Cover Page

INTERNATIONAL JOURNAL OF MULTIDISCIPLINARY EDUCATIONAL RESEARCH
ISSN:2277-7881(Print); IMPACT FACTOR :10.16(2026); IC VALUE:5.16; ISI VALUE:2.286
Peer-Peer Reviewed, Refereed & Open Access International Journal - As per UGC Norms
(Fulfilled Suggests Parameters of UGC by IJMER)
Volume:15, Issue:6(3), June 2026
Scopus Review ID: A2B96D3ACF3FEA2A
Article Received: Reviewed: Accepted
Publisher: Sucharitha Publication, India
Online Copy of Article Publication Available: www.ijmer.in

LOCATION-BASED SOCIAL MEDIA ANALYTICS USING HYBRID LDA–PLSA TOPIC MODELING AND USER INTEREST ANALYSIS

Ravi kumar Kuchipudi [1] and Dr V.V. Jaya Rama Krishnaiah[2]
[1]Research Scholar, Department of Computer Science and Engineering, Acharaya Nagarjuna University, Guntur
[2]Department of Computer Science and Engineering, Koneru Lakshmaiah Education Foundation, Vaddeswaram, India

Abstract:
The rapid growth of social networking platforms has transformed the way people communicate and share information across the world. Social media platforms generate massive amounts of data daily, creating opportunities for extracting valuable insights and understanding user behavior. Popular platforms such as Twitter and Facebook have become important sources of information where users express opinions, share experiences, and interact with communities. Social network analysis plays a significant role in identifying relationships, discovering patterns, and understanding interactions among users. This paper presents a geographical-based approach for analyzing Twitter and Facebook profiles and activities. The proposed system enables users to select specific geographic locations and collect social media data from those regions. By analyzing tweets, comments, and user interactions, the system identifies user interests and behavior patterns. The study focuses on applying big data analytics techniques to extract meaningful information and support effective social media analysis.
Keywords: Social Network Analysis, Topic Modeling, LDA, PLSA, Social Media Mining, Geographic Analysis, Topic Clustering, Big Data Analytics

1. INTRODUCTION

Social networks are web-based platforms that enable users to create public or semi-public profiles and interact with other users within an online environment. Social networking platforms such as Facebook, Twitter, LinkedIn, and Instagram have become an integral part of modern communication by allowing users to connect, share information, and express opinions on various topics. These topics may include politics, technology, entertainment, religion, education, and many other areas of daily life.

The rapid increase in the number of social media users has resulted in the generation of large volumes of data, creating opportunities for extracting meaningful insights. Although Facebook and Twitter serve similar communication purposes, their interaction patterns differ. Facebook primarily focuses on maintaining social relationships among family members and friends, while Twitter enables users to follow individuals and discussions based on shared interests. Both platforms provide users with different privacy and communication features that support information exchange and social interaction.

Social media data can be utilized to identify user interests, analyze behavioral patterns, and understand community trends. This study focuses on extracting location-based information from Twitter and Facebook platforms to analyze user activities across different geographic regions. The proposed approach collects and analyzes posts, tweets, and user interactions to identify significant topics and patterns. Keyword extraction techniques are employed to identify important terms from text data, while topic modeling approaches such as Latent Dirichlet Allocation (LDA) are used to determine user interests within specific locations. Furthermore, supervised and unsupervised learning methods are applied to cluster related information and generate network representations for social media analysis.

2. RELATED WORK

2.1 Data Retrieval

Data retrieval plays an important role in social media analysis because the quality of collected data directly influences the effectiveness of the analysis process. Various Application Programming Interfaces (APIs) are available to extract data from social networking platforms. In this study, API4J is utilized to collect data from Twitter and Facebook platforms. Twitter data are collected by retrieving tweets from users within specific geographical locations selected by the user. Similarly, Facebook data are obtained by extracting user posts, comments, and related information based on the selected geographic regions. The collected data serve as the input for further processing and analysis.

Cover Page

INTERNATIONAL JOURNAL OF MULTIDISCIPLINARY EDUCATIONAL RESEARCH
ISSN:2277-7881(Print); IMPACT FACTOR :10.16(2026); IC VALUE:5.16; ISI VALUE:2.286
Peer-Peer Reviewed, Refereed & Open Access International Journal - As per UGC Norms
(Fulfilled Suggests Parameters of UGC by IJMER)
Volume:15, Issue:6(3), June 2026
Scopus Review ID: A2B96D3ACF3FEA2A
Article Received: Reviewed: Accepted
Publisher: Sucharitha Publication, India
Online Copy of Article Publication Available: www.ijmer.in

2.2 Topic Identification

Topic identification is an important process for discovering popular and frequently discussed subjects in social media data. Latent Dirichlet Allocation (LDA) and Probabilistic Latent Semantic Analysis (PLSA) are commonly used topic modeling techniques. LDA is a probabilistic generative model used to identify hidden topics within a collection of documents, while PLSA is a statistical approach that discovers relationships between documents and terms. These techniques help identify trending topics and user interests from social media content. However, temporal information is often not fully considered, even though time-based information plays an important role in identifying evolving social media trends.

2.3 Keyword Extraction

Keyword extraction is a technique used to identify significant terms or phrases from text documents. Numerous supervised and unsupervised methods have been proposed for extracting informative keywords from textual data. Unsupervised approaches rely on statistical patterns and implicit relationships present within the document corpus, whereas supervised methods use pre-labeled datasets for training and classification purposes. Techniques such as KEA and GenEx are well-established supervised frameworks used for effective keyword extraction. Recently, advanced approaches based on neural networks and machine learning have also been introduced to improve the accuracy of identifying relevant keywords and phrases from large datasets.

2.4 Co-Occurrence Similarity

Co-occurrence similarity refers to the relationship between terms that frequently appear together within documents. This approach helps determine the association between words and identify significant keywords. If certain terms repeatedly occur together in similar contexts, they are likely to represent related concepts or topics. Co-occurrence similarity is commonly used to measure the strength of relationships between keywords and to improve topic identification. Techniques such as Co-Occurrence Double Checking (CODC) utilize occurrence frequency and text relationships to determine how closely terms or entities are connected within large datasets.

3. IMPLEMENTATION

The proposed system is designed to analyze social media data collected from Twitter and Facebook platforms based on geographical regions and user interests. The implementation process includes data extraction, preprocessing, topic modeling, keyword extraction, similarity measurement, outlier detection, and social network analysis. The overall workflow of the proposed approach is illustrated in Figure 3.1.

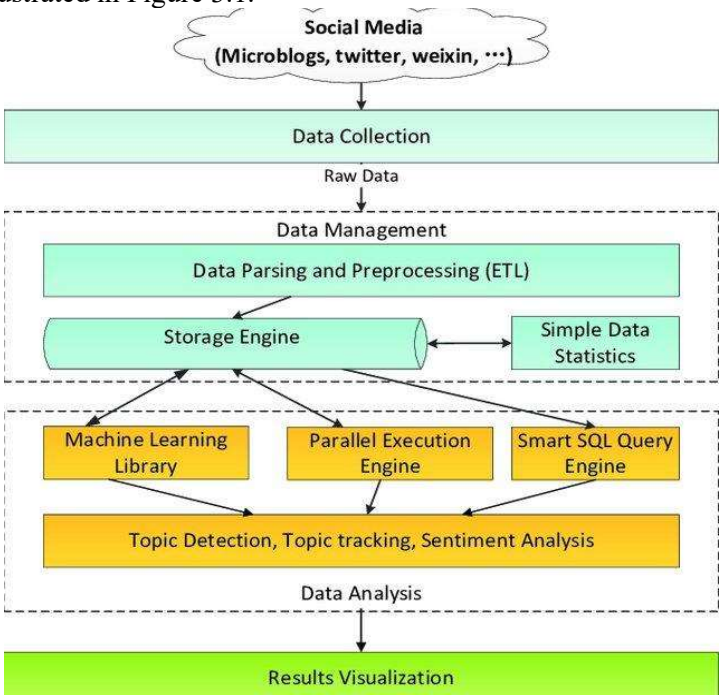

Figure 1: General Structure of the Proposed Approach

Cover Page

INTERNATIONAL JOURNAL OF MULTIDISCIPLINARY EDUCATIONAL RESEARCH
ISSN:2277-7881(Print); IMPACT FACTOR :10.16(2026); IC VALUE:5.16; ISI VALUE:2.286
Peer-Peer Reviewed, Refereed & Open Access International Journal - As per UGC Norms
(Fulfilled Suggests Parameters of UGC by IJMER)
Volume:15, Issue:6(3), June 2026
Scopus Review ID: A2B96D3ACF3FEA2A
Article Received: Reviewed: Accepted
Publisher: Sucharitha Publication, India
Online Copy of Article Publication Available: www.ijmer.in

3.1 Extraction of Twitter Profiles

The first stage of implementation involves selecting a target user community and geographic location for analysis. Users are required to provide a search keyword and select a geographic region from which tweets need to be collected. Information regarding various locations and region details is stored in a CSV file and utilized during data extraction.
Twitter provides public APIs that allow access to publicly available tweets and user profile information. In this work, the Twitter4J library is employed to establish communication with Twitter services. This library provides an efficient mechanism for integrating Twitter functionality into Java applications. Tweets are retrieved based on the user-defined search term and selected geographic area. Subsequently, user profile details and related tweet information are collected for further processing and analysis.

3.2 Extraction of Facebook Profiles

The Facebook profile extraction process is performed using third-party Java API components designed for social media data collection. The system retrieves user posts, comments, and related interactions based on selected geographic regions. Approximately 2000 posts collected from different users are considered as the training dataset for analysis. In addition, live news stream data are utilized to improve classification accuracy using multiple classifiers.

3.3 Preprocessing

Preprocessing is an essential step in improving data quality and reducing irrelevant information from collected social media content. Raw data collected from Twitter and Facebook may contain noisy information, duplicated entries, and unnecessary words. Therefore, preprocessing operations are performed to prepare the data for effective analysis.
The preprocessing stage includes:
• Removal of stop words
• Data filtering and cleaning
• Elimination of irrelevant terms
These operations reduce computational complexity and improve the quality of extracted information.

3.4 Topic Modeling

Topic modeling is employed to identify hidden semantic structures and meaningful patterns within social media text data. It is a text-mining technique used to discover frequently occurring themes and relationships among words in a document collection. The proposed system utilizes multiple topic modeling methods to identify topics from tweets and posts.
The topic modeling techniques used include:

- Latent Dirichlet Allocation (LDA)
- Probabilistic Latent Semantic Analysis (PLSA)

Although many irrelevant terms are removed during preprocessing, some terms and co-occurrence relationships may still exist within the generated graph. Therefore, only terms with high Quantitative Similarity (QS) values are retained while less significant terms are filtered.

3.4.1 Keyword Extraction

Keyword extraction is performed to identify important terms from social media posts and tweets. The process begins with collecting tweets and Facebook posts based on user-defined keywords. The extracted textual content forms the corpus used for analysis.

Cover Page

INTERNATIONAL JOURNAL OF MULTIDISCIPLINARY EDUCATIONAL RESEARCH
ISSN:2277-7881(Print); IMPACT FACTOR :10.16(2026); IC VALUE:5.16; ISI VALUE:2.286
Peer-Peer Reviewed, Refereed & Open Access International Journal - As per UGC Norms
(Fulfilled Suggests Parameters of UGC by IJMER)
Volume:15, Issue:6(3), June 2026
Scopus Review ID: A2B96D3ACF3FEA2A
Article Received: Reviewed: Accepted
Publisher: Sucharitha Publication, India
Online Copy of Article Publication Available: www.ijmer.in

For example, when the keyword "movies" is selected, the system collects all tweets and related posts containing the selected term. Twitter and Facebook APIs are used for retrieving online content and identifying relevant keywords.

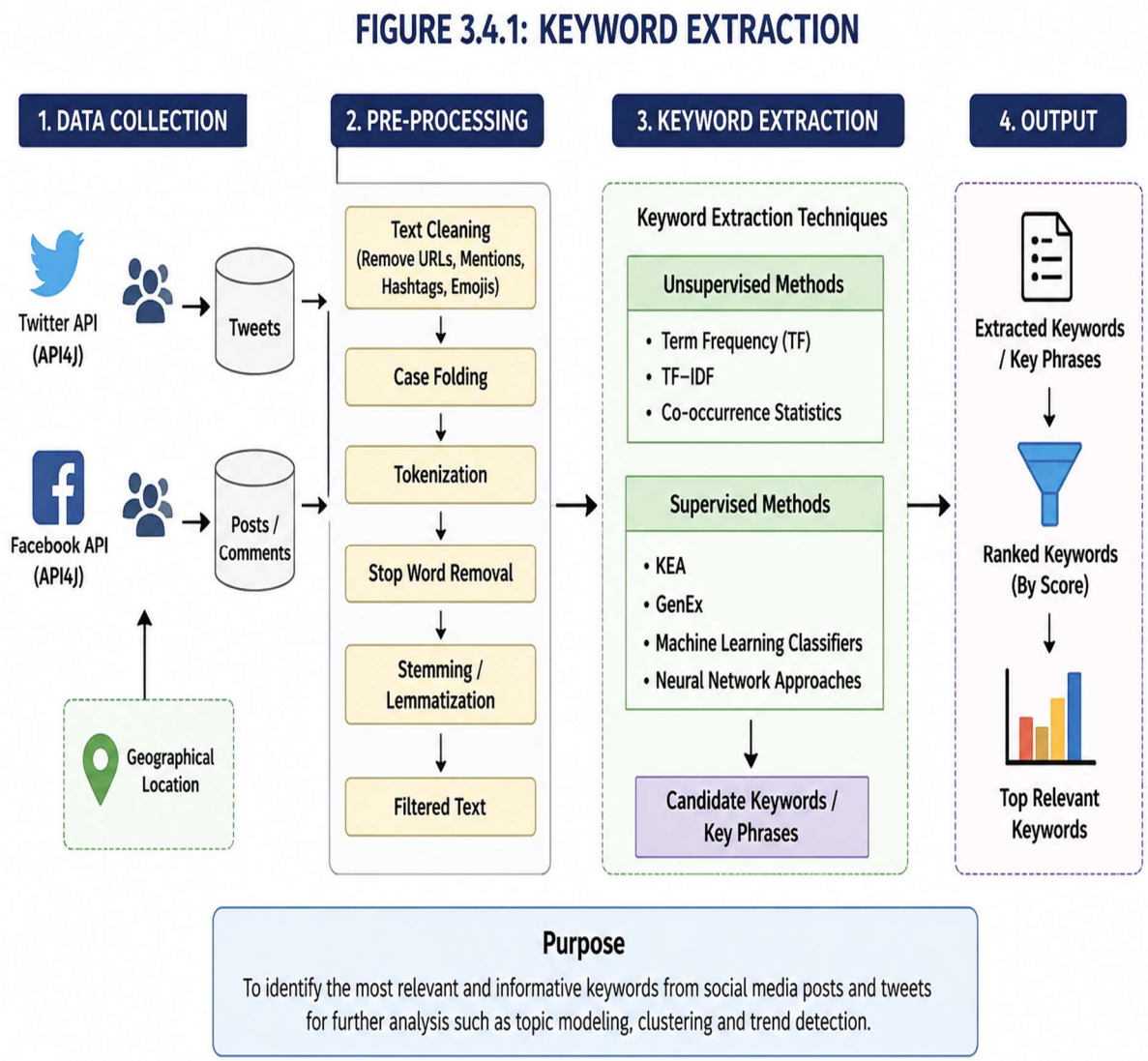

Figure 2: Keyword Extraction Process

3.4.2 Keyword Similarity and Topic Clustering

Keyword similarity analysis is used to identify relationships among terms extracted from Facebook and Twitter profiles. Similarity measures help understand how frequently keywords occur together and indicate common user interests and behavioral patterns. Related terms are grouped together using topic clustering techniques to identify significant themes within social media content.

3.4.3 Outlier Detection

Outlier detection is performed to identify unusual patterns or irregular observations within the dataset. The user can define a threshold value through the sigma parameter for identifying outliers. The output of this process includes a list of detected outliers along with graphical representations showing the distribution density of analyzed data samples.

Cover Page

INTERNATIONAL JOURNAL OF MULTIDISCIPLINARY EDUCATIONAL RESEARCH
ISSN:2277-7881(Print); IMPACT FACTOR :10.16(2026); IC VALUE:5.16; ISI VALUE:2.286
Peer-Peer Reviewed, Refereed & Open Access International Journal - As per UGC Norms
(Fulfilled Suggests Parameters of UGC by IJMER)
Volume:15, Issue:6(3), June 2026
Scopus Review ID: A2B96D3ACF3FEA2A
Article Received: Reviewed: Accepted
Publisher: Sucharitha Publication, India
Online Copy of Article Publication Available: www.ijmer.in

Traditional clustering methods may not efficiently process continuously changing social media content. Therefore, the proposed system employs a hybrid approach combining supervised and unsupervised learning techniques to process streaming data efficiently and generate meaningful models.

3.5 Social Network Analysis Based on User Interests

After clustering user interest data, social network analysis is performed to identify relationships among users and their areas of interest. The generated network graph represents the interactions and similarities among users. The system creates a URL associated with the generated graph and sends it to a specified email address, allowing users to access and visualize the network structure.

4. RESULTS

The proposed system was implemented to analyze social media data collected from Twitter and Facebook platforms based on geographic regions and user interests. Experimental analysis was performed on approximately 2000 user posts and tweets collected from different locations. The extracted data were processed using preprocessing techniques, topic modeling methods, and keyword extraction approaches to identify meaningful patterns and user interests.

The preprocessing stage successfully removed irrelevant terms, stop words, and noisy information from the collected social media content, thereby improving the quality of the dataset. The keyword extraction process effectively identified frequently occurring and relevant terms from tweets and Facebook posts. Keywords related to user interests, current trends, and regional discussions were extracted and ranked according to their significance.

The LDA and PLSA topic modeling techniques generated clusters of related topics from the extracted content. Similar topics and user interests were grouped together based on term similarity and co-occurrence relationships. The clustering process enabled the identification of major discussion themes across different geographic regions.

The generated social network graph provided a visual representation of relationships among users and their common interests. The results indicate that the proposed approach effectively supports location-based social media analysis and improves the identification of user interests and topic patterns.

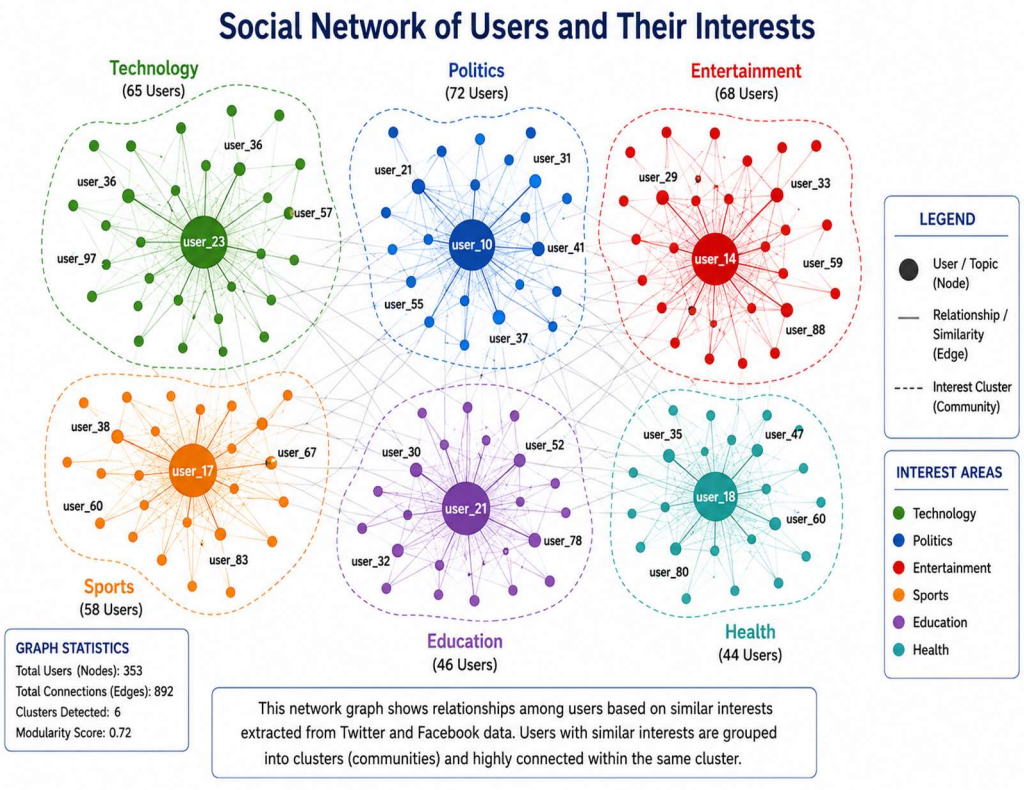

Figure 3: Graph Based on User Interested Area

Cover Page

INTERNATIONAL JOURNAL OF MULTIDISCIPLINARY EDUCATIONAL RESEARCH
ISSN:2277-7881(Print); IMPACT FACTOR :10.16(2026); IC VALUE:5.16; ISI VALUE:2.286
Peer-Peer Reviewed, Refereed & Open Access International Journal - As per UGC Norms
(Fulfilled Suggests Parameters of UGC by IJMER)
Volume:15, Issue:6(3), June 2026
Scopus Review ID: A2B96D3ACF3FEA2A
Article Received: Reviewed: Accepted
Publisher: Sucharitha Publication, India
Online Copy of Article Publication Available: www.ijmer.in

Figure 3 illustrates the network graph generated based on users' areas of interest extracted from Twitter and Facebook data. The graph represents relationships among users and their associated interests using interconnected nodes and edges. Each node represents a user or a topic of interest, while the edges indicate the connections or similarities among them. Users having similar interests are grouped together, forming clusters within the network. This graphical representation helps in understanding interaction patterns and identifying communities with common interests. The results demonstrate that users with similar preferences tend to form closely connected groups, allowing effective social network analysis.

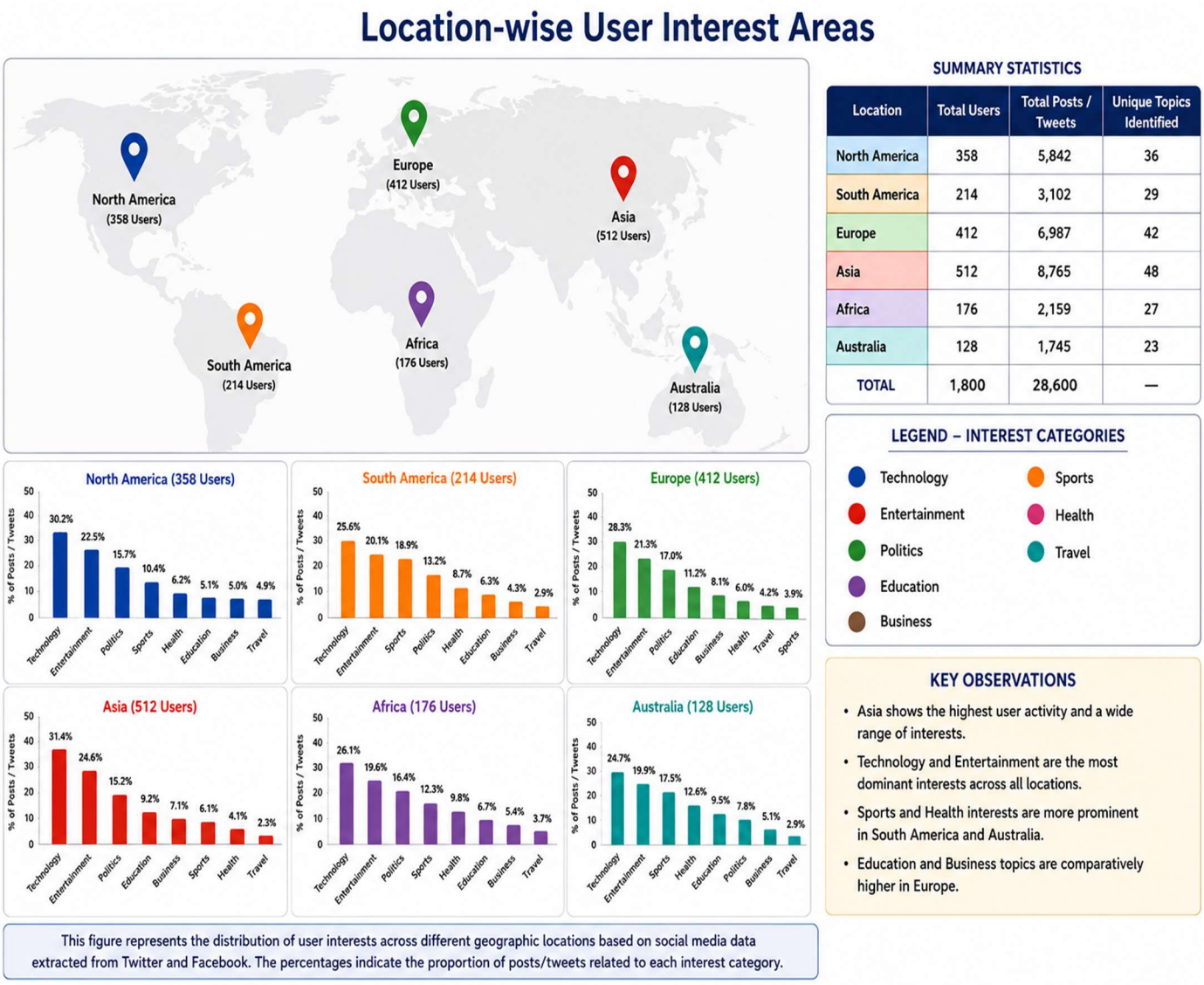

Figure 4: Location-wise User Interest Areas

Figure 4 presents the distribution of user interests across different geographic locations. The figure shows how users from different regions exhibit varying preferences and discussion topics based on their activities on social media platforms. The extracted tweets, comments, and posts are analyzed to identify frequently discussed subjects in each location. The results indicate that user interests differ across regions and can be grouped into distinct categories. This location-based analysis assists in understanding regional trends and identifying the most active topics within a particular geographic area.

Cover Page

INTERNATIONAL JOURNAL OF MULTIDISCIPLINARY EDUCATIONAL RESEARCH
ISSN:2277-7881(Print); IMPACT FACTOR :10.16(2026); IC VALUE:5.16; ISI VALUE:2.286
Peer-Peer Reviewed, Refereed & Open Access International Journal - As per UGC Norms
(Fulfilled Suggests Parameters of UGC by IJMER)
Volume:15, Issue:6(3), June 2026
Scopus Review ID: A2B96D3ACF3FEA2A
Article Received: Reviewed: Accepted
Publisher: Sucharitha Publication, India
Online Copy of Article Publication Available: www.ijmer.in

5. Conclusion

The rapid growth of social networking platforms has resulted in the generation of large volumes of user-generated data, creating opportunities for meaningful social media analysis. This paper presented a location-based approach for analyzing Twitter and Facebook data to identify user interests, trending topics, and interaction patterns across different geographic regions. The proposed system utilized social media APIs to collect data and applied preprocessing techniques to improve data quality before analysis.

Topic modeling techniques such as Latent Dirichlet Allocation (LDA) and Probabilistic Latent Semantic Analysis (PLSA) were employed to identify hidden topics and extract meaningful information from social media content. In addition, keyword extraction and co-occurrence similarity methods were used to discover relationships among terms and improve topic identification. The generated network graphs and clustering results provided insights into user communities and common interests.

The experimental results demonstrate that the proposed framework effectively supports social media analytics and user interest identification. Future work may focus on incorporating real-time streaming data and advanced machine learning techniques to improve prediction accuracy and scalability.

References

[1] D. M. Boyd and N. B. Ellison, "Social network sites: Definition, history, and scholarship," *Journal of Computer-Mediated Communication*, vol. 13, no. 1, pp. 210–230, 2007.
[2] T. Hofmann, "Probabilistic Latent Semantic Analysis," in *Proceedings of the Fifteenth Conference on Uncertainty in Artificial Intelligence*, Stockholm, Sweden, 1999, pp. 289–296.
[3] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent Dirichlet Allocation," *Journal of Machine Learning Research*, vol. 3, pp. 993–1022, Jan. 2003.
[4] M. Hu and B. Liu, "Mining and Summarizing Customer Reviews," in *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Seattle, WA, USA, 2004, pp. 168–177.
[5] C. C. Aggarwal, *Social Network Data Analytics*. Boston, MA, USA: Springer, 2011.
[6] J. Han, M. Kamber, and J. Pei, *Data Mining: Concepts and Techniques*, 3rd ed. Burlington, MA, USA: Morgan Kaufmann, 2011.
[7] H. Kwak, C. Lee, H. Park, and S. Moon, "What is Twitter, a Social Network or a News Media?" in *Proceedings of the 19th International Conference on World Wide Web (WWW)*, Raleigh, NC, USA, 2010, pp. 591–600.
[8] T. O'Reilly, "What Is Web 2.0: Design Patterns and Business Models for the Next Generation of Software," *Communications and Strategies*, vol. 65, pp. 17–37, 2007.
[9] S. Brin and L. Page, "The Anatomy of a Large-Scale Hypertextual Web Search Engine," *Computer Networks and ISDN Systems*, vol. 30, no. 1–7, pp. 107–117, 1998.
[10] I. H. Witten, G. W. Paynter, E. Frank, C. Gutwin, and C. G. Nevill-Manning, "KEA: Practical Automatic Keyphrase Extraction," in *Proceedings of the Fourth ACM Conference on Digital Libraries*, Berkeley, CA, USA, 1999, pp. 254–255.
[11] P. Turney, "Learning Algorithms for Keyphrase Extraction," *Information Retrieval*, vol. 2, no. 4, pp. 303–336, 2000.
[12] B. Pang and L. Lee, "Opinion Mining and Sentiment Analysis," *Foundations and Trends in Information Retrieval*, vol. 2, no. 1–2, pp. 1–135, 2008.