative on Social Media,” International Journal of Communication, Media, and Development Studies, vol. 3, no. 1, pp. 39-47, 2026.

[9] A. Ashraf, K. S. Gondal, and S. Ahmed, “Social Media and Political Polarization: Investigating the Role of Online Platforms in Shaping Public Opinion,” Journal of Development and Social Sciences, vol. 6, no. 1, pp. 705-717, 2025, doi: 10.47205/jdss.2025(6-I)61.

[10] M. Siddique and A. Tariq, “The Impact of Social Media on Political Polarization and Public Opinion in Democratic Societies,” Social Thought and Policy Review, vol. 3, no. 2, pp. 1-23, 2025.

[11] P. R. Choudhary, “Sentiment Analysis of Social Media Data for Public Opinion and Market Trend Prediction,” International Journal of Research Publications in Engineering, Technology and Management (IJRPETM), vol. 8, no. 2, pp. 11768-11771, 2025, doi: 10.15662/IJRPETM.2025.0802002.

[12] S. Swastiningsih, A. Aziz, and Y. Dharta, “The role of social media in shaping public opinion: a comparative analysis of traditional vs. digital media platforms,” The Journal of Academic Science, vol. 1, no. 6, pp. 620-626, 2024, doi: 10.59613/fm1dpm66.

[13] M. Petrova and A. Tapsoba, “Information and conflict: From the role of (social) media and public opinion to big data and forecasting,” Economic Policy, vol. 40, no. 124, pp. 845-878, 2025, doi: 10.1093/epolic/eiaf012.

[14] Z. Jiang, X. Li, S. Zhang, and X. Zhao, “A Weibo public opinion heat analysis and prediction model integrating BERT and X-means algorithms,” Computer Applications, vol. 45, no. 10, pp. 3138-3145, 2025, doi: 10.11772/j.issn.1001-9081.2024091371.

[15] Y. Peng and M. Wang, “Automatic Content Generation and Public Opinion Analysis of Intelligent Agents in Social Networks,” Yangtze Information & Communication, vol. 38, no. 12, pp. 141-143+147, 2025.

[16] D. Wang, F. Liu, and W. Lu, “Cross-Social Media Public Opinion Risk Perception: Construction and Implementation of a Theoretical Framework,” Journal of Information Science, vol. 43, no. 4, pp. 446-456, 2024, doi: 10.3772/j.issn.1000-0135.2024.04.006.

[17] A. Bany Mohammed, M. Al-Okaily, D. Qasim, et al., “Digital activism and public opinion: understanding the role of social media during the Gaza Conflict,” Journal of Islamic Marketing, vol. 16, no. 11, pp. 3366-3393, 2025, doi: 10.1108/JIMA-03-2024-0101.

[18] M. Hasanuddin, S. Khodijah, and C. A. Rizki, “Analysis of the Role of Social Media in Shaping Public Opinion on Social Issues,” Journal of Computer Science Artificial Intelligence and Communications, vol. 1, no. 1, pp. 7-11, 2024, doi: 10.64803/jocsaic.v1i1.1.

[19] E. J. Hasibuan, A. D. R. Putra, and A. S. Dirgantari, “The role of social media algorithms in shaping public opinion during political campaigns,” International Journal of Social and Human, vol. 1, no. 2, pp. 165-172, 2024.

[20] G. Alexander, “Role of social media influencers in shaping public opinion and consumer behavior in Greece,” International Journal of Communication and Public Relation, vol. 9, no. 1, pp. 13-26, 2024, doi: 10.47604/ijcpr.2269.

[21] M. Reveilhac, S. Steinmetz, and D. Morselli, “A systematic literature review of how and whether social media data can complement traditional survey data to study public opinion,” Multimedia Tools and Applications, vol. 81, no. 7, pp. 10107-10142, 2022, doi: 10.1007/s11042-022-12101-0.

[22] M. Zimmer and S. Logan, “Privacy concerns with using public data for suicide risk prediction algorithms: a public opinion survey of contextual appropriateness,” Journal of Information, Communication and Ethics in Society, vol. 20, no. 2, pp. 257-272, 2022, doi: 10.1108/JICES-08-2021-0086.

[23] D. M. Houston and A. Barone, “How the engagement of high-profile partisan officials affects education politics, public opinion, and polarization,” Education Finance and Policy, vol. 21, no. 1, pp. 97-120, 2026, doi: 10.26300/gw47-cc09.

[24] E. Holder and C. X. Bearfield, “Polarizing political polls: How visualization design choices can shape public opinion and increase political polarization,” IEEE Transactions on Visualization and Computer Graphics, vol. 30, no. 1, pp. 1446-1456, 2023, doi: 10.1109/TVCG.2023.3326512.

[25] T. Manju, A. K. Tarofder, and F. Azam, “Does Gender Moderate the Mediating Effects of Public Opinion on Relationships Between Media Polarization and Voter Intention? A Simple Question and a Complex Answer,” Revista de Gestao Social e Ambiental, vol. 18, no. 6, pp. 1-34, 2024, doi: 10.24857/rgsa.v18n6-020.

[26] S. Nasereddin, “Impact of social media platforms on international public opinion during the Israel war on Gaza,” Global Change, Peace & Security, vol. 35, no. 1, pp. 5-31, 2023, doi: 10.1080/14781158.2024.2415908.

[27] A. Afyare and M. A. H. Orey, “The Influence of Social Media on Political Discourse and Public Opinion,” Architecture Image Studies, vol. 6, no. 3, pp. 1634-1667, 2025, doi: 10.62754/ais.v6i3.506.

[28] Y. Sairambay, A. Kamza, Y. Kap, et al., “Monitoring public electoral sentiment through online comments in the news media: a comparative study of the 2019 and 2022 presidential elections in Kazakhstan,” Media Asia, vol. 51, no. 1, pp. 33-61, 2024, doi: 10.1080/01296612.2023.2229162.

[29] D. Zhang, “Data mining and big data in social media public opinion monitoring,” Media and Communication Research, vol. 4, no. 10, pp. 44-51, 2023, doi: 10.23977/mediacr.2023.041007.

[30] X. Chen, S. Duan, S. Li, et al., “A method of network public opinion prediction based on the model of grey forecasting and hybrid fuzzy neural network,” Neural Computing and Applications, vol. 35, no. 35, pp. 24681-24700, 2023, doi: 10.1007/s00521-023-08205-9.

[31] J. Zhang and Z. Li, “Prediction and Guidance of Negative Public Opinion Dissemination Based on a Sentiment Classification Algorithm,” Journal of ICT Standardization, vol. 13, no. 3, pp. 243-256, 2025, doi: 10.13052/jicts2245-800X.1331.