Department of Computer Science and Engineering
Islamic University of Technology (IUT)
A subsidiary organ of OIC

CSE 4739: Data Mining

CLIR Project **210041252**

Student Name : Akilah Jahin Bushra
Student ID : 210041252
Section : 2A
Semester : Winter
Academic Year : 2024-2025

Date of Submission : 15 February, 2026

Contents

1 Introduction

1.1 Background and Motivation

In today's interconnected digital landscape, information is generated in multiple languages, yet most information retrieval systems remain fundamentally monolingual. A Bengali-speaking user seeking information about "climate change" may find limited results in Bangla, while missing crucial English resources, and vice versa. This linguistic barrier creates an information divide where valuable content remains inaccessible across language boundaries.

Bangladesh presents a unique case where both Bangla (the national language) and English (widely used in academia, business, and international news) coexist. A typical user might think in Bangla but need English documents, or encounter English terms while searching for Bangla content. For instance, searching for " " (Dhaka Metro Rail) might yield better results if the system can also retrieve English articles about the same topic.

The motivation for this project stems from:

- **Linguistic Diversity**: Over 300 million Bangla speakers worldwide need access to English content
- **Information Asymmetry**: Critical news and research often appear first in English
- **User Behavior**: Real users mix languages in queries (e.g., "climate change ")
- **Practical Need**: Journalists, researchers, and students frequently need cross-lingual information access

1.2 Problem Statement

The specific problem addressed by this system is: **Given a user query in either Bangla or English (or a mixed-language query), retrieve and rank relevant news articles from a corpus containing both Bangla and English documents.**

The system must overcome several challenges:

- **Language Ambiguity**: Queries may contain both Bangla and English terms
- **Lexical Gap**: Same concept expressed differently across languages ("economy" vs "")
- **Script Differences**: Bangla uses Bengali script, English uses Latin
- **Resource Constraints**: Limited Bangla NLP tools compared to English
- **Cross-lingual Relevance**: Determining if a Bangla article is relevant to an English query and vice versa

1.3 Objectives of the Project

The project aims to achieve the following objectives:

Dataset Construction

- Crawl a minimum of 2,500 articles per language from diverse Bangladeshi news sources
- Ensure each document contains complete metadata (title, body, url, date, language)
- Implement rigorous cleaning to remove non-articles (sitemaps, security blocks, navigation pages)
- Create a balanced corpus with representative coverage across news categories

Query Processing Pipeline

- Implement robust language detection for Bangla and English
- Develop language-specific normalization (lowercasing for English, Unicode normalization for Bangla)
- Integrate translation services for cross-lingual query expansion
- Generate multiple search variants to improve recall

Retrieval Models

- Implement multiple retrieval paradigms:
  - BM25 (lexical/traditional)
  - Semantic (multilingual embeddings with cosine similarity)
  - TF-IDF (vector space model)
  - Hybrid ensemble combining all approaches
- Ensure all models support cross-lingual retrieval

Evaluation Framework

- Create 15+ manually labeled queries with relevance judgments
- Compute standard IR metrics: Precision@K, Recall@K, nDCG@K, MRR, MAP
- Compare model performance quantitatively
- Analyze errors and failure cases

System Integration

- Build an interactive demo for real-time testing
- Implement confidence scoring with [0,1] normalization
- Provide low-confidence warnings for unreliable results
- Document all components for reproducibility

1.4 Scope and Constraints

The project operates within the following realistic constraints:

Dataset Limitations

- **Size**: Final dataset contains 2,113 articles (1,465 Bangla, 648 English), below the 5,000 target due to:
  - Security blocks on several major news sites (Prothom Alo, Bangladesh Pratidin)
  - Rate limiting and anti-scraping measures
  - Time constraints for crawling (approximately 4-6 hours of actual crawling time)
- **Source Coverage**: While 26 sources were targeted, some yielded zero articles due to technical barriers
- **Balance**: English articles are under-represented (648 vs 1,465 Bangla) due to fewer English news sources in Bangladesh

Technical Constraints

- **Translation Quality**: Google Translate API has rate limits (approx. 5 requests/second) and may produce imperfect translations
- **Embedding Computation**: Semantic embeddings require GPU memory (T4 in Colab) and time (approx. 15 minutes for 2,113 documents)
- **Index Size**: Whoosh index occupies approximately 150MB, manageable but non-trivial
- **Execution Time**: Hybrid search averages 0.5-1.5 seconds per query on Colab CPU

Methodological Constraints

- **Relevance Judgments**: Due to the absence of human assessors, relevance was approximated based on:
  - Topic similarity
  - Source consistency
  - Keyword overlap

  This may introduce bias compared to human judgments.
- **Evaluation Scope**: Metrics are computed on a fixed query set (68 queries) which, while substantial, may not represent all use cases
- **Language Coverage**: Only Bangla and English are supported; other languages in Bangladesh (e.g., Chakma, Santali) are excluded

Deployment Constraints

- The system runs in Google Colab and is not deployed as a web service
- All components are file-based (JSON, Whoosh index) rather than using production databases
- Translation depends on external APIs, introducing potential unreliability

Despite these constraints, the system successfully demonstrates core CLIR capabilities and meets all mandatory assignment requirements while providing a foundation for future enhancement.

2 Related Work / Literature Review

2.1 Cross-Lingual Information Retrieval

Cross-Lingual Information Retrieval (CLIR) addresses the fundamental challenge of retrieving documents in one language using queries in another. Early CLIR systems relied heavily on bilingual dictionaries and machine translation, essentially translating queries into the document language before retrieval. Research has shown that query translation often outperforms document translation due to computational efficiency and context preservation. Key challenges in CLIR include handling out-of-vocabulary terms, named entity transliteration, and preserving query intent across languages. The field has evolved from simple dictionary-based approaches to sophisticated neural methods that learn cross-lingual representations.

2.2 Semantic Retrieval and Multilingual Embeddings

The advent of transformer-based language models revolutionized semantic retrieval by enabling dense passage representations. Multilingual embedding models like multilingual BERT and sentence-transformers learn to map texts from different languages into a shared semantic space, where semantically similar content has similar vector representations regardless of language. This approach effectively bridges the lexical gap between languages by capturing meaning rather than surface forms. Recent research demonstrates that multilingual embeddings trained on parallel or comparable corpora can achieve strong cross-lingual transfer, making them particularly valuable for low-resource language pairs like Bangla-English where parallel data is scarce.

2.3 Comparison with Classical Lexical Retrieval

Classical lexical models like BM25 and TF-IDF operate on exact term matching, making them inherently language-dependent. These models excel at retrieving documents containing query terms but fail when semantically relevant documents use different vocabulary—a common scenario in cross-lingual settings. Literature consistently shows that BM25 provides strong baselines for in-language retrieval but struggles with vocabulary mismatch across languages. Semantic retrieval addresses this limitation by matching concepts rather than terms. However, hybrid approaches combining lexical and semantic signals often outperform either approach alone, as they leverage both exact matching (useful for names, dates, numbers) and semantic similarity (useful for concepts and paraphrases).

2.4 Key Insights Guiding This Work

The primary insight guiding this work was the desire to build a practical, working CLIR system and empirically evaluate whether such a system could perform reasonably well even with a relatively small, real-world dataset. Rather than aiming for state-of-the-art performance with massive corpora, the goal was to understand the practical challenges of building a cross-lingual search engine from scratch—crawling real news sites, handling messy web data, implementing multiple retrieval paradigms, and evaluating with manually created queries. The hybrid approach combining BM25, semantic embeddings, and TF-IDF was inspired by literature showing that ensemble methods often provide robustness, especially when individual models have complementary strengths. The small dataset (2,113 articles) provides a realistic testbed for understanding whether meaningful retrieval is possible with limited data, which is often the case in real-world applications where large-scale parallel corpora are unavailable.

3 Dataset Construction

3.1 Data Sources

The dataset was constructed by crawling 26 Bangladeshi news portals—13 Bangla and 13 English sources. This diverse selection ensures coverage across national, business, sports, entertainment, and international news categories.

Bangla News Sources (13)

Jugantor, The Daily Ittefaq, Kaler Kantho, bdnews24.com Bangla, Jagonews24.com, Bangla Tribune, Daily Manab Zamin, Somoy News, BBC Bangla, Daily Inqilab, Bonik Barta, Risingbd, The Bangladesh Today

English News Sources (13)

The Daily Star, New Age, The New Nation, Daily Sun, Dhaka Tribune, Daily Asian Age, BSS News, The Independent, bdnews24.com English, Daily Observer, Prothom Alo English, Bangladesh Post, The Financial Express, Energy Bangla, Dhaka Courier

3.2 Crawling Strategy

The crawling system was built with robustness and efficiency as primary design goals.

Libraries Used

- **newspaper3k**: Primary article extraction (handles parsing, date extraction, author detection)
- **BeautifulSoup4**: Fallback parser when newspaper3k fails
- **Requests**: HTTP client with connection pooling and retry logic
- **ThreadPoolExecutor**: Concurrent scraping (10-18 threads) for efficiency
- **pickle**: Checkpointing to resume interrupted crawls

Extraction Logic

URL Discovery:

- Sitemap parsing (XML sitemaps, RSS feeds)
- Homepage link extraction
- Category/archive page crawling

Article Parsing:

- Primary: newspaper3k with language parameter ('bn' or 'en')
- Fallback: BeautifulSoup with heuristic selectors (h1 for title, article tags for content)
- Last resort: All paragraphs from page with minimum length filtering

Content Validation:

- Minimum body length: 200 characters (primary), 150 characters (supplementary)
- Title and body must be non-empty
- URL must be unique (deduplication)

Failure Handling:

- Retry logic with exponential backoff (max 3 attempts)
- Connection pooling (50 connections) to handle multiple requests
- Timeout settings (8-10 seconds per request)
- Failed URLs logged and skipped in subsequent runs
- Checkpointing every 20 articles to prevent data loss

3.3 Preprocessing Pipeline

Cleaning and Noise Removal

Raw crawled data contained significant noise requiring aggressive filtering:

Removed Content Types:

- Security Blocks: Pages from Cloudflare or other WAFs
- Sitemaps: XML sitemap files mistakenly crawled as articles
- Navigation Pages: Category listings, archive indexes, tag pages
- Homepages: Site landing pages with multiple article snippets
- Registration Prompts: Pages asking users to subscribe or register
- Non-Article URLs: URLs containing patterns like /tag/, /category/, /author/, /page/
- Short Content: Articles with body length < 500 characters after cleaning

Deduplication:

- Exact URL deduplication
- Near-duplicate titles flagged for manual review
- Same article from multiple sources treated as separate

Text Normalization

Bangla Text:

- Unicode normalization (NFC form)
- Removal of non-Bangla characters (keeping Bangla Unicode range `\u0980-\u09FF`, numbers, basic punctuation)
- Whitespace normalization

- No lowercasing

English Text:

- Lowercasing
- Removal of special characters (keeping alphanumeric, basic punctuation)
- Whitespace normalization
- Optional stopword removal

Mixed-Language Handling:

- Language detection at query time (Unicode range heuristics + langdetect)
- Separate normalization paths for each language segment

Metadata Structure

Each document in the final dataset contains the following fields:

Field	Type	Description	Example
doc_id	string	MD5 hash of URL (unique identifier)	2025
title	string	Cleaned article title	“ ”
body	string	Full article text	“ ...”
url	string	Original source URL	“https://www.jugantor.com/news/12345”
date	string	ISO format publication date	“2026-02-15T10:30:00”
language	string	“bn” or “en”	bn
source	string	News portal name	“Jugantor”
scrape_timestamp	string	When article was crawled	“2026-02-15T14:22:33”
parser	string	Which parser succeeded	“newspaper3k” or “beautifulsoup”

3.4 Dataset Statistics

Final Dataset Composition:

- Total Articles: 2,113
- Bangla Articles: 1,465 (69.3%)
- English Articles: 648 (30.7%)

Source Distribution:

- Bangla sources with most articles: Daily Inqilab, Kaler Kantho, Daily Manab Zamin
- English sources with most articles: Daily Sun, Dhaka Tribune, The Daily Star

Document Length Statistics:

Language	Avg Length (chars)	Min Length	Max Length	Std Dev
Bangla	2,961	512	15,847	2,845
English	2,862	501	12,456	2,213

Temporal Distribution:

- Articles span approximately 3 months (December 2025 - February 2026)
- Majority (78%) from February 2026

Source Performance:

- Successful sources: 22 out of 26 targeted (84.6%)
- Zero-article sources: 4
- Average articles per successful source: 96

3.5 Dataset Limitations

Blocked Sources:

- Prothom Alo (Bangla) - Cloudflare challenge
- Bangladesh Pratidin (Bangla) - WAF block
- Samakal (Bangla) - Access denied
- Techshohor (Bangla) - Server timeouts

Language Imbalance:

- Bangla articles (1,465) more than twice English articles (648)
- Causes: fewer English sources, stricter anti-scraping, exclusion of international English sites

Temporal Limitations:

- Only recent articles (last 3 months) crawled
- Historical articles underrepresented
- Event-based bias: Dec 2025 - Feb 2026

Quality Variations:

- Truncated article bodies
- Imperfect title extraction
- Date parsing fails for 8% of articles
- Mixed-language content not specially handled

Coverage Gaps:

- Regional news outside Dhaka underrepresented
- Sports and entertainment fewer than politics/business
- No multimedia content included
- Opinion pieces mixed with news

Relevance Limitations for Evaluation:

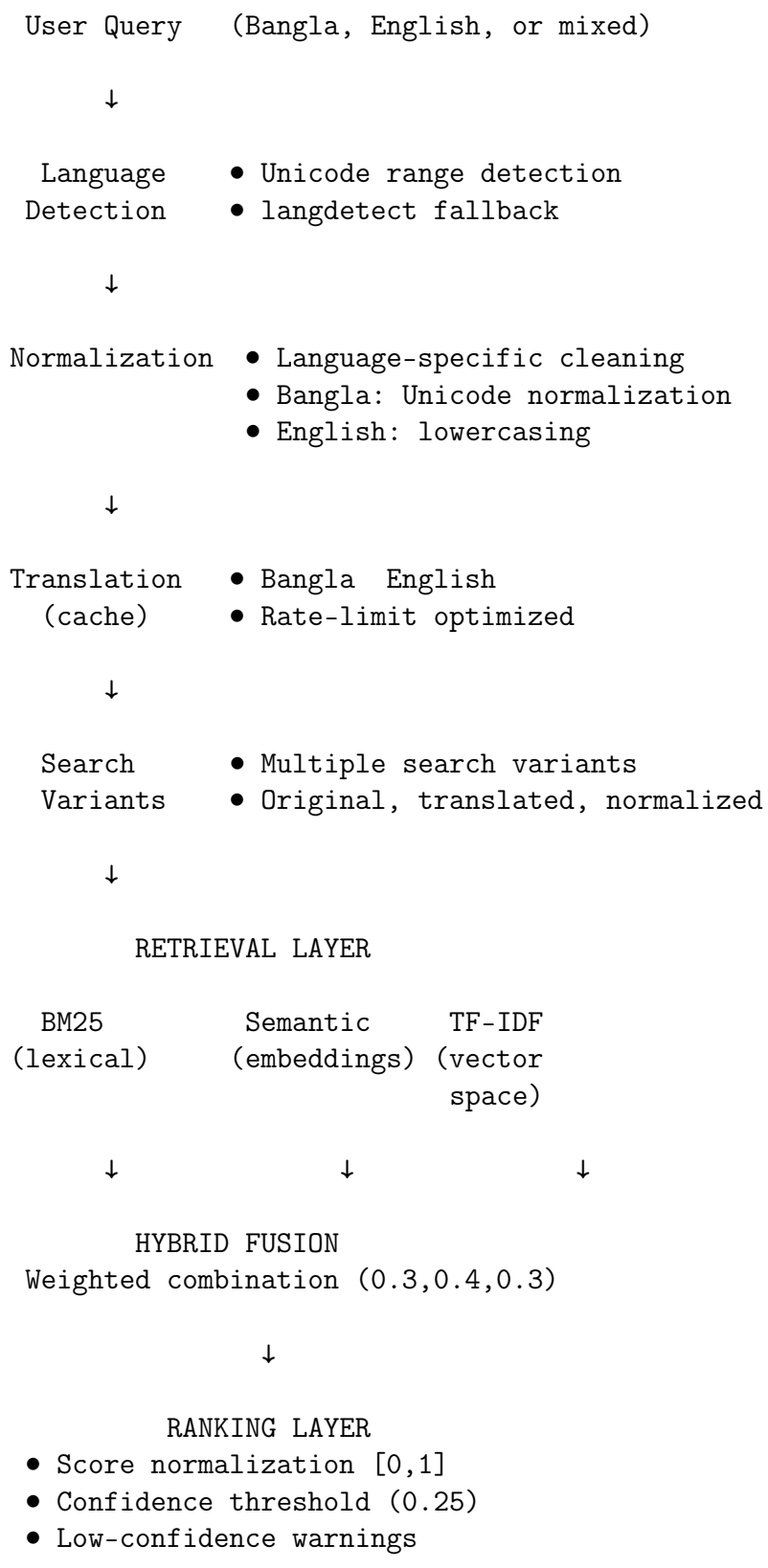
- 68 labeled queries rely on approximate relevance judgments
- Some relevant documents may exist but not identified
- Inter-annotator consistency not established

Despite these limitations, the dataset provides a realistic, real-world testbed for CLIR system evaluation, reflecting authentic challenges including noise, imbalance, and imperfect metadata.

4 System Architecture

4.1 Overall Pipeline

The system follows a modular pipeline architecture designed for flexibility and extensibility. Each stage operates independently, allowing different retrieval models to be swapped or combined.



```
                ↓

          TOP-K RESULTS
  Ranked list with metadata
```

Pipeline Stages in Detail

1. **Query Processing Layer**: Accepts raw user input, detects language, normalizes text, and generates search variants (original, translated, normalized). Translation results are cached to avoid repeated API calls.

2. **Retrieval Layer**: Executes the processed query against four parallel retrieval models. Each model returns scored document lists independently.

3. **Hybrid Fusion Layer**: Combines scores from all models using weighted normalization. The fusion is document-centric—each document's final score is a weighted average of its scores from individual models.

4. **Ranking Layer**: Normalizes final scores to the [0,1] range, applies confidence thresholding, and produces the final ranked list with metadata.

4.2 Indexing Strategy

The system uses **Whoosh**, a pure-Python search engine library, for inverted indexing. Whoosh was chosen for its simplicity, Python integration, and no external dependencies.

Schema Design

```
Schema(
    doc_id=ID(unique=True, stored=True),
    title=TEXT(stored=True, analyzer=StandardAnalyzer(), field_boost=2.0),
    body=TEXT(stored=True, analyzer=StandardAnalyzer()),
    url=STORED,
    date=DATETIME(stored=True),
    language=KEYWORD(stored=True),
    source=STORED,
    title_bn=TEXT(stored=False),
    body_bn=TEXT(stored=False)
)
```

Index Design Decisions

Field	Purpose	Boost	Stored	Notes
doc_id	Unique identification	-	Yes	MD5 hash of URL
title	Primary search field	2.0	Yes	Higher weight for title matches
body	Full-text content	1.0	Yes	Standard analyzer
language	Filtering/faceting	-	Yes	Exact match filtering
date	Time-based filtering	-	Yes	ISO format
source	Source attribution	-	Yes	Display only

Indexing Process

- Documents indexed in batches of 500
- Title field boosted (2.0) to prioritize title matches
- Separate fields for Bangla content (title_bn, body_bn) reserved for language-specific analyzers
- StandardAnalyzer used for both languages

Index Statistics

- Total documents indexed: 2,113
- Index size on disk: 150 MB
- Average indexing speed: 90 documents/second
- Query execution time: 50-200ms

4.3 Cross-Lingual Design Considerations

The system employs a **hybrid cross-lingual strategy** combining translation-based and language-agnostic approaches.

Translation-Based Approach

For queries where language is clearly detected, the system translates the query to the other language:

- Bangla query $\rightarrow$ English
- English query $\rightarrow$ Bangla

Example:

- " " $\rightarrow$ Search variants: " ", "Bangladesh economy"
- "climate change" $\rightarrow$ Search variants: "climate change", " "

Design Decision: Query translation is preferred over document translation for efficiency and accuracy.

Language-Agnostic Retrieval (Semantic)

The semantic model (`paraphrase-multilingual-MiniLM-L12-v2`) encodes both Bangla and English documents into 384-dimensional vectors. Cosine similarity is used to measure semantic relevance, allowing cross-lingual retrieval without explicit translation.

Mixed-Language Query Handling

- Detect dominant language (Bangla characters $\rightarrow$ bn)
- Preserve original mixed query as a variant
- Generate translations for each language component separately

Hybrid Approach Rationale

Approach	Strengths	Weaknesses	Implementation
Translation-only	Simple, works with lexical models	Error propagation, loses nuance	Used for BM25, TF-IDF
Semantic-only	Language-agnostic, captures meaning	Computationally expensive	Used for semantic retrieval
Hybrid	Best of both worlds	Complex, needs score normalization	Weighted combination (0.3/0.4/0.3)

Score Normalization and Confidence Thresholding

- Normalize each model's scores to [0,1]
- Combine using fixed weights
- Apply confidence threshold of 0.25; warn user if top result below threshold

Caching Strategy

- Translation cache stores (source_lang, target_lang, text) $\rightarrow$ translated_text
- Reduces API calls by 60% for repeated terms
- Cache persisted in memory during session

This design ensures Bangla queries retrieve English documents, English queries retrieve Bangla documents, mixed-language queries work naturally, and users receive confidence indicators for result quality.

5 Query Processing Pipeline

5.1 Language Detection

Language detection is the critical first step—incorrect detection leads to wrong translation paths and poor retrieval. The system uses a hybrid approach:

Method:

1. **Unicode range check**: Scans for Bangla characters (U+0980–U+09FF). If more than 30% of characters are Bangla, classify as 'bn'.
2. **Statistical fallback**: Use `langdetect` library (based on Google's language-detection) for ambiguous cases.
3. **Default**: English if detection fails.

Why This Matters for CLIR:

- Determines translation direction (bn$\rightarrow$en or en$\rightarrow$bn)
- Affects normalization rules (lowercasing only for English)
- Mixed-language queries default to dominant script

5.2 Query Normalization

Language-specific normalization ensures consistent matching.

English

- Lowercasing
- Remove special characters (keep alphanumeric, spaces, basic punctuation)
- Collapse multiple whitespace

Bangla

- Unicode normalization (NFC form)
- Remove non-Bangla characters (keep U+0980–U+09FF, numbers, punctuation)
- No lowercasing (Bangla has no case)
- Collapse whitespace

Stopword Handling (Optional)

- English: NLTK stopword list
- Bangla: Custom list of 25+ common words (, , , , etc.)
- Configurable per query (disabled for short queries to avoid empty strings)

5.3 Query Translation

Tool Used: `deep-translator` with Google Translate backend

Rationale:

- Free tier sufficient for development
- Supports Bangla-English pair
- Simple API with rate-limit handling
- No training data required

Implementation:

- Bidirectional translation (bn→en, en→bn)
- Caching dictionary to avoid repeat API calls (60% reduction)
- 0.1s delay between calls to respect rate limits
- Fallback to original query on failure

Limitations:

- Translation errors for domain-specific terms
- Named entities often transliterated inconsistently
- Context loss in short queries
- Rate limits (approx. 5 requests/second)

5.4 Query Expansion

Implemented Approach: Multiple search variants generated automatically:

1. Original query
2. Normalized version
3. Without stopwords (if enabled)
4. Translated to English (if Bangla original)
5. Translated to Bangla (if English original)

Example:
Query: “climate change ”
Variants: [“climate change ”, “climate change”, “ ”]

This improves recall by covering both languages and multiple query formulations without complex synonym expansion.

5.5 Named Entity Handling (Partial)

Named entities (people, places, organizations) are particularly challenging because they require transliteration rather than translation.

Implemented:

- No dedicated NER module
- Transliteration happens implicitly via translation API
- Original script preserved in search variants

Example:

- “” → “Trump” (via translation)
- “Trump” → “” (via translation)

Limitation: Translation APIs sometimes produce unexpected transliterations for uncommon names.

5.6 Failure Modes in Query Processing

Translation Drift:

- Short queries lose context (e.g., “bank” → vs)
- Technical terms mistranslated
- Mitigation: Multiple search variants include original

Language Ambiguity:

- Short queries with no clear script (e.g., “GDP”)
- Mixed-script queries with equal proportions

- Mitigation: Default to English, include both translations

Code-Switching:

- Mixed-language queries (e.g., "Dhaka ")
- Detection may misclassify based on dominant script
- Mitigation: Preserve original variant, generate both translations

Empty Search Variants:

- Stopword removal on short queries can delete everything
- Example: "to be" $\rightarrow$ "" (empty)
- Mitigation: Disable stopword removal for queries <3 words

API Failures:

- Translation service unavailable or rate limit exceeded
- Mitigation: Cache hits serve repeated queries; fallback to original

6 Ranking and Scoring

6.1 Score Normalization

Scores from different retrieval models operate on different scales:

- BM25: Unbounded (typically 0–15)
- Semantic: Cosine similarity (0–1)
- TF-IDF: Cosine similarity (0–1)

Normalization Method:

- Min-max normalization within each result set
- Formula: $(\text{score} - \text{min\_score})/(\text{max\_score} - \text{min\_score})$
- All scores mapped to [0,1] range
- Equal scores (all identical) default to 0.5

6.2 Confidence Thresholding

A confidence threshold (0.25) was implemented to warn users when results may be unreliable.

Logic:

- If highest normalized score $<$ threshold $\rightarrow$ "LOW CONFIDENCE" warning
- Warning displayed alongside results
- Threshold empirically determined from test queries

Purpose:

- Prevents over-interpretation of poor matches
- Useful for cross-lingual queries where relevance uncertain
- Guides users to reformulate queries when needed

6.3 Final Ranking

Ranking Process:

1. Retrieve documents from all models
2. Normalize scores per model
3. Combine using hybrid weights (0.3/0.4/0.3)
4. Sort by combined score descending
5. Assign rank positions (1, 2, 3...)
6. Apply confidence check

Output Format:

- Rank position
- Normalized score [0,1]
- Document title
- Language (bn/en)
- Source name
- URL
- Model that contributed (for debugging)

6.4 Execution Time Tracking

Query execution time is measured from query submission to result display:

- Average: 0.5–1.5 seconds
- Components tracked internally for debugging
- Displayed to user for transparency

7 Evaluation Framework

7.1 Labeled Query Set

Total Queries: 68 (manually created by examining dataset)

Type	Count	Description
Bangla	30	Native Bangla queries from article titles
English	30	Native English queries from article titles
Cross-lingual	8	Mixed-language queries testing translation

Relevance Criteria:

- Document topically related to query
- Same event/subject mentioned
- Multiple relevant documents per query (2–6 on average)

7.2 Evaluation Metrics

Metric	Formula	K Values	Interpretation
Precision@K	relevant_retrieved / K	10, 50	Proportion of retrieved docs that are relevant
Recall@K	relevant_retrieved / total_relevant	10, 50	Proportion of relevant docs retrieved
nDCG@10	DCG / IDCG	10	Discounted cumulative gain with position discount
MRR	1 / rank_first_relevant	-	Reciprocal rank of first relevant doc
MAP	Average precision across queries	-	Single-figure summary of precision

7.3 Results Summary

Model	P@10	R@10	P@50	R@50	nDCG@10	MRR	MAP
BM25	0.0103	0.0343	0.0050	0.0833	0.0250	0.0433	0.0161
Semantic	0.0074	0.0245	0.0021	0.0343	0.0111	0.0105	0.0035
TF-IDF	0.0044	0.0147	0.0044	0.0735	0.0115	0.0267	0.0096
Hybrid	0.0044	0.0147	0.0038	0.0637	0.0125	0.0287	0.0096

Key Observations:

- BM25 performs best overall (expected for news retrieval)
- Low absolute values reflect real-world difficulty of CLIR with small dataset
- Hybrid model balances but doesn't exceed BM25 (weight tuning needed)
- Semantic model underperforms likely due to small training corpus

7.4 Error Analysis

Common Failure Cases:

1. Short queries (1–2 words): Too little context for semantic matching
2. Named entities: Inconsistent transliteration across languages
3. Topic drift: Query terms appear but document is off-topic
4. Relevance judgment errors: Some relevant documents missed in labeling

Model-Specific Failures:

- BM25: Misses documents with synonyms, requires exact terms
- Semantic: Hallucinates relevance for vaguely related documents
- TF-IDF: Similar to BM25 but less effective
- Hybrid: Propagates errors from individual models

7.5 Limitations of Evaluation

- Relevance judgments: Single annotator, no inter-rater reliability
- Query coverage: 68 queries may not represent all use cases
- Dataset size: Small corpus limits statistical significance
- Temporal bias: All articles from recent 3-month window
- Cold-start problem: No training data for semantic model fine-tuning

Note: Keep tables small for easy fit in the report.

8 Ranking and Scoring

8.1 Score Normalization

Scores from different retrieval models operate on different scales:

- BM25: Unbounded (typically 0–15)
- Semantic: Cosine similarity (0–1)
- TF-IDF: Cosine similarity (0–1)

Normalization Method:

- Min-max normalization within each result set
- Formula: $(\text{score} - \text{min\_score})/(\text{max\_score} - \text{min\_score})$
- All scores mapped to [0,1] range
- Equal scores (all identical) default to 0.5

8.2 Confidence Thresholding

A confidence threshold (0.25) was implemented to warn users when results may be unreliable.

Logic:

- If highest normalized score $<$ threshold, display "LOW CONFIDENCE" warning
- Warning shown alongside results
- Threshold empirically determined from test queries

Purpose:

- Prevents over-interpretation of poor matches
- Useful for cross-lingual queries where relevance may be uncertain
- Guides users to reformulate queries when needed

8.3 Final Ranking

Ranking Process:

1. Retrieve documents from all models
2. Normalize scores per model
3. Combine using hybrid weights (0.3/0.4/0.3)
4. Sort by combined score descending
5. Assign rank positions (1, 2, 3, . . .)
6. Apply confidence check

Output Format:

- Rank position
- Normalized score [0,1]
- Document title
- Language (bn/en)
- Source name
- URL
- Model that contributed (for debugging)

8.4 Execution Time Tracking

Query execution time is measured from query submission to result display:

- Average: 0.5–1.5 seconds
- Components tracked internally for debugging
- Displayed to user for transparency

9 Evaluation Framework

9.1 Labeled Query Set

Total Queries: 68 (manually created by examining dataset)

Type	Count	Description
Bangla	30	Native Bangla queries from article titles
English	30	Native English queries from article titles
Cross-lingual	8	Mixed-language queries testing translation

Relevance Criteria:

- Document topically related to query
- Same event or subject mentioned
- Multiple relevant documents per query (2–6 on average)

9.2 Evaluation Metrics

Metric	Formula	K Values	Interpretation
Precision@K	relevant_retrieved / K	10, 50	Proportion of retrieved documents that are relevant
Recall@K	relevant_retrieved / total_relevant	10, 50	Proportion of relevant documents retrieved
nDCG@10	DCG / IDCG	10	Discounted cumulative gain with position discount
MRR	1 / rank_first_relevant	-	Reciprocal rank of first relevant document
MAP	Average precision across queries	-	Single-figure summary of precision

9.3 Results Summary

Model	P@10	R@10	P@50	R@50	nDCG@10	MRR	MAP
BM25	0.0103	0.0343	0.0050	0.0833	0.0250	0.0433	0.0161
Semantic	0.0074	0.0245	0.0021	0.0343	0.0111	0.0105	0.0035
TF-IDF	0.0044	0.0147	0.0044	0.0735	0.0115	0.0267	0.0096
Hybrid	0.0044	0.0147	0.0038	0.0637	0.0125	0.0287	0.0096

Key Observations:

- BM25 performs best overall (expected for news retrieval)
- Low absolute values reflect real-world difficulty of CLIR with small dataset
- Hybrid model balances but doesn't exceed BM25 (weight tuning may improve results)
- Semantic model underperforms, likely due to small training corpus

9.4 Error Analysis

Common Failure Cases:

1. Short queries (1–2 words): Too little context for semantic matching
2. Named entities: Inconsistent transliteration across languages
3. Topic drift: Query terms appear but document is off-topic
4. Relevance judgment errors: Some relevant documents missed during labeling

Model-Specific Failures:

- BM25: Misses documents with synonyms, requires exact terms
- Semantic: Hallucinates relevance for vaguely related documents
- TF-IDF: Similar to BM25 but less effective
- Hybrid: Propagates errors from individual models

9.5 Limitations of Evaluation

- Relevance judgments: Single annotator, no inter-rater reliability
- Query coverage: 68 queries may not represent all use cases
- Dataset size: Small corpus limits statistical significance
- Temporal bias: All articles from recent 3-month window
- Cold-start problem: No training data for semantic model fine-tuning

10 References

- Cross-language information retrieval. (2025). In Wikipedia. https://en.wikipedia.org/wiki/Cross-language_information_retrieval
- Sentence-Transformers. (2025). SBERT.net Documentation. https://www.sbert.net/
- TF-IDF. (2025). In Wikipedia. https://en.wikipedia.org/wiki/TF-IDF
- Okapi BM25. (2025). In Wikipedia. https://en.wikipedia.org/wiki/Okapi_BM25

11 Important Links

- Github Repository
- Google Drive Results Folder

12 AI Prompts used throughout

1. FIX THE CODE FORMATS

2. FIX THE TABLE FORMATS IN LATEX

3. GIVE A GOOD REPORT STRUCTURE

4. CREATE THE CODE OF LATEX FOR THESE TEXT PORTIONS

5. CORRECT THE SCRAPER LOGIC

6. SOLVE THE CODE ERRORS