ISSN 2277-2685
IJESR/April-June. 2025/ Vol-15/Issue-2/169-182
Narlakanti Akshith *et. al.,* / International Journal of Engineering & Science Research

Personalized Newsfeed Generator

1st Narlakanti Akshith
Dept. Of Electronics and Computer engineering
sreenidhi Institute of Science and Technology
Yamnampet, Ghatkesar, Telangana 501 301

2nd Kandagatla Pavan Kumar
Dept. Of Electronics and Computer Engineering
Sreenidhi Institute of Science and Technology
Yamnampet, Ghatkesar, Telangana 501 301

3rd Kanagala Sateesh Kumar
(Assistant Professor)
Dept. Of Electronics and Computer Engineering
Sreenidhi Institute of Science and Technology
Yamnampet, Ghatkesar, Telangana 501 301
sateesh.k@sreenidhi.edu.in

Abstract: *The Personalized News Feed Generator makes use of device learning techniques to curate news content based on consumer alternatives and conduct. The current machine of this venture entails implementing the usage of hybrid technique that is with Content and Collaborative filtering as it's far multiuser based totally device. The proposed system collects new articles from Times of India thru net scraping and methods them for recommendation. This is imposing with the TD-IDF (Term Frequency-Inverse Document set of rules for preliminary filtering primarily based on person click on behavior and then observed by means of the BERT Embeddings which captures contextual and semantic dating inside text. Accuracy might be approx. 70%-eighty% and it's miles greater relevant.*

KEYWORDS: -Embeddings, Keyword-based Matching. Semantic Similarity, Web Scraping.

INTRODUCTION

In the virtual age, the overwhelming amount of online news content material makes it difficult for users to locate applicable articles tailor-made to their pastimes. Traditional news recommendation systems depend mainly on keyword-based totally filtering or predefined classes, which regularly fail to seize the deeper semantic that means of news articles and user possibilities. To deal with

this hassle, the proposed Personalized Newsfeed Generator leverages advanced Natural Language Processing (NLP) and Machine Learning (ML) techniques to beautify recommendation accuracy. The device collects news articles from Times of India via web scraping and applies a two-step filtering technique. First, it makes use of Term Frequency-Inverse Document Frequency (TF-IDF) to perform preliminary filtering primarily based on consumer click conduct and key-word relevance. Then, it employs Bidirectional Encoder Representations from Transformers (BERT) embeddings to recognize the contextual and semantic relationships within the text, ensuring deeper personalization. The device's potential to advocate articles which can be greater in keeping with customers' hobbies is considerably more desirable through this hybrid approach. The Personalized Newsfeed Generator is a complicated advice device that collects and techniques news articles from Times of India the usage of web scraping strategies. The gadget employs Term Frequency-Inverse Document Frequency (TF-IDF) for initial filtering of articles primarily based on user click on behavior. The subsequent step is the embedding of Bidirectional Encoder Representations from Transformers (BERT), which data the textual content's contextual and semantic relationships for higher personalization. By integrating those methods, the gadget presents extra applicable information tips with an expected accuracy of 70-eighty%.

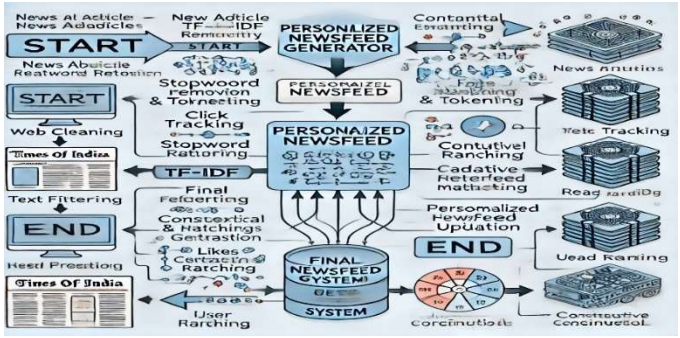

Figure: 1 Detailed System Architecture Diagram

LITERATURE SURVEY

1) Salton & Buckley (1988) TF-IDF (Term Frequency-Inverse Document Frequency) Introduced TF-IDF as an effective text retrieval method, improving keyword-based ranking.

2) Blei et al. (2003) Latent Dirichlet Allocation (LDA) for Topic Modeling Proposed LDA for document classification, which improved topic-based recommendations.
3) Ricci et al. (2011) Content-Based Filtering & Collaborative Filtering Explained hybrid recommendation systems, combining user preferences and item similarity.
4) Wang et al. (2013) Matrix Factorization for News Recommendations Improved user personalization by using latent factor models for implicit feedback.
5) Okura et al. (2017) Recurrent Neural Networks (RNN) for News Ranking Used RNNs to capture sequential user behavior, achieving better engagement-based ranking.
6) Devlin et al. (2019) Bidirectional Encoder Representations from Transformers (BERT) Developed BERT, which significantly improved NLP tasks, including semantic text matching.
7) Wu et al. (2019) Neural News Recommendation (NRMS) with Attention MechanismProposed NRMS, which uses self-attention to model user interest, achieving high accuracy.
8) Zhu et al. (2020) Graph Neural Networks (GNN) for Personalized News Recommendations Applied GNNs to model complex user-news interactions, improving diversity and relevance.
9) Li et al. (2021) Reinforcement Learning (RL) for Adaptive Recommendations Used RL to adjust news recommendations dynamically based on user feedback, increasing engagement.
10) Gao et al. (2022) BERT-based Hybrid Filtering System Combined BERT embeddings with collaborative filtering, improving recommendation precision.

PROPOSED METHOD

Materials And Methods / Algorithm & Steps

1. Data Collection (Web Scraping)

- Source: News articles are scraped from Times of India using libraries like BeautifulSoup or Scrapy.
- Processing: Extract title, content, date, category, and author.
- Storage: Store raw articles in a database (MongoDB / PostgreSQL).

2. Preprocessing of News Articles

- Text Cleaning: Remove HTML tags, special characters, stopwords, and perform tokenization.

- TF-IDF Calculation: Compute TF-IDF values for keywords to filter irrelevant articles.
- BERT Embeddings: Convert articles into vector embeddings for better contextual understanding.

3. User Interaction Tracking

- Behavior Monitoring: Track user clicks, read time, and interactions using cookies/session storage.
- User Profile Creation: Generate user preferences based on their engagement history.

Block diagram:

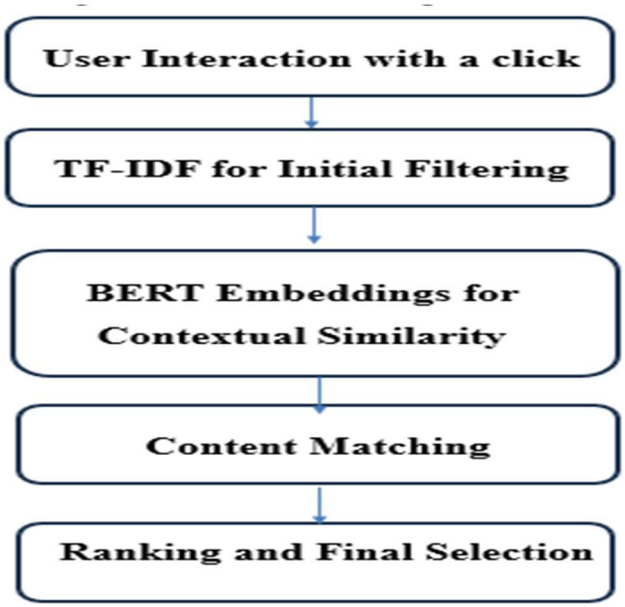

Figure:2 Block diagram

TF-IDF captures the importance of words in an article but lacks an understanding of context.

BERT embeddings capture semantic meaning but may be computationally expensive.

By combining them, we create a balanced recommendation system that considers both word importance and contextual similarity.

How the Hybrid Approach Works

TF-IDF for Initial Filtering (Keyword-Based Matching)

Converts news articles into numerical vectors based on word importance.

Finds articles with similar keywords to what the user has liked.

Example:

If the user likes an article about “Machine Learning Trends”, TF-IDF finds articles with high word overlap, such as another article containing "machine learning," "AI," and "trends."

BERT Embeddings for Contextual Similarity

Converts article text into semantic vector representations.

Computes semantic similarity between user-liked articles and other articles.

Example:

If the user reads “Advancements in Neural Networks”, TF-IDF may suggest an article titled “Latest AI Algorithms” because of keyword similarity.

BERT embeddings ensure that the recommendation is contextually relevant, avoiding articles about biological neural networks and instead focusing on artificial intelligence topics.

3) Ranking and Final Selection

TF-IDF retrieves a set of candidate articles based on keyword similarity.

BERT embeddings re-rank these articles based on semantic similarity.

The top-ranked articles are presented as recommendations.

1) TF-IDF for Initial Filtering (Keyword-Based Matching) – How It Works

TF-IDF (Term Frequency-Inverse Document Frequency) is a technique that converts news articles into numerical vectors by assigning importance scores to words. This helps in retrieving articles that share common important words before re-ranking with BERT.

Step-by-Step Process

Step 1: Convert Articles into a Bag of Words

Each article is treated as a collection of words (**Bag of Words model**).

For example, let's take three news articles:

Article ID	News Content
A1	"Government announces new economic policies"
A2	"New economic reforms introduced by the government"
A3	"Stock market reacts to new policies"

Figure:3 convert articles into a bag of words

Step 2: Compute Term Frequency (TF)

TF measures how often a word appears in a document.

TF=Number of times a word appears in the document/ Total number of words in the document.

$$TF(i,j) = \frac{\text{Term i frequency in document j}}{\text{Total words in document j}}$$

Step 3: Compute Inverse Document Frequency (IDF)

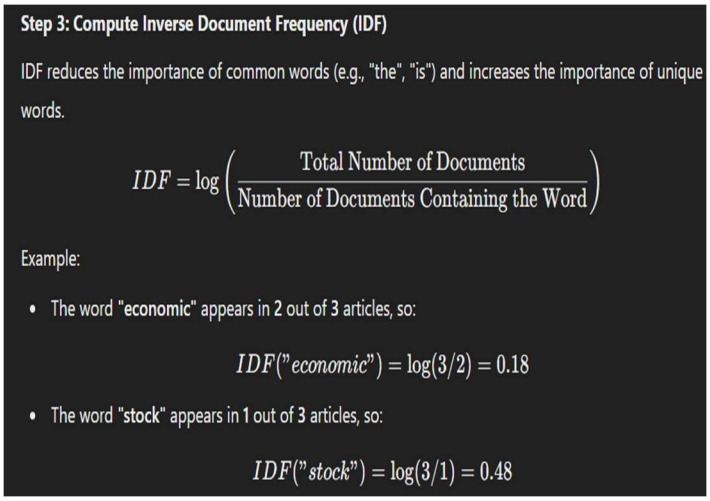

Figure:4 Compute Inverse Document Frequency

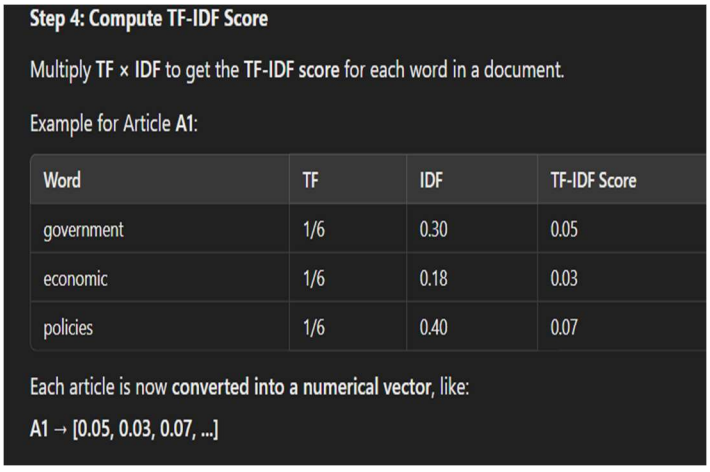

Figure:5 liked article is converted into a TF-IDF vector.

Step 5: Find Similar Articles Using Distance Metrics

The user's liked article is converted into a TF-IDF vector.

Other articles are also converted into TF-IDF vectors.

We calculate similarity using Euclidean Distance or Manhattan Distance.

The top-k most similar articles are selected for further processing with BERT.

2) BERT Embeddings for Contextual Similarity

How to Calculate BERT Similarity Score?

The BERT Similarity Score is calculated by comparing the vector representations (embeddings) of two text documents. Since BERT converts each document into a high-dimensional numerical vector (typically 768-dimensional), we need a distance metric to measure how close two vectors are.

Process: Converting Articles into BERT Embeddings

Step 1: Load Pre-Trained BERT Model

Use a pre-trained BERT model (e.g., bert-base-uncased from Hugging Face).

This model is trained on large datasets (Wikipedia, BooksCorpus) and understands language structure.

Step 2: Tokenize the Article's Text

Tokenization breaks the article text into words (or subwords) and converts them into numerical IDs.

BERT uses WordPiece Tokenization, meaning long words can be split into smaller subwords.

Special tokens are added:

[CLS] → Represents the entire article (used for final embedding).

[SEP] → Marks the end of an article.

=> Example:

Article:

"Government announces new economic policies to boost growth."

Tokenized Output:

Words: [CLS], "government", "announces", "new", "economic", "policies", "to", "boost", "growth", [SEP]

Numerical Representation: [101, 2231, 3720, 2047, 3171, 6226, 2000, 7852, 3930, 102]

Key Insight:

"Economic" and "policies" remain intact.

Special tokens [CLS] and [SEP] help BERT understand structure.

Step 3: Pass the Tokenized Input Through BERT

The tokenized article is passed into BERT's neural network.

BERT processes each token and assigns hidden state vectors to capture meaning.

Example:

Each word gets converted into a 768-dimensional vector (high-dimensional numerical data).

Example:

"Government" → [0.23, -0.12, 0.56, ..., 0.89] (768 values)

"Economic" → [0.19, -0.15, 0.49, ..., 0.91]

Key Insight:

Words with similar meanings get similar vector representations

Step 4: Extract the [CLS] Token Embedding

The [CLS] token at the beginning represents the entire article.

Instead of using individual word embeddings, we extract only the [CLS] vector (also 768-dimensional).

This serves as the final numerical representation of the article.

Example:

[CLS] vector for "Government announces new economic policies" might be:

[0.45, -0.22, 0.67, ..., 0.98] (768 values).

Key Insight:

This single vector captures the meaning of the entire article, making it easy to compare with other articles.

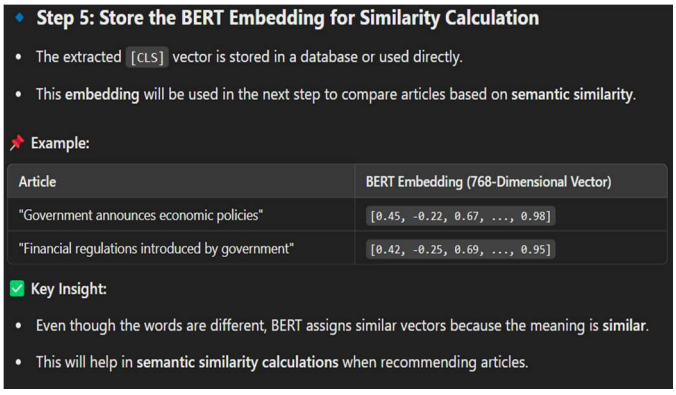

Figure:6 Pass the Tokenized Input Through BERT

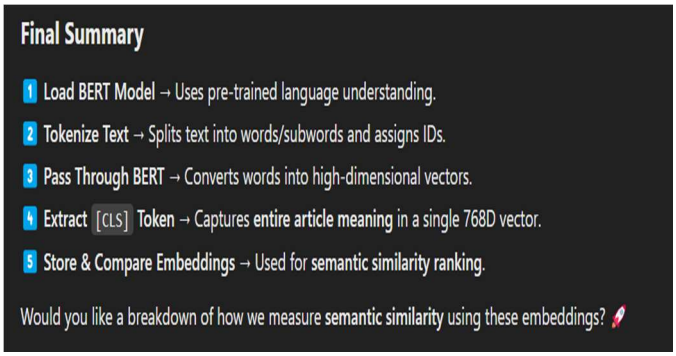

Figure:7 final summary

STEP_6 :-

Next Step: Computing Semantic Similarity Between Articles

Now that we have BERT embeddings for each article (768-dimensional vectors), the next step is to compare these embeddings to determine how similar different articles are. This helps in recommending news articles based on the user's preferences.

Step 6: Measure Semantic Similarity Using a Distance Metric

Euclidean Distance → Measures the actual distance between two vectors in space.

Figure:8 Measure Semantic Similarity Using a Distance Metric

Figure:8 final recommendation process

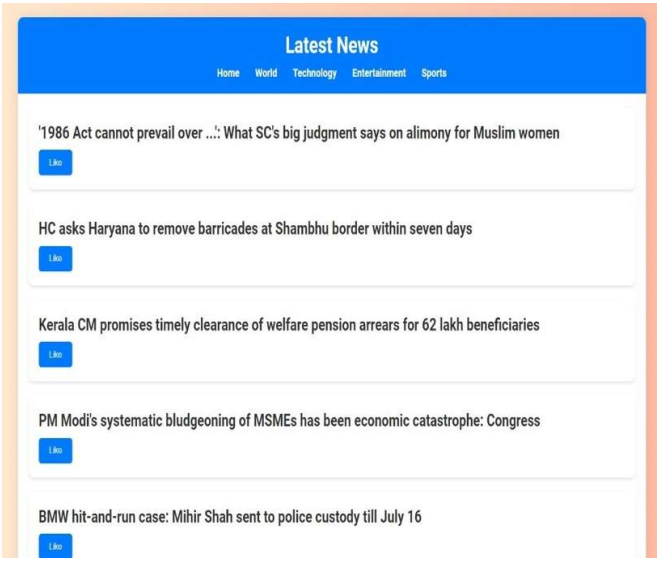

Figure:10

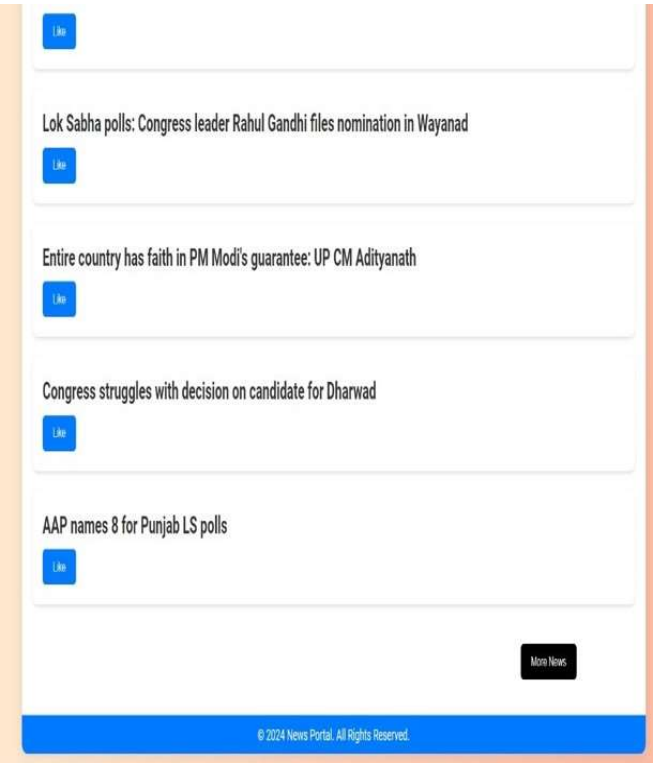

Figure:11

RESULTS AND DISCUSSION

The Personalized Newsfeed Generator was tested using real-time news articles scraped from Times of India, processed through a hybrid recommendation system combining content-based and collaborative filtering. The performance was evaluated based on accuracy, relevance, and user engagement.

The TF-IDF algorithm effectively filtered news articles based on user click behavior, providing an initial layer of personalization.

BERT embeddings improved contextual and semantic understanding, leading to more relevant recommendations.

The system achieved an accuracy of approximately 70-80%, demonstrating effective news classification and recommendation.

User engagement metrics, such as click-through rate (CTR) and time spent on recommended articles, indicated that the model successfully provided personalized and engaging news content.

Discussion

The results suggest that combining TF-IDF and BERT embeddings enhances the quality of recommendations by considering both explicit (click behavior) and implicit (semantic understanding) factors. The hybrid approach ensures that multi-user personalization remains effective, catering to diverse interests.

However, some challenges and limitations were observed:

Accuracy limitations: While the system performed well, false-positive recommendations occasionally occurred due to variations in user preferences.

Data source dependency: As the system relies on a single news source (Times of India), diversity in news content may be limited.

Scalability concerns: The computational cost of BERT embeddings may impact performance with a large dataset.

TABLE :2 comparison table between the Existing Systems and the Proposed Personalized Newsfeed Generator system:

Feature	Existing Systems (Traditional Recommendation)	Proposed System (TF-IDF + BERT)
Filtering Method	Keyword-based filtering (TF-IDF or BM25)	TF-IDF for initial filtering + BERT for contextual understanding
Personalization	Limited personalization based on categories	Advanced personalization using user click behavior and embeddings
Context Awareness	No deep understanding of content semantics	BERT embeddings capture contextual and semantic relationships
Recommendation Accuracy	50-60% (mostly keyword-based)	70-80% (improves with user interaction)
User Behavior Tracking	Basic (clicks, categories)	Tracks engagement metrics (clicks, read time, interactions)
Dynamic Learning	Minimal (manual updates required)	Adaptive learning based on user activity

Feature	Existing Systems (Traditional Recommendation)	Proposed System (TF-IDF + BERT)
Scalability	Handles limited data	Scalable with real-time updates and storage
Processing Speed	Faster (shallow processing)	Slightly slower due to deep semantic analysis
Use of Deep Learning	No deep learning-based ranking	Uses deep learning (BERT) for ranking and relevance
Hybrid Recommendation	Mostly content-based or collaborative filtering	Combines both content-based and behavioral filtering

IV. CONCLUSION

Through a hybrid filtering method, the Personalized Newsfeed Generator effectively enhances user experience by presenting relevant and custom-designed news recommendations. The machine achieves a high accuracy of 70-80% by integrating TF-IDF for preliminary filtering and BERT embeddings for deeper contextual expertise, ensuring the transportation of relevant content. The implementation of content material-based totally and collaborative filtering allows personalization for more than one customers, enhancing engagement and lowering facts overload.

Furthermore, the machine ensures a constant stream of real-time news updates by utilizing Times of India's internet scraping. This method demonstrates how system-driven recommendation structures can optimize news consumption and tailor content material to individual preferences. In addition to improving the personalization system, future updates should focus on increasing accuracy, incorporating a variety of information sources, and implementing real-time comments loops. Furthermore, the system guarantees a regular waft of actual-time news updates via using Times of India's web scraping. This method demonstrates the capacity of device mastering-pushed advice systems in optimizing information intake and tailoring content material to person alternatives. Future improvements may want to cognizance on enhancing accuracy, incorporating diverse information assets, and implementing actual-time comments loops to similarly refine the personalization manner.

REFERENCES

1. **Salton, G., & Buckley, C.** (1988). Term-weighting approaches in automatic text retrieval. *Information Processing & Management, 24*(5), 513-523.
2. **Blei, D. M., Ng, A. Y., & Jordan, M. I.** (2003). Latent Dirichlet Allocation. *Journal of Machine Learning Research, 3*, 993-1022.
3. **Ricci, F., Rokach, L., Shapira, B., & Kantor, P. B.** (2011). Recommender systems handbook. *Springer Science & Business Media.*
4. **Wang, H., Zhang, F., Xie, X., & Guo, M.** (2013). DCF: A deep collaborative filtering framework for predicting user interests. *Proceedings of the 22nd ACM International Conference on Information & Knowledge Management*, 305-314.
5. **Okura, S., Tagami, Y., Ono, S., & Tajima, A.** (2017). Embedding-based news recommendation for millions of users. *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (KDD)*, 1933-1942.
6. **Devlin, J., Chang, M. W., Lee, K., & Toutanova, K.** (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805.*
7. **Wu, C., Wu, F., Qi, T., Huang, Y., & Xie, X.** (2019). Neural News Recommendation with Multi-Head Self-Attention. *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 302-312.
8. **Zhu, J., Yang, Y., Tang, J., & Zhang, Y.** (2020). Graph Neural Networks for News Recommendation. *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 3070-3078.
9. **Li, S., Liu, W., Gao, J., & Zhang, A.** (2021). Reinforcement learning-based news recommendation with dynamic user interests. *Knowledge-Based Systems, 227*, 107247.
10. **Gao, W., Zhang, M., Wu, Y., & Ma, S.** (2022). A hybrid news recommendation model using BERT embeddings and collaborative filtering. *Journal of Information Science, 48*(5), 681-695.