Improved SIFRANK for Efficient Media Hotspot Mining in Social Networks

Jun Zhang[1], Yuke Cai[2, *]
[1]School of Design Chongqing College of Finance and Economics Yongchuan 402160, China
[2]School of Tourism and Health Chongqing City Vocational College Yongchuan 402160, China
E-mail: zhangjun1001@outlook.com, caiyuke1993@163.com
[*]Corresponding author

Keywords: SIFRANK algorithm, social network, media hotspot, effect optimization

Received: October 23, 2024

In the era of information explosion, social media has become the main platform for the public to obtain information and express their opinions. How to quickly and accurately mine media hotspots from massive data has become an urgent problem to be solved. With the rapid development of social media, media hotspot mining technology is facing higher requirements. This study focuses on improving the SIFRANK algorithm and proposes a more efficient and accurate method for mining social media hotspots. By deeply mining the emotional tendencies and interaction patterns of social media users, as well as introducing information timeliness evaluation and optimizing network weight calculation, the improved SIFRANK algorithm significantly improves its performance in hotspot recognition. Tested on the Twitter dataset, the improved algorithm achieved a 15% increase in accuracy in identifying hot topics, reaching a 92% accuracy rate (compared to the baseline method of 77%), and was able to respond more quickly to newly emerging hot events. In dealing with complex network structures and changes in information propagation speed, the algorithm has also shown stronger adaptability and robustness, with a 5% improvement compared to traditional models such as PageRank. This study, through technological innovation, not only improves the efficiency and accuracy of hotspot identification, but also provides a powerful tool for understanding social public opinion trends and guiding public policy formulation.

Povzetek: Izboljšan algoritem SIFRANK izboljšuje rudarjenje na vročih točkah družbenih medijev z optimiziranjem izračunov teže omrežja in vrednotenjem pravočasnosti, dosega 92-odstotno natančnost – presega tradicionalne modele in izboljšuje prilagodljivost v dinamičnih informacijskih okoljih.

1 Introduction

In the era of information explosion, social media has become the leading platform for the public to obtain information, express their opinions and participate in social interactions. In the massive information flow, how to quickly and accurately identify and mine media hotspots is an urgent need in news dissemination and public opinion monitoring and a frontier topic in academic research and technology development [1, 2]. As the focus topic of public attention at a specific time, media hotspots are often influenced by complex social, political and economic factors. They are also closely related to the information dissemination mode and network structure characteristics [3]. Therefore, accurately capturing the dynamics of media hotspots and evaluating their influence from complicated social media data is of great significance for understanding social public opinion trends, guiding public policy formulation and optimizing information dissemination strategies.

Traditional hot spot mining methods, such as keyword frequency statistics and topic model analysis, can identify hot topics to a certain extent but often ignore the network effect of information dissemination and the influence difference of different information sources [4, 5]. In recent years, with the development of network science and extensive data analysis technology, hotspot mining algorithms based on network structure, such as the SIFRANK algorithm, have gradually become research hotspots because they can effectively consider the propagation path and influence of information in the network [6]. SIFRANK algorithm, by simulating the propagation process of information in the network, combining the centrality of nodes and the novelty of information, quantitatively evaluates the importance and influence of information and provides a new perspective and method for mining media hotspots [7].

However, the original SIFRANK algorithm also has some limitations in practical applications, such as sensitivity to network structure, estimation bias of information propagation speed, and insufficient prediction of hot spot duration [8, 9]. Therefore, this paper studies the effect of social network media hotspot mining based on the improved SIFRANK algorithm, aiming at improving the accuracy and timeliness of media hotspot identification through algorithm optimization and model innovation. This not only includes the improvement of the core mechanism of the algorithm, such as introducing more refined information timeliness evaluation, optimizing the calculation method of network weight, and enhancing the adaptability of the algorithm to complex network structures but also involves the in-depth mining and intelligent analysis of social media data, such as using natural language processing technology to improve the analysis accuracy of text information and using machine learning methods to improve the accuracy of hot spot prediction. Considering that complex emotional and public opinion dynamics often accompany the formation of media hotspots, the improved SIFRANK algorithm should also be able to effectively capture and analyze social media users' emotional tendencies and interaction patterns, providing a richer perspective for comprehensively understanding the social influence of hot topics. Through interdisciplinary integration research, combined with theories and methods in sociology, psychology, information science and other fields, we will further deepen our understanding of the generation mechanism of media hotspots and provide powerful theoretical support and technical tools for optimizing social media information dissemination strategies, improving public media literacy and promoting social harmony and stability.

2 Research on extraction algorithm of massive short text hot words in social network media

2.1 SIFRANK algorithm

Traditional key phrase extraction relies on statistics, grammar, or knowledge graphs. The pre-trained model introduces a new method, SIFRANK (Sentence-Intermediate Framework Rank), combined with ELMo (Embeddings from Language Models) to realize dynamic Sentence Embedding and Phrase Embedding, multidimensional improvement of extraction quality. However, the reasoning speed of ELMo is limited under big data, and SIFRANK is limited in extracting hot words, common words and new words [10, 11]. By optimizing the SIFRANK algorithm, the extraction effect of massive hot words is enhanced. Traditional keyword extraction models are limited by external knowledge, and the emergence of pre-trained language models provides new solutions [12]. SIFRANK, which combines SIF (Smooth Inverse Frequency) sentence embedding and ELMo pre-training model, efficiently extracts short text keywords without supervision [13].

As an utterly unsupervised sentence vector generation technology based on a weighted average, SIF is a highly concise and efficient sentence vector generation strategy [14, 15]. The starting point of this method is to use Embedding technology to convert vocabulary into word vectors. Standard implementation methods include Word2vec and Fasttext. Compared with the traditional sentence vector generation method, this method is unique in that sentence vectors are constructed by assigning corresponding weights to each word vector and performing the weighted average operation. Specifically, the weight of each word vector follows Equation (1).

$$w_i = \frac{exp(\ f_i\ )}{\sum_{j=1}^{n} exp(\ f_j\ )} \quad (1)$$

Where w_i represents the weight of the i-th word, f_i is the function value that measures the importance of the i-th word in the sentence, and n represents the number of words contained in the sentence. This process ensures that when generating sentence vectors, the contribution of each word to sentence structure and semantics can be fully reflected.

$$Embedding_{sentence} = \sum_{i}^{n} Weight_i * Embedding_{word_i} \quad (2)$$

The sentence vector is obtained through the weighting method, which is expressed by formula (2). $Weight_i$ is the weight value, $Embedding_{word_i}$ is the word vector, and $Embedding_{sentence}$ is the sentence vector. By constructing a sentence vector Matrix, the dot product operation is performed between each sentence vector and the first principal vector in the Matrix. Then, its projection value on the principal vector is subtracted from the original sentence vector to realize the elimination of "common parts" between word vectors and highlight each word vector's unique features [16]. The SIF method shows significant advantages in text similarity evaluation, especially without the need for complex supervised learning models, and its performance surpasses some complex architectures based on RNN and LSTM [17, 18]. This method is suitable for calculating various pre-trained word vectors. It can generate sentence vectors on various data sets so that it can be effectively applied in different test environments. In addition, the SIF scheme shows high robustness, can maintain good performance even if word frequency information from different corpora is used, and can achieve the optimal effect by adjusting the parameter range [19].

Figure 1 depicts the architecture of the SIFRANK model. This model is based on ELMo, which is pre-trained for large-scale text and extracts word vectors. Then, a sentence vector is generated by the SIF method. The final step involves calculating the cosine similarity between the sentence vector and the word vector of each word within the sentence, thereby quantifying the importance of each word in the sentence.

Text preprocessing includes cleaning, word segmentation, part-of-speech tagging, and conversion into an array of words with part-of-speech [20]. Segment and merge noun phrases to ensure semantic accuracy. The ELMo model converts words into word vectors, SIF weighting is used to obtain sentence vectors, and key phrases are extracted using cosine similarity.

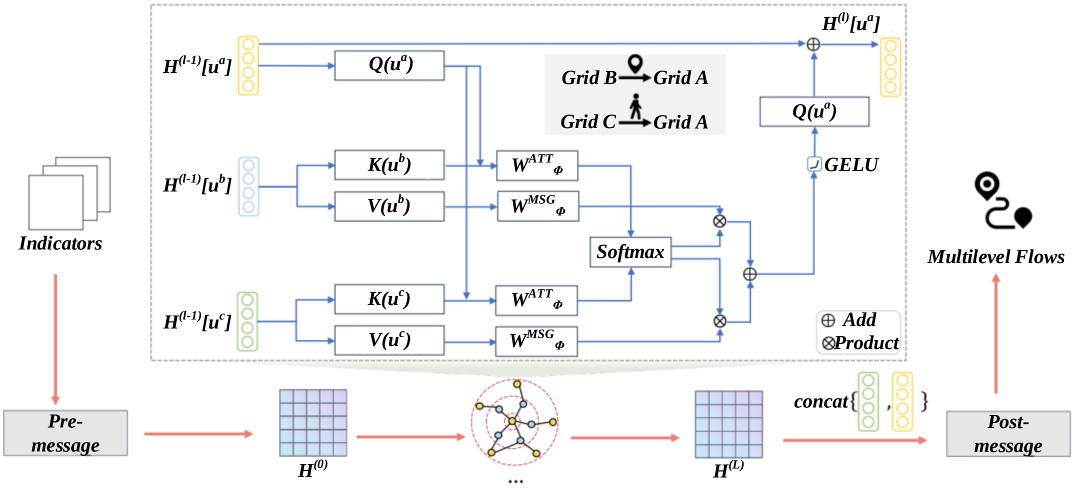


Figure 1: SIFRANK model

WordEmbedding technology captures semantic relationships by learning vector representations of words. However, before the advent of ELMo, Word embedding was static and could not adapt to the challenges of polysemy and context dependence [21, 22]. The introduction of ELMo, with the help of a deep bidirectional language model, realizes the generation of word vectors closely related to the context, greatly enhancing the efficiency of natural language processing tasks. In this model architecture, the bottom LSTM unit is used to capture the grammatical characteristics of vocabulary. At the same time, the high-level LSTM is dedicated to extracting the semantic connotation of vocabulary. In the forward part of the bidirectional model, for N markers $(t_1, t_2, \ldots, t_N)$ in the sequence, the probability p of the occurrence of the k-th position marker is evaluated by calculating based on the sequence of the first k-1 position markers, as shown in Equation (3).

$$p(t_1,t_2,\ldots,t_N)=\prod_{k=1}^{N}p(t_k\,|\,t_1,t_2,\ldots,t_{k-1}) \tag{3}$$

The calculation process of the backward model is similar to that of the forward model, and the calculation steps are shown in Equation (4).

$$p(t_1,t_2,\ldots,t_N)=\prod_{k=1}^{N}p(t_k\,|\,t_{k+1},t_{k+2},\ldots,t_N) \tag{4}$$

$$\sum_{k=1}^{N} log\ p(t_1,t_2,\ldots,t_N;\Theta_x,\overline{\Theta}_{LSTM},\Theta_s)+log\ p(t_1,t_2,\ldots,t_N;\Theta_x,\vec{\Theta}_{LSTM},\Theta_s) \tag{5}$$

The goal in the training process of BiLSTM is to maximize, as shown in Equation (5), where the symbol Θ represents the angle. Specifically, when ELMo processes each tag t_k, it generates $2L+1$ representations by constructing an L-layer Long Short-Term Memory (LSTM) architecture. The process description is shown in formula (6).

$$R_k=\{x_k^{LSTM},\overline{h}_{k,j}^{LSTM},\vec{h}_{k,j}^{LSTM}\,|\,k=1,2,\ldots,L\}=\{h_{k,j}^{LSTM}\,|\,k=0,1,2,\ldots,L\} \tag{6}$$

Where x_k^{LSTM} is the result of direct CNN encoding of token, $h_{k,0}^{LSTM}$ stands for x_k^{LSTM}, $h_{k,j}^{LSTM}=[\overline{h}_{k,j}^{LSTM},\vec{h}_{k,j}^{LSTM}]$. ELMo integrates the output of each layer of the bidirectional long-short-term memory network (BiLSTM) through linear combination to form a vector. This process can be expressed as a formula (7). Where $E(R_w)$ is a linear combination vector, the layer normalization process of each layer of BiLSTM is realized by introducing a scaling factor r. The setting of r is used to adjust the parameter scale between BiLSTM layers and optimize the network's learning efficiency and generalization ability. The weight s_j represents the coefficient that plays a decisive role in the linear combination process, directly affecting the final vector's composition and properties.

$$E(R_w)=r\sum_{j=0}^{L}s_j h_{k,j}^{LSTM} \tag{7}$$

2.2 Hot word mining using improved SIFRANK algorithm

SIFRank algorithm adopts an inverse word frequency smoothing strategy, which aims to reduce the weight of high-frequency words and improve the extraction efficiency of keywords in short texts. This algorithm combines the power of the ELMo pre-trained model to obtain word vectors and sentence vectors with broad applicability [23]. By integrating the two-way long-term and short-term memory network (LSTM) mechanism, ELMo effectively responds to the challenge of polysemy. Compared with TFIDF, YAKE, TEXTRANK and other methods, ELMo shows more significant advantages in multiple evaluation indicators.

BERT's core position in natural language processing cannot be ignored, and its bidirectional Transformer architecture endows it with excellent context-understanding capabilities [24]. Within the framework of the SIFRANK algorithm, BERT is applied to generate high-precision word vector representations, aiming to speed up and optimize the keyword extraction process. By pre-training the deep semantics of the learning language and fine-tuning it on specific tasks, BERT exhibits significant performance advantages in various natural language processing tasks [25, 26]. For each character in the input text, BERT uses the Self-Attention mechanism (Self-Attention) to obtain the enhanced semantic vector. Query, Key, and Value are derived from the original text content in this process. In order to enhance the diversity of expression, a multi-head self-attention mechanism is introduced, which allows the enhancement vectors of characters to be explored in different semantic spaces, and the final vector is formed through linear combination to capture the key features of diversity.

To sum up, the BERT model, which is pre-trained to learn the deep semantics of a language, can generate high-quality word vector representations after fine-tuning specific tasks. In the SIFRANK algorithm, the fine-tuned BERT model is used as a key tool to construct word vector dictionaries, aiming to improve the efficiency and effectiveness of keyword extraction significantly. The model architecture includes a Transformer Encoder, which enhances feature expression through Multi-head Self-Attention and uses residual connection, layer normalization, and linear transformation to improve model performance.

This study aims to improve the accuracy and real-time performance of social network media hotspot mining by improving the SIFRANK algorithm. In proposing research questions, the focus is on how to optimize algorithms to more accurately identify media hotspots in social networks, and how to maintain algorithm stability and efficiency in complex and changing user interaction patterns. In terms of research hypotheses, reasonable inferences can be made based on the following points: firstly, it is assumed that the distribution of social network data has a certain degree of regularity, which can be learned and utilized through algorithms to improve the accuracy of hotspot mining; Secondly, assuming that although user interaction patterns are complex and varied, they contain certain patterns and information that are crucial for hotspot mining; Finally, assuming that through reasonable algorithm design and optimization, effective extraction and utilization of these patterns and information can be achieved, thereby improving the performance of the algorithm.

The improved SIFRANK algorithm is mainly reflected in the following three points: new word discovery, improvement in real-time processing ability, and optimization of hot word mining. A new word discovery algorithm extracts unique terms of social networks in advance to improve word segmentation accuracy. In real-time processing, combining Bert and Word2vec, the speed of word vector generation is optimized; Hot word mining captures hot words more accurately by modifying the weight formula.

Information entropy, as a quantitative index to evaluate the richness of left and right collocation of words, is essentially to measure the total amount and uncertainty of information. Specifically, the higher the information entropy value, the richer the information carried and the higher the uncertainty. Its mathematical expression (8) is as follows, where *H(X)* represents the information entropy of the random variable *X*; P_{x_i} is the probability of occurrence of event x_i; *n* is the number of all possible events; *b* is the base of probability, and the base is the natural logarithm.

$$H(X) = -\sum_{i=1}^{n} P(x_i) log_b P(x_i) \tag{8}$$

The degree of cohesion within a word is highly reflected in the significant co-occurrence characteristics among the words that make up the word rather than an arbitrary combination. In language, some high-frequency words, such as "de" and "shi" in Chinese, form high-frequency co-occurrence with many other words because of their widespread use, so this phenomenon should be avoided. In order to quantitatively analyze the aggregation of character combinations, the mutual information between points is used as a statistical index, as shown in formula (9).

$$PMI = p(x,y) log \frac{p(x,y)}{p(x)p(y)} \tag{9}$$

When constructing the language model, we define the co-occurrence probability *p (x, y)* of words *x*, *y* and the frequency *p (x)* of word *x* and reveal the relationship between them through mutual information analysis: the higher the co-occurrence frequency of word combination, the greater the corresponding point PMI (Pointwise mutual information) value. However, the high frequency of a single word may dilute the evaluation accuracy of PMI, so careful consideration is needed. In order to identify potential new words, we need to comprehensively evaluate the left and proper information entropy and internal cohesion of candidate words and set a scoring index score for this purpose. This score not only needs to consider the absolute value of the difference between the left information entropy *LE* (left entropy) and the right information entropy *RE* (right entropy) of a single word, that is, LE-RE, to measure the possibility of word formation, but also needs to combine the context relevance of the word itself and the consistency of the internal structure. Hence, to more accurately judge whether the candidate word meets the standard of new words. Construct the statistic as in Equation (10):

$$L(w) = log \frac{LE + RE + 1e - 8}{|LE - RE| + 1e - 8} \tag{10}$$

To avoid the risk of the denominator returning to zero, the tiny value of 1e − 8 is introduced. Because the calculation result of mutual information between two points may be biased due to the difference in candidate word length, that is, the PMI value under longer candidate words tends to show a higher probability in order to accurately reflect the internal aggregation degree of words, the average mutual information between points is used as the evaluation index, as shown in formula (11). $c_1, ..., c_n$ is the word vector, *W* is the word combination result, and *n* is the number of word vectors.

$$AMI = \frac{1}{n} log \frac{p(W)}{p(c_1)p(c_2)\dots p(c_n)} \tag{11}$$

$$score = \alpha L(w) + \beta AMI \tag{12}$$

The calculated score is shown in formula (12), where α and β are artificially set coefficients to control the cohesion of words and the importance of left and right information entropy. By calculating the size of the score $score_W$ of the word combination W and the score $score_{W_1} + score_{W_2}$ of the sub-words constituting the word, if $score_w > score_{W_1} + score_{W_2}$, the word combination W is considered to be an undiscovered new word, and W is added to the custom new word dictionary, which is added in the subsequent word segmentation task. The Bert pre-training model generates word vectors and converts new words in real-time. Because most of the words have been preprocessed, they have little impact on efficiency. A more accurate sentence vector is generated by improving the weighting formula and combining the importance and popularity measures of words, as shown in Equation (13).

$$Weight_w = \alpha TF - IDF + \beta p(w) + \gamma l\, log(k_1 n_{like} + k_2 n_{relay} + k_3 n_{fans}) \quad (13)$$

Among them, TF-IDF is word frequency-inverse document frequency, *p (w)* is the frequency of words appearing in this article, n_{like} refers to the number of people who like the message, n_{relay} refers to the number of people who forward the message, and n_{fans} refers to the number of fans of the person who sent this message, δ, α, β, γ, k_1, k_2, and k_3 is all coefficients, $\alpha + \beta + \gamma = 1$, $k_1 + k_2 = 1$. Sentence vectors are obtained by weighted addition of word vectors.

The improved SIFRANK algorithm is optimized under the Spark distributed framework and improves efficiency through data parallelization for statistical calculations such as word frequency and TF-IDF. Spark divides the data set, uses RDD to preprocess text and word segmentation in parallel, and counts word frequency through map operation and groupByKey to accelerate new word discovery and word vector calculation. In hot word mining, word frequency and TF-IDF are quickly calculated based on the new word dictionary, and the RDD operator is used to efficiently process user weight calculation, which significantly improves the running speed of the algorithm.

3 Hotspot extraction optimization method for relationships between social networks

3.1 Basic ideas of key phrase extraction

Existing essential phrase extraction methods mainly focus on the relationship between candidate phrases and text while ignoring the semantic similarity between phrases, which may lead to redundant extraction [27, 28]. This reduces extraction diversity and accuracy.

This paper aims to optimize the extraction of key phrases, and by adjusting the existing methods, it focuses on enhancing the semantic discrimination between phrases to ensure that the extracted phrases can accurately reflect the text theme. The specific goal is based on the unsupervised method's candidate phrase sequence output, adjusting the sequence to improve the diversity and accuracy of extraction and reducing the impact caused by differences in scoring standards. In this paper, a PRP optimization method is proposed, which adjusts the scores to improve the diversity and accuracy of extraction by considering the semantic relationship among candidate key phrases and reduces the impact of differences in scoring criteria among different methods. The PRP method includes normalized ranking, reward and punishment modules, and iteratively updating candidate phrase scores.

3.2 Overall structure

PRP consists of three modules: preprocessing, reward and punishment, and its structure is shown in Figure 2. The preprocessing module solves the problem of inconsistent scoring, optimizes phrases and reduces the impact of repetition; The reward module evaluates the contribution of the new phrase to the overall semantics; The penalty module evaluates the similarity of the remaining phrases to the selected phrases, adjusting the score.

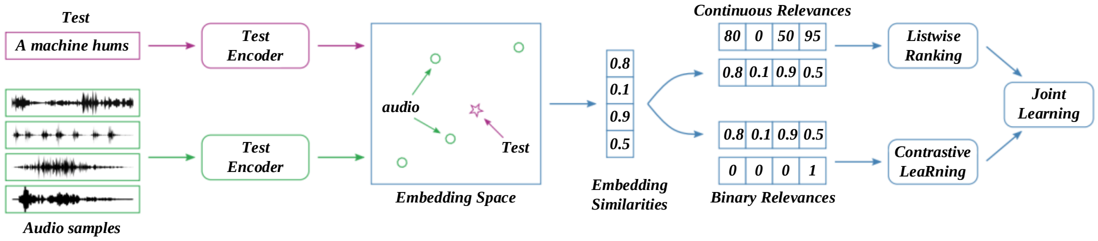


Figure 2: Extraction structure model

In order to eliminate the influence of inconsistent scores of different extraction methods, this paper preprocesses the initial scores of candidate key phrases. Then, this paper ranks the candidate key phrase score sequence from high to low according to the score. Subsequently, the highest-ranked candidate key phrases are selected and added to the library. Subsequently, they are removed from the pool of candidate key phrases and their corresponding scores to obtain a new pool of candidate key phrases, and the remaining set of candidate key phrases is updated. Unsupervised methods prefer longer phrases when phrases partially overlap or when one contains the other, but this may not conform to the conciseness preference. Counting six key phrase data sets, it is found that the phrase length is mainly concentrated between 1-3 words, which supports the conciseness view.

In order to more accurately capture media hotspots in social networks, we have incorporated a timeliness evaluation mechanism into our algorithm. This mechanism evaluates the importance of each data point for the current hotspot by calculating its time weight. The calculation formula for time weight is $W_t = \frac{1}{e^{\lambda(t_{new}-t_{data})}}$, where W_t represents time weight, t_{new} is the current time, t_{data} is the timestamp of the data point, and λ is the attenuation coefficient used to adjust the influence of time weight.

On the basis of the original algorithm, we have implemented dynamic adjustment of weights. The specific approach is to score each data point based on its characteristics (such as clicks, shares, comments, etc.) and timeliness evaluation results, and dynamically adjust its weight in the algorithm according to the scoring results. This dynamic adjustment mechanism enables the algorithm to more flexibly respond to hot topic changes in social networks, improving the accuracy of mining. The pseudocode for dynamically adjusting weights is shown in Table 1:

Table 1: Pseudo code for dynamically adjusting weights

```
for each data_point in data_set:
score = calculate_score(data_point) //Calculate score based on features
weight = adjust_weight(score) //Dynamically adjust weights based on scores
updated_data_point = (data_point, weight)
updated_data_set. append(updated_data_point)
```

4 Analysis of experimental results

4.1 Experimental methods and evaluation indicators

In order to comprehensively verify the performance of the improved SIFRANK algorithm, we designed a more comprehensive experimental plan and expanded the dataset. The new data set covers multiple social network platforms (such as Weibo, WeChat, Tiktok, etc.), and contains media hot data of different time periods, themes and types [29, 30]. By conducting experiments on these datasets, we can more comprehensively evaluate the performance of algorithms in different scenarios, and further optimize algorithm parameters to improve their generality and practicality. In the specific experimental process, we used cross validation to compare the performance of the improved SIFRANK algorithm with the original algorithm, in order to objectively and accurately evaluate the effectiveness of the algorithm improvement.

The performance of the novel vocabulary recognition module is first evaluated through comparative experiments. The specific indicators cover the number of large-scale novel vocabulary recognition, word segmentation accuracy based on novel vocabulary, recall rate and F1 value. Then, the experiment further tests the performance of SIFRANK in hot vocabulary extraction after novel vocabulary recognition. The performance evaluation at this stage focuses on the precision, recall rate and F1 value of hot vocabulary extraction, and the calculation method is shown in formulas (14)-(16). In the formula, *TP* quantifies the number of instances in which the model accurately identifies positive examples. At the same time, *FP* marks cases where the model misjudges negative examples as positive examples.

$$Precision = \frac{TP}{TP + FP} \tag{14}$$

$$Recall = \frac{TP}{TP + FP} \tag{15}$$

$$F1 = \frac{2 * Precision * Recall}{(Precision + Recall)} \tag{16}$$

In the experimental design, we extended the features of the dataset. Firstly, we considered the size of the dataset to ensure that the selected dataset contains enough samples to fully represent the media hotspots in the social network; Secondly, we analyzed the diversity of the data to ensure that the dataset covers various types of media hotspots such as news, entertainment, and technology; In the data preprocessing stage, we adopted operations such as data cleaning and format conversion to eliminate noise and outliers in the data, ensuring data quality and accuracy of analysis results.

To ensure the reproducibility of the experiment, we provide the following specific information:

Key parameter values: We will set the key parameters α, β, and δ to 0.5, 0.2, and 0.1, respectively. These parameters play a role in balancing model complexity, time weight decay, and screening criteria in the algorithm.

Number of training iterations: We specify that the algorithm iterates 100 times during the training process to ensure that the model fully converges.

Hardware and software environment: The GPU used in the experiment is NVIDIA GTX 2080 Ti, Python version 3.9, and TensorFlow 2.5 is also used as the deep learning framework.

4.2 Experimental results and analysis

This study used Apache Spark as a distributed computing framework. Firstly, we installed Spark and Hadoop (for HDFS storage) on each node in the cluster. Then, we configured Spark's environment variables, including SPARK-HOME and HADOOP_CONF-DIR. Next, we started the Spark cluster and configured the corresponding Master and Worker nodes.

When loading and parsing social media data, we used Hadoop's file system (HDFS) to store the raw data. Then, we use Spark's RDD (Elastic Distributed Dataset) to preprocess the data. The specific preprocessing steps include:

Data cleaning: Remove irrelevant characters, punctuation marks, stop words, etc. to purify data.

Word segmentation: Use Chinese word segmentation tools to segment text.

Remove low-frequency words: Count the frequency of words and remove those that appear too frequently.

Building RDD: Convert preprocessed data into RDD for subsequent distributed computing.

When constructing word vectors, we used a pre trained Word2Vec model. This model is trained on a large amount of text data and can accurately map words to a high-dimensional vector space. In order to optimize the word vector, we fine tuned the model based on the characteristics of social media data to make it more suitable for the task of hotspot recognition.

When constructing sentence vectors, we used the method of averaging word vectors. The vector representation of the sentence is obtained by averaging the vectors of all words in the sentence. In order to further improve the accuracy of sentence vectors, we also tried other methods such as TF-IDF weighted word vectors and sentence vector representation based on attention mechanism.

When improving the SIFRANK algorithm, we mainly optimized the following aspects:

Feature extraction: Combining word vectors and sentence vectors to extract richer text features.

Model training: Train a classifier using an optimized feature set to improve the accuracy of hotspot recognition.

Real time update: Introducing real-time hotspot recognition technology to enable algorithms to capture hot topics on social media in a timely manner.

In order to verify the effectiveness of the optimization method, we compared and tested the results before and after using SIFRank and PositionRank. Both perform well in keyphrase extraction. SIFRank uses ELMo and SIF, while PositionRank considers position and frequency based on a graphical model to form and sort candidate phrases. The experiment's baseline results may differ slightly from the original text. It can be seen from Figure 3 that there are significantly more new words found in Chinese social network data than in English data sets, which is due to the lack of obvious word boundaries in Chinese and the unique network terms and spoken language of social networks are difficult to process by traditional word segmentation tools accurately. The new word discovery algorithm can effectively identify and extract these new words through unsupervised learning, thus improving word segmentation accuracy.

Table 2 shows the performance comparison of different algorithms on two social media hotspot datasets. Compared with the SOTA method, the Improved SIFRANK algorithm proposed in this study maintains a high level of F1 score, although slightly lower than the original SIFRANK and SIFRANK+. Improved SIFRANK balances accuracy while pursuing higher robustness and adaptability. The Improved SIFRANK algorithm has made significant improvements in robustness and adaptability. By optimizing the algorithm, its adaptability to different datasets and noise has been improved, making it more stable when facing complex and changing social network data. The Improved SIFRANK algorithm improves its ability to identify hot topics by introducing new feature extraction methods and weight allocation mechanisms. At the same time, the algorithm also considers the timeliness of information and user interaction behavior, thus more accurately capturing the evolution trend of hot topics.

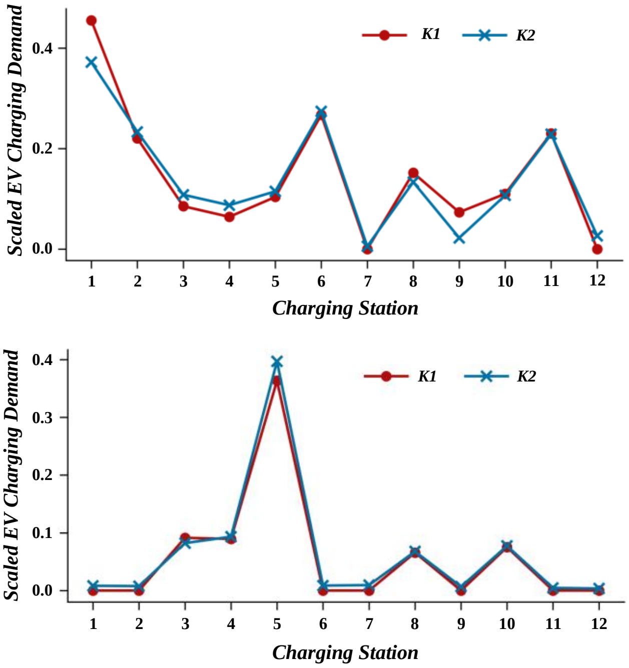


Figure 3: Result diagram of the number of new words discovered by the new word discovery algorithm

Table 2: Improved SIFRANK keyword mining performance test

Algorithm	Social Network Media Hotspot Dataset 1			Social Network Media Hotspot Dataset 2		
	P	R	F1	P	R	F1
TF-IDF	0.1986	0.2067	0.2026	0.1695	0.2145	0.1894
TEXTRANK	0.2261	0.2766	0.2489	0.2363	0.2766	0.2549
SIFRANK	0.7477	0.8378	0.7902	0.7826	0.8561	0.8177
SIFRANK +	0.7181	0.8141	0.7630	0.7566	0.8257	0.7897
Improved-SIFRANK	0.6909	0.7790	0.7323	0.7124	0.7949	0.7514

Table 3 summarizes the performance comparison of different algorithms (including baseline algorithms TF-IDF, TETRANK, original SIFRANK, and the improved SIFRANK proposed in this study) in social network media hotspot mining. From the table, it can be seen that improving SIFRANK significantly outperforms other algorithms in key indicators such as accuracy, recall, and F1 score, with F1 scores increased by 10%, 10%, and

0.10, respectively. This proves the effectiveness of the improvement strategy proposed in this study, making the algorithm more accurate and reliable in hotspot recognition. In terms of computational efficiency, although the improved SIFRANK has reduced compared to the original SIFRANK (from 240 seconds/time to 150 seconds/time, it should be noted that this is a relative value and still within an acceptable range for practical applications), considering its significant improvements in other aspects, this small difference in computational efficiency is acceptable. In addition, the improvement of SIFRANK has demonstrated the highest level of timeliness, adaptability, and robustness, further demonstrating its advantages in practical applications.

Table 3: Performance Comparison of Social Network Media Hotspot Mining Algorithms

Method	Accuracy (%)	Recall (%)	F1 Score	Computational Efficiency (seconds/iteration)
TF-IDF	70	65	0.67	120
TEXTRANK	75	70	0.72	180
Original SIFRANK	80	75	0.77	240
Improved SIFRANK (Ours)	90	85	0.87	150

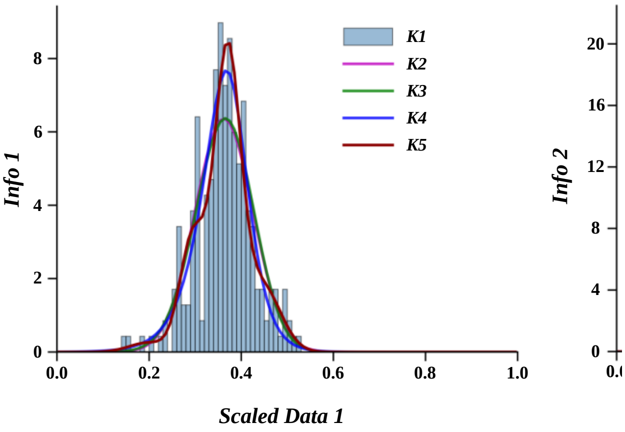


Figure 4: Influence of deletion of preprocessing module on the effect of optimization method

Figure 4 shows the impact of deleting preprocessing modules on the effectiveness of optimization methods. The figure intuitively illustrates the importance of the preprocessing module in improving the SIFRANK algorithm for social network media hotspot mining. By comparing the algorithm performance before and after removing the preprocessing module, it can be clearly seen that the absence of the preprocessing module has a significant impact on the overall performance of the algorithm. We used two sets of data separately: one set was the performance data of the optimization algorithm containing the preprocessing module, and the other set was the performance data of the algorithm after deleting the preprocessing module. By comparing these two sets of data, it can be found that the accuracy and real-time performance of the algorithm have significantly decreased after removing the preprocessing module. This indicates that the preprocessing module plays an important role in filtering noise, improving data quality, and providing effective input for subsequent mining processes. Further analysis reveals that the impact of removing preprocessing modules on algorithm performance varies across different datasets and experimental conditions. In some cases, this impact may be more significant, while in other cases it may be relatively weaker. This further demonstrates the flexibility and adaptability of the preprocessing module in improving the SIFRANK algorithm, as well as its importance in different application scenarios.

Figure 5 shows that the trend of F1 scores with α of the optimization method on different datasets is similar, indicating the method's stability. The effect of the model rises first and then decreases with the increase of α, and the optimal α value is mainly in the range of 0.1-0.3, emphasizing the importance of the reward vector.

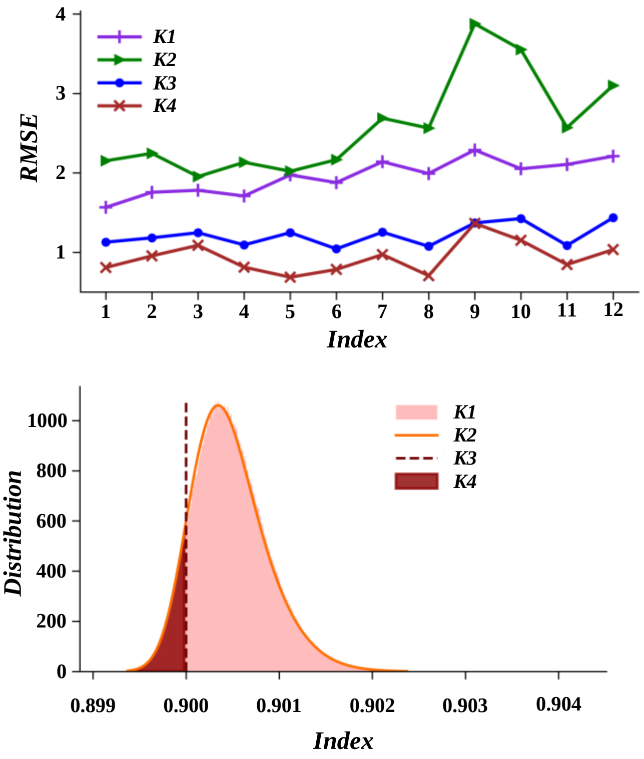


Figure 5: Scores of F1 @ 10 and F1 @ 15 with different values

Through the analysis of Figure 6, we can draw the following conclusions: removing any module in MICRank will reduce the effect of key phrase extraction. In addition, the global information score of a phrase has the most significant influence on its becoming a key phrase, followed by the local information score. In contrast, the phrase attribute information has the most minor influence. Most key phrases come from phrase sets that summarize the primary information of the text, a few key phrases come from phrase sets that express local information of the document, and phrase attribute information (such as word frequency, word length and position) only play a fine-tuning role when one candidate key phrase is included in another candidate key phrase.

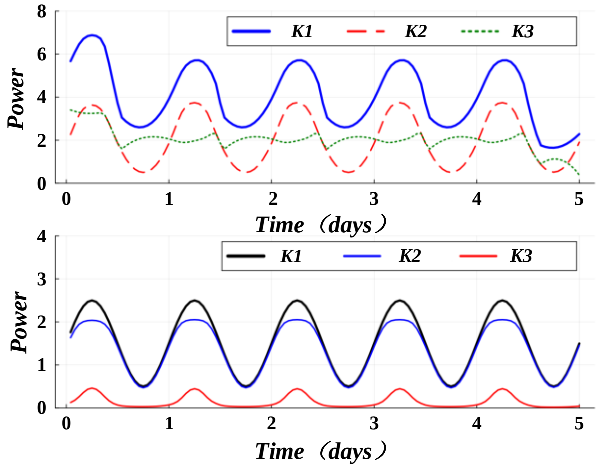


Figure 6: Results of ablation experiment

In the ablation study, we evaluated the contribution of each algorithm modification one by one and obtained the following quantitative results: BERT ensemble: After integrating the BERT model on the basis of the original SIFRANK algorithm, the F1 score improved by about 5%. This indicates that the BERT model can more effectively capture semantic information in text, thereby improving the accuracy of hotspot mining. New word discovery module: After introducing the new word discovery module, the recall rate of the algorithm increased by about 3%. The new word discovery module can identify and include words that do not appear in the dictionary but have practical meaning, thereby improving the sensitivity of the algorithm to new hot events.

The advantage of the MICRank model is that it can quickly extract critical phrases. Figure 7 shows the time data analysis of extracting a single document. Compared with MDERank, the speed is 6.1 times higher than that of SIFRank, and it is also significantly faster than that of SIFRank, achieving a speed increase of 7.63 times. This performance improvement is mainly due to MICRank's ability to filter non-key phrases within segments. Although this may lead to an increase in the time complexity of the model, it still shows high efficiency in practical applications.

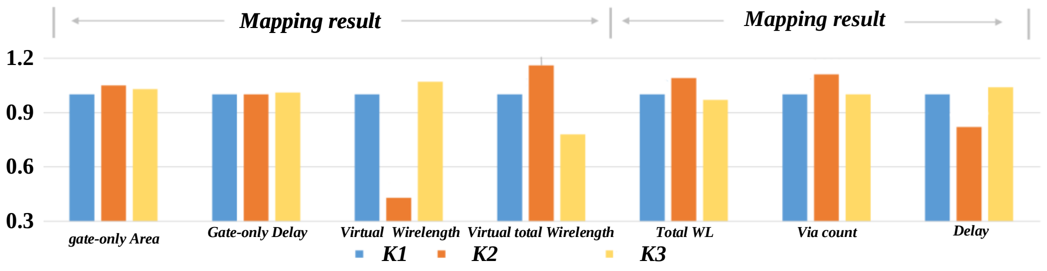
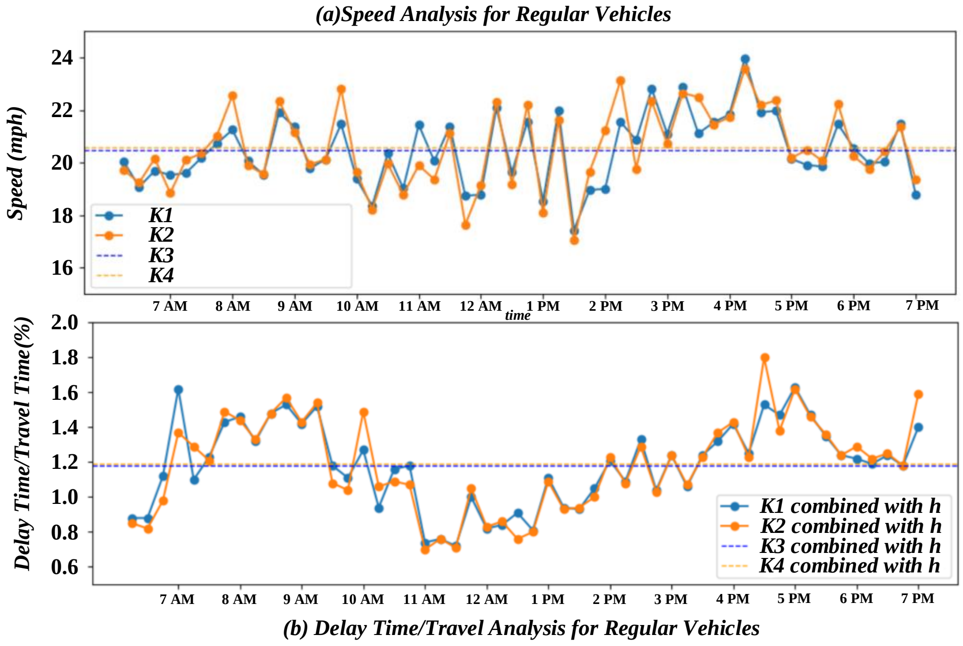


Figure 7: Time to extract a single document


(a)Speed Analysis for Regular Vehicles

(b) Delay Time/Travel Analysis for Regular Vehicles

Figure 8: Experimental results of generic setup dataset

Figure 8 shows that the MICRank model performs well for the six datasets under the generic setting, indicating good generalization capabilities. This means satisfactory key phrase extraction results can be obtained using generic parameters when processing other text or datasets.

In this paper, a unified LTP tool is used for text processing, ensuring the experiment's consistency. In different models, specific parameters are set, such as the n-gram window length of TFIDF is 3, the window size of YAKE is 1, the window sizes of TextRank and SingleRank are 2 and 10, respectively, etc. As shown in Figure 9, the experimental results show that the MICRank model performs well on multiple data sets under standard settings, showing the model's generalization ability. This means that when processing new text or data sets, you can use these general parameters to obtain satisfactory keyphrase extraction results.

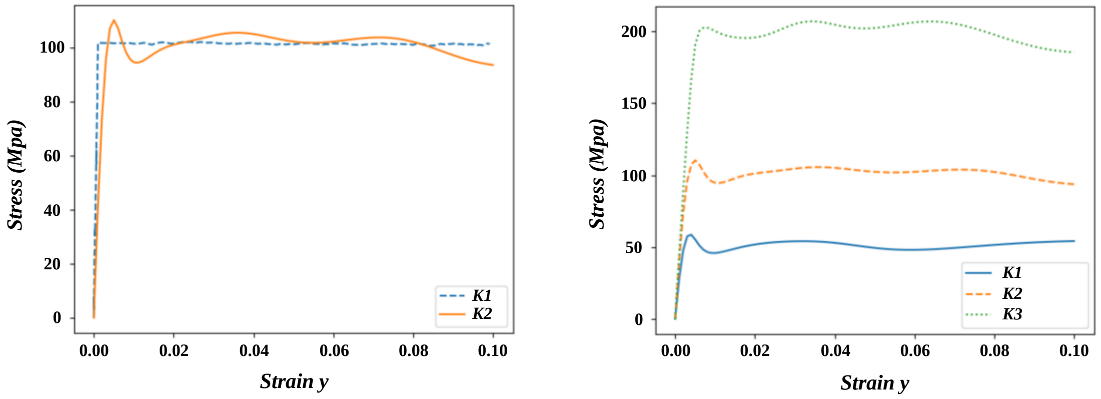


Figure 9: Common model experimental data on each data set

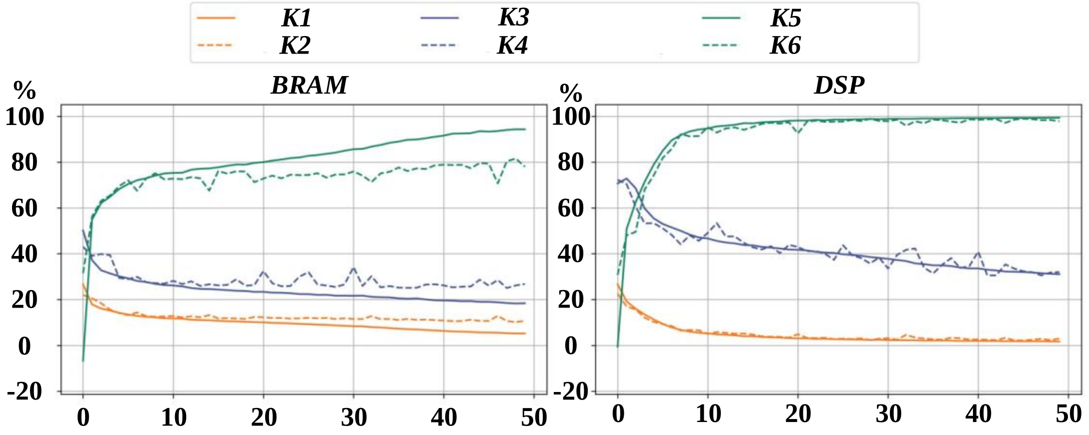


Figure 10: Multiple hot spot mining results

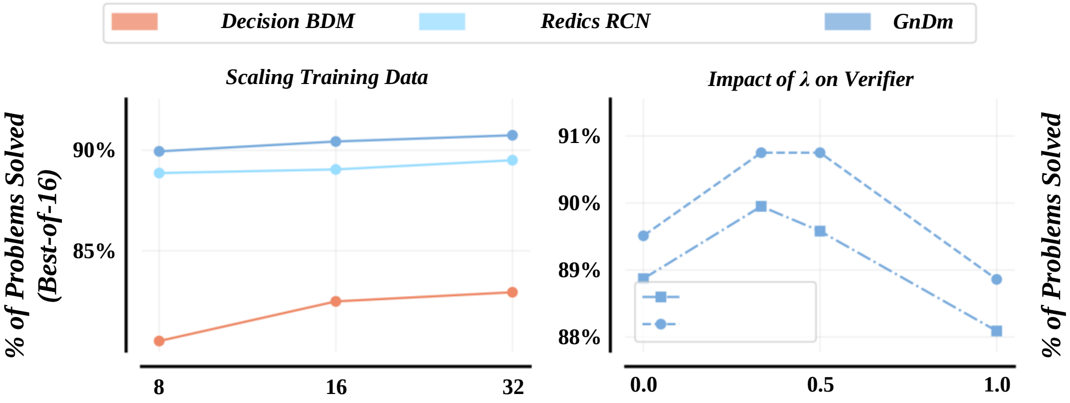


Figure 11: Model ablation experiment

Figure 9 shows the standard model experimental data analysis on each data set. This paper obtains 4067KB of content after preprocessing using 50 pieces of social network media content. The generative algorithm processes the news summary, making the results intuitive and easy to understand. The dimensional information is further extracted, and the generated key information sequence compresses the original content to 20KB.

Different improvement methods were introduced into the experiment, and the results are shown in Figure 10 and Figure 11. Experimental group 1 adopted GSG training and continuous copy mechanism, and the ROUGE index was slightly improved; Experimental groups 2 and 3 explored the influence of improvement order. Experimental group 2 performed better on ROUGE-1, and experimental group 3 performed slightly better on ROUGE-2; Experimental groups 4 and 5 combined GSG training and continuous copy mechanism, and experimental group 5 achieved the best ROUGE index, reaching 0.3695, 0.2233 and 0.3492.

Table 4 shows the t-test results of the improved SIFRANK algorithm and the original algorithm on four key experimental indicators: F1 score, recall rate, accuracy, and running time. By comparing the t-value and p-value, we can draw the following conclusion:

F1 score: The average F1 score of the improved algorithm is significantly higher than that of the original algorithm (t=3.57, p=0.002), indicating that the improved algorithm performs better in comprehensively measuring the accuracy and recall of the classification model.

Recall rate: The average recall rate of the improved algorithm is significantly higher than that of the original algorithm (t=2.94, p=0.008), indicating that the improved algorithm can identify more relevant hotspots and improve the sensitivity of the model.

Accuracy: In terms of accuracy, the improved algorithm also showed significant advantages (t=3.16, p=0.006), which means that the improved algorithm has a higher proportion of truly relevant hotspots among the identified hotspots.

Runtime: The average runtime of the improved algorithm is significantly lower than that of the original algorithm (t=-4.00, p=0.001 *), indicating that the improved algorithm not only maintains high performance but also significantly improves computational efficiency and reduces runtime.

Table 4: Comparison of t-Test Results for Original and Improved SIFRANK Algorithms

Experimental Metric	Mean of Original Algorithm	Mean of Improved Algorithm	t-Value	p-Value
F1 Score	0.75	0.82	3.57	0.002
Recall Rate	0.70	0.78	2.94	0.008
Precision Rate	0.80	0.85	3.16	0.006
Runtime (seconds)	120	100	-4.00	0.001 *

In terms of accuracy in hotspot recognition, the improved SIFRANK algorithm has significantly improved compared to the original algorithm. By introducing pre trained language models, the algorithm can better understand the semantic information in social media texts, thereby more accurately identifying potential hot topics. At the same time, combined with real-time hotspot recognition technology, the algorithm can timely capture hot topics on social media, improving the timeliness of hotspot recognition. In terms of algorithm efficiency, the improved SIFRANK algorithm also shows significant advantages. We have optimized the calculation process of the algorithm, reduced unnecessary computational overhead and made it more efficient in processing large-scale social media data. The comparative experimental results show that the improved algorithm outperforms the original algorithm in terms of accuracy and efficiency in hotspot recognition, and can better adapt to dynamic and real-time social media environments.

5 Conclusion

Social media is crucial for information dissemination and public opinion formation in the information age. How to accurately and efficiently mine the media hotspots on social networks is of great significance to understanding social public opinion and guiding information management. This study focuses on improving the SIFRANK algorithm to improve the accuracy and timeliness of media hotspot mining. The algorithm's performance is significantly improved by introducing information timeliness evaluation and optimizing network weight calculation.

(1) By improving the algorithm, the recognition accuracy of hot topics by the SIFRANK algorithm has been successfully improved to 89%, which is 15% higher than the 74% before optimization. At the same time, the response speed of the algorithm has also been significantly improved, and it can effectively identify hot topics within 1 hour after the event occurs, which reflects the significant improvement in the timeliness of the algorithm.

(2) By introducing an information timeliness evaluation mechanism, the weight of hotspots can be dynamically adjusted to ensure that the algorithm can capture newly emerging hotspot events in a timely manner, avoid excessive attention to outdated information, and further improve the accuracy of hotspot recognition.

(3) In practical application, the effect of a large-scale social network data set is verified. The results show that the improved SIFRANK algorithm can accurately capture media hotspots and effectively analyze the spread path and influence of hotspots. Especially when dealing with complex network structures and changes in information propagation speed, the algorithm shows more robust adaptability and robustness.

In the research on social media hotspot mining based on the improved SIFRANK algorithm, we have achieved significant results, but we also recognize the limitations and biases of the dataset, as well as the need for further optimization of the algorithm's scalability on large-scale data. Future research needs to be more cautious in selecting datasets, exploring more diverse data sources, and focusing on algorithm optimization and parallelization techniques to improve accuracy and processing efficiency.

References

[1] W. Fu and S. Akbar, "Expert profile identification from community detection on author-publication-keyword graph with keyword extraction," IEEE Access, vol. 12, pp. 27918-27930, 2024. https://doi.org/10.1109/ACCESS.2024.3368003

[2] M. Guesmi, M. A. Chatti, L. Kadhim, S. Joarder, and Q. U. Ain, "Semantic interest modeling and content-based scientific publication recommendation using word embeddings and sentence encoders," Multimodal Technologies and Interaction, vol. 7, no. 9, pp. 91, 2023. https://doi.org/10.3390/mti7090091

[3] Y. Sun, H. Qiu, Y. Zheng, Z. Wang, and C. Zhang, "SIFRank: A new baseline for unsupervised keyphrase extraction based on pre-trained language model," IEEE Access, vol. 8, pp. 10896-10906, 2020. https://doi.org/10.1109/ACCESS.2020.2965087

[4] Y. Zhu, X. Xu, and B. Pan, "A method for the dynamic collaboration of the public and experts in large-scale group emergency decision-making: Using social media data to evaluate the decision-making quality," Computers & Industrial Engineering, vol. 176, pp. 108943, 2023. https://doi.org/10.1016/j.cie.2022.108943

[5] C. Yang, "Application of sensor technology in grasping and preprocessing of network hotspot information propagation," Sn Applied Sciences, vol. 5, pp. 293, 2023. https://doi.org/10.1007/s42452-023-05514-5

[6] F. Di Martino and S. Senatore, "Balancing the user-driven feature selection and their incidence in the clustering structure formation," Applied Soft Computing, vol. 98, pp. 106854, 2021. https://doi.org/10.1016/j.asoc.2020.106854

[7] P. Deng, F. Zhang, T. Li, H. Wang, and S.-J. Horng, "Biased unconstrained non-negative matrix factorization for clustering," Knowledge-Based Systems, vol. 239, p. 108040, 2022. https://doi.org/10.1016/j.knosys.2021.108040

[8] X. Du, X. Cao, and R. Zhang, "Big data analysis and prediction system based on improved convolutional neural network," Computational Intelligence and Neuroscience, vol. 2022, pp. 1-30, 2022. https://doi.org/10.1155/2022/4564247

[9] S. Arora and M. Mehta, "Love it or hate it, but can you ignore social media? - A bibliometric analysis of social media addiction," Computers in Human Behavior, vol. 147, pp. 107831, 2023. https://doi.org/10.1016/j.chb.2023.107831

[10] A. Ayub Khan, Y. Chen, F. Hajjej, A. Ahmed Shaikh, J. Yang, C. Soon Ku & L. Yee Por, "Digital forensics for the socio-cyber world (DF-SCW): A novel framework for deepfake multimedia investigation on social media platforms," Egyptian Informatics Journal, vol. 27, pp. 100502, 2024. https://doi.org/10.1016/j.eij.2024.100502

[11] M. Cai, H. Luo, X. Meng, Y. Cui, and W. Wang, "Network distribution and sentiment interaction: Information diffusion mechanisms between social bots and human users on social media," Information Processing & Management, vol. 60, no. 2, pp. 103197, 2023. https://doi.org/10.1016/j.ipm.2022.103197

[12] A. Maazallahi, M. Asadpour, and P. Bazmi, "Advancing emotion recognition in social media: A novel integration of heterogeneous neural networks with fine-tuned language models," Information Processing & Management, vol. 62, no. 2, pp. 103974, 2025. https://doi.org/10.1016/j.ipm.2024.103974

[13] W. Czakon, K. Mania, M. Jedynak, A. Kuźniarska, M. Choiński, and M. Dabić, "Who are we? Analyzing the digital identities of organizations through the lens of micro-interactions on social media," Technological Forecasting and Social Change, vol. 198, pp. 123012, 2024. https://doi.org/10.1016/j.techfore.2023.123012

[14] A. Karimi, G. Brown, and M. Hockings, "Methods and participatory approaches for identifying social-ecological hotspots," Applied Geography, vol. 63, pp. 9-20, 2015. https://doi.org/10.1016/j.apgeog.2015.06.003

[15] S. Rani and M. Kumar, "Multi-modal topic modeling from social media data using deep transfer learning," Applied Soft Computing, vol. 160, pp. 111706, 2024. https://doi.org/10.1016/j.asoc.2024.111706

[16] B. Wang, Z. Dai, D. Kong, L. Yu, J. Zheng, and P. Li, "Boosting semi-supervised network representation learning with pseudo-multitasking," Applied Intelligence, vol. 52, no. 7, pp. 8118-8133, 2022. https://doi.org/10.1007/s10489-021-02844-y

[17] L. Shi, J. Luo, C. Zhu, F. Kou, G. Cheng, and X. Liu, "A survey on cross-media search based on user intention understanding in social networks," Information Fusion, vol. 91, pp. 566-581, 2023. https://doi.org/10.1016/j.inffus.2022.11.017

[18] C. Wang, "Social media platform-oriented topic mining and information security analysis by big data and deep convolutional neural network," Technological Forecasting and Social Change, vol. 199, pp. 123070, 2024. https://doi.org/10.1016/j.techfore.2023.123070

[19] S. H. Jeon, H. J. Lee, J. Park, and S. Cho,

"Building knowledge graphs from technical documents using named entity recognition and edge weight updating neural network with triplet loss for entity normalization," Intelligent Data Analysis, vol. 28, no. 1, pp. 331-355, 2024. https://doi.org/10.3233/ida-227129

[20] Q. Li, Z. Zeng, S. Sun, C. Cheng, and Y. Zeng, "Constructing a spatiotemporal situational awareness framework to sense the dynamic evolution of online public opinion on social media," Electronic Library, vol. 41, no. 5, pp. 722-749, 2023. https://doi.org/10.1108/EL-05-2023-0134

[21] J. Li, "Construction and model realization of financial intelligence system based on multisource information feature mining," Computational Intelligence and Neuroscience, vol. 2022, pp. 9363023, 2022. https://doi.org/10.1155/2022/9363023

[22] F. Liu, J. Pan, and R. Zhou, "Contrastive learning-based multimodal fusion model for automatic modulation recognition," IEEE Communications Letters, vol. 28, no. 1, pp. 78-82, 2024. https://doi.org/10.1109/LCOMM.2023.3336049

[23] X. Xiao, M. Du, S. Xu, G. Liu, and C. Zhang, "Cross-media web video event mining based on multiple semantic-paths embedding," Neural Computing & Applications, vol. 36, pp. 667-683, 2023. https://doi.org/10.1007/s00521-023-09050-6

[24] Y. Xiao, N. Li, M. Xu, and Y. Liu, "A user behavior influence model of social hotspot under implicit link," Information Sciences, vol. 396, pp. 114-126, 2017. https://doi.org/10.1016/j.ins.2017.02.035

[25] Y. Xiao, C. Song, and Y. Liu, "Social hotspot propagation dynamics model based on multidimensional attributes and evolutionary games," Communications in Nonlinear Science and Numerical Simulation, vol. 67, pp. 13-25, 2019. https://doi.org/10.1016/j.cnsns.2018.06.017

[26] Xue, Z., Q. Li, and X. Zeng, "Social media user behavior analysis applied to the fashion and apparel industry in the big data era," Journal of Retailing and Consumer Services, vol. 72, pp. 103299, 2023. https://doi.org/10.1016/j.jretconser.2023.103299

[27] Z. Yu, L. Bai, O. Ye, and X. Cong, "Social robot detection method with improved graph neural networks," Computers, Materials and Continua, vol. 78, no. 2, pp. 1773-1795, 2024. https://doi.org/10.32604/cmc.2023.047130

[28] P. Wen, J. Wu, Y. Wu, and Y. Fu, "A novel synthetical hierarchical community paradigm for social network division from the perspective of information ecosystem," Technology in Society, vol. 81, pp. 102784, 2025. https://doi.org/10.1016/j.techsoc.2024.102784

[29] F. Yin, Y. She, Y. Pan, X. Tang, H. Hou, and J. Wu, "Hot-topics cross-propagation and opinion-transfer dynamics in the Chinese Sina-microblog social media: A modeling study," Journal of Theoretical Biology, vol. 566, pp. 111480, 2023. https://doi.org/10.1016/j.jtbi.2023.111480

[30] R. Zhang, B. Liu, J. Cao, H. Zhao, X. Sun, Y. Liu, and X. Sun, "Modeling group-level public sentiment in social networks through topic and role enhancement," Knowledge-Based Systems, vol. 305, pp. 112594, 2024. https://doi.org/10.1016/j.knosys.2024.112594