ISSN 2706-655X Vol. 8, Issue 2: 30-36, DOI: 10.25236/IJFET.2026.080205

Design and Implementation of a Deep Learning-Based Intelligent Analysis and Early Warning System for Weibo Social Hot Topics

Xia Jiayue[1,a], Du Shiping[1,b,*], Lang Yuxin[1,a], Ma Yuechen[1,a], Zhang Yu[1,a], Fan Yiqi[1,a]

[1]University of Science and Technology Liaoning, Anshan, China
[a]2677399204@qq.com, [b]1034182681@qq.com
**Corresponding author*

Abstract: *To address the complex nonlinear characteristics of information dissemination in Weibo social hot topics, as well as the shortcomings of traditional monitoring methods—such as insufficient semantic understanding, experience-dependent feature engineering, and lagging warnings—this paper designs and implements a learning-based intelligent analysis and early warning system. First, a deep semantic representation model for Weibo text is constructed based on BERT to handle issues of short, noisy, and contextually complex text. Second, a hotspot identification and sentiment evolution analysis method is developed by integrating HDBSCAN density clustering with BiLSTM-Attention. Third, a quantitative hotspot calculation model combining user features, content features, and dissemination features is proposed, along with a four-level early warning mechanism. Finally, the effectiveness and practicality of the system are verified using real Weibo data. This system provides technical support for online public opinion governance.*

Keywords: *Deep learning; Weibo public opinion; Hot topic analysis; Early warning system; BERT; BiLSTM-Attention*

1. Introduction

As one of China's most influential social media platforms, Weibo boasts over 250 million daily active users, making it a core arena for the generation, fermentation, and dissemination of social hot topics. Unlike traditional media, Weibo information dissemination is characterized by decentralization, instantaneous fission, and emotion-driven nature; an ordinary post can evolve into a social hot topic sweeping the entire internet within hours. Typical cases such as the "Hebei barbecue restaurant assault incident" in 2023 and the "Xiaomi SU7 car accident incident" in 2024 demonstrate that the evolution of Weibo public opinion has high uncertainty and complex nonlinear characteristics. How to identify potential hot topics in real time from massive amounts of Weibo data, accurately grasp the evolution of public sentiment, and establish a scientific early warning mechanism has become a key issue that urgently needs to be addressed in the field of social governance[1-2].

Traditional public opinion monitoring methods mainly rely on keyword matching and simple statistical indicators, such as threshold triggers for reposts, comments, and likes. These methods suffer from three inherent drawbacks: first, insufficient semantic understanding, failing to recognize implicit expressions and sarcastic language; second, feature engineering relying on human experience, making it difficult to capture high-dimensional nonlinear features; and third, lagging early warning mechanisms, often only identifying issues after a public opinion crisis has erupted. In recent years, breakthroughs in deep learning technology have provided new solutions to these problems. Models such as BERT (Bidirectional Encoder Representations from Transformers), BiLSTM (Bi-directional Long Short-Term Memory), and graph attention networks have demonstrated superior performance in tasks such as text representation, sentiment analysis, and rumor detection[3-4].

The research objective of this paper is to design and implement an end-to-end intelligent analysis and early warning system for Weibo social hot topics. Specifically, the main contributions of this study include: (1) constructing a deep semantic representation model for Weibo text based on BERT to solve the representation problems of short, noisy, and complex contexts in Weibo text; (2) designing a hotspot identification and sentiment evolution analysis method that integrates HDBSCAN density

ISSN 2706-655X Vol. 8, Issue 2: 30-36, DOI: 10.25236/IJFET.2026.080205

clustering and BiLSTM-Attention sentiment analysis[5]; (3) proposing a hotspot calculation model that comprehensively considers user characteristics, content characteristics, and dissemination characteristics, as well as a four-level early warning mechanism; and (4) verifying the effectiveness and practicality of the system based on real Weibo data[6].

2. Overview of Related Technologies

2.1 Application of Deep Learning in Text Representation

Text representation is a fundamental task of natural language processing, and its core goal is to convert unstructured text into a vector form that can be processed by computers. Early representation methods include one-hot encoding, bag-of-words model, and TF-IDF. These methods suffer from the curse of dimensionality and semantic gap. The proposal of static word vector models such as Word2Vec and GloVe has achieved low-dimensional embedding of semantic space, but it cannot solve the problem of polysemy[7-8].

In 2018, Google's BERT model completely changed the technological paradigm of natural language processing. BERT adopts the Transformer bidirectional encoder structure and, through two pre-training tasks---Masked Language Model (MLM) and Next Sentence Prediction (NSP)---can capture the dynamic semantics of words in context. In Weibo sentiment analysis, BERT can effectively handle complex linguistic phenomena such as new internet slang, emoticons, and ironic expressions. Research by Zhang Shih-hao et al. shows that the BXpre model, which integrates BERT for Weibo sentiment analysis, achieves a 90.88% prediction relevance rate, a 12.71 percentage point improvement over the traditional LSTM method[9].

2.2 Weibo Hot Topic Identification Methods

Hot topic identification aims to automatically discover important events that are forming or have already erupted from massive Weibo streams. Traditional methods are mainly based on TF-IDF vectorization and LDA topic models, clustering by calculating text similarity. The limitation of these methods is that they ignore text semantic similarity and require a pre-defined number of topics[10].

Density-based clustering algorithms show unique advantages in topic identification. HDBSCAN (Hierarchical Density-Based Spatial Clustering of Applications with Noise), as an improved version of DBSCAN, can automatically determine the number of clusters, identify clusters of arbitrary shapes, and classify noise points as unclassified data. He Yuanzheng et al.'s research, using the "Xiaomi SU7 car accident incident" as the research object, employed BERT semantic representation and HDBSCAN clustering algorithm to perform multi-dimensional feature segmentation of Weibo public opinion events, achieving good results.

2.3 Weibo Sentiment Analysis Technology

Sentiment analysis is used to determine the emotional tendency expressed by Weibo text, usually divided into three categories: positive, negative, and neutral. In public opinion monitoring, sentiment analysis can not only reflect the public's attitude towards events, but also predict the evolution of public opinion.

The BiLSTM-Attention model is currently a commonly used architecture for sentiment analysis. BiLSTM captures the contextual dependencies of text through two LSTM layers, forward and backward, while the Attention mechanism assigns weights to different words, highlighting the influence of sentiment words. Cheng Zhecheng's research uses a Bi-LSTM sentiment analysis method that integrates an improved attention mechanism to achieve real-time public opinion monitoring of trending topics on Weibo. Graph Attention Network (GAT) is used to learn the deep semantic features of Weibo comments, achieving an accuracy of 92.6% after integrating user features.

2.4 Current Status of Public Opinion Early Warning Mechanism Research

The core of public opinion early warning lies in constructing a scientific heat evaluation index system and early warning level classification standards. Existing research mainly constructs heat indicators from three dimensions: content features (text length, keyword strength), user features (number of followers, authentication type, historical behavior), and dissemination features (time series data of forwarding, commenting, and liking). Cheng Weidong et al. proposed an early prediction model of public opinion heat based on Bi-GRU, achieving an accuracy rate of 92.41%, and found a strong correlation between content heat and the heat of users' historically published content through ablation experiments. Lin Shaofu et al.'s patent proposed a sentiment early warning method using XGBoost multi-feature fusion, calculating public opinion early warning values from the perspective of multi-dimensional features such as forwarding, commenting, liking, and sentiment polarity.

3. Overall System Design

3.1 System Architecture Design

The Weibo Social Hotspot Intelligent Analysis and Early Warning System designed in this paper adopts a layered architecture, which is divided into a data acquisition layer, a data storage layer, an algorithm analysis layer, and a business application layer from bottom to top. The system architecture is shown in Figure 1.

The data acquisition layer builds a distributed crawler system based on the Scrapy framework, acquiring data through two methods: Weibo open API and simulated web page requests. To address Weibo's anti-crawling mechanisms, the system maintains a dynamic IP proxy pool and a cookie pool, and designs an adaptive request frequency control algorithm.

The data storage layer adopts a hybrid storage architecture. Raw Weibo data (posts, comments, user information) is stored in an Elasticsearch cluster, utilizing its inverted index mechanism to support full-text retrieval; structured feature data is stored in MySQL; vectorized representation results are stored in a Redis cache for real-time analysis.

The algorithm analysis layer is the core of the system, which includes a text representation module, a topic clustering module, a sentiment analysis module, and a heat calculation module. Each module is deployed in a microservice manner and communicates asynchronously through a message queue.

The business application layer provides a visual interface, early warning push, and decision support services to different user groups.

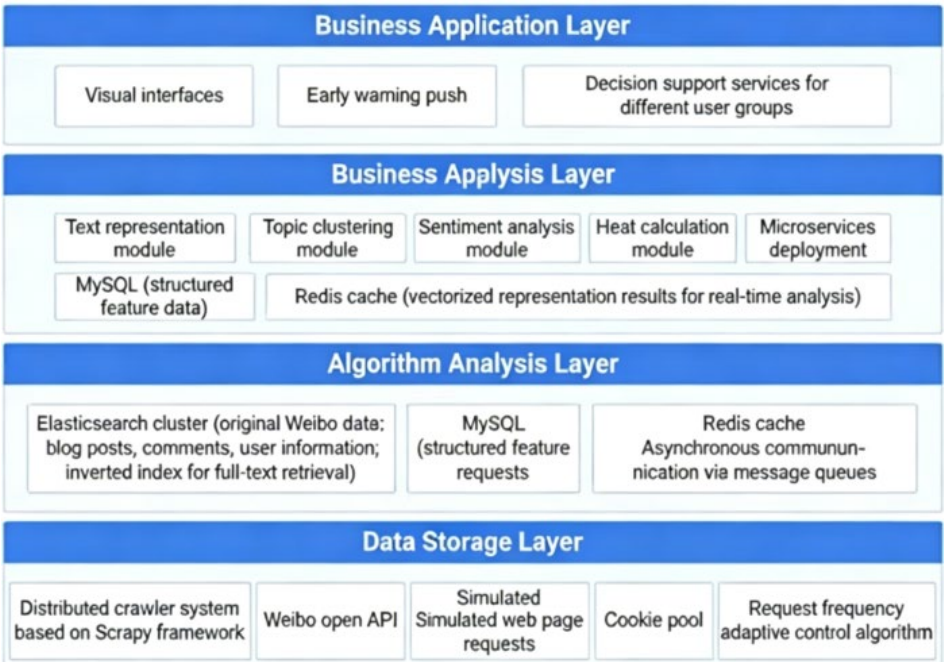

Figure 1: Overall System Architecture

3.2 System Functional Module Division

The system is divided into six core functional modules, and the functional descriptions of each module are shown in Table 1.

ISSN 2706-655X Vol. 8, Issue 2: 30-36, DOI: 10.25236/IJFET.2026.080205

Table 1: Core Function Modules

Module Name	Main Functions	Key Technologies
Data Acquisition Module	Weibo Posts, Comments, User Information Acquisition	Scrapy distributed crawler, IP proxy pool
Data Preprocessing Module	Text Cleaning, Word Segmentation, Stop Word Removal, Emoji Conversion	Jieba word segmentation, regular expressions, Expression conversion
Text Representation Module	Weibo Text Vectorization, Semantic Representation	BERT pre-trained model, Fine-tuning
Hotspot Identification Module	Automatic Discovery and Tracking of Hot Topics, Classification and Topic Extraction	HDBSCAN density clustering, Category and theme extraction
Emotion analysis module	Comment Sentiment Tendency Classification, Evolutionary Analysis	BiLSTM-Attention, Sentiment Dictionary
Early Warning and Judgment Module	Popularity Calculation, Ranking, Early Warning Push	Multi-feature Fusion, XGBoost classification

4. Implementation of Core System Modules

4.1 BERT-based Weibo text representation module

Weibo texts are characterized by their short length (usually no more than 140 characters), high noise levels (including @users, #topics#, and URL links), and strong contextual dependence, posing special requirements for text representation. This paper uses the BERT-base-Chinese pre-trained model as the basic architecture for text representation.

In the data preprocessing stage, the original Weibo text is first cleaned: HTML tags and URL links are removed, "@username" is replaced with the general tag "[USER]", and emoticons are converted into corresponding text descriptions. A custom dictionary is constructed to assist in word segmentation for common new internet words and abbreviations in Weibo.

The specific process of text representation is as follows: the preprocessed text is input into the BERT model, passing through a 12-layer Transformer encoder to obtain a 768-dimensional context representation for each token. The output vector corresponding to the [CLS] tag is taken as the semantic representation of the entire Weibo post. To address the large volume of Weibo comment data, model distillation is employed to compress the representation dimension, improving computational efficiency while maintaining accuracy.

Experiments compared the performance of Word2Vec, BERT, and RoBERTa representation methods on Weibo sentiment classification. The results showed that BERT-base achieved an F1 score of 89.7%, significantly outperforming Word2Vec's 76.3%, validating the advantages of pre-trained models in Weibo text representation.

4.2 Hot Topic Recognition Module Based on HDBSCAN

The goal of hot topic identification is to automatically discover new topics from the continuously flowing Weibo data stream and track the evolution of topics. This paper adopts the strategy of "sliding window + incremental clustering" to achieve real-time topic detection.

First, Weibo data is sliced into segments with a 5-minute time window, and all posts within the window are represented using BERT vectorization. Then, the HDBSCAN algorithm is used to perform density clustering on the vectors. The core advantage of HDBSCAN is that it can automatically identify clusters of arbitrary shapes without pre-setting the number of clusters and effectively filter out noise points (posts that do not belong to any topic).

For each identified cluster, a combination of TF-IDF and TextRank is used to extract topic words. Specifically, the high-frequency words of all posts within the cluster are counted, and the TF-IDF value is calculated by comparing it with the background corpus to select candidate words with high discriminative power; then, the co-occurrence relationship between words is calculated using the TextRank algorithm, and finally 3-5 topic words are determined as topic tags.

ISSN 2706-655X Vol. 8, Issue 2: 30-36, DOI: 10.25236/IJFET.2026.080205

In the topic tracking stage, cosine similarity is used to calculate the semantic similarity between new window topics and historical topics. When the similarity exceeds the threshold (experimentally determined to be 0.65), it is judged as a continuation of the same topic; otherwise, it is considered as a new topic. Table 2 shows examples of hot topics identified by the system within a certain time window.

Table 2: Example of Hotspot Identification Results

Topic ID	**Keywords**	**Number of blog posts**	**Number of participating users**	**Start time**
T001	Xiaomi SU7, car accident, Autonomous driving	2345	1892	2024-04-15 14:30
T002	May Day holiday, holiday arrangements, complaints	5678	4321	2024-04-15 15:45
T003	Huawei P70, press conference, chips	1890	1567	2024-04-15 16:20

4.3 Sentiment Analysis Module Based on BiLSTM-Attention

The sentiment analysis module is used to determine the sentiment tendency of Weibo comments and track the evolution of sentiment over time. The structure of the BiLSTM-Attention model designed in this paper is as follows:

The input layer receives a sequence of text vectors represented by BERT; the BiLSTM layer consists of two LSTM networks, forward and backward, which capture the past and future contextual information of the text, respectively, and output the hidden state at each time step; the Attention layer calculates the attention weights at each time step to highlight the importance of sentiment words; the fully connected layer and the Softmax layer output the sentiment category probability.

The model training uses the cross-entropy loss function, the optimizer is Adam, and the learning rate is set to 0.001. The training data comes from 50,000 manually annotated Weibo comments, with annotation categories of positive, negative, and neutral. The Kappa value for annotation consistency test is 0.83.

Experimental results show that the BiLSTM-Attention model achieves an accuracy of 91.2% on the test set, which is higher than the 86.5% of the standard LSTM and 88.7% of TextCNN. Figure 2 shows the curve of the proportion of negative sentiment in a hot topic event over time. It can be seen that the negative sentiment reaches its peak 2 hours after the event breaks out, and then gradually declines.

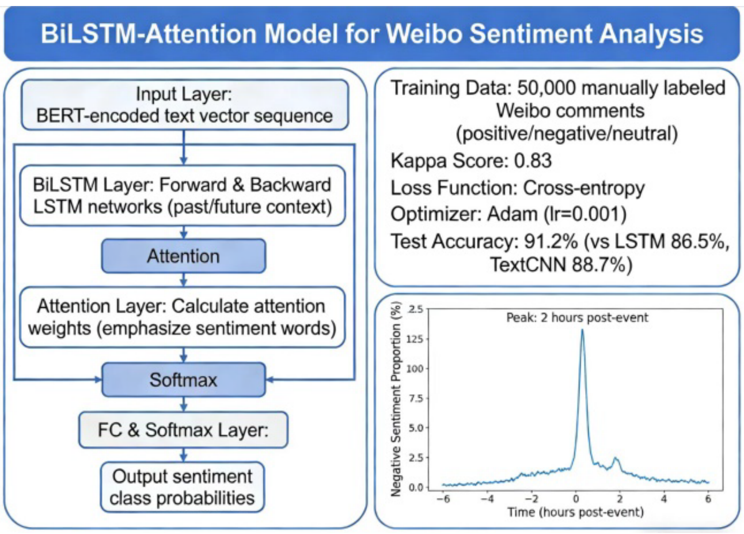

Figure 2: Emotional evolution curve of hot events

4.4 Integration of multi-feature heat calculation and early warning module

Public opinion heat calculation is the foundation of the early warning mechanism. The heat index (HI) proposed in this paper comprehensively considers content characteristics, user characteristics, and dissemination characteristics. The calculation formula is:

ISSN 2706-655X Vol. 8, Issue 2: 30-36, DOI: 10.25236/IJFET.2026.080205

$$HI = \alpha \cdot C_{content} + \beta \cdot C_{user} + \gamma \cdot C_{propagation}$$

Among them, content heat ($C_{content}$) is calculated based on the semantic distance between the blog post and the topic center; user heat (C_{user}) considers factors such as the number of blogger's fans, certification type, and influence of historical posts; dissemination heat ($C_{propagation}$) integrates the temporal variation characteristics of forwards, comments, and likes, and draws on the prediction method based on the second derivative of prior popularity in the BXpre model. The weight coefficients α, β, and γ are determined by the analytic hierarchy process (AHP) and are 0.3, 0.3, and 0.4, respectively.

Based on the heat calculation, this paper establishes a four-level early warning mechanism, and the early warning level classification standard is shown in Table 3.

Table 3: Classification Criteria for Public Opinion Early Warning Levels

Early warning level	Heat index range	Sentiment tendency	Response suggestions
Level 1 (Blue)	[0, 30)	No obvious tendency	Continue to observe
Level 2 (Yellow)	[30, 60)	Neutral or slightly negative	Pay attention to the situation
Level 3 (Orange)	[60, 85)	Obviously negative	Initiate consultation
Level 4 (Red)	[85, 100]	Strongly negative and rapidly rising	Immediate action

The early warning trigger adopts a dual mechanism: when the heat index exceeds the threshold and the proportion of negative sentiment exceeds 50%, the system automatically generates an early warning report and pushes it to relevant management personnel through three methods: SMS, email, and system message.

5. Experiment and Evaluation

5.1 Experimental Data and Environment

The experimental data comes from real data collected from Weibo's public API from January to June 2024, covering various topics such as society, entertainment, and technology. The dataset contains 1.2 million original blog posts, 8.5 million comments, and 450,000 user information entries. Twenty public events that have become social hotspots were selected as test cases, including "Hebei Yanjiao Explosion", "Xiaomi SU7 Car Accident", and "May Day Holiday Adjustment Controversy".

The experimental environment was configured as follows: Intel Xeon Gold 6248R processor, 256GB RAM, NVIDIA A100 GPU with 40GB VRAM, Ubuntu 20.04 operating system, PyTorch 1.12 deep learning framework, Elasticsearch 7.15 database, and MySQL 8.0 database.

5.2 System Performance Evaluation

The system performance evaluation is conducted from three dimensions: hotspot identification accuracy, sentiment classification accuracy, and early warning timeliness.

Hotspot identification accuracy is evaluated using precision, recall, and F1 score. Baseline methods are LDA topic model and K-means clustering. Experimental results are shown in Table 4.

4: Performance Comparison of Hotspot Identification

Method	Accuracy (%)	Recall (%)	F1 value (%)
LDA + K-means	67.3	71.2	69.2
Word2Vec + DBSCAN	78.6	75.4	77.0
BERT + HDBSCAN (This article)	89.2	86.7	87.9

The confusion matrix of the sentiment classification module on the test set shows that the accuracy of positive sentiment classification is 92.3%, the accuracy of negative sentiment classification is 90.8%, and the accuracy of neutral sentiment classification is 89.5%. The proportion of negative sentiment misclassified as positive sentiment is only 3.2%, indicating that the model can effectively identify critical comments.

The timeliness evaluation of early warning adopts the indicator of "advance time", which is the difference between the system warning time and the peak time of public opinion. In 20 test cases, the average early warning time of the system was 47 minutes, with 7 cases exceeding 1 hour in advance. Compared with the keyword threshold method (with an average lead time of 22 minutes), the timeliness of early warning has been significantly improved.

5.3 Case Analysis

This article analyzes the "Xiaomi SU7 car accident incident" in April 2024. The incident occurred at 14:23 on April 15th. The system first detected related Weibo posts (12 in total) at 14:35, with a popularity index of 12.3, triggering a blue alert. At 15:10, the number of related posts increased to 234, the popularity index rose to 42.7, and the negative sentiment ratio was 58%, triggering an orange alert. At 15:50, the topic became a trending topic on Weibo, with the number of posts exceeding 2000, the popularity index reaching 78.5, and the negative sentiment ratio reaching 82%. The system automatically generated a public opinion analysis report and pushed it to the regulatory authorities. Post-event review showed that the system's alert time was about 45 minutes earlier than manual monitoring.

6. Conclusion

This paper designs and implements a deep learning-based intelligent analysis and early warning system for social hot topics on Weibo. The system uses the BERT model to realize the deep semantic representation of Weibo text, automatically identifies hot topics through HDBSCAN density clustering, uses BiLSTM-Attention network for sentiment classification, and constructs a heat calculation model that integrates multiple features and a four-level early warning mechanism. Experimental results show that the system has achieved good results in hot topic identification, sentiment analysis and early warning timeliness, and can provide effective technical support for network public opinion supervision.

Acknowledgements

Liaoning University of Science and Technology Undergraduate Innovation and Entrepreneurship Project in 2026

References

[1] CHEN T T, XU R F, HE Y L, et al. A sentiment analysis method for microblog based on BERT and BiLSTM-Attention[C]//Proceedings of 2020 IEEE International Conference on Progress in Informatics and Computing (PIC). Shanghai: IEEE, 2020: 50-55.

[2] LI Y Z, XIE Y C, WANG X F, et al. Early warning of public opinion on social networks based on deep learning[J]. IEEE Access, 2020, 8: 2098-2009.

[3] Zhang Shih-hao, Li Xin, Zhao Xin-yang, et al. Weibo public opinion heat analysis and prediction model integrating BERT and X-means algorithm [J]. Computer Applications, 2025, 45(10): 3138-3145.

[4] He Yuanzheng, You Huanxiang, Liu Pingyu, et al. Evolution mechanism of Weibo hot topics and public opinion guidance strategy under sudden public events - based on deep learning and sentiment analysis [R]. National Undergraduate Innovation Training Program, 2025.

[5] Lin Shaofu, Ren Dongliang. Sentiment early warning method based on public figure public opinion knowledge graph and XGBoost multi-feature fusion: CN112905800B [P]. 2025-01-14.

[6] Sarfraz A, Ahmad A, Zeshan F. Synews: a synergy-based rumor verification system[J]. Social Network Analysis and Mining, 2024, 14: 57.

[7] Cheng Zhecheng. Research and implementation of Weibo public opinion monitoring model based on deep learning [J]. Computer Era, 2023(11): 124-126,130.

[8] Zhu Y, Wang G, Huang X, et al. Enhancing real-time rumor detection on Weibo through user and content feature integration[J]. Computing and Informatics, 2025, 44(5): 1123-1140.

[9] Cheng Weidong. Early prediction and empirical study of network public opinion information heat based on deep learning [J]. Information Science, 2024, 42(2): 148-156.

[10] Li Haojie. Multimodal Brand Tracking and Event Detection on Social Media [R]. Dalian University of Technology, 2025.