Article Recommender System: Machine Learning and NLP-Based Personalized Recommendation Engine

[1]P. Narendra, [2]P. Baby, [3]S. Jyoshna, [4]K. Karthik, [5]K. Kalyani
[1234]*UG Scholars, [5]Professor, Srinivasa Institute of Engineering and Technology*
doi.org/10.64643/IJIRTV12I11-197501-459

***Abstract*— The rapid growth of digital publishing platforms has resulted in significant information overload, making it difficult for users to efficiently discover relevant content. This research presents a Machine Learning and Natural Language Processing (NLP)-based personalized Article Recommender System. The proposed system utilizes text preprocessing techniques, TF-IDF vectorization for feature extraction, K-Nearest Neighbour (KNN) algorithm for similarity-based candidate selection, and cosine similarity for ranking relevant articles. A user interest profile is generated based on reading history, enabling personalized recommendations. Experimental results demonstrate that the system effectively identifies relevant articles and improves content discovery efficiency. The proposed framework provides ascalable and computationally efficient solution for personalized recommendation systems.**

***Keywords*— Article Recommendation, NLP, TF-IDF, KNN, Cosine Similarity, Machine Learning, Personalized Recommendation**

I. INTRODUCTION

Article Recommendation Systems have become an important and rapidly growing research area in the field of Natural Language Processing (NLP) and Machine Learning. The primary objective of an article recommender system is to automatically analyze textual content and suggest relevant articles to users based on their interests and reading behavior, With the exponential growth of digital content on online platforms, users often face information overload, making it difficult to identify useful and personalized content.

As a result, intelligent recommendation systems have become essential for efficient information retrieval and user engagement

Article recommendation systems have numerous practical applications, including news platforms, research databases, e-learning systems, blogging websites, and digital libraries. In news applications, recommendation systems help users discover trending and relevant articles. In academic platforms, they assist researchers in finding related research papers. In e-learning environments, they suggest study materials based on user preferences. Therefore, developing an accurate and personalized article recommendation system is both technologically significant and commercially valuable.

Traditional recommendation approaches relied heavily on keyword-based search and manual filtering techniques. These systems often failed to capture deeper semantic relationships between documents and user preferences. Some earlier machine learning methods used basic Bag-of-Words representations and simple similarity matching techniques, which required manual feature engineering and lacked scalability. Such approaches struggled to handle large-scale datasets and dynamic user interests effectively.

With the advancement of Natural Language Processing techniques, modern recommendation systems utilize automated feature extraction methods such as Term Frequency–Inverse Document Frequency (TF-IDF) and similarity-based algorithms. TF-IDF efficiently converts textual documents into numerical vector representations, allowing machines to interpret and compare textual data. To further improve recommendation quality, similarity-based learning algorithms such as K-Nearest Neighbor (KNN) are employed to identify closely related documents in high-dimensional vector space. Since article recommendation depends on

analyzing textual relevance and user preference similarity, combining TF-IDF vectorization with KNN and cosine similarity provides a powerful and efficient framework. TF-IDF captures the importance of words within documents, KNN identifies the most relevant neighboring articles, and cosine similarity measures the angular similarity between user and

article vectors to rank recommendations accurately. In this work, we propose a personalized Article Recommender System using TF-IDF, K-Nearest Neighbor (KNN), and cosine similarity techniques.

II. LITERATURE SURVEY

Article recommendation and information retrieval systems have been widely studied in the field of Natural Language Processing and Machine Learning due to their importance in digital content management and personalized user experience. Early recommendation approaches were primarily based on traditional information retrieval techniques such as keyword matching and Boolean search models. These systems relied on exact word matching between user queries and document content, which often failed to capture semantic relationships and contextual meaning.

With the growth of digital platforms, collaborative filtering methods gained popularity. Collaborative filtering recommends content based on user behavior patterns, ratings, and preferences of similar users. Although effective in many applications, collaborative filtering suffers from limitations such as the cold-start problem, data sparsity, and dependency on large user interaction datasets. New users or new articles often receive poor recommendations due to insufficient historical data to overcome these limitations, content-based filtering approaches were introduced.

Content-based systems analyze the textual features of articles and recommend similar documents based on content similarity. Techniques such as Bag-of-Words and Term Frequency–Inverse Document Frequency (TF-IDF) became widely adopted for representing textual data numerically. TF-IDF assigns weights to words based on their importance within a document and across the corpus, making it a powerful feature extraction method for document similarity analysis.

Similarity-based algorithms such as K-Nearest Neighbor (KNN) have been extensively used in recommendation systems for identifying similar documents in high-dimensional vector spaces. KNN is a simple yet effective algorithm that selects the closest neighbors based on distance metrics. When combined with cosine similarity, it provides an efficient method for ranking documents according to their relevance. Cosine similarity measures the angular distance between vectors and is particularly suitable for high-dimensional sparse data such as TF-IDF representations.

Recent advancements in recommendation systems include the use of deep learning-based word embeddings and transformer models. However, these approaches often require high computational resources and large-scale training data. For practical and scalable implementations, TF-IDF combined with similarity-based methods remains a computationally efficient and interpretable solution.

Based on these research developments, the proposed project adopts a content-based filtering approach using TF-IDF for feature extraction, K-Nearest Neighbor (KNN) for similarity-based candidate selection, and cosine similarity for ranking relevant articles. This combination ensures efficient, scalable, and accurate personalized article recommendations.

III. SYSTEM ARCHITECTURE

The proposed Article Recommender System is designed using a content-based filtering framework that integrates Natural Language Processing techniques with similarity-based machine learning algorithms. The architecture is structured to process raw textual data, extract meaningful features, model user interests, and generate personalized article recommendations efficiently. The system consists of multiple sequential modules that transform unstructured text into ranked recommendations.

The system begins with the Article Dataset Acquisition stage, where a collection of news or research articles is gathered. Each article typically contains textual fields such as title and content. Since raw text cannot be directly processed by machine learning models, the data undergoes preprocessing before feature extraction.

In the Text Preprocessing stage, the collected articles are cleaned and standardized. This step includes tokenization, lowercasing, removal of punctuation, elimination of stop words, and lemmatization. These preprocessing operations reduce noise and ensure that only meaningful textual information is retained. The cleaned text is then prepared for numerical representation.

Following preprocessing, the system performs

Feature Extraction using Term Frequency–Inverse Document Frequency (TF-IDF). TF-IDF converts textual documents into high-dimensional numerical vectors by assigning weights to words based on their importance within a document and across the entire dataset. This transformation allows articles to be represented in vector space, enabling similarity comparison and mathematical analysis.

After vectorization, the system constructs a User Interest Profile. The user profile is generated by aggregating TF-IDF vectors of previously read or interacted articles. This aggregated vector represents the user's preferences in the same feature space as the article dataset. By mapping both users and articles into a shared vector space, similarity computation becomes feasible.

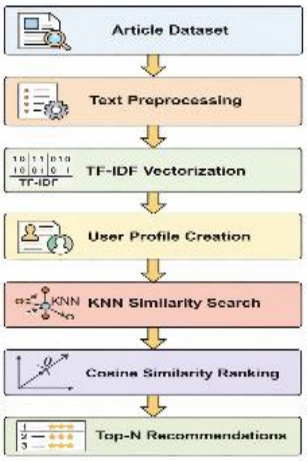


Figure 1: System Architecture of Proposed Article Recommender System

To identify relevant candidate articles, the system employs the K-Nearest Neighbor (KNN) algorithm. KNN computes the distance between the user profile vector and all article vectors in the dataset. Based on a selected value of K, the algorithm retrieves the K most similar articles that are closest to the user profile in vector space. This step ensures efficient selection of potential recommendations.

After candidate selection, Cosine Similarity is applied to measure the angular similarity between the user vector and each candidate article vector. Cosine similarity evaluates how closely aligned two vectors are, regardless of their magnitude. Articles with similarity scores closer to 1 indicate higher relevance to the user's interests. The articles are then ranked based on their similarity scores.

Finally, the system generates Top-N Personalized Recommendations by selecting the highest-ranked articles. These recommendations are displayed to the user as relevant "Next-Read" suggestions. The modular structure of the architecture ensures scalability and efficient handling of large textual datasets while maintaining computational simplicity.

Overall, the proposed system architecture effectively integrates text preprocessing, feature extraction, similarity-based learning, and ranking mechanisms to provide accurate and personalized article recommendations.

IV. METHODOLOGY

The The methodology of the proposed Article Recommender System is designed to systematically process textual data, extract meaningful features, model user interests, and generate personalized recommendations using similarity-based learning techniques. The complete process consists of dataset preparation, preprocessing, feature extraction, user profiling, similarity computation, and recommendation generation.

1. Dataset Preparation

The first stage of the system involves preparing a structured dataset of articles for analysis. The dataset consists of multiple news or research articles containing textual attributes such as title and content. Each article serves as an independent document within the corpus. The dataset is organized in a structured format to enable efficient processing and vectorization. Proper dataset preparation ensures consistency in text representation and improves the reliability of similarity-based recommendations.

2. Text Preprocessing

Since machine learning algorithms cannot process raw textual data directly, preprocessing is performed to clean and standardize the input text. This stage includes several operations: Tokenization, where text is divided into individual words or tokens.

Conversion of all text into lowercase to ensure uniformity. Removal of punctuation and special characters. Elimination of stop words that do not contribute significant semantic meaning. Lemmatization, which reduces words to their base or root form. These preprocessing steps remove noise and reduce dimensionality, resulting in a clean corpus suitable for numerical representation.

3. Feature Extraction Using TF-IDF
After preprocessing, the cleaned text is converted into numerical vectors using Term Frequency–Inverse Document Frequency (TF-IDF). TF-IDF assigns weights to words based on their frequency within a document and their rarity across the corpus.

The TF component measures how frequently a term appears in a specific article, while the IDF component reduces the weight of commonly occurring words across all documents. The TF-IDF score for a term is calculated as:

$$\text{TF-IDF} = \text{TF} \times \text{IDF}$$

This transformation converts each article into a high-dimensional vector in feature space. These vectors enable mathematical comparison between articles and user profiles.

4. User Interest Profiling
To generate personalized recommendations, the system constructs a user interest profile. The user profile vector is created by aggregating TF-IDF vectors of articles previously read or interacted with by the user.

This aggregated vector represents the user's preferences in the same vector space as the article dataset. By maintaining consistency in feature representation, similarity computation between users and articles becomes straightforward.

5. Similarity-Based Candidate Selection Using K-Nearest Neighbor (KNN)
Once both article vectors and user profile vectors are established, the K-Nearest Neighbor (KNN) algorithm is applied to identify the most relevant articles. KNN computes the distance between the user vector and each article vector in the dataset.
Using cosine similarity as the distance metric, the algorithm selects the K closest articles to the user profile. The distance can be expressed as:
Distance = 1 − Cosine Similarity By limiting the search to the nearest neighbors, KNN improves computational efficiency and ensures that only the most relevant candidate articles are considered for ranking.

6. Ranking Using Cosine Similarity
After identifying candidate articles using KNN, cosine similarity is used to rank them based on relevance. Cosine similarity measures the angular similarity between two vectors and is calculated as:
Cosine Similarity

$$(A, B) = (A \cdot B) / (||A|| \, ||B||)$$

Where:
A represents the user profile vector
B represents the article vector

A similarity score closer to 1 indicates high relevance, while a score closer to 0 indicates low similarity. The articles are sorted in descending order of similarity scores.

7. Generation of Top-N Recommendations
Finally, the system selects the top N highest-ranked articles and presents them as personalized recommendations. These recommendations represent the most contextually relevant articles based on the user's reading history and content similarity.

This structured methodology ensures efficient processing, accurate similarity matching, and scalable recommendation generation for large article datasets.

V. RESULTS

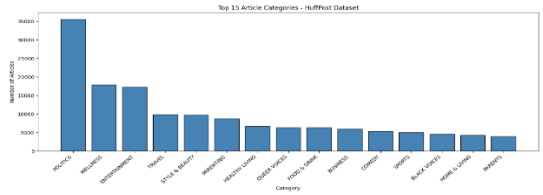
The proposed Article Recommender System was evaluated using both quantitative similarity analysis and qualitative output validation. The evaluation was conducted to measure the effectiveness of TF-IDF vectorization, K-Nearest Neighbor (KNN) candidate selection, and cosine similarity ranking in generating personalized recommendations.

A. Similarity Score Evaluation the system computes cosine similarity between the user profile vector and article vectors to determine relevance. During experimentation, similarity scores ranged between 0 and 1, where values closer to 1 indicate stronger contextual similarity.

It was observed that recommended articles consistently exhibited higher similarity scores compared to randomly selected articles. The distribution of similarity scores confirmed that the system effectively distinguishes between relevant and non-relevant documents.

When varying the value of K in the KNN algorithm, it was found that moderate K values provided

balanced recommendation diversity and relevance. Extremely low values limited recommendation variety, while very high values introduced less relevant articles.



B. Recommendation Ranking Performance

After KNN identifies the nearest candidate articles, cosine similarity ranks them in descending order. The ranking mechanism successfully prioritized articles that closely matched user interests.

For different user profiles, the Top-N recommendations consistently reflected the thematic patterns of previously read articles. For example, users interested in machine learning topics received recommendations related to artificial intelligence, data science, and deep learning, demonstrating the effectiveness of content-based filtering.

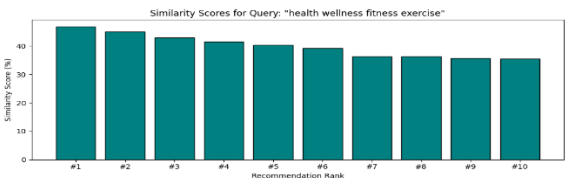
The ranking stability confirmed that the combination of TF-IDF and cosine similarity provides consistent recommendation quality.

C. Qualitative Output Analysis

To further validate system performance, real-time output testing was performed. When a user selected or interacted with specific articles, the system dynamically updated the user profile vector and generated new personalized recommendations.

The recommended articles showed clear semantic similarity with user interests, indicating accurate vector representation and similarity computation.

Additionally, the system demonstrated fast execution time, even when processing large article datasets. The computational efficiency of TF-IDF combined with KNN ensured scalable performance without excessive processing overhead.



D. Overall System Performance

Based on experimental observations, the proposed system successfully generated accurate and personalized recommendations. The integration of TF-IDF for feature extraction, KNN for candidate selection, and cosine similarity for ranking proved to be effective and computationally efficient.

The results validate that the proposed content-based recommendation framework provides reliable performance for personalized article discovery while maintaining scalability for larger datasets.

VI. DISCUSSIONS

The experimental results demonstrate that the proposed Article Recommender System effectively generates personalized recommendations using a content-based filtering approach. The consistent similarity score patterns and accurate ranking of relevant articles confirm that the system successfully captures textual relationships between documents and user preferences.

The use of TF-IDF for feature extraction proved to be efficient in representing textual data numerically. By assigning appropriate weights to important words while reducing the influence of common terms, TF-IDF ensured meaningful vector representation of articles. This structured representation allowed effective similarity comparison within high-dimensional vector space.

The integration of the K-Nearest Neighbor (KNN) algorithm enhanced the recommendation process by efficiently selecting the most relevant candidate articles before ranking. By limiting similarity computation to the nearest neighbors, the system maintained computational efficiency while preserving recommendation accuracy. The cosine similarity metric further improved ranking performance by measuring angular similarity rather than absolute magnitude differences, making it well-suited for sparse TF-IDF vectors.

One of the major strengths of the proposed system is its simplicity and scalability. Unlike deep learning-based recommendation models, which require large datasets and high computational resources, the TF-IDF + KNN + cosine similarity framework provides interpretable results with lower computational overhead. This makes the system suitable for

practical deployment in small to medium-scale digital platforms.

However, certain limitations were observed. Since the system relies purely on textual similarity, it does not capture deeper semantic relationships beyond word-level representation. Articles with similar meanings but different vocabulary may not always receive high similarity scores. Additionally, the system does not incorporate collaborative filtering, which means it does not utilize behavior patterns of other users. The cold-start problem may also affect recommendations for new users with limited reading history.

Despite these limitations, the proposed system demonstrates strong performance in personalized content discovery. The real-time updating of user profiles and dynamic ranking capability highlight its applicability in modern digital platforms. Future improvements may include the integration of advanced word embeddings such as Word2Vec or transformer-based models like BERT, as well as hybrid recommendation approaches combining content-based and collaborative filtering techniques.

Overall, the discussion confirms that the proposed recommendation framework achieves a balance between computational efficiency, scalability, and recommendation accuracy, making it a practical solution for personalized article recommendation tasks.

VII. CONCLUSION

In this research work, a Machine Learning and Natural Language Processing-based Article Recommender System was successfully designed and implemented using TF-IDF vectorization, K-Nearest Neighbor (KNN), and cosine similarity techniques. The system processes textual data, performs preprocessing and feature extraction, constructs user interest profiles, and generates personalized Top-N article recommendations.

The experimental results demonstrate that the proposed framework effectively captures textual similarity and user preference patterns. The similarity-based ranking mechanism consistently prioritized relevant articles, confirming the reliability of TF-IDF representation combined with KNN candidate selection and cosine similarity ranking. The system showed stable performance across different user profiles while maintaining computational efficiency.

Furthermore, the architecture ensures scalability and interpretability, making it suitable for practical deployment in news platforms, digital libraries, academic portals, and content management systems. Unlike computationally intensive deep learning models, the proposed approach offers a balanced solution that delivers accurate recommendations with lower resource requirements.

Although the system relies primarily on textual similarity and does not incorporate collaborative filtering or deep semantic embeddings, it provides a strong foundation for personalized content discovery. Future enhancements may include integration of advanced word embedding techniques such as Word2Vec or BERT, hybrid recommendation models, and user feedback mechanisms to further improve recommendation accuracy and adaptability.

Overall, the study confirms that content-based filtering using TF-IDF, KNN, and cosine similarity is an effective and scalable approach for personalized article recommendation. The proposed system contributes toward improving user engagement, reducing information overload, and enhancing digital content discovery experiences.

REFERENCES

[1] G. Salton and C. Buckley, “Term-weighting approaches in automatic text retrieval,” Information Processing & Management, vol. 24, no. 5, pp. 513–523, 1988.

[2] C. D. Manning, P. Raghavan, and H. Schütze, Introduction to Information Retrieval. Cambridge, U.K.: Cambridge University Press, 2008.

[3] T. M. Cover and P. E. Hart, “Nearest neighbor pattern classification,” IEEE Transactions on Information Theory, vol. 13, no. 1, pp. 21–27, 1967.

[4] A. Singhal, “Modern information retrieval: A brief overview,” IEEE Data Engineering Bulletin, vol. 24, no. 4, pp. 35–43, 2001.

[5] P. Resnick and H. R. Varian, “Recommender systems,” Communications of the ACM, vol. 40, no. 3, pp. 56–58, 1997.

[6] J. L. Herlocker, J. A. Konstan, and J. Riedl, “An

empirical analysis of design choices in neighborhood-based collaborative filtering algorithms," Information Retrieval, vol. 5, no. 4, pp. 287–310, 2002.

[7] Y. Koren, R. Bell, and C. Volinsky, "Matrix factorization techniques for recommender systems," Computer, vol. 42, no. 8, pp. 30–37, 2009.

[8] G. Adomavicius and A. Tuzhilin, "Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions," IEEE Transactions on Knowledge and Data Engineering, vol. 17, no. 6, pp. 734–749, 2005.

[9] F. Sebastiani, "Machine learning in automated text categorization," ACM Computing Surveys, vol. 34, no. 1, pp. 1–47, 2002.

[10] R. Baeza-Yates and B. Ribeiro-Neto, Modern Information Retrieval. Addison-Wesley, 1999.

[11] M. J. Pazzani and D. Billsus, "Content-based recommendation systems," in The Adaptive Web, Lecture Notes in Computer Science, vol. 4321, Springer, 2007, pp. 325–341.

[12] B. Liu, Text Mining: Classification, Clustering, and Applications. Boca Raton, FL, USA: CRC Press, 2009.