Kurdistan Journal of Applied Research (KJAR)
Print-ISSN: 2411-7684 | Electronic-ISSN: 2411-7706

Original Article

A Multilingual Hybrid News Recommendation Framework for Educational Web Portals

Przha Barzan Mohialdeen [a] *, **Sarkar Hasan Ahmed** [b]

[a] Information Technology Department, Technical College of Informatics, Sulaimani Polytechnic University, Sulaymaniyah, Iraq.
[b] Computer Network Department, Technical College of Informatics, Sulaimani Polytechnic University, Sulaymaniyah, Iraq.

Submitted: 22 July 2025
Revised: 21 August 2025
Accepted: 05 September 2025

*** Corresponding Author:**
przha.barzan.tci@spu.edu.iq

Keywords: Multilingual news recommendation, educational web portals, Hybrid recommendation algorithms, Sentence-BERT, Term Frequency – Inverse Document Frequency, Best Matching 25.

How to cite this paper: P. B. Mohialdeen, S. H. Ahmed, "A Multilingual Hybrid News Recommendation Framework for Educational Web Portals", KJAR, vol. 10, no. 2, pp: 212-228, Dec 2025, **doi: 10.24017/science 2025.2.14**

Abstract: University web portals increasingly serve as vital platforms for academic information sharing, yet effective news recommendation in resource-constrained, multilingual environments remains challenging due to limited labeled data, sparse user profiles, and linguistic diversity. This study presents a modular hybrid news recommendation framework tailored for educational web portals in low-resource settings. The approach integrates lexical methods, specifically Term Frequency–Inverse Document Frequency (TF–IDF) and Best Match 25 (BM25), with semantic retrieval based on Sentence-BERT (SBERT), combined with unsupervised clustering for topical diversification and a fuzzy-logic fusion layer to integrate heterogeneous similarity signals. A publicly available multilingual dataset of 1,389 university news articles was collected via a custom crawler, and a Flask-based API was implemented for real-time recommendation. Evaluation relies on an automatic hybrid ground truth generated by fusing SBERT, TF–IDF, and BM25 signals. On the ground truth subset, the hybrid model attains Precision@5 = 0.96 and NDCG@5 = 0.945, outperforming SBERT (Precision@5 = 0.93; NDCG@5 = 0.859), with improvements shown to be statistically significant (paired t-test on NDCG@5, $p < 1e\text{-}5$). Clustering enhances thematic diversity (entropy 1.697 vs. 0.032), reducing concentration on repeated announcements. Multilingual experiments demonstrate consistently high precision across Arabic, Kurdish, and English but reveal substantially lower recall for underrepresented languages, highlighting dataset imbalance and representation challenges. Fusion weights were tuned on a validation split to balance precision and recall while mitigating the dominance of any single signal across languages and content types. The proposed framework provides an interpretable and extensible solution for multilingual academic news recommendation in scenarios where interaction data are scarce, offering a practical foundation for future work on language-aware preprocessing, human validation of labels, and supervised re-ranking.

1. Introduction

In today's world of abundant information, news recommendation systems are essential for helping users navigate vast volumes of online content and remain updated on topics of interest [1, 2]. This function is particularly important for university websites, as it enables students and academics to access relevant information and tailor their educational experiences to their specific needs [3]. Despite their potential, implementing effective news recommendation systems in low-resource and multilingual environments remains a significant challenge [4, 5]. One major limitation is the scarcity of labeled data, which many machine learning–based systems require for accurate model construction [6].

In the context of higher education institutions in the Kurdistan Regional Government of Iraq, university websites host multilingual resources and news, yet no labeled datasets exist for this content [7]. Another difficulty is the absence of user profiles, which play a crucial role in many recommendation systems to enable personalized results [8]. Collecting extensive user data is often not feasible in academic contexts, either because of privacy concerns or technological limitations [9, 10].

To address these challenges, this paper introduces a multilingual dataset comprising 1,389 news articles in English, Arabic, and Kurdish, collected from academic websites in the Kurdistan Regional Government using a custom web crawler. In addition, it proposes a hybrid news recommendation system tailored for resource-constrained and multilingual educational environments. The system leverages content-based filtering techniques, including TF–IDF, BM25, and SBERT embeddings. Unsupervised K-Means clustering is employed for automatic article categorization, while fuzzy logic is applied to integrate heterogeneous signals through a robust fusion mechanism that enhances overall performance. By combining these strategies, the proposed framework mitigates the limitations of individual methods and enables effective multilingual news recommendation in contexts where user profiles and labeled datasets are limited.

The remainder of this paper is organized as follows: section 2 reviews related work, section 3 presents the proposed approach, section 4 reports the experimental results, section 5 discusses the findings, and section 6 concludes with future research directions.

2. Related Works

Recommender systems have been widely applied in educational and news portals to match users with the most relevant resources, such as courses, research topics, and timely news items. Ensuring high relevance and personalization in these environments directly affects learning outcomes and user engagement. Educational portals pose additional challenges: short texts (titles), domain terminology, and multilingual or low-resource languages (for example, Kurdish and dialectal Arabic), all of which complicate both lexical matching and dense-embedding approaches.

Javaji and Sarode [11] proposed Multi-BERT, a hybrid method that fuses SBERT and Robustly Optimized BERT Approach (RoBERTa) by treating sentences as tokens to capture intra- and inter-sentence relations. The approach generates sentence and document embeddings, fuses them via sentence tokenization to enable cross-sentence interactions, and ranks items by combined similarity over the fused vectors. On a Goodreads children's subset (using genre overlap as a proxy for relevance), Multi-BERT reports Precision@5 = 0.9413, Precision@10 = 0.7889, and Precision@25 = 0.7621; reported baselines in the same study include SBERT (Precision@5 = 0.7563, Precision@10 = 0.7764, Precision@25 = 0.7294) and Term Frequency – Inverse Document Frequency (TF-IDF) (Precision@5 = 0.8164, Precision@10 = 0.8128, Precision@25 = 0.7877). The paper reports Precision@K only and notes that the relative advantage narrows or reverses at larger retrieval sizes, which the authors attribute to empirical and scaling trade-offs.

Building on dense embeddings, Juarto and Girsang [12] combined SBERT sentence/document embeddings with a neural collaborative filtering (NCF) supervised ranking layer for news recommendation. SBERT produces dense semantic representations that feed into the NCF ranker, which learns to score user–item pairs from interaction data. Their experiments report large gains (Precision ≈ 99.14%, Recall ≈ 92.48%, F1 ≈ 95.69%) on news variants, illustrating the effectiveness of pairing semantic retrieval with supervised ranking when sufficient interaction data and fine-tuning resources are available.

In contrast to dense-embedding pipelines, lexical and probabilistic methods remain strong, interpretable baselines, especially for short texts and low-resource languages. Yunanda *et al.* [13] implemented a practical news recommender using Term Frequency-Inverse Document Frequency (TF-IDF) and cosine similarity on a Microsoft News (MIND) subset: the titles are preprocessed (case folding, stopword removal, stemming), TF-IDF vectors are computed for titles, and each user is represented by a history vector (a 70/30 train/test split). Cosine similarity among user history vectors is used to identify candidate readers, and the most frequently clicked items among candidates form the top-10 recommendations. On 5,000 titles and 5,286 users (≥40 clicks), the system achieved a Hit Rate@10 of 80.77% (1,281 hits / 1,586 test users), showing that TF-IDF plus cosine similarity on title-level signals can yield strong practical hit rates given sufficient implicit click histories. The study does not report other common metrics such as Precision@K, Recall@K, F1, Normalized Discounted Cumulative Gain (NDCG), Mean Average Precision (MAP), or Mean Reciprocal Rank (MRR).

Complementing these term level methods, BM25 is widely used as a robust lexical retrieval baseline, with detailed theoretical and practical analyses available in the probabilistic relevance literature [14]. Wang and Yuan [15] introduce an interest level aware BM25 variant that partitions a user's history into term sets (e.g., must, constrained select, select) and adjusts BM25's term frequency contribution so then terms from higher interest sets boost the ranking more strongly. Evaluated on Tsinghua University course selection logs (test set: 300 students), their study reports results using SAT@k, a binary relevance metric indicating the percentage of users for whom at least one satisfactory course was found within the top-k recommendations. Specifically, they report SAT@1 = 44%, SAT@3 = 70%, SAT@5 = 89%, and SAT@10 = 93%. While these SAT@k scores are not directly comparable with rank-aware metrics like Precision@K, Recall@K, F1-score, or NDCG, the findings demonstrate that the task-aware weighting of BM25 can produce highly satisfactory top k recommendations in educational portals, motivating the use of interest-weighted term signals in our hybrid fusion.

Domain- and language-specific preprocessing is crucial for morphologically rich languages. Accordingly, Kazemifard [16] developed an emotional Arabic news recommender that combines TF-IDF content representations with collaborative filtering to strengthen item signals and incorporate emotion labels to filter and re-rank recommendations. The system applies Arabic normalization and stemming to mitigate morphological variation and integrates emotion information, sourced from user feedback and external analysis, into the final rankings. In a small user study, the hybrid approach achieved roughly 86% precision, 87% recall, and an F1 of about 86%; auxiliary emotion-detection components achieved high accuracy (~90%) with EEG-based signals but performed poorly (~42%) with generic tone analyzers, underscoring the need for Arabic-specific emotion models and careful preprocessing.

Alotaibi *et al*. [17] propose a content-based recommender for Arabic books that represents documents with TF-IDF vectors and compares them using cosine similarity, difflib. SequenceMatcher, and semantic methods (Doc2Vec for long texts; BERT for short texts/titles). Their corpus comprises 250 books from the Alshamela library across five genres; evaluation was performed using three human annotators who rated pairwise similarity on a 1–5 scale. Alotaibi *et al*. [17] report that cosine similarity on full-text TF-IDF vectors most closely matched human judgments, whereas semantic similarity on English-translated titles (used because of Arabic BERT tokenization issues) and SequenceMatcher on Arabic titles produced more variable outcomes. The study does not report standard IR metrics (Precision@K, Recall@K, F1, NDCG, Hit Rate), which limits direct quantitative benchmarking against other recommender studies.

In addition to education and news-focused recommenders, hybrid systems in e-commerce illustrate the effectiveness of combining content-based and collaborative filtering strategies. Yin and Zhang [18] propose a hybrid recommender for men's apparel that leverages textual product features, user interactions, and transformer-based embeddings (T5 and BERT) to generate contextually rich representations. Sentiment analysis of product reviews is incorporated to capture nuanced user preferences, and a weighted scoring mechanism determines the final recommendations. The evaluation shows that the hybrid system outperforms baseline methods such as FastText, Doc2Vec, and Word2Vec, achieving a similarity score of 94.76%, highlighting the potential of integrating multiple signals and advanced embeddings in hybrid recommendation frameworks.

Despite progress across domains, there have been no studies that have investigated recommendation systems in the Kurdish language. Existing studies in Kurdish Natural Language Processing (NLP) instead focus on creating varied datasets and tackling foundational tasks such as fake news detection, sentiment analysis, and text/news classification as described by Badawi *et al.* [19], referred to as entity recognition as described by Abdullah *et al.* [20], and script-aware text classification for Arabic-script languages, including Kurdish as previously described by Abdullah *et al.* [21]. These works highlight advances in resource creation and model adaptation but confirm the absence of recommendation-specific research for Kurdish. Table 1 provides an overview of the prior research on recommendation systems, including their datasets, methods, and reported performance.

Table 1: Summary of previous studies related to recommendation systems.

Ref	Technique	Language	Dataset	Results
[11]	Multi-BERT — fuse SBERT + RoBERTa via sentence tokenization; ranking by combined similarity	Multilingual (evaluated on Goodreads children's subset)	Goodreads—children's subset	P@5 / P@10 / P@25 — Multi-BERT: 0.9413 / 0.7889 / 0.7621; SBERT: 0.7563 / 0.7764 / 0.7294; TF-IDF: 0.8164 / 0.8128 / 0.7877. Only Precision@K reported; advantage narrows at larger K.
[12]	SBERT embeddings → Neural Collaborative Filtering (NCF) supervised ranker	English	News datasets	Precision ≈ 99.14%; Recall ≈ 92.48%; F1 ≈ 95.69%.
[13]	TF-IDF (title level) + cosine similarity; candidate selection via similar users then popular items	English	MIND subset: 5,000 titles; 5,286 users (≥40 clicks); 70/30 split	HitRate@10 = 80.77% (1,281 hits / 1,586 test users).
[15]	Interest-level aware BM25 (partition history; weight term contributions)	English	Tsinghua course selection logs (test set: 300 students)	SAT@1 = 44%; SAT@3 = 70%; SAT@5 = 89%; SAT@10 = 93% (binary satisfaction metric).
[16]	Emotion-aware hybrid: TF-IDF + collaborative filtering; emotion labels for reranking; Arabic preprocessing	Arabic	Small user study / thesis dataset	Hybrid: Precision ≈ 86%; Recall ≈ 87%; F1 ≈ 86%. Emotion detection: EEG ≈ 90%; generic tone ≈ 42%.
[17]	TF-IDF full-text vectors; cosine similarity; SequenceMatcher; semantic methods (Doc2Vec/BERT)	Arabic (books); English (titles)	250 books (Alshamela library);	Cosine TF-IDF (full text) achieved the highest agreement with human pairwise similarity judgments (representation agreement); no rank-aware recommender metrics were reported.
[18]	TF and BERT-derived SBERT	English.	Amazon Review Dataset	94.76% similarity scores

3. Materials and Methods

This section details the steps of creating a dataset and the recommendation system, including article categorization, hybrid ground truth generation, recommendation serving, and comprehensive evaluation.

3.1 System Architecture

The overall architecture of the proposed framework is modular and scalable, consisting of five main components: (1) Data Collection Module, (2) Data Storage Layer, (3) Article Categorization Module, (4) Recommendation Engine, and (5) Evaluation and Analytics Module. The system is designed for extensibility, allowing for the integration of additional universities, languages, or recommendation algorithms as needed. Figure 1 illustrates the processes and key stages of the study.

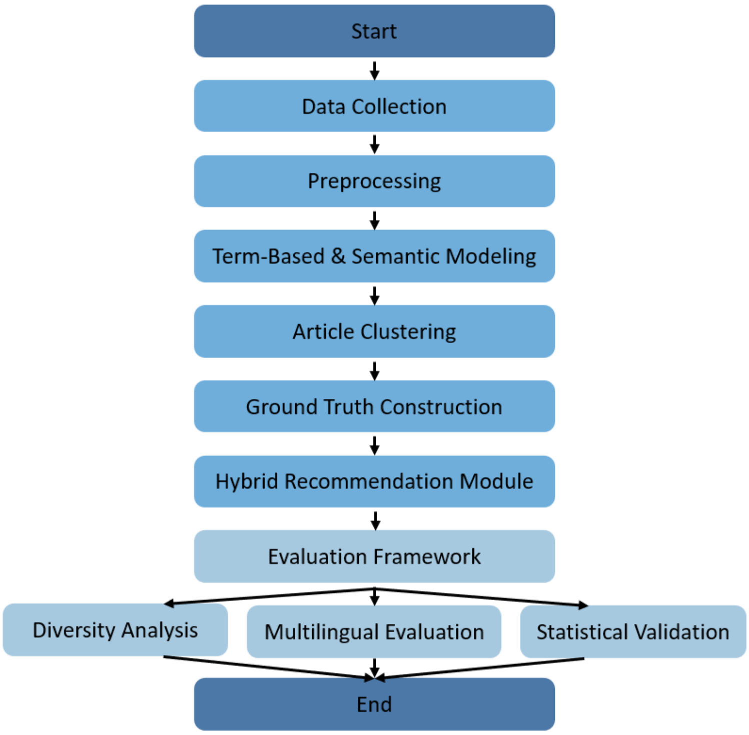

Figure 1: System architecture of the proposed multilingual hybrid recommendation framework.

3.2 *Dataset Creation*

A multilingual dataset was compiled from the public news pages of six Kurdistan university portals: Sulaimani Polytechnic University, University of Sulaimani, University of Garmian, University of Goizha, Qaiwan International University, and Koya University. Standard dataset-splitting and evaluation practices from news-recommendation benchmarks such as MIND were followed [22]. A purpose-built crawler was implemented in Python 3.8, using requests (v2.26.0) and BeautifulSoup4 (v4.11.1) for static pages, and Selenium (v3.141.0, headless Chrome) for dynamic content. Supporting packages included mysql-connector-Python for database interaction, Python-dotenv for configuration management, and lxml for robust parsing. Each portal was configured individually with listing URLs, prioritized CSS selectors for titles and article links, pagination rules, and a Selenium flag to indicate whether JavaScript rendering was required. The crawler handled paginated lists by following site-specific selectors up to a configurable limit. For each news item, the title and link were extracted using fallback logic if the primary selectors failed. The full article text was retrieved from the article URL using prioritized content selectors (e.g., post-content, entry-content, article p); where full content was unavailable, the first few paragraphs were stored. Adjustable delays (2–3 s) were introduced between requests to limit server load.

Each article record included:

- id (unique identifier)
- title (headline)
- content (cleaned text)
- link (article URL)
- language (en, ar, ku)
- url_hash (SHA-256 deduplication key)
- source (originating portal)

- created_at / updated_at (timestamps)
- additional experimental fields: category, embedding vector, train/test split flag, and unsupervised cluster ID. The full dataset and schema are released open-source for reproducibility.

3.3 *Preprocessing*

To ensure data quality and to match the system workflow (Figure 1), a preprocessing pipeline was applied after collection. Articles with missing or trivial titles/content were discarded during extraction. Duplicate records were prevented by computing SHA-256 hashes of URLs and enforcing uniqueness constraints. Unicode inconsistencies and boilerplate characters were normalized, and trivial tokens were excluded [23]. Language identification was performed heuristically by counting script-specific characters, mapping each record to English (en), Arabic (ar), or Kurdish (ku). Ambiguous or mixed-script records were filtered out.

This deduplication and normalization step yielded 1389 unique articles, comprising 604 English, 283 Arabic, and 502 Kurdish available on the GitHub repository (https://github.com/przhaB/Multilingual-News-Dataset). Figure 2 illustrates language proportions, and table 2 presents representative records with multilingual coverage. The repository contains:

- dataset.csv — full dataset (1,389 records, UTF-8 encoding preserved for Arabic/Kurdish scripts).
- schema.txt — textual schema description of released fields.
- crawler/ source code with export and reproducibility scripts.

Table 2: Sample records from the multilingual news dataset.

Id	Title	Language	Content	Source	Created date
101	New Research Center Opens in SPU	En	The Sulaimani Polytechnic University inaugurated a new research center…	Sulaimani Polytechnic University	2023-03-05
542	*بەرەوەری خوێندنی زانستە یەکەم لە زانکۆی سلێمانی*	Ku	*زانکۆی سلێمانی کۆنگرەیەکی زانستی بڕیاردا...*	University of Sulaimani	2023-04-21
870	*ورشة عمل عن الذكاء الاصطناعي*	Ar	*أقيمت ورشة عمل حول تطبيقات الذكاء الاصطناعي برعاية كلية...*	University of Garmian	2023-05-17

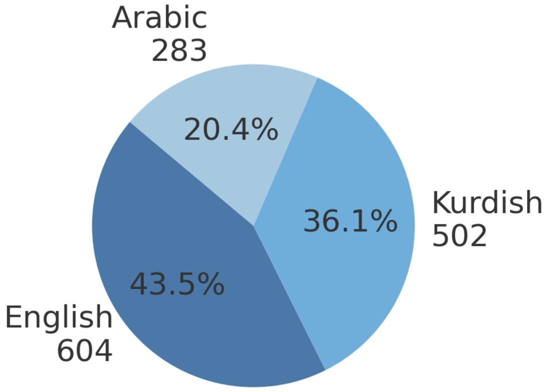

Figure 2: Language distribution in the multilingual university news dataset.

3.4 Hybrid Recommendation Module

The recommendation engine, illustrated in figure 1, is implemented as a hybrid module that integrates multiple similarity signals to generate robust and interpretable recommendations. Rather than relying on a single technique, the module combines TF–IDF lexical similarity, BM25 probabilistic relevance, and SBERT semantic embeddings [24].

TF–IDF captures token-level overlap and domain-specific terminology that may not always be represented in dense embeddings. BM25 provides a probabilistic relevance formulation, accounting for term frequency saturation and inverse document frequency weighting, which is particularly effective for short to medium-length news texts. SBERT, in contrast, encodes articles into contextualized embeddings, enabling semantic matching beyond surface-level token overlap, which is a critical feature for multilingual content.

Each similarity score is normalized to the [0,1] interval using min–max scaling. The final hybrid score for a query candidate pair is computed as a weighted sum of the normalized signals, as expressed in equation (1):

$$S_{\{hybrid\}} = w_{\{SBERT\}\backslash cdot}S^{\{norm\}}_{\{SBERT\}} + w_{\{TF!-!IDF\}\backslash cdot}S^{\{norm\}}_{\{TF!-!IDF\}} + w_{\{BM25\}\backslash cdot}S^{\{norm\}}_{\{BM25\}} \tag{1}$$

subject to the constraint:

$$w_{\{SBERT\}} + w_{\{TF!-!IDF\}} + w_{\{BM25\}} = 1 \tag{2}$$

Equation (2) ensures that the contributions of the individual similarity components are properly balanced and interpretable, preventing any single signal from dominating the hybrid score.

The weighting parameters are optimized on a validation split to balance precision and recall. To address cases where similarity signals diverge, fuzzy inference rules are employed. For instance, situations in which SBERT embeddings indicate high semantic similarity but lexical overlap remains minimal are adjusted to reduce false positives, ensuring that each component contributes proportionally to its reliability across different languages and content types.

By combining lexical, probabilistic, and semantic signals in this weighted manner, the hybrid module outputs a ranked list of articles for users. As shown in section 4, this approach outperforms individual components, particularly in multilingual settings where token overlap and embedding quality vary across languages.

3.5 Article Categorization through Clustering

To assign thematic categories to news articles, titles were vectorized using TF–IDF (unigrams, stop words removed) and clustered with K-Means (k = 5, random state = 42) using the Scikit-learn library [25]. Table 3 presents the clustering parameters for the five semantic labels used in diversity analysis. While clustering sweeps later explored a wider range of *k* for optimal internal cluster quality, a *k* of 5 was chosen for manual semantic labeling to align with broad thematic categories. Cluster labeling via manual inspection is an accepted human-in-the-loop practice for assigning semantically meaningful tags to clusters when automated labelers are unreliable [26]. In this work, each cluster was manually inspected, and a meaningful label was assigned based on the predominant theme of the articles in the cluster. The resulting five categories are: Research, Education, Events, Student Life, and Technology. These labels were stored in the database and later used when analyzing recommendation diversity.

Table 3: Clustering methodology parameters.

Step	Parameter/Value
Vectorization	TF–IDF (unigrams, stop-words removed)
Clustering	Clustering
Number of clusters	k = 5
Stability	Single run, random_state = 42
Output	Categories mapped to 5 semantic labels (Research, Education, Events, Student Life, Technology)

3.6 Clustering Setup and Evaluation

To explore latent thematic structures in the multilingual news dataset and identify the optimal number of clusters, clustering experiments were conducted separately for English, Arabic, and Kurdish articles. Titles were first normalized using the preprocessing pipeline described in Section 3.2 and transformed into TF–IDF vector representations. MiniBatchKMeans was used as the clustering algorithm with controlled random seeds to ensure reproducibility.

Internal clustering quality was assessed using three complementary metrics: the Silhouette coefficient [27], the Davies–Bouldin index [28], and the Adjusted Rand Index (ARI) [29]. These measures respectively quantify intra-cluster cohesion, inter-cluster separation, and agreement between clusterings across repeated runs.

Candidate numbers of clusters k∈{2,…,12}k \in \{2, \dots, 12\}k∈{2,…,12} were evaluated. Each configuration was applied to stratified subsamples of 100 articles per language, and performance scores were averaged across 10 independent runs to mitigate the effects of initialization. Table 6 summarizes the clustering outcomes, reporting the mean Silhouette and Davies–Bouldin scores (± standard deviation) along with the number of samples and features for each setting.

A consistent trend of monotonic improvement was observed as k increased. For all three languages, Silhouette values rose steadily with additional clusters, while Davies–Bouldin indices decreased, indicating that higher values of k yielded improved cluster separation and compactness without introducing fragmentation.

For English articles, performance stabilized at k = 10, achieving a Silhouette score of 0.87 and a Davies–Bouldin index of 0.31, representing the optimal balance between cohesion and separation. The Arabic and Kurdish datasets required finer granularity, with optimal values reached at k = 12. At this point, Silhouette scores were 0.82 for Arabic and 0.88 for Kurdish, accompanied by Davies–Bouldin indices of 0.40 and 0.33, respectively.

The low standard deviations across runs confirmed the stability of the clustering outcomes, while no singleton or degenerate clusters were observed at the chosen k. The bold rows in Table 6 highlight the selected values of k used for downstream analyses, along with the top TF–IDF terms characterizing each cluster. The reported No. samples values correspond to the 100-item subsamples used during the sweep, while the final cluster assignments and downstream analyses were validated on the full per-language corpora.

This systematic sweep complements the manual k = 5 labeling by providing quantitative evidence for cluster quality and robustness, ensuring that both semantic interpretability and internal consistency are considered when analyzing multilingual article collections.

3.7 Hybrid Ground Truth Generation

Accurate evaluation of the recommendation framework requires a reliable ground truth set of relevance pairs, indicating whether a candidate article is relevant to a given query article. In the absence of manually labeled data, a hybrid automatic ground truth was constructed using three complementary similarity measures computed in parallel: SBERT semantic similarity, TF–IDF lexical similarity, and BM25 probabilistic scoring. SBERT similarity was computed using the all-MiniLM-L6-v2 model (Sentence Transformers v2.2.0), applying cosine similarity between embeddings of query and candidate articles. TF–IDF similarity was calculated independently on preprocessed English, Arabic, and Kurdish text, including Unicode normalization, case folding, punctuation removal, and language-specific stopword removal, considering unigrams and bigrams with a minimum document frequency of 2 and a maximum of 5000 features. BM25 similarity was computed using BM25Okapi with tokenization into lowercased terms. Each method retrieved the top 50 candidates per query, and the union of these candidate lists formed the ground truth candidate pool, ensuring that no single similarity measure dominated the selection. The similarity scores from the three measures were normalized using min–max scaling and combined using a weighted sum to produce a single fused score for each query–candidate pair, as defined in equation 3:

$$s_{combined} = w_{sbert} \cdot s_{sbert}^{norm} + w_{tfidf} \cdot s_{tfidf}^{norm} + w_{bm25} \cdot s_{bm25}^{norm} \quad (3)$$

subject to:

$$w_{sbert} + w_{tfidf} + w_{bm25} = 1 \tag{4}$$

Weight values and threshold parameters were optimized on a validation split (50% of query articles, random seed = 42) to maximize the F1 score on the precision–recall curve. Candidate pairs exceeding the optimized threshold were labeled as relevant.

This sample illustrates that document pairs with relatively strong agreement across the methods typically exceed the decision threshold and are labeled Relevant, whereas pairs with weaker cross-method alignment fall below the cut-off and are consequently assigned a Non Relevant label. It should be emphasized that these examples represent automatically generated labels derived from algorithmic similarity fusion rather than human annotations. This strategy resembles distant supervision methods where labels are derived automatically from surrogate signals [30].

Each ground truth entry contains the query and candidate IDs, raw and normalized scores for each similarity measure, the fused score, the binary relevance label, and the provenance of the candidate within the unioned pool. This hybrid approach leverages the semantic, lexical, and probabilistic signals in a complementary manner. TF–IDF provides token-level details not always captured by SBERT embeddings, while multilingual preprocessing ensures consistent treatment across English, Arabic, and Kurdish. The combined weighted fusion improves coverage, reduces bias toward any single similarity measure, and produces a robust ground truth set suitable for evaluation using metrics such as Precision@K, Recall@K, F1 score, and NDCG@K.

4. Results

This section presents the evaluation results of the proposed multilingual hybrid news recommendation framework. It reports the performance of individual content-based filtering algorithms, the hybrid recommendation engine, the impact of clustering on recommendation diversity, and the system’s multilingual capabilities. In addition, the statistical significance of observed differences is examined, and the evaluation process is critically analyzed to ensure the reliability and validity of the findings.

4.1 Performance of Content-Based Filtering Algorithms

Table 4 reports the performance of the three individual content-based filtering algorithms (TF–IDF, BM25, SBERT) across four evaluation metrics: Precision@5, Recall@5, F1-score@5, and NDCG@5, calculated over the full dataset.

Table 4: Performance of individual content-based filtering algorithms on the full dataset (without ground truth filtering).

Algorithm	Precision@5	Recall@5	F1-score@5	NDCG@5
TF-IDF	0.75	0.23	0.27	0.81
BM25	0.75	0.20	0.25	0.79
SBERT	0.75	0.22	0.27	0.859

To complement the tabular presentation, the same results are visualized in figure 3. This grouped bar chart facilitates easier comparison among the algorithms across the four metrics. The chart shows that while all three methods achieve identical precision values (0.75**)**, they differ slightly in recall, F1, and NDCG. SBERT yields the highest NDCG (0.859), reflecting superior ranking quality, while TF–IDF demonstrates a marginal advantage in recall compared with BM25.

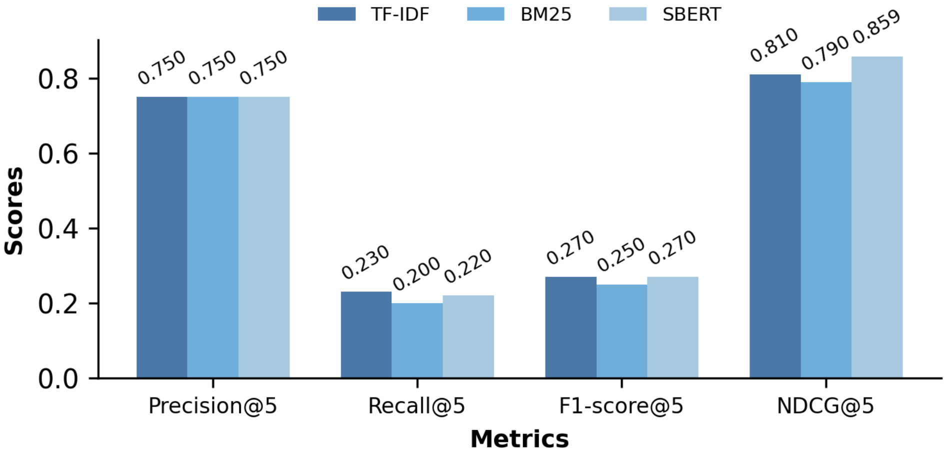

Figure 3: Performance of the individual content-based filtering algorithms (TF–IDF, BM25, SBERT) across four evaluation metrics.

4.2 Ground Truth Illustration

To provide greater clarity regarding the structure of the automatically generated ground truth dataset, a representative sample of labeled article pairs is presented in table 5. Each row corresponds to a query article paired with a candidate article, accompanied by their similarity scores from the three underlying methods (SBERT, TF–IDF, and BM25), the fused score obtained through weighted combination, and the final binary label (*Relevant* or *Non-Relevant*) assigned according to the learned threshold.

Table 5: Example entries of automatically generated ground truth.

Query ID	Candidate ID	SBERT score	TF–IDF score	BM25 score	Combined score	Label
101	205	0.82	0.31	0.55	0.67	Relevant
101	318	0.41	0.22	0.33	0.32	Non-Relevant
112	298	0.77	0.46	0.59	0.68	Relevant
120	450	0.59	0.28	0.41	0.43	Non-Relevant
135	278	0.88	0.39	0.62	0.74	Relevant

This sample illustrates that document pairs with relatively strong agreement across the methods typically exceed the decision threshold and are labeled *Relevant*, whereas pairs with weaker cross-method alignment fall below the cut-off and are consequently assigned a *Non-Relevant* label. It should be emphasized that these examples represent automatically generated labels derived from algorithmic similarity fusion rather than human annotations.

4.3 Article Clustering Results

Table 6 presents the clustering outcomes for English, Arabic, and Kurdish news articles across candidate cluster numbers k ranging from 2 to 12. The results report means Silhouette and Davies–Bouldin scores (± standard deviation) averaged across ten runs to ensure robustness.

A consistent trend of monotonic improvement was observed ask increased. For all three languages, Silhouette values rose steadily with the addition of clusters, while Davies–Bouldin indices decreased. This indicates that higher values of k enhanced both cluster cohesion and separation without fragmenting the data into overly small or unstable groups.

The English dataset exhibited the most rapid gains, with cluster quality improving sharply from k = 5 onwards. Performance stabilized at k = 10, where the Silhouette score reached 0.87 and the Davies–Bouldin index dropped to 0.31. This configuration provided the optimal balance between compact, well-separated clusters and interpretability.

By contrast, the Arabic and Kurdish datasets required finer granularity to achieve similar quality levels. Arabic articles continued to benefit from additional clusters up to k = 12, where the Silhouette score plateaued at 0.82 and the Davies–Bouldin index reached 0.40. The Kurdish dataset showed the strongest overall separation, achieving a Silhouette score of 0.88 and a Davies–Bouldin index of 0.33 at k = 12. These results suggest that the richer morphological variation and diversity of topics in Arabic and Kurdish text necessitate more clusters to adequately capture latent themes compared to English.

Consistency across runs was confirmed by the low standard deviations, which rarely exceeded 0.01. This stability indicates that the clustering outcomes were not sensitive to initialization, strengthening confidence in the reproducibility of the results. Furthermore, no singleton or degenerate clusters were observed at the chosen k, underscoring the robustness of the procedure.

The bold rows in table 6 highlight the selected k values used for downstream analyses. For each language, the corresponding top TF–IDF terms per cluster were also extracted, providing interpretable thematic labels that support subsequent qualitative exploration of the news corpora.

Table 6: Clustering sweep results (silhouette ± std, Davies–Bouldin ± std).

Languages	K	Silhouette	Davies-Bouldin	No. samples	No. features
English	2	0.173±0.001	1.951±0.650	100	107
English	3	0.255±0.005	1.947±0.044	100	107
English	4	0.342±0.002	1.665±0.012	100	107
English	5	0.436±0.002	1.420±0.004	100	107
English	6	0.526±0.003	1.219±0.001	100	107
English	7	0.615±0.003	1.007±0.032	100	107
English	8	0.700±0.002	0.854±0.005	100	107
English	9	0.785±0.001	0.701±0.013	100	107
English	**10**	**0.870±0.000**	**0.308±0.000**	**100**	**107**
Arabic	2	0.169±0.004	1.042±0.060	100	107
Arabic	3	0.248±0.005	1.470±0.277	100	107
Arabic	4	0.325±0.006	1.577±0.095	100	107
Arabic	5	0.408±0.004	1.466±0.039	100	107
Arabic	6	0.492±0.004	1.271±0.030	100	107
Arabic	7	0.579±0.002	1.087±0.014	100	107
Arabic	8	0.662±0.003	0.908±0.019	100	107
Arabic	9	0.746±0.000	0.731±0.000	100	107
Arabic	10	0.814±0.000	0.428±0.000	100	107
Arabic	11	0.815±0.009	0.460±0.080	100	107
Arabic	**12**	**0.821±0.001**	**0.402±0.086**	**100**	**107**
Kurdish	2	0.172±0.003	1.384±0.359	100	98
Kurdish	3	0.253±0.008	1.787±0.091	100	98
Kurdish	4	0.348±0.009	1.671±0.080	100	98
Kurdish	5	0.435±0.008	1.472±0.004	100	98
Kurdish	6	0.528±0.001	1.173±0.011	100	98
Kurdish	7	0.615±0.002	0.997±0.008	100	98
Kurdish	8	0.701±0.004	0.863±0.003	100	98
Kurdish	9	0.786±0.002	0.680±0.030	100	98
Kurdish	10	0.869±0.000	0.305±0.000	100	98
Kurdish	11	0.878±0.005	0.271±0.014	100	98
Kurdish	**12**	**0.882±0.008**	**0.332±0.120**	**100**	**98**

The bold rows indicate the k selected for downstream use and top TF-IDF terms per cluster chosen, and the complete sweep CSV is provided as supplementary material. The 'No. samples' values reported in table 6 correspond to the 100-item subsamples used during the sweep. Final cluster assignments and downstream analyses were validated on the full per-language corpora.

4.4 Performance of the Hybrid Recommendation Engine

Using weighted averaging, the hybrid recommendation engine combines SBERT, TF-IDF, and BM25. Table 7 compares the hybrid model to the best-performing individual algorithm SBERT on the subset of articles with ground-truth data.

Table 7: Performance comparison between the hybrid model and SBERT on the ground truth subset.

Algorithm	Precision@5	Recall@5	F1-score@5	NDCG@5
Hybrid	0.96	0.19	0.22	0.945
SBERT	0.93	0.16	0.19	0.859

4.5 Impact of Clustering on Recommendation Diversity

The impact of clustering on the diversity of news recommendations was assessed by analyzing the distribution of recommended articles across thematic clusters and the entropy of said distributions. The framework was evaluated under two conditions: with and without clustering. When clustering was applied, the recommendations spanned five distinct clusters, resulting in an entropy score of 1.697. In contrast, the absence of clustering led to recommendations being concentrated in only two clusters, with a substantially lower entropy score of 0.032. To further interpret these clusters, representative terms were extracted from the article titles associated with each group. As summarized in table 8, the clusters correspond to meaningful university-related themes such as workshops, announcements, publications, institutional agreements, and cultural activities.

Table 8: Thematic interpretation of clusters with representative terms.

Cluster	Representative keywords	Interpreted theme
0	workshop, training, faculty, session, student	Academic workshops & training
1	conference, announcement, ceremony, rector, opening	University events & announcements
2	journal, publication, Scopus, award, research	Research outputs & publications
3	memorandum, collaboration, agreement, partner, cooperation	Institutional partnerships & MoUs
4	festival, culture, exhibition, seminar, community	Cultural & community engagement

As illustrated in figure 4, clustering enabled the recommendations to be distributed across all identified thematic groups, whereas the absence of clustering led to a concentration on announcements (Cluster 1). By broadening coverage across the clusters, the entropy increased significantly, confirming that clustering enhances recommendation diversity. This improvement is crucial for reducing redundancy and mitigating echo-chamber effects in academic news recommendations.

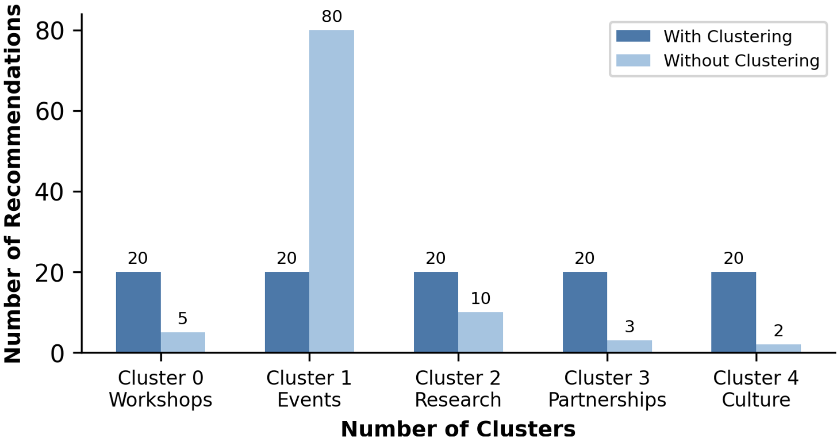

Figure 4: Distribution of recommendations with and without clustering, demonstrating the impact of clustering on recommendation diversity.

4.6 Evaluation of Multilingual Support

The multilingual evaluation indicates a consistent pattern across languages. Precision is high for all languages (Kurdish = 0.946, English = 0.854, Arabic = 0.983), while recall is substantially lower, particularly for Kurdish (0.085) and Arabic (0.066). This demonstrates that although the retrieved items are generally correct, a significant portion of relevant content remains un-retrieved in low-resource languages.

Several factors contribute to this outcome. First, dataset imbalance exists, as the Arabic and Kurdish subsets are considerably smaller than the English subset, limiting coverage and reducing retrieval diversity. Second, morphological and orthographic variation is pronounced in both languages, with complex morphology, optional diacritics, and multiple script conventions complicating tokenization and decreasing the effectiveness of TF–IDF and BM25 matching. Third, embedding coverage varies; multilingual embedding models such as multilingual Bidirectional Encoder Representations from Transformers (mBERT), Cross-lingual Language Model – RoBERTa (XLM-R), and multilingual SBERT variants perform strongly for high-resource languages, but their representation quality is weaker for Kurdish and dialectal Arabic, restricting retrieval performance in these languages, consistent with broader observations in low-resource NLP.

These findings have practical implications. The **r**ecommendations for Arabic and Kurdish are highly precise but often omit other relevant documents, potentially reducing content coverage and limiting perceived utility. The results underscore the structural challenges inherent in multilingual recommendation within resource-imbalanced and linguistically complex environments.

4.7 Multilingual Performance Results

For fair cross-language comparison, approximately 300 articles in each language were used to test the system (English = 300, Kurdish = 300, Arabic = 283). Precision@5, Recall@5, and NDCG@5 were calculated using balanced subsets for each language. A summary of the results is presented in table 9.

Table 9: Multilingual recommendation performance by language

Language	Precision@5	Recall@5	NDCG@5	Sample size
Kurdish	0.946	0.085	0.745	300
English	0.854	0.268	0.789	300
Arabic	0.983	0.066	0.893	283

The results are computed on balanced evaluation subsets of ~300 articles per language (English = 300, Kurdish = 300, Arabic = 283) to enable comparability across high- and low-resource languages. The whole dataset sizes for reference are English = 604, Kurdish = 502, and Arabic = 283.

4.8 Statistical Significance Testing

To rigorously assess the differences in recommendation performance between the hybrid model and the best-performing individual content-based algorithm (SBERT), a paired t-test was conducted on the NDCG@5 scores. This statistical approach ensures that observed improvements are not attributable to random variation and enhances the reliability of the findings.

The analysis was performed on a random sample of 200 articles with available ground truth data. The results, summarized in table 10, indicate that the hybrid model significantly outperforms the SBERT baseline.

Table 10: Results of the statistical significance test comparing the hybrid and SBERT models.

Metric	Hybrid mean	SBERT mean	t-statistic	p-value	Effect size (Cohen's d)	Sample size
NDCG@5	0.9450	0.8595	6.1850	0.000001	0.3322	200

5. Discussion

The results demonstrate that a hybrid approach combining SBERT, TF-IDF, and BM25 with clustering and a fuzzy-fusion layer significantly improves top-k recommendation performance on a multilingual, low-interaction academic news corpus (Precision@5 = 0.96; NDCG@5 = 0.945; paired t-test on NDCG@5: p = 0.00001).

The system architecture integrates SBERT for semantic sentence embeddings, TF-IDF and BM25 for token-level matching, an unsupervised clustering stage for topical diversification, and a fuzzy-fusion module to combine signals into final rankings. The hybrid configuration outperformed SBERT alone (SBERT: Precision@5 = 0.93; NDCG@5 = 0.859). The methodological details of the automatic ground truth generation are described in **s**ection 3.5, and the statistical outcomes are reported in table 10. These findings align with the prior work showing that combining lexical and semantic signals improves top-k ranking quality [11, 12, 14].

TF-IDF and BM25 acted as stable, interpretable baselines (Precision@5 ≈ 0.75). Token-level and probabilistic retrieval methods often remain competitive for short texts and in resource-constrained settings [13-15]. The higher performance reported in some prior studies typically reflects access to interaction histories or supervised re-ranking models; when such signals are unavailable, lexical plus semantic fusion provides a practical alternative for improved top-k performance [12, 13, 15].

Clustering substantially increased thematic coverage and diversity of recommendations. Entropy rose from 0.032 without clustering to 1.697 with clustering. Clustering redistributed the recommendations away from announcement-concentrated results toward a broader mix of research, events, student life, and technology topics. Table 8 summarizes this shift. This outcome supports hybrid-recommender literature advocating multi-perspective retrieval and clustering to reduce redundancy and broaden coverage in the recommendation lists [11, 14]. Such hybrid recommender strategies have been widely observed to outperform individual embedding-based models in broader recommendation research [31].

The language-specific results show notable disparities. English items achieved higher recall while Kurdish and Arabic exhibited much lower recall (Kurdish 0.085; Arabic 0.066). Contributing factors include dataset imbalance, orthographic variation, and a complex morphology that degrades token-based matching and reduces the effective quality of off-the-shelf multilingual encoders for under-resourced languages. Prior Arabic-focused studies recommend careful normalization and auxiliary signals to mitigate such issues [16]. Comparable work for Kurdish highlights the importance of tailored preprocessing and representation choices [17]. Addressing language-specific preprocessing and expanding minority-language data are therefore priorities.

A central methodological limitation is the reliance on the automatically generated ground truth labels produced by the fused SBERT, TF–IDF, and BM25 signals (Section 3.6). The evaluation framework relied exclusively on these automatically generated binary relevance labels (Relevant / Non-Relevant), which served as the benchmark for metrics including Precision@5, Recall@5, F1 score, and NDCG@5. No manual validation of the automatic labels was performed in the present study. This limitation is important because automatic labeling may introduce biases in evaluation—for instance, inflating measured precision or under-representing unretrieved relevant items. In multilingual settings, cultural references, idiomatic expressions, and orthographic variation can influence relevance judgments in ways that automatic methods cannot fully capture. Consequently, absolute metric values should be interpreted with caution, though the main comparative finding—that the hybrid configuration outperforms SBERT under the same evaluation protocol—remains valid. Future work should therefore include manual validation using independent annotators across English, Arabic, and Kurdish, with conflicts adjudicated to establish a gold standard and inter-annotator agreement quantified using metrics such as Cohen's κ. Such human annotation will provide a more reliable benchmark, enable the quality of automatic labels to be quantified, and enhance the robustness and external validity of evaluation outcomes.

The evaluation should also extend beyond precision and recall. User-centered studies that measure diversity, novelty, and user trust will provide a fuller picture of real-world utility in academic portals. Incorporating direct user feedback enables the assessment of subjective satisfaction and perceived relevance. Combining such studies with controlled offline validation will clarify whether the high-precision, low-recall pattern persists in operational use and how it affects user satisfaction.

Practical recommendations and next steps are the following. First, expand and re-balance the corpus to increase the coverage for Kurdish and Arabic. Second, implement language-specific normalization and tokenization pipelines (for example, stemming, diacritics handling, and script normalization) that have demonstrated benefits for Arabic-script languages [16, 17]. Third, explore semi-supervised or supervised re-ranking when interaction data become**s** available to capture **the** additional gains reported in supervised pipelines [12]. Fourth, integrate collaborative, explainable, or feedback-driven modules into the modular fusion architecture to improve personalization and user trust [14].

The evaluation framework in this study relied on automatically generated binary relevance labels (Relevant / Non-Relevant), introduced in section 3.5 (Hybrid Ground Truth Generation). These labels were derived through a hybrid fusion of SBERT semantic similarity, TF–IDF lexical overlap, and BM25 scores, and they served as the benchmark for evaluation metrics including Precision@5, Recall@5, F1 score, and NDCG@5. No manual validation of the automatic labels was performed in the present study. Consequently, all reported results are based exclusively on the hybrid automatic ground truth. The motivation and protocol for future manual validation work will involve employing independent annotators across English, Arabic, and Kurdish, with conflicts adjudicated to establish gold standard labels and inter-annotator agreement assessed. While hybrid labeling provides a scalable foundation, the absence of human validation represents a limitation. In multilingual settings, cultural references, idiomatic expressions, and orthographic variation can influence relevance judgments in ways that automatic methods cannot capture. Future human annotation will therefore provide a more reliable gold standard and enable the quality of the automatic labels to be quantified, enhancing the robustness of the evaluation outcomes.

In summary, relative to the studies cited in section 2 [11, 17], the hybrid approach attains competitive top-k precision and improved ranking quality in a multilingual, low-interaction academic-news setting. Observed differences with some prior reports are attributable primarily to dataset composition, supervision level, and language resources rather than to deficiencies of the fusion strategy. The framework provides a practical, modular foundation for low-resource academic portals with a clear roadmap to strengthen language support, validation, and user alignment.

6. Conclusions

This work presents the development and evaluation of a multilingual hybrid news recommendation system created for educational web portals that operate under resource limitations and across multiple languages. The design combines unsupervised clustering, fuzzy logic, traditional content-based approaches such as TF-IDF and BM25, semantic embeddings with SBERT, and a custom web crawler. Bringing these elements together helps the system cope with challenges around scarce data, the absence of user profiles, and the complexity of working with more than one language. The experiments showed that clustering broadened the range of recommendations and that the hybrid model generally achieved higher accuracy and ranking quality than any of the individual algorithms tested.

There are, however, some important limitations. The evaluation is based on a small set of metrics and on automatically generated ground truth, both of which can introduce bias. Future research should expand the dataset to encompass more languages, incorporate expert or crowdsourced validation into the ground truth process, and leverage behavioral data and explicit user feedback to achieve more adaptive personalization. It is also important to look beyond precision and recall to user-oriented criteria such as diversity, novelty, and trust, ideally measured through user studies.

Overall, the system provides a foundation for further research on personalized recommendations and multilingual information retrieval in academic settings. Because the framework and dataset have been built to be extensible, other researchers or institutions can adapt them for a variety of domains, including education, civic information, and healthcare services.

Author contribution: Przha Barzan Mohialddin: Original article, Data curation, Writing –original draft. **Sarkar Hasan Ahmed:** Supervision, Validation, Visualization, Writing –review & editing.

Data availability: Data will be available upon reasonable request by the authors**.**

Conflicts of interest: The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Funding: The authors did not receive support from any organization for conducting the study.

References

[1] A. Iana, M. Alam, and H. Paulheim, "A survey on knowledge-aware news recommender systems," *Semantic Web*, vol. 15, no. 1, pp. 21–82, 2024. doi: 10.3233/SW-222991.

[2] J. Bobadilla, F. Ortega, A. Hernando, and A. Gutiérrez, "Recommender systems survey," *Knowledge-Based Systems*, vol. 46, pp. 109–132, 2013. doi: 10.1016/j.knosys.2013.03.012.

[3] S. S. Kundu, D. Sarkar, P. Jana, and D. K. Kole, "Personalization in education using recommendation system: An overview," in *Computational Intelligence in Digital Pedagogy*. Singapore: Springer, 2020, pp. 85–111, doi: 10.1007/978-981-15-5258-8_5.

[4] A. Agbeyangi and H. Suleman, "Advances and challenges in low-resource-environment software systems: A survey," *Informatics*, vol. 11, no. 4, p. 90, 2024. doi: 10.3390/informatics11040090.

[5] J. Lin et al., "How can recommender systems benefit from large language models: A survey," *ACM Transactions on Information Systems*, vol. 43, no. 2, pp. 1–47, 2025. doi: 10.1145/3678004.

[6] Z. Chen, W. Gan, J. Wu, K. Hu, and H. Lin, "Data scarcity in recommendation systems: A survey," *ACM Transactions on Recommender Systems*, vol. 3, no. 3, pp. 1–31, 2025. doi: 10.1145/3639063.

[7] K. M. Awlla, H. Veisi, and A. A. Abdullah, "Sentiment analysis in low-resource contexts: BERT's impact on Central Kurdish," *Language Resources and Evaluation*, vol. 59, no. 1, pp. 1–31, 2025. doi: 10.1007/s10579-024-09720-2.

[8] S. Bansal, K. Gowda, and N. Kumar, "Multilingual personalized hashtag recommendation for low-resource Indic languages using graph-based deep neural networks," *Expert Systems with Applications*, vol. 236, p. 121188, 2024. doi: 10.1016/j.eswa.2023.121188.

[9] Y. Ge et al., "A survey on trustworthy recommender systems," *ACM Transactions on Recommender Systems*, vol. 3, no. 2, pp. 1–68, 2025. doi: 10.1145/3652891.

[10] E. Purificato, L. Boratto, and E. W. De Luca, "User modeling and user profiling: A comprehensive survey," *arXiv preprint*, arXiv:2402.09660, 2024. doi: 10.48550/arXiv.2402.09660.

[11] S. R. Javaji and K. Sarode, "Multi-BERT for Embeddings for Recommendation System," *arXiv preprint* arXiv:2308.13050, 2023. doi: 10.48550/arXiv.2308.13050.

[12] B. Juarto and A. Suganda Girsang, "Neural collaborative with sentence BERT for news recommender system," *International Journal on Informatics Visualization*, vol. 5, no. 4, p. 448, 2021. doi: 10.30630/joiv.5.4.678.

[13] G. Yunanda, D. Nurjanah, and S. Meliana, "Recommendation System from Microsoft News Data using TF IDF and Cosine Similarity Methods," *Building of Informatics, Technology and Science*, vol. 4, no. 1, 2022. doi:10.47065/bits.v4i1.1670.

[14] S. E. Robertson and H. Zaragoza, "The probabilistic relevance framework: BM25 and beyond," *Foundations and Trends® in Information Retrieval*, vol. 3, no. 4, pp. 333–389, 2009. doi: 10.1561/1500000019

[15] X. Wang and F. Yuan, "Course Recommendation by Improving BM25 to Identify Students' Different Levels of Interests in Courses," International Conference on New Trends in Information and Service Science, 2009. doi:10.1109/NISS.2009.104.

[16] M. Kazemifard, "Emotional Arabic News Recommender System," M.Sc. Thesis, 2017.

[17] S. Alotaibi, and M. B. Khan, "Development of the recommender system of Arabic books based on the content similarity," *International Journal of Computer Science and Network Security*, vol. 22, no. 8, 2022. doi: https://doi.org/10.22937/IJCSNS.2022.22.8.23

[18] C. Yin and Z. Zhang, "A Study of sentence similarity based on the All-minilm-l6-v2 model with 'Same Semantics, Different Structure' after fine tuning," Second International Conference on Image, Algorithms and Artificial Intelligence (ICIAAI), 2024, pp. 677–684, Atlantis Press. doi: 10.2991/978-94-6463-540-9_69.

[19] S. Badawi, A. M. Saeed, S. A. Ahmed, P. A. Abdalla, and D. A. Hassan, "Kurdish news dataset headlines (KNDH) through multiclass classification," *Data in Brief*, vol. 48, p. 109120, 2023. doi: 10.1016/j.dib.2023.109120.

[20] A. A. Abdullah et al., "NER-RoBERTa: Fine-tuning RoBERTa for named entity recognition (NER) within low-resource languages," *arXiv preprint*, arXiv:2412.15252, 2024. doi: 10.48550/arXiv.2412.15252.

[21] A. A. Abdullah, A. H. Gandomi, T. A. Rashid, S. Mirjalili, L. Abualigah, M. Živković, and H. Veisi, "The role of orthographic consistency in multilingual embedding models for text classification in Arabic-script languages," *arXiv preprint*, arXiv:2507.18762, 2025. doi: 10.48550/arXiv.2507.18762.

[22] H. Wu, F. Dai, R. Lv, H. Dong, T. Su, Z. Liu, Y. Yang, Y. Jiang, and Z. Wang, "MIND: A large-scale dataset for news recommendation," in Proc. 58th Annual Meeting of the Association for Computational Linguistics (ACL), 2020, pp. 3597–3606. Available: https://aclanthology.org/2020.acl-main.331/

[23] C. D. Manning, P. Raghavan, and H. Schütze, Introduction to Information Retrieval. Cambridge, UK: Cambridge University Press, 2008. doi:10.1017/CBO9780511809071

[24] R. Burke, “Hybrid recommender systems: Survey and experiments,” *User Modeling and User-Adapted Interaction*, vol. 12, no. 4, pp. 331–370, 2002. doi: 10.1023/A:1021240730564

[25] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and É. Duchesnay, “Scikit-learn: Machine learning in Python,” *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011. Available: https://jmlr.org/papers/v12/pedregosa11a.html

[26] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent Dirichlet Allocation," *Journal of Machine Learning Research*, vol. 3, pp. 993–1022, 2003. doi: 10.1162/jmlr.2003.3.4-5.993.

[27] P. J. Rousseeuw, "Silhouettes: A graphical aid to the interpretation and validation of cluster analysis," *Journal of Computational and Applied Mathematics*, vol. 20, pp. 553-65 1987. doi:10.1016/0377-0427(87)90125-7.

[28] D. L. Davies and D. W. Bouldin, "A cluster separation measure," IEEE Transactions on *Pattern Analysis and Machine Intelligence*, vol. PAMI-1, no. 2, pp. 224--227, 1979. doi:10.1109/TPAMI.1979.4766909.

[29] L. Hubert and P. Arabie, "Comparing partitions," *Journal of Classification*, vol. 2, no. 1, pp. 193–218, 1985. doi: 10.1007/BF01908075.

[30] M. Mintz, S. Bills, R. Snow, and D. Jurafsky, “Distant supervision for relation extraction without labeled data,” in Proceedings of the 47th Annual Meeting of the Association for Computational Linguistics, Singapore, 2009, pp. 1003–1011. doi: 10.3115/1690219.1690287

[31] S. Zhang, L. Yao, A. Sun, and Y. Tay, “Deep learning-based recommender system: A survey and new perspectives,” *ACM Computing Surveys*, vol. 52, no. 1, pp. 1–38, 2019. doi: 10.1145/3285029