IARJSET ISSN (O) 2393-8021, ISSN (P) 2394-1588

International Advanced Research Journal in Science, Engineering and Technology

Impact Factor 8.311 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 13, Issue 10, October 2026

DOI: 10.17148/IARJSET.2026.131051

Big Data Based Social Media Sentiment Analysis and User Behavior Mining

Sheeba S[1], Dileepa D[2], Gagan R[3], Varun Gowda G P[4], Vikram R T[5]

Assistant Professor, Dept. of CSE Don Bosco Institute of Technology, Bengaluru, India[1]

Dept. of CSE Don Bosco Institute of Technology, Bengaluru, India[2-5]

Abstract: Social media platforms generate a continuous stream of user-generated content in the form of posts, comments, reactions, reviews, and other interactions. Although this information can provide useful insights into public opinion and user behavior, its large volume, high generation rate, and unstructured nature make conventional processing approaches difficult to apply. This paper presents a Big Data-based framework for analyzing social media content by combining distributed data processing, Natural Language Processing (NLP), and machine learning techniques. Apache Hadoop and Apache Spark are used to support large-scale data storage and processing, while NLP techniques are applied to prepare the collected textual data for analysis. User behavior is studied using parameters such as posting frequency, interaction levels, activity periods, and engagement patterns. For sentiment classification, the textual data is represented using Term Frequency-Inverse Document Frequency (TF-IDF), followed by classification using the Multinomial Naive Bayes algorithm. The resulting data is presented through charts and analytical reports to make the identified trends easier to interpret. The system classifies social media content into positive, negative, and neutral categories and provides behavioral information that can be used for further analysis.

Keywords: user behavior mining, sentiment analysis, social media, computer big data analysis, user experience.

I. INTRODUCTION

Social media has become an important source of user-generated information. Platforms such as Facebook, Instagram, X, and other online communities receive large numbers of posts, comments, reviews, likes, shares, and reactions every day. These interactions contain information about user opinions, preferences, interests, and behavior. As the amount of such content continues to increase, extracting useful information from it becomes increasingly difficult.

A major challenge associated with social media data is its volume and unstructured nature. New information is generated continuously, while textual content can contain incomplete sentences, symbols, hyperlinks, emojis, abbreviations, and other forms of noise. Traditional data processing methods are not well suited for managing such datasets at scale. Big Data frameworks provide a way to distribute storage and computation across multiple systems, allowing large datasets to be processed more efficiently.

Apache Hadoop provides distributed storage through the Hadoop Distributed File System (HDFS), while Apache Spark supports distributed and parallel data processing. Together, these technologies can be used to build a scalable environment for social media analytics.

Natural Language Processing is also important when analyzing textual social media content. One of its major applications is sentiment analysis, which determines the general opinion expressed in a piece of text. Depending on the application, text can be classified as positive, negative, or neutral. Such classification can help identify general public response, customer opinions, and changes in sentiment over time.

Apart from sentiment, the activities performed by users provide another source of useful information. User behavior mining examines factors such as posting frequency, interaction counts, activity periods, and engagement patterns. Analyzing these characteristics can help identify highly active periods, differences in user participation, and general engagement trends.

This work combines sentiment analysis and user behavior mining within a common Big Data framework. The system uses Hadoop for distributed storage, Spark for large-scale processing, NLP techniques for preparing textual information, TF-IDF for feature representation, and Multinomial Naive Bayes for sentiment classification. The resulting information is presented using graphical visualizations so that the observed patterns can be interpreted more easily.

© IARJSET This work is licensed under a Creative Commons Attribution 4.0 International License

IARJSET ISSN (O) 2393-8021, ISSN (P) 2394-1588

International Advanced Research Journal in Science, Engineering and Technology

Impact Factor 8.311 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 13, Issue 10, October 2026

DOI: 10.17148/IARJSET.2026.131051

II.RELATED WORK

Several studies have investigated the use of Big Data technologies, NLP, and machine learning for analyzing social media information.

proposed a Big Data-based approach for social media user behavior mining and sentiment analysis. Their work considered large-scale user interactions and behavioral trends and also explored the use of distributed computing for recommendation-related applications. The approach was able to handle large datasets, although its implementation required considerable computational resources and a relatively complex infrastructure[1]. Investigated sentiment analysis of social media comments using NLP techniques. Their approach included text preprocessing, feature extraction, and machine learning-based classification. The system classified comments into positive, negative, and neutral categories. The study also highlighted the importance of dataset quality and preprocessing, since errors or inconsistencies at these stages can affect the resulting classification performance[2]. Presented a Big Data analytics framework for obtaining information from social media platforms.

The study focused on processing high-volume data and identifying trends in real time. Although the approach supported large-scale analysis, interpreting highly unstructured text remained a challenge. Problems such as sarcasm and ambiguity can make it difficult to determine the intended sentiment of a social media post[3].

proposed a CNN-based deep learning approach for sentiment classification of Twitter data. [4]Their method was able to capture more complex patterns in textual data and provided improved prediction performance compared with some traditional machine learningapproaches.

Data mining based methods for classifying electricity user behavior characteristics have more accurate classification results [5]. Mao Qun took Android as the research object and explored user behavior patterns through analysis of similar Android applications [6]. Shi Houbing believed that during the operation of the network, managers must collect massive amounts of data in real-time and record all their behavior[7]. Liu Yi planned to use interactive visualization analysis methods to study and analyze the online behavior of university student users. He obtained the distribution of each attribute on the Internet through the analysis of three dimensions of users, topic categories, and time series, so as to find the difference and correlation of user behavior on the Internet [8]. With the rapid development of information technology, the content on social media is becoming increasingly rich. Jain P K believed that sentiment analysis in online comments has attracted many institutions, and sentiment analysis has become a hot topic in fields such as machine learning and natural language processing. Currently, sentiment analysis based on deep learning has achieved good results [9]. One of the most common applications of deep learning algorithms is sentiment analysis. Compared with previous methods, Pandian A P's research provided a better and more efficient automatic feature extraction technique [10]. Khan R believed that social media plays a crucial role in the current world of lockdowns and further social imbalances among people [11]. Alqaryouti O utilized an aspect based sentiment analysis hybrid method that integrates domain dictionaries and rules to analyze comments received by entity intelligent applications [12].

However, deep learning models generally require larger labeled datasets and greater computational resources for training and deployment.Most of these approaches concentrate primarily on sentiment analysis or user behavior analysis. The work presented in this paper combines the two within a single Big Data framework. This provides both sentiment information and behavioral information from the same social media dataset

Table I: Summary of Related Work

Ref.	Approach	Accuracy/Performance	Limitations
[1]	Big Data Analytics + Social Media User Behavior Mining	Efficient handling of large-scale social media datasets	High computational cost, complex infrastructure requirements
[2]	NLP-based Sentiment Analysis using Text Preprocessing + ML Classification	Good sentiment classification performance (Positive/Negative/ Neutral)	Comparative review study; no dedicated deployed real-time system
[3]	Big Data Analytics for Social Media Insights + Trend Analysis	Real-time trend identification and scalable processing	Difficulty handling highly unstructured text and sarcasm detection

IARJSET ISSN (O) 2393-8021, ISSN (P) 2394-1588

International Advanced Research Journal in Science, Engineering and Technology

Impact Factor 8.311 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 13, Issue 10, October 2026

DOI: 10.17148/IARJSET.2026.131051

[4]	Deep Learning-based Sentiment Analysis using CNN	Higher sentiment prediction accuracy compared to traditional ML models	Requires large labeled datasets and higher computational resources
[5]	Proposed System: Hadoop + Spark + NLP + TF-IDF + Multinomial Naive Bayes	Approx. % sentiment classification accuracy with user behavior analysis	Accuracy can be improved using advanced deep learning models

II. METHODOLOGY

The proposed framework processes social media information through a sequence of stages. The major stages are data collection, preprocessing, distributed storage and processing, user behavior mining, sentiment analysis, and result visualization.

A. Dataset Collection

The first stage is the collection of social media data. Data can be obtained from platforms such as Facebook, Instagram, and X using their respective APIs. The collected information may include posts, comments, timestamps, likes, shares, and other interaction-related attributes.

Since social media platforms continuously generate new information, the collection process needs to accommodate both the volume and rate at which data is produced. The raw dataset obtained at this stage is passed to the preprocessing phase before further analysis is performed.

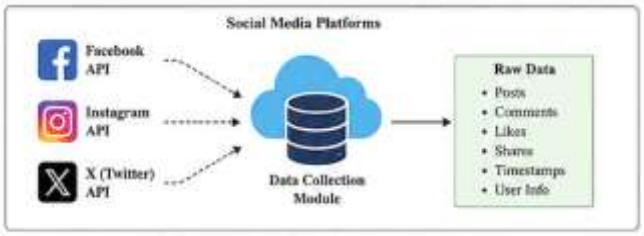


Fig. 1 Data collection from social media platforms.

B. Data Preprocessing

Raw social media data generally contains noise and irrelevant information. Directly applying machine learning algorithms to such data can reduce the quality of the resulting analysis. Preprocessing is therefore used to convert the raw content into a cleaner and more consistent representation.

The dataset is first examined for missing values and duplicate records. Incomplete entries can be removed where appropriate, while duplicate records are eliminated to prevent the same information from influencing the analysis more than once.

Text cleaning is then performed to remove elements that are not required for the selected analysis. These may include URLs, special characters, punctuation, numbers, emojis, mentions, and unnecessary spaces.

The remaining text is converted to lowercase so that different capitalization of the same word is not treated as separate features.

Stop-word removal is subsequently applied. Common words such as "is", "the", "and", and "of" generally provide limited information for sentiment classification and can therefore be removed. Tokenization divides the text into individual words or tokens, providing a suitable representation for subsequent processing.

Depending on the requirements of the dataset, stemming or lemmatization can also be used to normalize related word forms. After preprocessing, the text is converted into numerical features using Term Frequency-Inverse Document Frequency (TF-IDF). TF-IDF assigns weights to terms according to their occurrence within the dataset and helps represent the importance of individual words for machine learning.

IARJSET ISSN (O) 2393-8021, ISSN (P) 2394-1588

International Advanced Research Journal in Science, Engineering and Technology

Impact Factor 8.311 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 13, Issue 10, October 2026

DOI: 10.17148/IARJSET.2026.131051

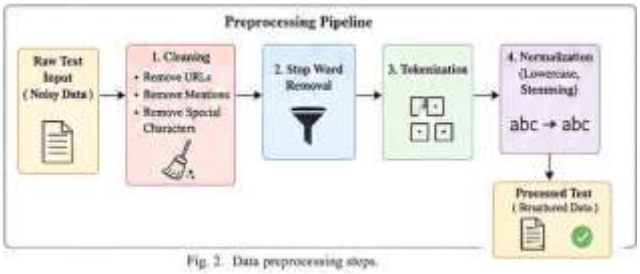


Fig. 2. Data preprocessing steps.

C. Big Data Storage and Processing

After preprocessing, the structured data is stored and processed using Big Data technologies. Conventional storage and processing approaches become difficult to use as the dataset grows in size and complexity. A distributed architecture therefore provides a more suitable approach for handling the collected information.

The Hadoop Distributed File System (HDFS) is used for distributed storage. HDFS divides large datasets into blocks and distributes these blocks across multiple computing nodes. This allows large amounts of data to be stored while providing fault tolerance and improved availability.

Apache Spark is used as the processing engine. Spark supports parallel processing and in-memory computation, allowing analytical operations to be performed efficiently on large datasets.

It can be used for operations such as feature processing, behavioral analysis, and sentiment classification.

The combination of Hadoop and Spark provides separate but complementary functions. Hadoop manages distributed storage, whereas Spark performs the computational operations on the stored data. This arrangement allows the framework to accommodate increasing amounts of social media information while maintaining an efficient processing workflow.

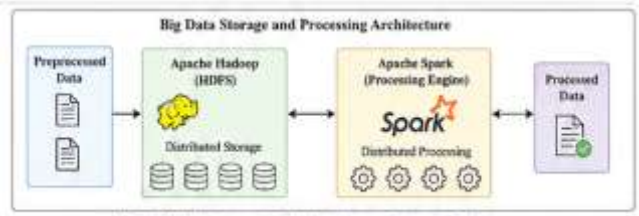


Fig. 3. Big data storage using Hadoop and processing using Spark.

D. User Behavior Mining

User behavior mining examines how users interact with social media platforms. Information such as posts, comments, likes, shares, and timestamps can be used to determine patterns in user activity and engagement.

Several behavioral features are extracted from the processed dataset. These include posting frequency, number of interactions, engagement levels, time of activity, and participation patterns. Timestamp information can be analyzed to determine when users are most active.

For analysis, user activity can be grouped into time periods such as morning, afternoon, evening, and night. Measures such as the number of posts per day, number of comments, likes received, and sharing frequency provide additional indicators of user engagement.

The extracted features are organized into structured behavioral indicators. Data mining techniques can then be applied to identify patterns within the activity data. This makes it possible to determine periods of higher activity, differences in engagement, and the participation levels of users.

The behavioral results are represented through graphs and charts. These visualizations provide a simpler way to examine activity distributions and engagement patterns and can complement the sentiment information generated by the classification module.



Fig. 4. User behavior mining and pattern analysis.

IARJSET ISSN (O) 2393-8021, ISSN (P) 2394-1588

International Advanced Research Journal in Science, Engineering and Technology

Impact Factor 8.311 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 13, Issue 10, October 2026

DOI: 10.17148/IARJSET.2026.131051

E. Sentiment Analysis

The sentiment analysis module determines the general emotional orientation of social media text. The classification used in this work consists of three categories: positive, negative, and neutral.

The analysis begins with the preprocessed textual dataset. Since machine learning algorithms require numerical input, the cleaned text is converted into feature vectors using TF-IDF. The resulting vectors represent the importance of words within the dataset.

The Multinomial Naive Bayes classifier is then used to perform sentiment classification. The model learns the relationship between the extracted textual features and the corresponding sentiment categories from the training data. Once trained, the classifier can assign a sentiment category to previously unseen text.

The use of TF-IDF together with Multinomial Naive Bayes provides a relatively simple machine learning pipeline for text classification. It also allows the textual processing stage to be combined with the distributed Big Data processing environment.

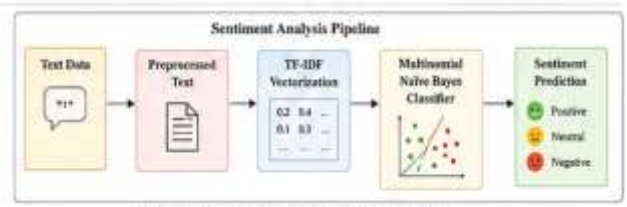


Fig. 5. Sentiment analysis using TF-IDF and Naive Bayes.

F. Result Visualization

The final stage of the framework is the visualization of the analytical results. Raw social media data and numerical outputs can be difficult to interpret directly, particularly when the dataset contains a large number of records. Graphical representation makes the resulting patterns easier to understand.

User behavior results can be represented using bar charts, pie charts, and activity-distribution graphs. These visualizations can show posting frequency, engagement levels, and the distribution of activity across different periods of the day.

The sentiment module produces corresponding visualizations for positive, negative, and neutral content. Sentiment distribution graphs allow the overall opinion within the analyzed dataset to be examined more easily.

The combined visual outputs can reveal peak activity periods, engagement patterns, dominant sentiment categories, and other trends. Analytical reports generated from these results can be used to support interpretation and further decision-making.



Fig. 6. Visualization of analysis results.

III. RESULTS AND DISCUSSION

The developed framework was used to analyze social media information from two perspectives: user behavior and sentiment. The processing pipeline handled the collected data, extracted interaction-related characteristics, and classified textual information into positive, negative, and neutral categories.

A. User Behavior Analysis Results

The user behavior module examined activity using timestamps, posting frequency, and engagement-related information. The purpose was to determine how activity was distributed across different periods.

The analysis showed that user participation was not uniform throughout the day. Certain time intervals contained higher levels of activity, indicating periods of increased engagement. Bar charts were used to represent posting activity, while pie charts were used to show the distribution of activity across the selected time categories.

IARJSET ISSN (O) 2393-8021, ISSN (P) 2394-1588

International Advanced Research Journal in Science, Engineering and Technology

Impact Factor 8.311 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 13, Issue 10, October 2026

DOI: 10.17148/IARJSET.2026.131051

These observations provide a simple way to identify periods in which users are more active. Such information can be useful when studying engagement patterns and determining suitable periods for content delivery or further analysis.



B. Sentiment Analysis Results

The sentiment classification module used TF-IDF to represent textual data and Multinomial Naive Bayes to classify the resulting feature vectors. The classifier assigned social media content to positive, negative, or neutral categories.

The model was evaluated using accuracy, precision, recall, F1-score, and a confusion matrix. The obtained classification accuracy was approximately 60%. This indicates that the model was able to identify general sentiment patterns, although its performance was moderate.

Some classification errors were observed in cases involving ambiguous wording, sarcasm, and context-dependent expressions. These characteristics are common in social media text and are difficult to capture using a relatively simple text classification approach.



C. Visualization Results

The system generated graphical representations of the analytical results. The main outputs included:

- Bar charts representing daily posting activity.
- Pie charts showing the distribution of user activity across different periods.
- Sentiment distribution graphs representing positive, negative, and neutral content.
- Analytical reports summarizing the observed patterns.

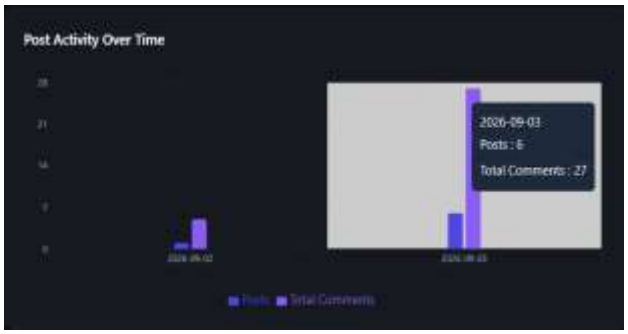
The visual outputs provide a more accessible representation of the processed information and make it easier to compare activity and sentiment trends.

IARJSET ISSN (O) 2393-8021, ISSN (P) 2394-1588

International Advanced Research Journal in Science, Engineering and Technology

Impact Factor 8.311 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 13, Issue 10, October 2026

DOI: 10.17148/IARJSET.2026.131051



D. Discussion

The results indicate that a combination of Big Data technologies and NLP techniques can provide a practical framework for analyzing large collections of social media data. Hadoop provides distributed storage for the dataset, while Spark supports the processing required for behavioral and sentiment analysis.

The user behavior module was able to identify differences in activity across time periods and provide information about engagement patterns. This information adds another dimension to the analysis beyond sentiment classification.

The sentiment classifier produced an accuracy of approximately 60%. While this demonstrates that TF-IDF and Multinomial Naive Bayes can identify broad sentiment patterns, the result also shows that there is scope for improvement. Social media language frequently contains slang, abbreviations, sarcasm, emojis, and expressions whose meaning depends heavily on context.

More advanced approaches such as CNN, LSTM, or Transformer-based models could be investigated in future work to improve sentiment classification. Such approaches may provide better handling of complex linguistic patterns, although they would also require additional computational resources and suitable training data.

Overall, combining user behavior mining and sentiment analysis provides a broader view of social media activity. Instead of examining only the sentiment of individual posts, the framework also considers how users interact and when activity occurs.

IV. CONCLUSION

This paper presented a Big Data-based framework for social media sentiment analysis and user behavior mining. The framework combines distributed storage and processing with NLP and machine learning techniques to analyze large volumes of social media information.

Apache Hadoop was used to provide distributed storage, while Apache Spark supported large-scale data processing. The preprocessing stage removed unnecessary elements from the collected textual data and prepared it for machine learning. TF-IDF was then used to convert text into numerical feature representations, and Multinomial Naive Bayes was applied to classify the content into positive, negative, and neutral sentiment categories.

The user behavior component examined posting frequency, engagement, interaction patterns, and activity across different periods of the day. These results provided additional information about how users participate on social media platforms. Charts and graphical reports were used to present both behavioral and sentiment-related findings in an interpretable form. The sentiment classifier achieved an approximate accuracy of 60%, showing that the selected approach can identify general sentiment patterns but still has limitations. Ambiguous language, sarcasm, slang, emojis, and contextual expressions remain challenging for the classification model.

Future work can focus on improving the classification stage using advanced deep learning and Transformer-based models. Other possible extensions include multilingual sentiment analysis, real-time processing, and fake-news detection. These additions could broaden the capabilities of the framework and improve its usefulness for large-scale social media analytics.

IARJSET ISSN (O) 2393-8021, ISSN (P) 2394-1588

International Advanced Research Journal in Science, Engineering and Technology

Impact Factor 8.311 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 13, Issue 10, October 2026

DOI: 10.17148/IARJSET.2026.131051

ACKNOWLEDGMENT

The authors would like to express their sincere gratitude to Sheeba S, Assistant Professor, Department of Computer Science and Engineering, Don Bosco Institute of Technology, Bengaluru, for her guidance and valuable suggestions throughout this work. Her feedback and technical support contributed significantly to the development and completion of the project.

The authors also thank the Department of Computer Science and Engineering, Don Bosco Institute of Technology, Bengaluru, for providing the infrastructure and resources required to carry out the work. The support and cooperation received from faculty members and peers are also gratefully acknowledged.
Finally, the authors thank their families for their continued encouragement and support throughout the development of this work.

REFERENCES

[1] V. S. Vedhika Srinidhi et al., "Utilizing Big Data Analytics for Social Media Insights," 2024.
[2] K. Srinidhi et al., "Sentiment Analysis of Social Media Comments using Natural Language Processing," 2024.
[3] "Social Media Big Data and Pattern Mining." IEEE Research Paper, 2024.
[4] "Sentiment Analysis of Telecommunication Providers on Twitter Using CNN," 2023.
[5] Fei Fei, Weng Liguo, Shou Ting, et al. Research on Classification of Behavioral Characteristics of Power Users Based on Data Mining [J]. Automation Technology and Applications, 2023, 42 (11): 101-104.
[6] Mao Qun, Wang Weiwei, You Feng, Zhao Ruilian, Li Zheng. User behavior pattern mining and reuse methods for Android applications [J]. Computer Applications, 2022, 42 (7): 2155-2161.
[7] Shi Houbing. Mobile communication user behavior analysis under core network data mining [J]. Communication Power Supply Technology, 2023, 40 (9): 145-147.
[8] Liu Yi, Gao Ming, Tian Liangliang, et al. Interactive visualization analysis of network user behavior based on web log mining [J]. Journal of Yan'an University: Natural Science Edition, 2023, 42 (3): 78-85.
[9] Jain P K, Saravanan V, Pamula R. A hybrid CNN-LSTM: A deep learning approach for consumer sentiment analysis using qualitative user-generated contents[J]. Transactions on Asian and Low-Resource Language Information Processing, 2021, 20(5): 1-15.
[10] Pandian A P. Performance evaluation and comparison using deep learning techniques in sentiment analysis[J]. Journal of Soft Computing Paradigm, 2021, 3(2): 123-134.
[11] Khan R, Shrivastava P, Kapoor A, et al. Social media analysis with AI: sentiment analysis techniques for the analysis of twitter covid-19 data[J]. J. Crit. Rev, 2020, 7(9): 2761-2774.
[12] Alqaryouti O, Siyam N, Abdel Monem A, et al. Aspect-based sentiment analysis using smart government review data[J]. Applied Computing and Informatics, 2024, 20(1/2): 142-161.
[13] Mehta P, Pandya S. A review on sentiment analysis methodologies, practices and applications[J]. International Journal of Scientific and Technology Research, 2020, 9(2): 601-609.
[14] Kumar S, De K, Roy P P. Movie recommendation system using sentiment analysis from microblogging data[J]. IEEE Transactions on Computational Social Systems, 2020, 7(4): 915-923.
[15] Isnain A R, Supriyanto J, Kharisma M P. Implementation of K-Nearest Neighbor (K-NN) Algorithm For Public Sentiment Analysis of Online Learning[J]. IJCCS (Indonesian Journal of Computing and Cybernetics Systems), 2021, 15(2): 121-130.
[16] Kondapally, M., Kumar, K.N. and Mohan, C.K., 2024, June. Object Detection in Transitional Weather Conditions for Autonomous Vehicles. In 2024 International Joint Conference on Neural Networks (IJCNN) (pp. 1-8). IEEE.
[17] Koniammal, R., Ramesh, G.P. and Naidu, G.G., 2023, July. Design a Multiband Frequency of Partial Defected Ground for Hexagonal Fractal Antenna. In 2023 International Conference on Data Science and Network Security (ICDSNS) (pp. 1-7). IEEE.
[18] Qing W Z D .Keyword Extraction from Scientific Research Projects Based on SRP-TF-IDF[J].Electronic Journal: English edition, 2021, 030(004):P.652-657.
[19] Li Y, Shen Y, Chen L, et al. Zebra: When temporal graph neural networks meet temporal personalized PageRank[J]. Proceedings of the VLDB Endowment, 2023, 16(6): 1332-1345.