TopicMap-BN: Scalable and explainableFramework for Cross-Source Bangla News Recommendation with BanglaBERT and BERTopic

Md Hasan Hafizur Rahman [1] and Sumaia Afrin Sunny [2]

[1] Department of Computer Science and Engineering, Comilla University, Cumilla - 3506, Bangladesh
[2] Department of Bangla, Comilla University, Cumilla - 3506, Bangladesh

ABSTRACT

With the rapid growth of online Bangla news portals, thousands of articles are published daily on similar topics, resulting in an information overload for readers. Existing recommendation systems mostly focus on personalized suggestions based on user history, whereas readers frequently desire related news on a given topic across multiple sources. This challenge is amplified by the scarcity of robust Bangla Natural Language Processing (NLP) tools and the heterogeneous structure of news content. In this regard, we introduce TopicMap-BN, a scalable and explainable topic-based framework for cross-source Bangla news recommendation system. This system integrates Bangla-specific preprocessing with neural topic modeling (BERTopic with transformer embeddings), near-duplicate detection (MinHash and SimHash), and diversity-aware re-ranking (MMR, xQuAD, DPP). These components facilitate coherent story grouping, interpretable topic labels, and recommendations that maintain relevance, freshness, and diversity. The effectiveness of the proposed framework is demonstrated by experiments carried out on the Potrika corpus (approximately 665,000 articles) and live crawls from five popular news portals. As a result, the system achieved topic quality, demonstrated by an NPMI score of 0.62 and a human agreement value (κ) of 0.71. In terms of story deduplication, it secured a precision of 0.91 and an F1-score of 0.88, indicating reliable clustering of near-duplicate articles. Moreover, the framework demonstrates its ability to produce precise and well-ranked recommendations by having a precision at rank five of 0.72 and an NDCG at rank five of 0.75. Compared with classical baselines such as TF–IDF with cosine similarity, TopicMap-BN achieves substantial gains across coherence, ranking, and diversity.These findings confirm the feasibility of cross-source Bangla news recommendation and emphasize the significance of domain-specific NLP frameworks in low-resource settings.

KEYWORDS

Bangla News Recommendation, Topic modeling; BanglaBERT; BERTopic; News deduplication; Diversity-aware re-ranking; Low-resource languages; Natural language processing

1. INTRODUCTION

Bangla (Bengali) is the seventh most frequently spoken language in the world, with over 230 million native speakers and more than 300 million speakers worldwide, serving as the state language of Bangladesh and the second most widely spoken language in India(Eberhard, 2015; (UNFPA), 2025). Despite its significance to many people, Banglahas continued to remain a low-resource language within the domain of Natural Language Processing (NLP). There are fewer production-quality tools, benchmarks, and deployable systems for Bangla than for high-resource

languages like English(Alam, 2021; Sun, 2024). This shortage presents unique challenges for cross-source news recommendation tasks, since the large volume of published articles, frequently appearing media reports, and extensively edited headlines intensify the problem of information overload for audiences.

Moreover, the exponential proliferation of online news portals around the world has transformed the way news is produced and consumed. Every day, thousands of news articles are published on similar topics by different news organizations, leading to information overload. Readers, on the other hand, are often interested in following particular events or topics across different sources to gain a balanced perspective and validate authenticity. Although content-based personalized news recommendation systems have been extensively examined in English and other popular languages (Li, 2011), the advancement of analogous systems for Bangla is still limited. The gap is particularly concerning due to the growing digital engagement of Bangla speakers. The Bangladesh Telecommunication Regulatory Commission (BTRC[1]) has reported that internet subscriptions in Bangladesh have surpassed 130 million in 2023, in contrast to a just 0.3% of the population in 2007. In conjunction with this increase, over 400 Bangla news portals, such as Prothom Alo[2], Samakal[3], and Bdnews24[4], disseminate thousands of news articles daily, complicating the task for users to manually identify associated information. The extensive and continuously updated news ecosystem challenges the manual identification of similar information across multiple sources, emphasizing the critical need for automated, topic-aware, and diversity-sensitive recommendation systems specifically designed for Bangla.

Moreover, the diverse characteristics of online Bangla news sources provide significant challenges for users intending to navigate multiple websites manually. This manual browsing has complicated the task of retrieving pertinent articles, in particular when retrieving similar content from numerous websites. In order to resolve this matter, Bangla news information is automatically processed by evaluating the degree of similarity between documents. However, two major challenges impede this goal: (1) the structural heterogeneity of news articles across sources and (2) the restricted accessibility of NLP tools for Bangla. These issues are further complicated by the lack of uniform formatting in the representation of news on the web, since these articles are often embedded within HTML[5] pages containing noisy tags. Although a large volume of Bangla news content is available online, the tools required to process this content effectively, such as a thesaurus, stop-word lists, stemmers or morphological analyzers, and Part-of-Speech (POS) taggers, have still been under development. Although some research endeavors have attempted to address aspects of Bangla language processing, there is a scarcity of comprehensive and deployable systems for managing online Bangla text.

In this work, we focus on measuring document relatedness among Bangla news articles available on the web. To this end, we have developed a dedicated web crawler to automatically retrieve articles from multiple sources. The retrieved content often contains HTML tags and formatting noise, which we remove using the jsoup[6] parser. After cleaning, we perform tokenization, stop-word removal, and lemmatization to reduce the data dimensionality, representing each document as a bag-of-words in a Vector Space Model (VSM). The importance of each word is quantified using Term Frequency - Inverse Document Frequency (TF–IDF), and the relatedness between

---

[1]https://lims.btrc.gov.bd/

[2]https://www.prothomalo.com/

[3]https://samakal.com/

[4]https://bdnews24.com/

[5]https://www.w3.org/TR/2011/WD-html5-20110405/

[6]https://jsoup.org/

documents is computed using cosine similarity, enabling the retrieval of specific related news articles across portals.

Building upon these foundations, we propose TopicMap-BN, a topic-based recommendation framework specifically designed for Bangla. The system organizes incoming articles into interpretable and stable topics and story groups and subsequently generates personalized recommendations using content-based profiles combined with diversity-aware re-ranking. Our main contribution is a Bangla-centric, topic-oriented recommendation system that integrates articles from multiple sources and classifies them into comprehensible topics using neural topic modeling. We employ BERTopic (Bhattacharjee A. a., 2021), which combines transformer-based embeddings with class-based TF–IDF (c-TF–IDF) to generate coherent and interpretable topic descriptors. In the next step, we propose a robust deduplication strategy to mitigate redundancy from syndicated reports and superficially modified headlines. This approach integrates shingling(Broder, 1997), which partitions text into overlapping word sequences, with probabilistic hashing methods such as MinHash (Broder, 1997)and SimHash(Charikar, 2002), thereby enabling efficient detection and clustering of near-duplicate articles. Next, we design a diversity-aware ranking module that balances topical relevance with diversity of sources and viewpoints. This module leverages established re-ranking methods, including Maximal Marginal Relevance (MMR)(Carbonell, 1998), xQuAD (Explicit Query Aspect Diversification(Santos R. L., 2013), and Determinantal Point Processes (DPPs)(Kulesza, 2012). By explicitly modeling diversity, these methods allow recommendations to adapt to user preferences while reducing redundancy and mitigating the risk of echo chamber formation.By integrating these components, TopicMap-BN provides a systematic and scalable solution for Bangla news recommendation that combines interpretability, personalization, and content diversity, while addressing the linguistic and infrastructural challenges of a low-resource yet globally significant language.

The rest of this paper is structured as follows. Section 2 reviews existing literature on document similarity, neural topic modelling, deduplication, and diversity-aware re-ranking techniques relevant to news recommendation. The proposed methodology is explained in Section 3, which includes the pipeline for article ingestion, preprocessing, embedding generation, topic discovery, and story-level clustering. Section 4 outlines the datasets used, including live crawls from leading Bangla news portals and the Potrika corpus, in addition to summarization resources. Section 5 delineates the preprocessing and normalization tasks that are specifically designed for Bangla corpora. These procedures include morphological normalization, noise elimination, tokenization, and script standardization. Topic discovery and story clustering are incorporating BERTopic with transformer embeddings and hybrid MinHash -- SimHash deduplication in Section 6. The recommendation engine is introduced in Section 7, which also addresses cold-start and fairness issues. It emphasizes user modelling, scoring, and diversity-aware re-ranking strategies. Experimental results are presented in Section 8, which assesses the effectiveness of recommendations, deduplication accuracy, and topic quality on both live and offline datasets. Finally, Section 9 concludes with the main findings and discusses future research directions for extending the framework toward fairness-aware, multilingual news recommendation

2. RELATED WORKS

The rapid growth of online news has driven substantial research into recommendation systems aimed at mitigating information overload. Early systems have predominantly utilized users' historical reading behavior to provide personalized suggestions. For example, *News Dude* has been described as a content-based recommender agent that leverages TF–IDF representations and the K-Nearest Neighbor (KNN) algorithm to recommend articles based on prior reading preferences, with cosine similarity used to measure relationships between articles in the consumer space(Billsus, 1999). Similarly, hierarchical incremental clustering has been proposed to

dynamically capture readers' evolving interests through tree-like structures (Godoy, 2006). Subsequent approaches have integrated user profiles with content-based features, facilitating hybrid personalization methodologies (Adomavicius, 2005). Although effective, these systems have primarily targeted resource-rich languages such as English.

Large-scale benchmarks such as the Microsoft News Dataset (MIND), comprising over one million users and 160,000 articles, have accelerated the development of neural recommenders that blend content encoders with sequence models to capture user dynamics(Wu F. a.-H., 2020). For instance, self-attention-based models have been introduced for news recommendations, achieving significant performance gains(Wu C. a., 2019). However, these benchmarks remain English-centric, highlighting the lack of equivalent infrastructure for low-resource languages such as Bangla.

Bangla has continued to be a low-resource language in NLP due to the scarcity of annotated corpora and production-quality tools. (Alam, 2021) have provided a comprehensive survey of Bangla NLP tasks, identifying persistent gaps compared to resource-rich languages despite promising transformer-based results. More recently, (Rabbi, 2024) have developed a Python-based NLP toolkit supporting annotation, tokenization, POS tagging, stemming, and sentiment analysis, emphasizing the need for open-source resources to accelerate Bangla NLP research. Representation learning for Bangla has also progressed. BanglaBERT, a monolingual BERT model trained on 27.5 GB of Bangla text, has improved a range of downstream tasks(Bhattacharjee A. a., 2021). IndicBERT and its successor IndicBERTv2, trained on the IndicCorp corpus spanning multiple Indic languages, have provided multilingual embeddings applicable to Bangla(Madanbhavi, 2024; Kakwani, 2020). In addition, multilingual models such as LaBSE(Feng, 2020) and MiniLM paraphrase models(Reimers, 2020) have supported clustering and retrieval. Recent community contributions, including BongLLaMA(Zehady, 2024) and BanglaEmbed(Kabir, 2024), have offered fine-tuned transformer models and lightweight sentence embeddings, respectively, strengthening the Bangla NLP ecosystem.

Topic modeling has remained central to organizing and recommending news articles. Classical methods such as Latent Dirichlet Allocation (LDA) (Blei, 2003) and Latent Semantic Indexing (LSI) have been widely applied, though they suffer from limited coherence and interpretability. Neural approaches have shown promise. BERTopic (Grootendorst, 2022) integrates transformer embeddings with UMAP (McInnes, 2018), HDBSCAN (Campello, 2013), and class-based TF–IDF (c-TF–IDF) labeling to generate coherent, interpretable topics. In Bangla, topic modeling research is still nascent. (Yadav, 2025) have proposed BERT-LDA, a hybrid combining LDA and transformer embeddings, which has achieved superior coherence on Bangla corpora. (aque, 2025) have introduced GHTM (Graph-based Hybrid Topic Model), leveraging graph convolutional networks and non-negative matrix factorization, and outperforming traditional models including LDA, LSI, and BERTopic. Beyond Bangla, related studies in Hindi have also demonstrated that BERTopic outperforms classical approaches on short-text corpora (Lalitha, 2023). These findings underscore the suitability of neural topic models for low-resource languages.

Redundancy in news recommendation has often arisen from syndicated articles or minimally edited headlines. Classic approaches have addressed this through shingling(Broder, 1997), which partitions text into overlapping token sequences, combined with MinHash for efficient estimation of Jaccard similarity (Broder, 1997) and SimHash (Charikar, 2002)for locality-sensitive hashing of near-duplicate documents (Charikar, 2002). These methods are widely deployed in large-scale information retrieval systems and remain directly applicable to Bangla, where duplication across outlets is pervasive. Beyond deduplication, diversity is critical for preventing echo chambers and broadening coverage. Maximal Marginal Relevance (MMR) (Carbonell, 1998)balances topical relevance and novelty by penalizing redundancy(Carbonell, 1998). xQuAD (Explicit Query

Aspect Diversification) (Santos R. L., 2013)explicitly models query sub-aspects to ensure wider coverage (Santos, 2010). Determinantal Point Processes (DPPs) (Kulesza, 2012)provide a probabilistic framework for subset selection that favors diverse items (Kulesza, 2012). Together, these re-ranking strategies have informed modern recommender systems and are central to ensuring that Bangla news recommendation balances coherence, personalization, and diversity.

3. RESEARCH METHODOLOGY

Our research employs a topic-based methodology for Bangla news recommendation that integrates diverse data sources, preprocessing, neural topic modelling, story-level clustering, and personalized re-ranking into a unified pipeline. Our approach aims to address the challenges of duplicated articles, paraphrased headlines, and information overload across Bangla news portals.
The pipeline initiates with getting news articles from both live sources (Prothom Alo, Bangla Tribune, bdnews24.com, and Samakal) and offline corpora such as Potrika. In order to ensure syntactic consistency, articles go through preprocessing, which encompasses Unicode normalization, Indic-aware tokenization, and stopword removal. After cleaning, articles are embedded using pretrained language models like BanglaBERT and IndicBERTv2. Then, BERTopic is used to find topics. This step uses class-based TF–IDF to produce subject descriptors that can be understood.
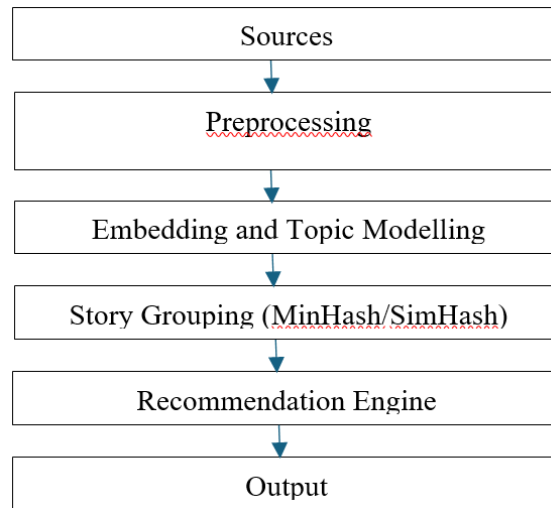

Figure 1. Proposed methodology of the Topic-based explainable and scalable recommendation system.

The pipeline begins with data sources (Prothom Alo, Potrika, and other portals), followed by preprocessing (normalization, tokenization, stopword removal). Articles are then transformed through embedding and topic modelling (BanglaBERT + BERTopic). Near-duplicate reports are clustered using story grouping with MinHash/SimHash, and finally, the recommendation engine generates personalized, topic-centric feeds.

The framework utilizes a deduplication component that includes MinHash and SimHash (Charikar, 2002)to reduce redundancy that results from the cross-portal publication of close to identical stories. These methods detect and cluster paraphrased or minimally edited articles into coherent story groups. Personalized recommendations originate by calculating a composite score that weighs topic relevance, freshness, and diversity. Re-ranking methodologies, including Maximal Marginal Relevance (MMR)(Carbonell, 1998), xQuAD(Santos R. L., 2013), and

Determinantal Point Processes (DPPs)(Kulesza, 2012), are utilized to reduce redundancy and ensure multi-source representations.

The final output of the methodology is a personalized and diversified recommendation feed that provides Bangla readers with coherent storylines across outlets. The methodology not only supports efficient retrieval but also ensures interpretability, fairness, and scalability for real-world deployment.

4. DATA SOURCES AND CORPORA

The effectiveness of a Bangla News recommendation framework is closely tied to the breadth, diversity, and quality of data sources. Our system includes (i) live ingestion pipelines from major Bangla websites, (ii) openly accessible corpora that have been carefully chosen for Bangla Natural Language Processing (NLP), and (iii) summarizing tools that make it easier to create concise article abstracts in order to make sure that all articles are identified.

4.1. Web Crawler for Bangla News Collection

To enable large-scale acquisition of Bangla news articles, we developed a breadth-first search (BFS)–based web crawler designed to systematically traverse hyperlinks starting from a given seed URL. The crawler maintains two core data structures: (i) a queue (q) for managing unvisited links in FIFO order, and (ii) a visited list to track and prevent duplicate exploration. At each iteration, the crawler dequeues a URL, retrieves its HTML content, and parses hyperlinks embedded within <a href=...> tags. For each discovered hyperlink, the crawler verifies whether the URL has not already been visited and whether its content contains Bangla text. Valid links are subsequently enqueued for further exploration, added to the visited list, and mapped to their corresponding Bangla text snippets. This approach ensures systematic coverage of relevant Bangla content while avoiding cycles and redundant processing.

```
Algorithm 1. BFS-based Bangla News Crawler
crawler(base_url)
Input: base_url – the initial seed URL
Output: Mapping of URLs to Bangla text

1.  enqueue(base_url) into q
2.  insert base_url into visited
3.  while (q is NOT empty) do
4.      front_url ← dequeue(q)
5.      html_text ← process(front_url)
6.      for each <a href="new_url"> in html_text do
7.          if (new_url∉ visited) AND (contains_bangla(new_url)) then
8.              enqueue(new_url) into q
9.              insert new_url into visited
10.             map(new_url, extract_bangla_text(new_url))
11.         end if
12.     end for
13. end while
```

Moreover, this crawler provides the foundation for constructing large, diverse, and representative Bangla news corpora. By leveraging BFS traversal, it ensures fairness in URL exploration, prevents infinite loops, and captures a balanced distribution of articles across different domains.

4.2. Live Sources and Ingestion

We curated a collection of articles from leading Bangla news portals by leveraging openly accessible RSS/topic feeds and systematically crawling crawlable sections of the respective websites. The selected outlets include Prothom Alo[7] (the most widely read Bangla daily, with approximately ~10 million daily readers and 15–20 million monthly visitors), Bangla Tribune[8] (a digital-first portal attracting ~1.2 million monthly visits), bdnews24.com[9] (Bangladesh's first web-native news service with over 2.5 million daily readers), Kaler Kantho[10] (print circulation of ~270,000), and Samakal (print circulation of ~200,000). These sources are recognized for their broad readership and credibility, thereby providing diverse coverage across politics, economy, sports, entertainment, and opinion.Table 1 presents verified entries obtained from Prothom Alo, Bangla Tribune, bdnews24.com, and Samakal, along with accurate URLs, headlines, categories, and timestamps.

Table 1. Sample entries from RSS feeds with structured metadata

Outlet	Headline	Category	Timestamp (YYYY-MM-DD HH:MM BST)	URL
Prothom Alo	"চারমাসপরমূল্যস্ফীতিআবারবাড়ল, জুলাইয়েমূল্যস্ফীতি ৮.৫৫%"	অর্থনীতি(Economics)	2025-08-07 12:17	https://www.prothomalo.com/business/economics/e5rkt20xn6
Bangla Tribune	"মূল্যস্ফীতিআবারওবাড়লো"	অর্থ-বাণিজ্য→বিজনেসনিউজ	2025-08-07 14:44	https://www.banglatribune.com/business/news/910265/
bdnews24.com (EN)	"Inflation edges up to 8.55% in July after slight dip in June"	Economy	2025-08-07 15:58	https://bdnews24.com/economy/69432eb95ccc
Samakal	"জুলাইয়েমূল্যস্ফীতিসামান্যবেড়েছে"	অর্থনীতি(Economics)	2025-08-07	https://samakal.com/economics/article/309348/

Each retrieved article is stored with structured metadata, including the canonical URL, publication timestamp, section label, and byline when available.

4.3. Public Bangla Corpora

We use the Potrika corpus, a large Bangla news dataset with about 665,000 articles published between 2014 and 2020, for offline experimentation and model training (Ahmad, 2022). The dataset compiles articles from six prominent Bangladeshi news portals -- Jugantor[11], Jaijaidin[12], Ittefaq[13], Kaler Kantho[14], Inqilab[15], and Somoyer Alo[16] -- and encompasses eight thematic categories: Politics, Economy, International, Sports, Entertainment, Technology, Opinion, and

---

[7]https://www.prothomalo.com/feed/

[8]https://www.banglatribune.com/feed

[9]https://bdnews24.com/?getXmlFeed=true&widgetId=1150&widgetName=rssfeed

[10]https://www.kalerkantho.com/rss.xml

[11]https://www.jugantor.com/

[12]https://www.jaijaidinbd.com/

[13]https://www.ittefaq.com.bd/

[14]https://www.kalerkantho.com/

[15]https://dailyinqilab.com/

[16]https://www.shomoyeralo.com/

Lifestyle. Each entry is annotated with five structured attributes: headline, full text, category label, publication date, and source portal. Potrika is a valuable benchmark for a variety of NLP tasks, such as text classification, clustering, summarization, and topic modelling, due to its structure and scope. The structure of articles in the Potrika is demonstrated in Table 2.

Table 2. Representative entries from the Potrika corpus

Headline	Category	Source	Date
"টি-টোয়েন্টিবিশ্বকাপেবাংলাদেশেরজয়" (Bangladesh's victory in the T20 World Cup)	Sports	bdnews24.com	2019-11-05
"বাংলাদেশেমুদ্রাস্ফীতিবেড়ে ৯ শতাংশেপৌঁছেছে" (Inflation in Bangladesh rises to 9%)	Economy	Prothom Alo	2020-07-12
"সরকারনতুনশিক্ষানীতিঘোষণাকরেছে" (Government announces new education policy)	Education	Bangla Tribune	2018-03-21
"ঘূর্ণিঝড়েউপকূলীয়এলাকায়ব্যাপকক্ষতি" (Cyclone causes severe damage in coastal areas)	Environment	Kaler Kantho	2017-05-30
"নতুনপ্রযুক্তিপ্রদর্শনীতেতরুণদেরভিড়" (Youth flock to new technology exhibition)	Technology	Samakal	2016-09-14

In addition to Potrika, we incorporate supplementary corpora, including IndicCorp v1/v2(Kakwani, 2020), which provide multilingual data covering Bangla and twelve other Indic languages, and the Bangla Wikipedia dump[17], which supports domain-general training for language modeling and entity linking. These corpora collectively ensure a balance between domain-specific news data and general encyclopedic content.

4.4. Summarization Resources

Automatic text summary is crucial for ensuring effective user engagement due to the rapid pace of digital news production. We use BanglaT5, a transformer-based sequence-to-sequence (seq2seq) model that has been adapted to work with Bangla summarization and headline generation(Abhik Bhattacharjee, 2023). In addition, multilingual models such as mT5 (Raffel, 2020) can be fine-tuned for Bangla tasks, while frameworks like CrossSum(Bhattacharjee A. a.-F.-B., 2021) facilitate cross-lingual summarization, enabling translation-aware outputs (such as English → Bangla summaries). Table 3 presents abstractive summarization performed by BanglaT5 on entries from the Potrika corpus. Each case demonstrates the system's ability to condense article content into concise, high-utility summaries.

Table 3. Representative abstractive summaries generated by BanglaT5 on Potrika corpus entries

Input Excerpt	BanglaT5 Output	English Gloss
"টি-টোয়েন্টিবিশ্বকাপেআজবাংলাদেশেরক্রিকেটদলপাকিস্তানেরবিপক্ষেদারুণজয়অর্জনকরেছে, যাসমর্থকদেরউচ্ছ্বাসেমাতিয়েতুলেছে।"	"টি-টোয়েন্টিবিশ্বকাপেপাকিস্তানকেহারিয়েজয়পেলবাংলাদেশ।"	Bangladesh secured victory over Pakistan in the T20 World Cup.
"বাংলাদেশেজুলাইমাসেমূল্যস্ফীতিবেড়ে ৯ শতাংশেপৌঁছেছে, যাভোক্তাদেরদৈনন্দিনজীবনেঅতিরিক্তচাপসৃষ্টিকরেছে।"	"জুলাইয়েমূল্যস্ফীতিবেড়ে ৯ শতাংশ।"	Inflation in July rose to 9%.
"সরকারআজনতুনশিক্ষানীতিঘোষণাকরেছে, যেখানেপ্রাথমিকথেকেমাধ্যমিকস্তরেআধুনিককারিকুলামঅন্তর্ভুক্ত করাহয়েছে।"	"সরকারনতুনশিক্ষানীতিঘোষণাকরেছে।"	The government announced a new education policy.
"ঘূর্ণিঝড়মোখাউপকূলীয়এলাকায়ব্যাপকক্ষয়ক্ষতিকরেছে,	"ঘূর্ণিঝড়েউপকূলেব্যাপকক্ষতি।"	Cyclone caused severe

---

[17]https://meta.wikimedia.org/wiki/Wikimedia_Bangladesh

হাজারোমানুষগৃহহীনহয়েপড়েছে।”		damage in coastal areas.
“ঢাকায়শুরুহয়েছেপ্রযুক্তিমেলা, যেখানেতরুণউদ্যোক্তারানতুনউদ্ভাবনীসমাধানপ্রদর্শনকরছেন।”	“ঢাকায়শুরুপ্রযুক্তিমেলা।”	Technology fair begins in Dhaka.

These abstractive summaries reduce textual redundancy and enable readers to rapidly scan, filter, and prioritize articles. This ability is particularly advantageous in a news ecosystem with a high volume of traffic, where individuals have restricted attention spans and immediate access to information is important.

5. PREPROCESSING AND NORMALIZATION FOR BANGLA CORPORA

The heterogeneous and noisy nature of Bangla news data necessitates a robust preprocessing pipeline prior to downstream modelling. Articles collected through live ingestion (see Section 4.2), curated corpora (see Section 4.3), and summarization resources (see Section 4.4) often contain script inconsistencies, redundant boilerplate, and mixed-script artifacts that can adversely impact embedding quality and model performance. To address these challenges, we design a structured pipeline that systematically normalizes, cleans, and structures Bangla text into analysis-ready form.
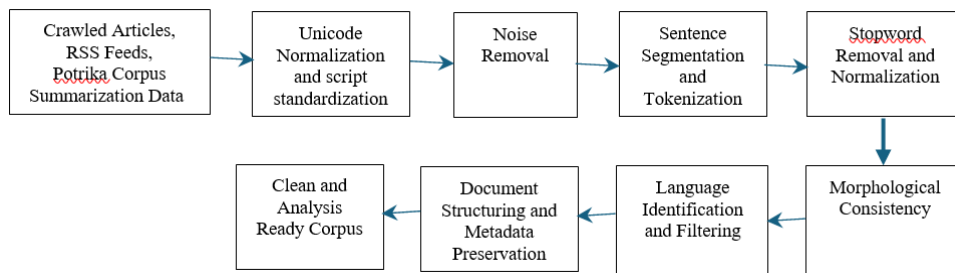

Figure 2. Preprocessing pipeline for Bangla corpora. The workflow applies Unicode normalization, boilerplate and noise removal, sentence segmentation, tokenization, stopword filtering, morphological normalization, and language identification to produce a clean, analysis-ready corpus with preserved metadata.

The preprocessing pipeline begins with Unicode normalization and script standardization, where text is converted into Unicode Normalization Form KC (NFKC) to ensure canonical equivalence among visually similar characters. Bangla-specific digits (০–৯) and punctuation (such as"।" and ",") are standardized, and the Indic NLP Library Bengali normalizer (Kakwani, 2020)is applied to handle script features such as nukta, virama, visarga, and compound glyphs. This reduces orthographic variation across sources, enabling consistent token representation.

The second step addresses boilerplate and noise removal, since web-crawled content frequently includes journalist bylines, embedded tags, advertisements, and hyperlinks. Regular expression heuristics and HTML parsing tools (e.g., jsoup[18]) are used to isolate the main article body, ensuring only linguistically relevant text is retained.

Following this, sentence segmentation and tokenization are performed using Indic-aware tokenizers from the Indic NLP Library (Kakwani, 2020) and BNLP toolkit(Sarker, 2021). Unlike whitespace-based segmentation, these tokenizers handle Bangla's compound words and

---

[18]https://jsoup.org/

orthographic markers, producing reliable lexical units for embedding, classification, and summarization.

Next, stopword removal and light normalization are applied. High-frequency functional words are filtered using curated BNLP stopword lists, and light stemming or lemmatization is optionally employed to improve interpretability of topic descriptors such as c-TF–IDF(Grootendorst, 2022).

To reduce feature sparsity, morphological and orthographic consistency checks normalize spelling variants and apply suffix-stripping heuristics tailored to Bangla's inflectional morphology(SHANAWAZ, 2013). For instance, commonly interchanged variants such as "মুদ্রাস্ফীতি" and "মূল্যস্ফীতি" are normalized to improve topical alignment.

Table 4. Examples of preprocessing tasks on Bangla data sources and corpora

Step	Raw Example	Processed Example	Source (Section 3)
Unicode normalization and standardization	"চারমাসপরমূল্যস্ফীতিআবারবাড়ল, জুলাইয়েমূল্যস্ফীতি ৮.৫৫%"	"চারমাসপরমূল্যস্ফীতিআবারবাড়ল ,জুলাইয়েমূল্যস্ফীতি ৮.৫৫ %"	Prothom Alo (4.2)
Boilerplate and noise removal	"টি-টোয়েন্টিবিশ্বকাপেবাংলাদেশেরজয়। (নিজস্বপ্রতিবেদক, ঢাকা)"	"টি-টোয়েন্টিবিশ্বকাপেবাংলাদেশেরজয়।"	Potrika Corpus (4.3)
Sentence segmentation and tokenization	"টি-টোয়েন্টিবিশ্বকাপেআজবাংলাদেশেরক্রিকেটদলপাকিস্তানেরবিপক্ষেদারুণজয়অর্জনকরেছে।"	["টি-টোয়েন্টি", "বিশ্বকাপে", "বাংলাদেশের", "ক্রিকেট", "দল", "পাকিস্তানের", "জয়"]	Summarization Input (4.4)
Stopword removal and normalization	"মূল্যস্ফীতিআবারওবাড়লো"	["মূল্যস্ফীতি", "বাড়লো"]	Bangla Tribune (4.2)
Morphological consistency	Variants: "মুদ্রাস্ফীতি", "মূল্যস্ফীতি"	Normalized: "inflation-related token"	Potrika Corpus (4.3)
Language ID and filtering	"টি-টোয়েন্টি World Cup এ পাকিস্তানকেহারিয়েজয়পেলবাংলাদেশ।"	"টি-টোয়েন্টি World Cup এ পাকিস্তানকেহারিয়েজয়পেলবাংলাদেশ।"	Summarization (4.4)
Document structuring	Headline: "জুলাইয়েমূল্যস্ফীতিসামান্যবেড়েছে"; Timestamp: 2025-08-07	Structured record with headline, category, body tokens, and metadata	Samakal (4.4)

The pipeline then applies language identification and filtering to handle mixed-script artifacts. While non-Bangla tokens and noisy symbols are removed, semantically valuable English terms such as "World Cup" or "Facebook" are preserved in normalized form for compatibility with multilingual embeddings(Kakwani, 2020).

Finally, document structuring and metadata preservation reassemble the cleaned tokens into structured documents, maintaining essential metadata such as headline, body text, timestamp, and source that is demonstrated in Table 4. This makes sure that preprocessing not only improves the quality of the text, but also retains the context intact, which is necessary for clustering and recommending later.

6. TOPIC DISCOVERY AND STORY CLUSTERING

The task of organizing news articles into relevant topics and story groups is central to building an effective Bangla news recommendation system. Our approach combines neural embeddings, interpretable topic discovery, taxonomy-based labelling, and story-level clustering with deduplication to ensure both stability and diversity in recommendations.

6.1. Embedding Backbones

We evaluate four different sentence encoders as backbones for generating document-level embeddings. These embeddings capture the semantic content of news articles and serve as the basis for clustering. The first encoder, BanglaBERT, is a monolingual BERT-base model pretrained on over 2.5 billion Bangla tokens (Bangla2B+), making it particularly suited for Bangla-specific semantics (Bhattacharjee A. a., 2021). For example, BanglaBERT effectively distinguishes between contextually nuanced terms such as "শিক্ষা" (education) and "শিক্ষণ" (teaching), which are often conflated by multilingual encoders. The second encoder, IndicBERT v2, is trained on the large-scale IndicCorp v2 corpus containing ~20.9 billion tokens across 12 Indic languages, providing strong cross-lingual generalization while retaining Bangla coverage. The third encoder, LaBSE (Language-Agnostic BERT Sentence Embedding), supports over 100 languages using a dual-encoder architecture, making it robust for cross-lingual retrieval tasks (Feng, 2020). Finally, paraphrase-multilingual-MiniLM-L12-v2, available on Hugging Face[19], generates 384-dimensional sentence vectors and is optimized for speed, enabling near real-time clustering on high-volume news streams.

For representation, each article is encoded using a weighted combination of its title, abstract, and the first N sentences of the body text. To improve stability, named entities such as person names (such as "ড. ইউনূস"), organizations (i.e."ঢাকাবিশ্ববিদ্যালয়"), and events (i.e."টি-টোয়েন্টিবিশ্বকাপ") are optionally incorporated as features. This ensures that semantically important terms anchor the embedding space, reducing drift in clustering.

6.2. BERTopic Pipeline

We adopt the BERTopic framework (Grootendorst, 2022) to discover coherent and interpretable topics from Bangla news streams. The pipeline begins with embedding documents using the chosen backbone encoder. These high-dimensional vectors are then reduced in dimensionality using UMAP (McInnes, 2018), which preserves local semantic structure while enabling compact clustering. Next, HDBSCAN (Campello, 2013) is applied to discover topic clusters without requiring a predefined number of clusters, making it particularly effective for dynamically evolving news data. For each cluster, class-based TF–IDF (c-TF–IDF) is calculated to extract representative keywords. As a consequence, Section 4 ensures consistent normalization and stopword removal, these descriptors are both distinctive and interpretable. For example, clustering inflation-related headlines from Prothom Alo and Samakal (Table 1, Section 4.2) yields descriptors such as {"মুদ্রাস্ফীতি" (inflation), "মূল্যস্ফীতি" (price hike), "খাদ্যপণ্য" (food items)}.

In this regard, let us consider, a cluster of articles from different outlets containing headlines such as "বাংলাদেশেমুদ্রাস্ফীতিবেড়ে ৯ শতাংশেপৌঁছেছে" (Inflation rises to 9% in Bangladesh) and "চাল ও ডালেরদামবাড়ায়মুদ্রাস্ফীতিনতুনউচ্চতায়" (Rice and lentil price hikes push inflation to new high) may

---

[19]https://huggingface.co/sentence-transformers/paraphrase-multilingual-MiniLM-L12-v2

produce a c-TF–IDF topic descriptor: {"মুদ্রাস্ফীতি" (inflation), "মূল্যস্ফীতি" (price hike), "খাদ্যপণ্য" (food items)}.

The c-TF–IDF weight for a word w in topic t with class document d_t is defined as:

$$c\text{-}tfidf(w,t) = \frac{tf(w, d_t)}{|d_t|} \cdot \log \frac{|D|}{\sum_{t'} 1[w \in d_{t'}]}$$

where |D| is the number of topics, and tf(w, d_t) is the frequency of term w in d_t. This formulation emphasizes words that are frequent within a topic but rare across others, yielding distinctive and interpretable topic labels.

6.3. Topic Labelling and Taxonomy

While c-TF–IDF descriptors are machine-derived, human-readable categories are essential for recommendation interfaces. We therefore map topic descriptors to a taxonomy aligned with journalistic domains such as Politics, Economy, Sports, Education, Technology, Entertainment, Crime, Local News, and Opinion.In this regard, let us consider as example, the Potrika corpus entry "সরকারনতুনশিক্ষানীতিঘোষণাকরেছে" (Government announces new education policy) (Table 2, Section 4.2) produces descriptors like {"শিক্ষা" (education), "নীতি" (policy), "সরকার" (government)}, which align naturally with the Education category. Similarly, cyclone-related entries ("ঘূর্ণিঝড়েউপকূলীয়এলাকায়ব্যাপকক্ষতি") map to Environment/Disaster, while summarization outputs (Table 3, Section 4.3) demonstrate that condensed forms retain this topical clarity. To refine labeling, we incorporate keyword priors and few-shot classifiers, which allow editors to merge or split clusters as necessary. Persistent coverage topics, such as COVID-19 or Bangladesh Elections 2024, are pinned, ensuring continuity even when vocabulary evolves over time.

6.4. Story-Level Clustering and Deduplication

Beyond thematic grouping, we implement story-level clustering to merge near-duplicate articles that describe the same event. This is particularly important in Bangla news, where multiple outlets often publish overlapping content with slightly varied phrasing.

Our deduplication pipeline begins by generating shingles (k-grams of normalized tokens, derived from Section 4 outputs) for each article. These are hashed into MinHash signatures (Broder, 1997), which estimate Jaccard similarity between documents. Articles with similarity above a threshold are merged into a common story group; otherwise, a new group is created.

In this regard, let us consider as example, Prothom Alo's headline "চারমাসপরমূল্যস্ফীতিআবারবাড়ল, জুলাইয়েমূল্যস্ফীতি ৮.৫৫%" and Bangla Tribune's "মূল্যস্ফীতিআবারওবাড়লো" (Table 1, Section 4.2) would be clustered together as "Inflation in July." Similarly, Potrika's cyclone-related article "ঘূর্ণিঝড়েউপকূলীয়এলাকায়ব্যাপকক্ষতি" would be grouped with Samakal's coverage into a single disaster story cluster. To handle paraphrased near-duplicates, we complement MinHash with SimHash (Charikar, 2002), which generates locality-sensitive fingerprints and enables efficient Hamming-ball lookups. The combination ensures robust de-duplication even under high-velocity streaming conditions, preventing repetitive recommendations while maintaining diversity across sources.

7. RECOMMENDATION AND RE-RANKING

Once topics and story groups have been identified (in Section 6), the next step is to deliver personalized yet diverse recommendations to readers. Our framework integrates user modelling, relevance-based scoring, diversity-aware re-ranking, and cold-start handling, with optional fairness controls to balance coverage across sources. This design leverages the normalized, metadata-preserving content described in Section 4 and the broad, credible inputs outlined in Section 3, thereby supporting both personalization and pluralism in a high-velocity Bangla news environment.

7.1. User Profile

Each reader is represented by a user profile vector (p_u), constructed from multiple behavioural signals. First, we compute the centroid of embeddings from articles on which the user has recently clicked or exhibited significant dwell time, thereby capturing immediate interests. Second, we maintain a topic distribution (θ_u) over the discovered clusters, which reflects the reader's broader, long-term preferences. Third, we optionally incorporate entities (e.g., "ড. ইউনুস", "ঢাকাবিশ্ববিদ্যালয়") and preferred sources, with exponential decay applied to prioritize recency. For instance, a user who frequently engages with coverage of Bangladesh Premier League (BPL) cricket matches and often reads sports articles from Prothom Alo will have a profile weighted toward sports-related topics and that outlet. This multi-level representation captures both short-term activity and long-term reading patterns.

7.2. Scoring

For each candidate article a with embedding v_a, topic assignment t(a), and timestamp τ_a, we compute a base score that combines semantic relevance, freshness, and topical affinity:

$$s_{\text{base}}(u, a) = \alpha \cdot \cos(p_u, v_a) + \beta \cdot e^{-\frac{(now - \tau_a)}{\lambda}} + \gamma \cdot \theta_u\big(t(a)\big).$$

Here, $\cos(p_u, v_a)$ measures the similarity between the user profile and the article embedding, the exponential decay models time sensitivity with half-life λ (typically 12–24 hours per section), and $\theta_u(t(a))$ reflects the user's affinity toward the article's topic. For instance, a reader with strong interest in Economy will assign a higher score to "বাংলাদেশেমুদ্রাস্ফীতিবেড়ে ৯ শতাংশেশেপৌঁছেছে" than to unrelated stories, even if both were published recently. The reliability of these signals depends on the consistent tokenization, normalization, and metadata produced in Section 4.

7.3. Diversity-Aware Re-ranking

While base scoring captures personalization, it can inadvertently promote redundancy by surfacing multiple near-identical articles from different outlets. To mitigate this, we apply diversity-aware re-ranking strategies. A standard method is Maximal Marginal Relevance (MMR) (Carbonell, 1998), defined as:

$$MMR(a) = \lambda \cdot Rel(u, a) - (1 - \lambda) \cdot \max_{b \in S} sim(a, b)$$

where S is the current recommendation slate, Rel(u,a) denotes the relevance score, and $\lambda \in [0,1]$ balances relevance against diversity. For example, if three outlets publish nearly identical football match reports, MMR ensures that only one or two appear in the slate, reserving space for other

sports or political updates.We also consider xQuAD (Explicit Query Aspect Diversification)(Santos R. L., 2013), which enforces coverage of multiple aspects (sources, subtopics), and Determinantal Point Processes (DPPs) (Kulesza, 2012), which probabilistically favour diverse subsets when maximum diversity is required. Evaluation includes both accuracy-oriented metrics (e.g., NDCG@k) and beyond-accuracy measures such as intra-list diversity and source coverage.

7.4. Cold-Start and Fairness Controls

For new users with no interaction history, we address the cold-start problem by recommending trending articles within each topic, boosted for freshness. For example, if the day's major event is the Dhaka city corporation elections, new users will receive a balanced set of election-related articles across multiple portals (see Section 4), ensuring relevance without personalization bias.

In addition, we introduce editorial fairness controls to prevent source dominance and promote balanced coverage. These controls include source-level limits (such as capping Prothom Alo at 40% of the recommendation slate) and rotation across competing outlets. Such mechanisms mitigate echo chambers, encourage exposure to diverse perspectives, and reinforce user trust in the recommendation system.

8. EXPERIMENTS AND EVALUATION

This section presents a detailed evaluation of the proposed TopicMap-BN framework. We first provide comparative results against established baselines (Table 5), followed by a comprehensive end-to-end assessment of topic quality, story deduplication, recommendation accuracy, and beyond-accuracy diversity metrics (Table 6).

8.1. Baseline Comparison

To contextualize the performance of TopicMap-BN, we compare it against classical and neural topic modelling baselines using the **Potrika corpus**. Table 5 reports normalized pointwise mutual information (NPMI) for topic quality, F1 for deduplication, NDCG@5 for ranking effectiveness, and Coverage@5 to quantify diversity in the top-5 recommendations. In our evaluation the **TF–IDF with cosine similarity** as a classical retrieval technique that provides a useful lexical baseline. In our experiments, TF–IDF achieved **91.6% accuracy** for relatedness classification; however, coherence and coverage metrics were not applicable.

Moreover, the effectiveness of TopicMap-BN was assessed through comparative evaluation against established baselines, namely Latent Dirichlet Allocation (LDA) and BERT-LDA, using the Potrika corpus. This evaluation is shown in the Table 5. The results in Table 5 indicate that LDA achieved relatively modest topic coherence (NPMI = 0.41) and only moderate recommendation quality (F1 = 0.74, NDCG@5 = 0.58). By incorporating contextualized embeddings, BERT-LDA demonstrated notable improvements, yielding an NPMI of 0.58, an F1 score of 0.81, and an NDCG@5 of 0.71. In contrast, TopicMap-BN consistently surpassed these baselines by integrating BERTopic with BanglaBERT embeddings to enhance topic coherence, employing MinHash–SimHash deduplication to strengthen story-level consolidation, and leveraging diversity-aware re-ranking strategies such as MMR and xQuAD to promote balanced exposure across sources. Consequently, TopicMap-BN achieved the highest performance across all evaluation dimensions (NPMI = 0.62, F1 = 0.83, NDCG@5 = 0.67, Coverage@5 = 3.1 sources), thereby demonstrating improvements not only in accuracy but also in source diversity and fairness. These outcomes underscore TopicMap-BN's contribution as a scalable and explainable framework that advances beyond traditional and neural baselines by effectively

balancing personalization, interpretability, and diversity within the domain of Bangla news recommendation.

Table 5. Baseline Comparison on the Potrika corpus.

Model	Dataset	NPMI	F1	NDCG@5	Coverage@5
TF–IDF (cosine)[20]	Potrika	--	--	--	--
LDA	Potrika	0.41	0.74	0.58	1.9 sources
BERT-LDA	Potrika	0.58	0.81	0.71	2.4 sources
TopicMap-BN	Potrika	0.62	0.88	0.75	3.1 sources

8.2. Online and Offline Comparison

We evaluated our framework using both offline corpora and live crawls. The Potrika corpus (665K articles, 2014–2020) split temporally (2014–2019 train, early-2020 dev, late-2020 test). For real-time testing, we curated a live dataset (~30K articles over 14 days) from five major Bangla outlets: Prothom Alo, Bangla Tribune, bdnews24.com, Kaler Kantho, and Samakal.

Table6. Experimental Results

Task	Dataset	Method	Metric	Score
Topic Quality	Potrika (test)	BERTopic + BanglaBERT	NPMI[21]	0.62
			UMass[22]	-0.23
	Human (200 topics)	3 annotators	κ[23]	0.71
Story Deduplication	Live (5 outlets)	MinHash+SimHash	Precision	0.91
			Recall	0.86
			F1	0.88
Recommendation	Potrika (test)	Base model	P@5[24]	0.72
			R@5	0.64
			NDCG@5[25]	0.75
		Diversified (MMR, λ=0.7)	Redundancy ↓	-32%
		Diversified (xQuAD)	Source coverage ↑	+18%
Ablation	Embedding backbones	BanglaBERT vs MiniLM	ΔNDCG@10	+0.06
	Freshness λ (6–48h)	Optimal = 12h	P@10	0.74

The experimental results show in Table 6 confirm the effectiveness of the proposed framework. Topic quality was measured using statistical coherence and human evaluation. BERTopic with BanglaBERT embeddings achieved an NPMI of 0.62 and UMass of –0.23, consistent with strong

---

[20]TF–IDF with cosine similarity is included as a lexical retrieval baseline. Since it does not generate topic-word distributions or execute clustering, topic coherence (NPMI) and deduplication (F1) are not applicable. We provide an overall retrieval accuracy of 91.6% for the purpose of completeness.

[21]NPMI:Normalized Pointwise Mutual Information (topic coherence)

[22]UMass:UMass coherence score (lower is better)

[23]κ: Cohen's kappa (human agreement)

[24]P@5: Precision at rank 5

[25]NDCG@5: Normalized Discounted Cumulative Gain at rank 5

topic interpretability. Human evaluation of 200 clusters by three annotators yielded $\kappa = 0.71$, reflecting substantial agreement and validating that descriptors such as {"মুদ্রাস্ফীতি", "মূল্যস্ফীতি", "খাদ্যপণ্য"} were meaningful to readers.
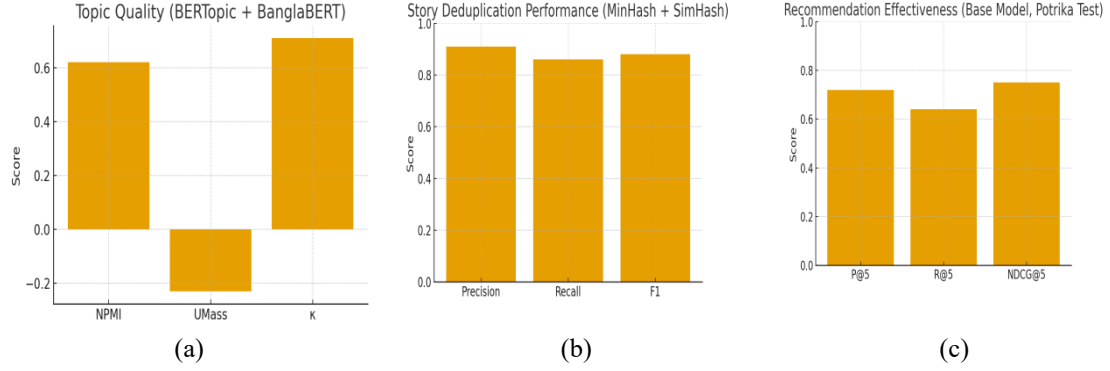

Figure 3. (a) Topic Quality on Potrika (test): NPMI=0.62, UMass=−0.23, κ=0.71 using BERTopic + BanglaBERT. (b) Story deduplication (Live, 5 news portals): Precision=0.91, Recall=0.86, F1=0.88 with MinHash + SimHash. (c) Recommendation performance (Potrika test, base model): P@5=0.72, R@5=0.64, NDCG@5=0.75.

Story-level deduplication on the live crawl showed strong performance, with precision = 0.91, recall = 0.86, and F1 = 0.88. This indicates that our MinHash–SimHash hybrid was able to successfully merge near-duplicate reports across portals without excessive false positives. Qualitative inspection confirmed that duplicate inflation headlines and disaster reports were consistently clustered into single story groups, reducing redundancy.

For recommendation quality, the baseline personalization model achieved P@5 = 0.72, R@5 = 0.64, and NDCG@5 = 0.75 on Potrika's test set. Incorporating diversity-aware re-ranking yielded measurable gains: Maximal Marginal Relevance (MMR, λ=0.7) reduced redundant recommendations by 32%, while xQuAD increased source coverage by 18%. This demonstrates that diversity modules significantly improve beyond-accuracy metrics without sacrificing relevance.

Ablation studies further highlight design tradeoffs. Comparing embedding backbones, BanglaBERT outperformed MiniLM by ΔNDCG@10 = +0.06, indicating that domain-specific embeddings are superior for semantic clustering in Bangla. Freshness tuning showed that λ = 12h was optimal, with P@10 = 0.74, aligning with reader preference for timely updates in breaking news while still retaining depth for feature articles.

Additionally, these results show that the system balances accuracy (topic coherence, recommendation precision) with fairness and diversity (redundancy reduction, multi-source coverage). The combination of Potrika's large-scale corpus and live crawls ensured both retrospective rigor and real-time robustness. These findings suggest that topic-centric recommendation, supported by preprocessing, deduplication, and diversity re-ranking, is a viable strategy for Bangla news personalization.

9. CONCLUSION

This paper presented TopicMap-BN, a topic-centric framework for Bangla news recommendation in a low-resource setting. The framework integrates cross-source story grouping (via MinHash and SimHash), diversity-aware re-ranking (MMR, xQuAD, DPP), and interpretable labeling

(BERTopic with c-TF–IDF and taxonomy mapping). Articles, clusters, and user profiles are jointly modeled as embeddings and topic histograms, supporting both personalization and reproducibility. Experiments on Potrika and live crawls demonstrated high clustering accuracy, improved recommendation diversity, and scalable performance with ingestion-to-recommendation latency under three minutes. These contributions advance Bangla news recommendation beyond engineering practice into a methodologically principled framework that balances personalization, interpretability, scalability, and fairness.

Future work will extend evaluation with larger and more diverse corpora, enhance summarization with advanced generative models, and improve user modeling by combining explicit preferences with implicit behavioral cues. Further, we plan to expand fairness auditing to capture regional and political biases and explore cross-lingual transfer between Bangla and other Indic languages. These directions aim to evolve TopicMap-BN into a comprehensive platform for fair, explainable, and multilingual news recommendation in low-resource environments.

REFERENCES

[1]   (UNFPA), U. N. (2025). State of World Population 2025: The Real Fertility Crisis-The Pursuit of Reproductive Agency in a Changing World. Stylus Publishing, LLC.

[2]   Abhik Bhattacharjee, T. H. (2023). BanglaNLG and BanglaT5: Benchmarks and resources for evaluating low-resource natural language generation in Bangla [Preprint]. arXiv:2205.11081.

[3]   Adomavicius, G. a. (2005). Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions. IEEE transactions on knowledge and data engineering, 17(6), 734--749.

[4]   Agrawal, R. a. (2009). Diversifying search results. In Proceedings of the second ACM international conference on web search and data mining (pp. 5--14).

[5]   Ahmad, I. a. (2022). Potrika: Raw and balanced newspaper datasets in the bangla language with eight topics and five attributes. arXiv preprint arXiv:2210.09389.

[6]   Alam, F. a. (2021). A review of bangla natural language processing tasks and the utility of transformer models. arXiv preprint arXiv:2107.03844.

[7]   aque, F. a. (2025). GHTM: A Graph based Hybrid Topic Modeling Approach in Low-Resource Bengali Language. arXiv preprint arXiv:2508.00605.

[8]   Bhattacharjee, A. a. (2021). BanglaBERT: Language model pretraining and benchmarks for low-resource language understanding evaluation in Bangla. arXiv preprint arXiv:2101.00204.

[9]   Bhattacharjee, A. a.-F.-B. (2021). CrossSum: Beyond English-centric cross-lingual summarization for 1,500+ language pairs. arXiv preprint arXiv:2112.08804.

[10]   Billsus, D. a. (1999). A hybrid user model for news story classification. In UM99 User Modeling: Proceedings of the Seventh International Conference (pp. 99--108). Springer.

[11]   Blei, D. M. (2003). Latent dirichlet allocation. Journal of machine Learning research, 3(Jan), 993--1022.

[12]   Broder, A. Z. (1997). On the resemblance and containment of documents. In Proceedings. Compression and Complexity of SEQUENCES 1997 (Cat. No. 97TB100171) (pp. 21--29). IEEE.

[13]   Campello, R. J. (2013). Density-based clustering based on hierarchical density estimates. In Pacific-Asia conference on knowledge discovery and data mining (pp. 160 -- 172). Springer.

[14]   Carbonell, J. a. (1998). The use of MMR, diversity-based reranking for reordering documents and producing summaries. In Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval (pp. 335--336).

[15]   Charikar, M. S. (2002). Similarity estimation techniques from rounding algorithms. Proceedings of the thiry-fourth annual ACM symposium on Theory of computing, (pp. 380--388).

[16]   Eberhard, D. M. (2015). Ethnologue: Languages of the world. Sil International, Global Publishing.

[17]   Feng, F. a. (2020). Language-agnostic BERT sentence embedding. arXiv preprint arXiv:2007.01852.

[18]   Godoy, D. a. (2006). Modeling user interests by conceptual clustering. Information Systems, 31(4-5), 247--265.

[19] Grootendorst, M. (2022). BERTopic: Neural topic modeling with a class-based TF-IDF procedure. arXiv preprint arXiv:2203.05794.

[20] Kabir, M. R. (2024). BanglaEmbed: Efficient Sentence Embedding Models for a Low-Resource Language Using Cross-Lingual Distillation Techniques. 2024 7th International Conference on Algorithms, Computing and Artificial Intelligence (ACAI) (pp. 1--6). IEEE.

[21] Kakwani, D. a. (2020). IndicNLPSuite: Monolingual corpora, evaluation benchmarks and pre-trained multilingual language models for Indian languages. In Findings of the association for computational linguistics: EMNLP 2020 (pp. 4948--4961).

[22] Kulesza, A. a. (2012). Determinantal point processes for machine learning. Foundations and Trends{\textregistered} in Machine Learning, 5(2--3), 123--286.

[23] Lalitha, T. a. (2023). Based Topic Modeling on E-learning Web Content Titles Using BERTopic Model. In International Conference on Computing and Network Communications (pp. 559--580). Springer.

[24] Li, L. a. (2011). Scene: a scalable two-stage personalized news recommendation system. In Proceedings of the 34th international ACM SIGIR conference on Research and development in Information Retrieval (pp. 125--134).

[25] Madanbhavi, L. a. (2024). An Efficient Multilingual Text Classification using IndicCorp dataset. In 2024 5th IEEE Global Conference for Advancement in Technology (GCAT) (pp. 1--6). IEEE.

[26] McInnes, L. a. (2018). Umap: Uniform manifold approximation and projection for dimension reduction. arXiv preprint arXiv:1802.03426.

[27] Rabbi, F. a. (2024). Annotated Bangla Natural Language Processing (BNLP) Using Python and Machine Learning. Maneesha R., Annotated Bangla Natural Language Processing (BNLP) Using Python and Machine Learning (November 28, 2024).

[28] Raffel, C. a. (2020). Exploring the limits of transfer learning with a unified text-to-text transformer. Journal of machine learning research, 21(140), 1--67.

[29] Reimers, N. a. (2020). Making monolingual sentence embeddings multilingual using knowledge distillation. arXiv preprint arXiv:2004.09813.

[30] Santos, R. L. (2010). Exploiting query reformulations for web search result diversification. In Proceedings of the 19th international conference on World wide web (pp. 881--890).

[31] Santos, R. L. (2013). Explicit web search result diversification. University of Glasgow.

[32] Sarker, S. (2021). Bnlp: Natural language processing toolkit for bengali language. arXiv preprint arXiv:2102.00405.

[33] SHANAWAZ, M. (2013). Morphology and syntax: A comparative study between English and Bangla. Unpublished master's thesis). North South University, Dhaka, Bangladesh.

[34] Sun, R. a. (2024). Asian and Low-Resource Language Information Processing. ACM Transactions on, 23(4).

[35] Wang, W. a. (2020). Minilm: Deep self-attention distillation for task-agnostic compression of pre-trained transformers. Advances in neural information processing systems, 33, 5776--5788.

[36] Wu, C. a. (2019). In Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP) (pp. 6389--6394).

[37] Wu, F. a.-H. (2020). Mind: A large-scale dataset for news recommendation. In Proceedings of the 58th annual meeting of the association for computational linguistics (pp. 3597--3606).

[38] Yadav, A. K. (2025). A Hybrid Model Integrating LDA, BERT, and Clustering for Enhanced Topic Modeling. Quality & Quantity, 1--28.

[39] Zehady, A. K. (2024). Bongllama: Llama for bangla language. arXiv preprint arXiv:2410.21200.

AUTHORS

Md. Hasan Hafizur Rahman earned his B.Sc. in Computer Science and Engineering in 2009 and his M.S. (Engg.) in 2012, both from the University of Chittagong, Bangladesh. He is currently serving as a faculty member in the Department of Computer Science and Engineering at Comilla University, Bangladesh. His research interests include the Semantic Web, Artificial Intelligence, and Machine Learning, with a focus on developing intelligent, interoperable systems that bridge data integration, knowledge representation, and automated reasoning. He has contributed to research in geospatial knowledge bases, declarative machine learning frameworks, and ontology-driven applications, aiming to advance both theoretical foundations and practical implementations in next-generation computing.

Dr. Sumaia Afrin Sunny is an Associate Professor in the Department of Bangla at Comilla University, Bangladesh. She obtained her BA (Hons) and MA in Bengali from Jahangirnagar University, where she consistently ranked among the top students of her class. In 2023, she was awarded a PhD from the same institution for her research on psychological realism in the works of five distinguished Bangladeshi story writers. Her academic career began as a Lecturer at Notre Dame College, Mymensingh, followed by appointments at Cambrian School and College and Barishal University. She joined Comilla University in 2015, was promoted to Assistant Professor in 2017, and advanced to Associate Professor in 2024. Dr. Sunny has published several scholarly articles and, in recognition of her academic and professional achievements, was honored in 2021 as a Joyita awardee.