International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 13 Issue: 01 | Jan 2026 www.irjet.net p-ISSN: 2395-0072

Recommendation System Using NLP-Based Similarity Modeling and Clustering Techniques

Nakibinge Gavin[1]

[1]*Department of Information and Communication Technology,Marwadi University, India*

Abstract - *The rapid growth of digital scholarly repositories has resulted in an overwhelming volume of academic publications, making it increasingly difficult for researchers to efficiently discover relevant research papers. Conventional keyword-based search mechanisms often fail to capture semantic relationships between documents and require significant manual effort. To address this challenge, this paper presents an intelligent research paper recommendation system based on Natural Language Processing (NLP), similarity modeling, clustering, and topic modeling techniques. The proposed system utilizes the arXiv dataset comprising over 250,000 research papers across multiple scientific domains. Titles and abstracts are preprocessed and represented using Term Frequency–Inverse Document Frequency (TF-IDF) vectors. Cosine similarity is employed to compute semantic similarity between research papers. Dimensionality reduction techniques such as Principal Component Analysis (PCA) and t-Distributed Stochastic Neighbor Embedding (t-SNE) are applied for visualization and analysis of high-dimensional data. Furthermore, K-Means clustering is used to group similar papers, and Latent Dirichlet Allocation (LDA) is applied to extract dominant topics from each cluster. A Flask-based web application deployed on Hugging Face Spaces provides an interactive interface for searching and recommending research papers. Experimental results demonstrate that the system effectively identifies semantically related research articles and organizes them into meaningful clusters, thereby improving the efficiency and quality of academic literature discovery.*

Keywords: Research Paper Recommendation, Natural Language Processing, TF-IDF, Cosine Similarity, Clustering, Topic Modeling, Information Retrieval, NLP

1. INTRODUCTION

The exponential growth of digital research repositories such as arXiv and Google Scholar has resulted in a massive increase in scholarly publications across all scientific domains. While this growth promotes knowledge dissemination, it also creates significant challenges for researchers attempting to identify relevant literature efficiently. Traditional keyword-based search mechanisms rely primarily on lexical matching and often fail to capture semantic relationships between research papers, leading to incomplete or irrelevant results [1].

Recommender systems have emerged as an effective solution for managing information overload by providing personalized content suggestions. In academic research, recommendation systems can assist scholars by identifying papers closely aligned with their research interests and reducing the time required for manual literature surveys [2]. Content-based recommendation approaches analyze textual information such as titles and abstracts to compute similarity between documents without relying on user interaction data, thereby avoiding the cold-start problem [3].

Recent advances in Natural Language Processing (NLP) have enabled the use of statistical and semantic techniques such as TF-IDF, cosine similarity, clustering, and topic modeling to enhance document similarity analysis [4]. Motivated by these developments, this paper proposes an NLP-based research paper recommendation system that integrates similarity modeling, clustering, and topic modeling techniques. A web-based implementation deployed on Hugging Face Spaces demonstrates the practical applicability of the proposed system.

2. Related Work

Several studies have explored academic paper recommendation using collaborative filtering, content-based filtering, and hybrid approaches. Collaborative filtering methods rely on user behaviour data such as citations, downloads, or reading history, but they often suffer from cold-start and sparsity issues when user data is limited [5].

Content-based approaches utilize textual features of research papers to recommend similar documents. TF-IDF combined with cosine similarity has been widely adopted for document similarity analysis due to its simplicity and interpretability [1]. Topic modelling techniques such as Latent Dirichlet Allocation (LDA) have been used to uncover latent thematic structures in academic corpora [2]. More recent studies employ deep learning models such as Word2Vec, Doc2Vec, and BERT for semantic representation; however, these methods require higher computational resources and complex training pipelines [6].

In contrast, this work focuses on a scalable and interpretable NLP-based recommendation system using classical machine learning techniques that provide effective results while maintaining computational efficiency.

3. Dataset Description

The proposed system utilizes the arXiv dataset obtained from Kaggle, which contains metadata for approximately 250,000 research papers across multiple scientific domains [7]. Each record includes the paper title, abstract, authors, subject categories, and unique identifiers. Titles and abstracts are used as the primary textual features in this study, as they provide concise summaries of research contributions and are commonly employed in content-based recommendation systems [3].

Table – 1: Summary of the arXiv Dataset Used

Parameter	Description
Dataset Source	arXiv (Kaggle)
Number of Papers	250,000
Text Fields Used	Title, Abstract
Domains Covered	Computer Science, Physics, Mathematics, Engineering
Language	English

Table 1 summarizes the key characteristics of the dataset used for experimentation.

4. Proposed Methodology

4.1 Text Preprocessing

Text preprocessing is applied to titles and abstracts to reduce noise and improve feature quality. The preprocessing steps include tokenization, conversion to lowercase, removal of stop words, and elimination of punctuation and special characters.

4.2 Feature Extraction Using TF-IDF

The cleaned textual data is transformed into numerical feature vectors using Term Frequency–Inverse Document Frequency (TF-IDF). TF-IDF effectively captures the importance of terms within documents relative to the entire corpus [1].

4.3 Similarity Computation

Cosine similarity is used to compute semantic similarity between TF-IDF vectors. Papers with higher cosine similarity scores are considered more semantically related and are ranked higher in the recommendation list [4].

4.4 Dimensionality Reduction

Principal Component Analysis (PCA) is applied to reduce the dimensionality of the TF-IDF matrix while preserving 95% of the variance. t-Distributed Stochastic Neighbor Embedding (t-SNE) is further used to project high-dimensional vectors into a two-dimensional space for visualization and qualitative analysis [8].

4.5 Clustering and Topic Modelling

K-Means clustering is employed to group research papers into clusters based on semantic similarity. Latent Dirichlet Allocation (LDA) is applied to extract dominant topics and keywords from each cluster, improving interpretability and thematic understanding [2].

5. System Architecture

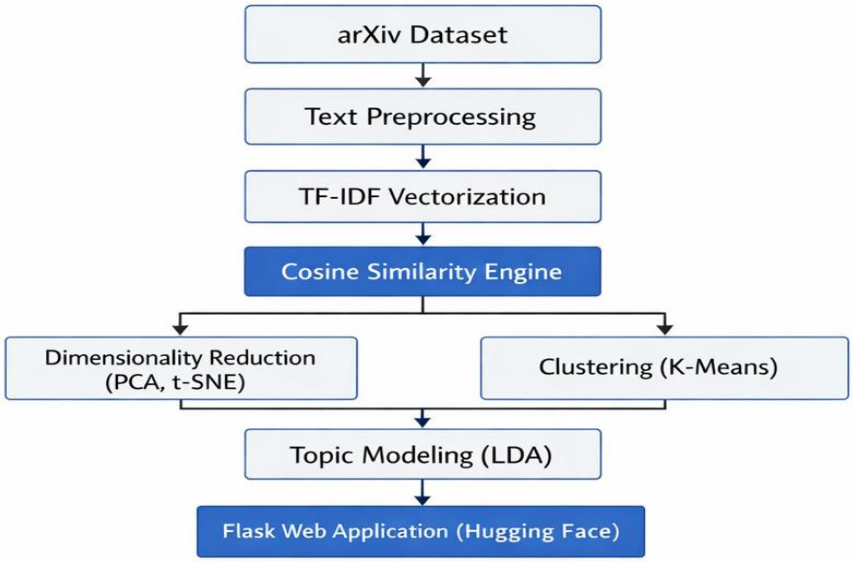

Fig. 1: Architecture of the Proposed Research Paper Recommendation System

Fig. 1 illustrates the overall architecture of the proposed system. The process begins with the arXiv dataset, followed by text preprocessing and TF-IDF vectorization. Cosine similarity is used to compute document similarity. Dimensionality reduction techniques support visualization, while clustering and topic modeling organize papers into thematic groups. The recommendation results are delivered through a Flask-based web application deployed on Hugging Face Spaces.

6. Recommendation Process Flow

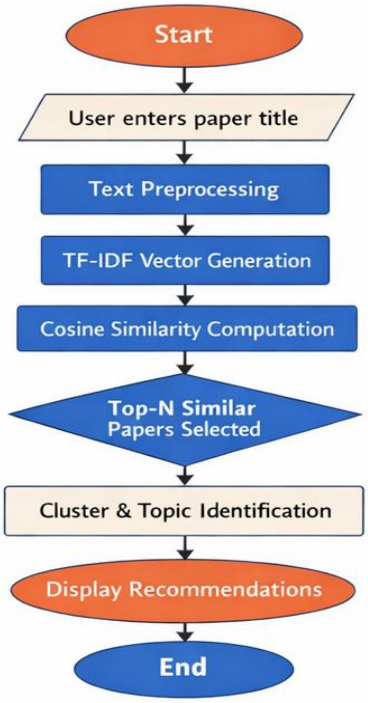

Fig. 2: Flowchart of Research Paper Recommendation Process

Fig. 2 presents the flowchart of the recommendation process. The user inputs a research paper title, which undergoes preprocessing and vector generation. Similarity computation identifies the top-N related papers, which are mapped to clusters and topics before being displayed to the user.

7. Performance Evaluation

7.1 Precision@K

Precision@K measures the relevance of the top-K recommended papers and is defined as:

$$\text{Precision@K} = \frac{\text{Number of relevant papers in top-K}}{K}$$

Relevance is determined based on subject category overlap and semantic similarity. Experimental results indicate high Precision@5 and stable Precision@10 values, demonstrating that relevant papers are ranked prominently.

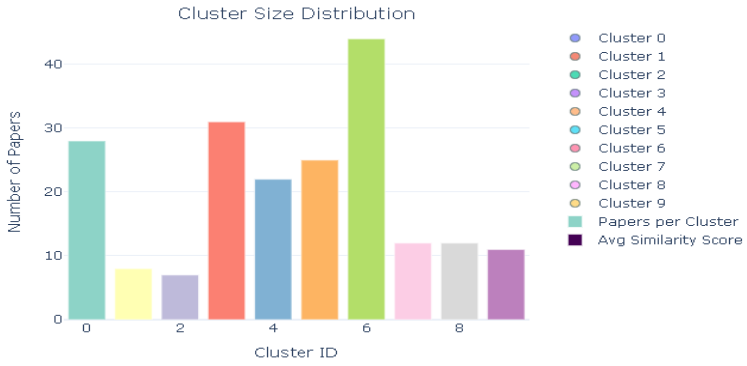

Chart -1: Cluster size distribution

7.2 Similarity Score Analysis

Cosine similarity scores are analyzed to assess semantic closeness between recommended papers and the input paper. The system consistently produces high similarity scores for recommended papers, confirming the effectiveness of TF-IDF-based similarity modeling.

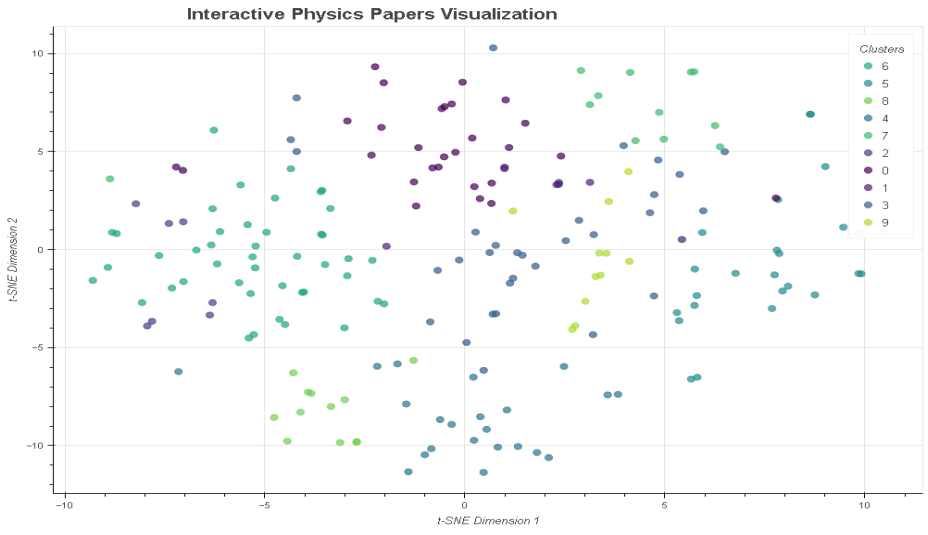

Chart -2: Interactive physic paper visualization

7.3 Qualitative Evaluation

Cluster coherence and topic consistency are examined qualitatively. Papers within the same cluster share similar research themes, and LDA-generated keywords accurately represent cluster topics. t-SNE visualization further confirms clear separation between clusters.

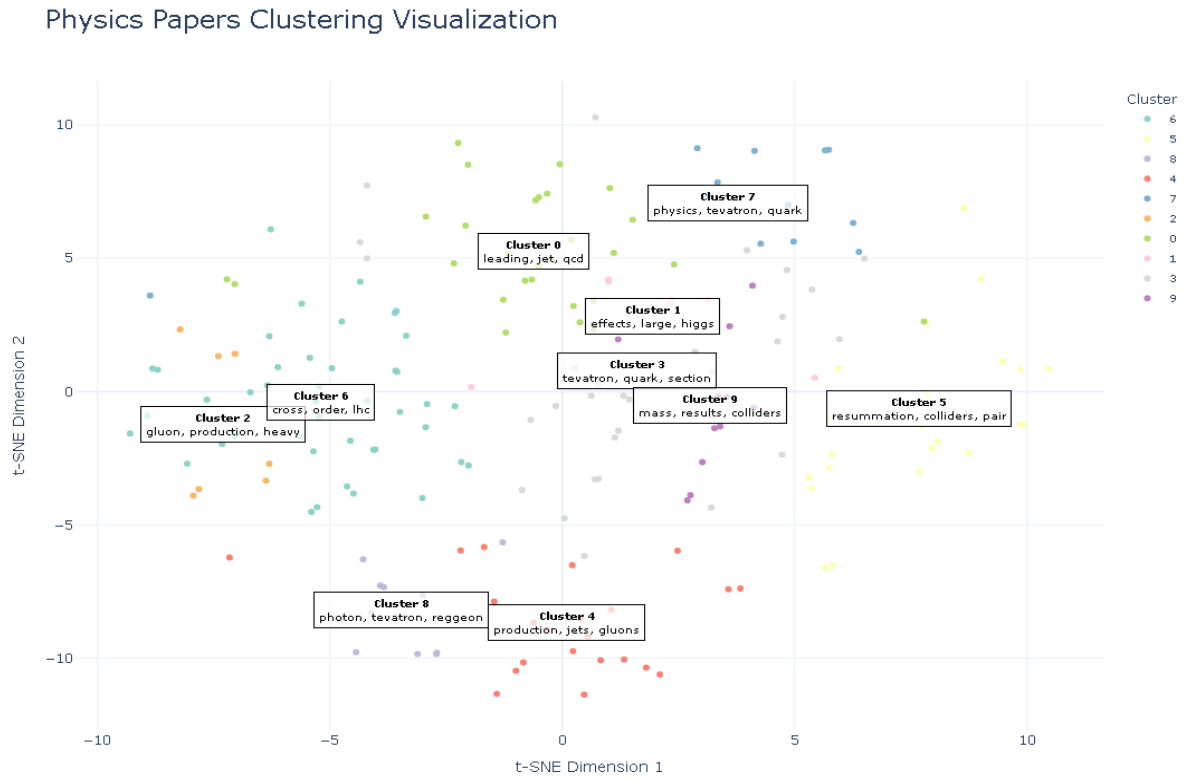

Chart -3: physics paper clustering visualization

Table – 2: Performance Evaluation Summary

Metric	Observation
Precision@5	High relevance
Precision@10	Consistent performance
Similarity Scores	Strong semantic alignment
Cluster Coherence	Well-defined topics
System Response Time	Fast

8. Web Application Implementation

A web-based application is developed using the Flask framework and deployed on Hugging Face Spaces. The application enables users to search for research papers by title and receive recommendations based on semantic similarity. Paper metadata such as titles, authors, and identifiers are displayed to enhance usability.

9. Conclusion and Future Work

This paper presented a professional NLP-based research paper recommendation system integrating similarity modelling, clustering, and topic modelling techniques. The system effectively addresses the challenge of academic literature discovery by

identifying semantically related research papers and organizing them into meaningful clusters. Experimental evaluation demonstrates strong recommendation relevance and clustering quality.

Future work includes integrating deep learning-based embeddings such as BERT, incorporating user profiling and collaborative filtering, and extending the system to support real-time and cross-domain recommendations.

References

[1] J. Ramos, "Using TF-IDF to Determine Word Relevance in Document Queries," Proc. First Instructional Conf. on Machine Learning,2003.

[2] D. Blei, A. Ng, and M. Jordan, "Latent Dirichlet Allocation," Journal of Machine Learning Research, 2003.

[3] C. Manning, P. Raghavan, and H. Schütze, Introduction to Information Retrieval, Cambridge University Press, 2008.

[4] G. Salton and C. Buckley, "Term-Weighting Approaches in Automatic Text Retrieval," Information Processing & Management,1988.

[5] J. Bobadilla et al., "Recommender Systems Survey," Knowledge-Based Systems, 2013.

6] J. Devlin et al., "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," NAACL, 2019.

[7]CornellUniversity,"arXivDataset,"Kaggle,2023.

[8] L. van der Maaten and G. Hinton, "Visualizing Data using t-SNE," Journal of Machine Learning Research, 2008.

BIOGRAPHIES

a final-year undergraduate student pursuing a Bachelor's degree in Information and Communication Technology. This research paper is based on his final year project, which focuses on designing and implementing an NLP-based research paper recommendation system.