Comparative Performance Analysis of BM25 and Vector Space Model for Document Retrieval in Gujarati News Corpora

Shreya M. Kapadia
Department of Information and Communication Technology, VNSGU, Surat

Payal D. Joshi
Department of Information and Communication Technology, VNSGU, Surat

ABSTRACT
Retrieving relevant information from Gujarati news articles is a challenging task because of the limited availability of computational resources and language-processing tools for Gujarati, despite the rapid growth of digital news content. In this study, an information retrieval–based framework for Gujarati news document retrieval is proposed using the GSF-2009 corpus released under the FIRE evaluation initiative. Two classical retrieval models, BM25 and the Vector Space Model (VSM), are employed to retrieve and rank documents relevant to user-defined event-oriented queries. Experimental evaluation is performed using both short and descriptive Gujarati queries. For the short query "ગુજરાતમાં ભારે વરસાદ", VSM demonstrates better performance with Recall = 0.7 and F1-score = 0.8, whereas BM25 records Recall = 0.3 and F1-score = 0.5. In contrast, for the descriptive query "ગુજરાતમાં ભારે વરસાદના કારણે અનેક જિલ્લાઓમાં પૂર જેવી સ્થિતિ", BM25 outperforms VSM with Precision = 1.0, Recall = 0.7, and F1-score = 0.8, whereas VSM achieves Precision = 0.8, Recall = 0.5, and F1-score = 0.6. The results indicate that VSM performs more effectively for short keyword-based queries, while BM25 achieves better retrieval effectiveness for long and context-rich queries.

Explicit event detection is not performed in this study; however, event-oriented retrieval is effectively supported through retrieval of documents associated with real-world events. The proposed framework provides an effective baseline for Gujarati news retrieval and supports further research in event-oriented retrieval for low-resource Indic languages.

Keywords
Gujarati News Retrieval, Information Retrieval, BM25, Vector Space Model (VSM), GSF-2009 Corpus, Event-Oriented Retrieval.

1. INTRODUCTION
The rapid growth of digital media platforms and online news sources has generated an enormous volume of textual data. As a result, retrieving relevant and meaningful information from large-scale news repositories has become a significant challenge for modern information access systems. News articles continuously report real-world events such as elections, floods, political developments, economic changes, accidents, and social incidents, making efficient document retrieval essential for extracting event-specific information from extensive news corpora. The exponential growth of digital news content has transformed information retrieval into a critical research problem for large-scale text analytics. Therefore, identifying and retrieving contextually relevant documents has become increasingly important for news analysis, information access, and decision-making applications.

Information Retrieval (IR) techniques play an important role in identifying and ranking documents that satisfy a user's information need from large text collections. According to Salton et al. [1], the primary objective of information retrieval is to retrieve documents relevant to a user's information need. Classical retrieval models such as BM25 and the Vector Space Model (VSM) are widely used for ranking documents based on query relevance and textual similarity. These models remain effective for large-scale document retrieval tasks because of their simplicity, computational efficiency, and strong retrieval performance.

Despite significant progress in information retrieval research for high-resource languages such as English, comparatively limited work has been carried out for Gujarati language processing. Gujarati is spoken by more than 55 million people worldwide; however, low-resource languages continue to face significant challenges because of the limited availability of annotated corpora, lexical resources, and computational tools. These limitations create challenges in developing efficient Gujarati news retrieval systems and highlight the need for further research in Gujarati Natural Language Processing (NLP) and information retrieval.In this context, event-oriented news retrieval serves as an effective benchmark for evaluating retrieval effectiveness on Gujarati news corpora, although explicit event detection is not the primary objective of this study. In news corpora, events generally correspond to real-world incidents such as natural disasters, political announcements, protests, or social developments. Event-related queries require retrieval models to identify documents that are not only topically relevant but also contextually associated with a specific event or situation. Therefore, event-oriented retrieval provides an effective experimental setting for evaluating the performance of retrieval models on Gujarati news corpora.

The Gujarat Samachar FIRE 2009 (GSF-2009) corpus released under the Forum for Information Retrieval Evaluation (FIRE) initiative is utilized in this study. The dataset provides a suitable corpus for evaluating Gujarati information retrieval in a low-resource language setting. To evaluate retrieval effectiveness, the same event-oriented queries are processed using BM25 and the Vector Space Model (VSM), which are classical information retrieval models used for document retrieval and ranking. BM25 follows a probabilistic retrieval approach that incorporates term frequency, inverse document frequency, and document length normalization for ranking relevant documents. In contrast, VSM represents documents and queries as TF–IDF vectors and computes similarity using cosine similarity for document ranking. A comparative analysis of

these models helps evaluate their effectiveness in retrieving Gujarati news documents under different query conditions.

The main contributions of this study are summarized as follows:

- Construction of an inverted index for efficient Gujarati news retrieval.
- Comparative implementation and evaluation of BM25 and the Vector Space Model (VSM) for Gujarati document retrieval.
- Experimental analysis using short and descriptive Gujarati queries for evaluating retrieval effectiveness under different query conditions.

2. RELATED WORK

Classical information retrieval models continue to play a significant role in event-oriented news retrieval. The Vector Space Model (VSM), proposed by Salton et al. [1], represents documents and queries using TF–IDF weighting and cosine similarity, and remains a strong baseline for document retrieval tasks. Probabilistic ranking functions such as BM25 have consistently demonstrated improved performance over TF–IDF–based models for ad-hoc retrieval tasks, particularly for news corpora with varying document lengths.

Information retrieval from news articles has been widely studied in the context of topic detection and tracking. Early work by Allan et al. [2] and Yang et al. [3] laid the foundation for event-oriented retrieval by modelling events as emerging topics in news streams using similarity-based methods. These studies established benchmark approaches for both online and retrospective event analysis.

For Indian languages, the Forum for Information Retrieval Evaluation (FIRE) has played an important role in advancing multilingual information retrieval research. Basaka [4] explored similarity-based approaches for Indian language news retrieval as part of the FIRE EDNIL shared task, demonstrating the applicability of IR-based techniques in low-resource language settings. Singh et al. [5] proposed a semantico-syntactic framework for event mention detection in Hindi, highlighting the importance of linguistic features for Indian languages.

Recent efforts have focused on developing large-scale Indic news datasets and retrieval benchmarks. Mirashi et al. [6] introduced the L3Cube-IndicNews dataset to support news analysis across multiple Indian languages, while Haq et al. [7] proposed IndicIRSuite, providing standardized datasets and retrieval models for Indic-language IR evaluation. In parallel, survey studies on neural event extraction have documented the transition from traditional statistical models to deep learning and transformer-based approaches [8], [9]. Although neural methods achieve strong performance, their dependence on large annotated datasets makes classical information retrieval approaches suitable for low-resource languages such as Gujarati.

2.1 Research Gap Analysis

Most existing information retrieval research focuses on English and other widely used languages, while limited work has been carried out for Gujarati news retrieval. In particular, comparative analysis of BM25 and the Vector Space Model (VSM) for Gujarati event-oriented news retrieval has not been widely explored.

This work attempts to address this research gap by applying classical information retrieval models to Gujarati news retrieval using the GSF-2009 dataset.

3. METHODOLOGY

Event-oriented retrieval of Gujarati news articles is performed using classical Information Retrieval (IR) techniques within the proposed framework. In this framework, user queries are treated as representations of real-world events such as floods, political developments, social incidents, and economic activities. Relevant Gujarati news documents are retrieved from the news corpus using classical information retrieval models for document retrieval and ranking models.

The overall workflow of the proposed framework consists of dataset collection, text preprocessing, inverted index construction, and document retrieval using BM25 and the Vector Space Model (VSM). The retrieved documents are ranked according to their relevance to the user query and are comparatively analyzed under different query conditions.

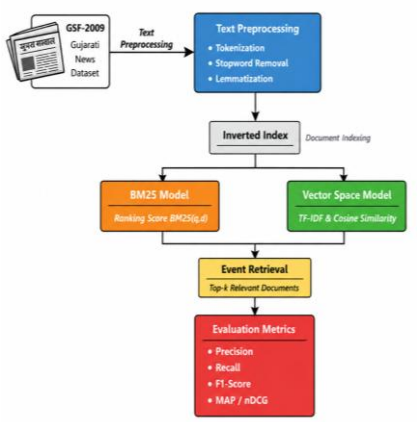

Fig 1: Overall architecture of the proposed IR-based framework for Gujarati news retrieval using preprocessing, inverted indexing, BM25, and the Vector Space Model (VSM)

Figure 1 illustrates the workflow of the proposed Gujarati news retrieval framework. The process begins with the GSF-2009 Gujarati news dataset, which contains unstructured news articles collected from multiple domains. The news documents undergo several preprocessing operations to normalize the Gujarati text and improve document consistency.

After preprocessing, an inverted index is constructed to support efficient document retrieval. The indexed documents are then processed using BM25 and the Vector Space Model (VSM), which retrieve and rank documents according to their relevance to the user query. Finally, retrieval effectiveness is comparatively analyzed using short and descriptive Gujarati event-oriented queries.

The following subsections describe each stage of the proposed methodology in detail.

3.1 Dataset Description

The Gujarat Samachar FIRE 2009 (GSF-2009) corpus released under the Forum for Information Retrieval Evaluation (FIRE) initiative is utilized in this study. The dataset provides a suitable corpus for evaluating Gujarati information retrieval in a low-resource language setting. The corpus contains Gujarati news articles collected from multiple domains such as politics, natural disasters, elections, social incidents, and economic developments.

The dataset consists of unstructured Gujarati textual documents and is widely used for Gujarati information retrieval research. The diversity of news topics within the corpus enables effective evaluation of retrieval models under realistic event-oriented query conditions.

Table 1 presents the overview of the GSF-2009 dataset used in this study.

Table 1. Overview of the GSF-2009 dataset used for evaluating Gujarati news retrieval in a low-resource language setting

Attribute	Description
Dataset Name	Gujarat Samachar FIRE 2009 (GSF-2009)
Language	Gujarati
Number of Documents	Approximately 32,000 – 35,000 news articles
Data Type	Unstructured News Articles
Domains Covered	Politics, Elections, Natural Disasters, Social Incidents, Economic Developments
Source	FIRE (Forum for Information Retrieval Evaluation)
Nature of the Dataset	Unstructured textual data
Dataset Category	Low-resource language dataset
Purpose of Usage	Evaluation of event-oriented Gujarati news retrieval using information retrieval models
Supported Retrieval Models	BM25 and Vector Space Model (VSM)
Research Application	Event-oriented Gujarati document retrieval

3.2 Indexing

Information Retrieval (IR) systems generally consist of two major pipelines: indexing and document retrieval. Indexing is an important stage in the proposed retrieval framework because efficient document retrieval depends on properly processed and organized textual data. In this work, the Gujarati news corpus is first preprocessed to remove noisy and irrelevant textual components. The cleaned and normalized text is then converted into an inverted index structure for efficient term-based searching.

The following subsections describe the preprocessing and indexing stages in detail.

3.2.1 Text Preprocessing

Text preprocessing is performed to improve the quality of document representation and retrieval effectiveness. Gujarati news articles contain punctuation symbols, linguistic variations, stopwords, and redundant textual components that may negatively affect retrieval performance. Therefore, preprocessing operations are applied before indexing.

The preprocessing pipeline includes text cleaning, tokenization, stopword removal, and lemmatization. These operations help generate normalized textual representations for efficient document retrieval.

Table 2 presents the preprocessing operations applied in the proposed framework.

Table 2: Presents the preprocessing operations applied to Gujarati news documents in the proposed framework.

Steps	Description
Text Cleaning	Removal of punctuation marks, numerals, special characters, and unnecessary symbols from the Gujarati text.
Tokenization	Splitting Gujarati text into individual word-level tokens.
Stopword Removal	Elimination of frequently occurring Gujarati stopwords using a predefined stopword list.
Lemmatization	Conversion of words into their root or base forms to reduce morphological variations.

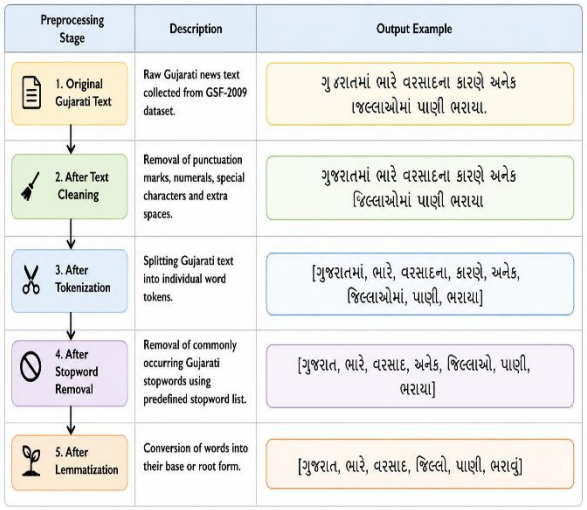

Fig 2: Example of Gujarati news text before preprocessing and the transformed output after each preprocessing stage

The preprocessing operations reduce textual noise and improve term consistency within the corpus. The generated normalized tokens are subsequently used for inverted index construction and document retrieval.

3.2.2 Inverted Index Construction

An inverted index is constructed from the preprocessed Gujarati news corpus to support efficient document retrieval. The inverted index maps each unique term to the list of documents in which the term appears, along with its corresponding term frequency. This indexing mechanism significantly reduces search complexity and enables fast retrieval of relevant documents.

The inverted index structure is formally represented as:

term→{(docID,tf)}

where:

- docID represents the unique document identifier
- tf represents the frequency of the term in the corresponding document

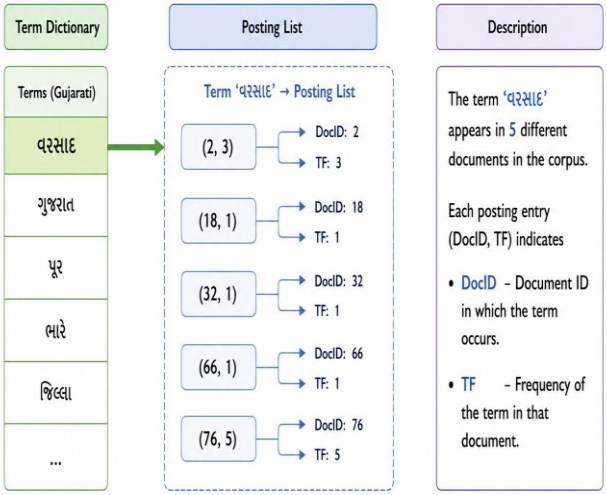

Fig 3: Vector representation of the inverted index showing the posting list generated for the Gujarati term "વરસાદ" with corresponding document IDs and term frequencies

Figure 3 illustrates a sample posting list generated from the inverted index for the Gujarati term "વરસાદ". Each posting entry stores the document identifier and the number of occurrences of the term within that document.

The inverted index enables direct access to documents containing the query terms without scanning the entire corpus. This indexing structure forms the foundation for both BM25 and Vector Space Model (VSM) retrieval operations implemented in the proposed retrieval framework.

3.3 Document Retrieval Models

Document retrieval is performed after the indexing stage using two classical Information Retrieval (IR) models, namely BM25 and the Vector Space Model (VSM). These models are used to retrieve and rank Gujarati news documents according to their relevance to user-defined event-oriented queries.

In the retrieval stage, the user query is processed and matched against the indexed Gujarati news corpus. Both retrieval models compute relevance scores between the query and documents, after which the documents are ranked according to their relevance scores. The top-ranked documents are then retrieved as the final output.

BM25 retrieval model follows a probabilistic retrieval approach that incorporates term frequency, inverse document frequency, and document length normalization for ranking relevant documents. In contrast, VSM represents documents and queries as TF–IDF vectors and computes cosine similarity for document ranking.

The retrieval effectiveness of BM25 and VSM is comparatively analyzed in the Results and Analysis section. The behavior of both retrieval models under different query conditions is examined, and their effectiveness for Gujarati news retrieval tasks is evaluated.

The following subsections describe the BM25 and Vector Space Model (VSM) retrieval techniques used in the proposed framework.

3.3.1 BM25 Model

BM25 (Best Matching 25) is a probabilistic retrieval and ranking model widely used in information retrieval systems. The model estimates document relevance using term frequency, inverse document frequency, and document length normalization.

The BM25 score for a document d with respect to query q is calculated as:

$$\mathbf{BM25(q,\ d) = \Sigma\ [\ IDF(t) \times (\ tf(t,d) \times (k1 + 1)\ )\ /\ (\ tf(t,d) + k1 \times (1 - b + b \times |d|\ /\ avgdl\ )\ )\ ]}$$

Where:

- tf(t,d) = frequency of term t in document d
- |d| = length of document d
- avgdl = average document length in the corpus
- k1, b = tuning parameters (usually k1 = 1.5, b = 0.75)

The IDF (Inverse Document Frequency) is calculated as:

$$\mathbf{IDF(t) = \log(\ (N - df(t) + 0.5)\ /\ (df(t) + 0.5)\ )}$$

Where:

- N = total number of documents
- df(t) = number of documents containing term t

BM25 effectively retrieves documents for descriptive and context-rich queries because it incorporates document length normalization and probabilistic relevance scoring.

3.3.2 Vector Space Model

The Vector Space Model (VSM) represents documents and queries as vectors in a high-dimensional term space using TF–IDF weighting. In this model, document relevance is determined by measuring the similarity between the query vector and document vectors using cosine similarity.

VSM is widely used as a baseline retrieval approach because of its simplicity and effectiveness in keyword-based document retrieval tasks.

The TF-IDF weight is calculated as:

$$\mathbf{TF\text{-}IDF(t,d) = tf(t,d) \times \log(N\ /\ df(t))}$$

Where:

- tf(t,d) = frequency of term t in document d
- df(t) = number of documents containing term t
- N = total documents

To measure similarity between a query and a document, cosine similarity is used:

$$\mathbf{Similarity\ \ (q,\ d) = (q \cdot d)\ /\ (|q| \times |d|)}$$

Where:

- q · d = dot product of query and document vectors
- |q|, |d| = magnitude (length) of vectors

VSM performs effectively for short keyword-based queries because retrieval primarily depends on direct lexical similarity between query terms and document content. In this work, VSM

serves as a baseline retrieval model for comparative evaluation against BM25.

4. RESULTS AND ANALYSIS

The retrieval effectiveness of BM25 and the Vector Space Model (VSM) is evaluated using Gujarati event-oriented queries on the GSF-2009 dataset.

The experiments are conducted to analyze the behavior of both document retrieval models under different query conditions,specifically short keyword-based queries and long descriptive queries. Retrieval performance is evaluated using standard Information Retrieval (IR) evaluation metrics, namely Precision, Recall, and F1-score.

Precision measures the proportion of retrieved documents that are relevant to the given query, whereas Recall measures the proportion of relevant documents successfully retrieved from the corpus. The F1-score represents the harmonic mean of Precision and Recall and provides a balanced evaluation of retrieval effectiveness.

The evaluation metrics are formally defined as follows:

Precision = Relevant Retrieved Documents / Total Retrieved Documents

Recall = Relevant Retrieved Documents / Total Relevant Documents

F1-Score = 2 × (Precision × Recall) / (Precision + Recall)

Two Gujarati event-oriented queries representing the same real-world event are used for evaluation:

- Short Query: “ગુજરાતમાં ભારે વરસાદ”
- Descriptive Query: “ગુજરાતમાં ભારે વરસાદના કારણે અનેક જિલ્લાઓમાં પૂર જેવી સ્થિતિ”

Table 3 presents the query-wise comparison of BM25 and VSM using Precision, Recall, and F1-score for Gujarati news retrieval .

Table 3. Query-wise comparison of BM25 and VSM using Precision, Recall, and F1-score for Gujarati news retrieval.

Query	Model	Precision	Recall	F1-Score
ગુજરાતમાં ભારે વરસાદ	**BM25**	1.0	0.3	0.5
	VSM	1.0	0.7	0.8
ગુજરાતમાં ભારે વરસાદના કારણે અનેક જિલ્લાઓમાં પૂર જેવી સ્થિતિ	**BM25**	1.0	0.7	0.8
	VSM	0.8	0.5	0.6

Table 3 shows that the experimental results demonstrate that both BM25 and VSM are capable of retrieving relevant Gujarati news documents for event-oriented queries. For the short query “ગુજરાતમાં ભારે વરસાદ”, VSM achieves better retrieval effectiveness with a Recall value of 0.7 and an F1-score of 0.8, indicating stronger performance for concise keyword-based searches. In contrast, BM25 records a lower Recall value of 0.3 and an F1-score of 0.5 for the same query.

For the more descriptive query “ગુજરાતમાં ભારે વરસાદના કારણે અનેક જિલ્લાઓમાં પૂર જેવી સ્થિતિ”, BM25 outperforms VSM with Precision = 1.0, Recall = 0.7, and F1-score = 0.8. The improved performance of BM25 can be attributed to its probabilistic ranking mechanism and document length normalization, which make it more effective for handling long and context-rich queries. Meanwhile, VSM achieves Precision = 0.8, Recall = 0.5, and F1-score = 0.6 for the descriptive query.

Based on these retrieval patterns, Table 4 summarizes the comparative strengths of both retrieval models in terms of query handling, contextual relevance, and retrieval effectiveness.

Table 4. Analytical Comparison of BM25 and VSM for Gujarati News Retrieval

Analysis Criteria	BM25 Performance	VSM Performance	Observation
Short Query Performance	Good	Very Good	VSM performs effectively for keyword-based queries
Descriptive Query Performance	Excellent	Moderate	BM25 handles contextual and long queries more effectively
Retrieval Strategy	Probabilistic ranking	Vector similarity matching	Both models use different retrieval mechanisms
Query Context Handling	Better	Limited	BM25 captures contextual relevance more efficiently
Keyword-based Retrieval	Effective	Highly Effective	VSM performs well for direct term matching
Overall Retrieval Effectiveness	Better for complex queries	Better for simple queries	Model performance depends on query type

Table 4 is analytically derived from the experimental results presented in Table 3. The observations in Table 4 are based on the comparative evaluation of Precision, Recall, and F1-score obtained for BM25 and VSM using short and descriptive Gujarati event-oriented queries. Based on these retrieval patterns, Table 4 summarizes the analytical comparison of BM25 and VSM based on the experimental results obtained from Table 3, highlights the retrieval behavior of both models under different query conditions and demonstrates their respective strengths for Gujarati news retrieval tasks.

As illustrated in Fig 4, the comparative analysis shows the retrieval performance of BM25 and VSM for the short Gujarati query “ગુજરાતમાં ભારે વરસાદ”.

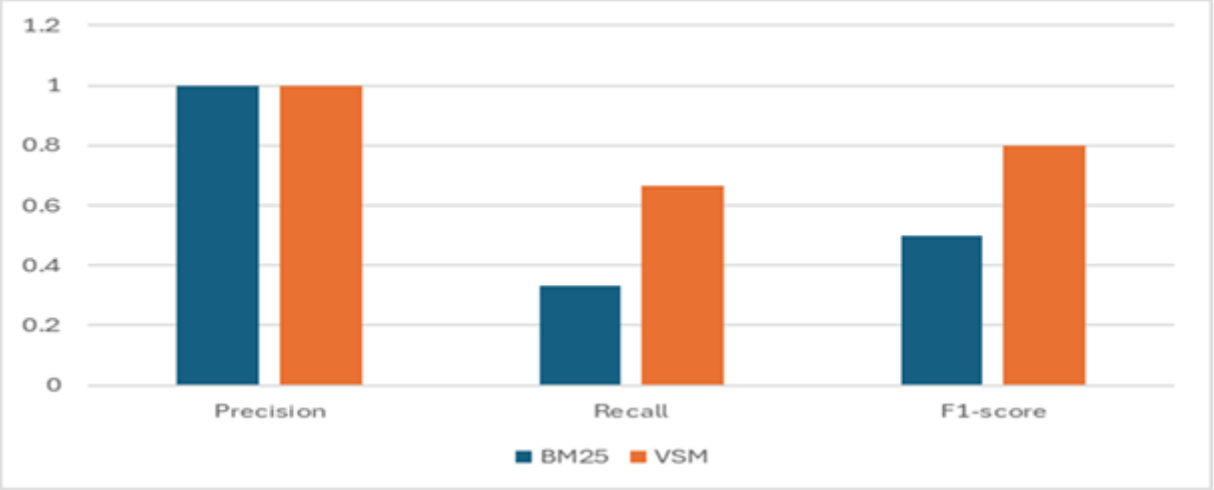

Fig 4: Comparative retrieval performance of BM25 and Vector Space Model (VSM) for the short event-oriented Gujarati query "ગુજરાતમાં ભારે વરસાદ"

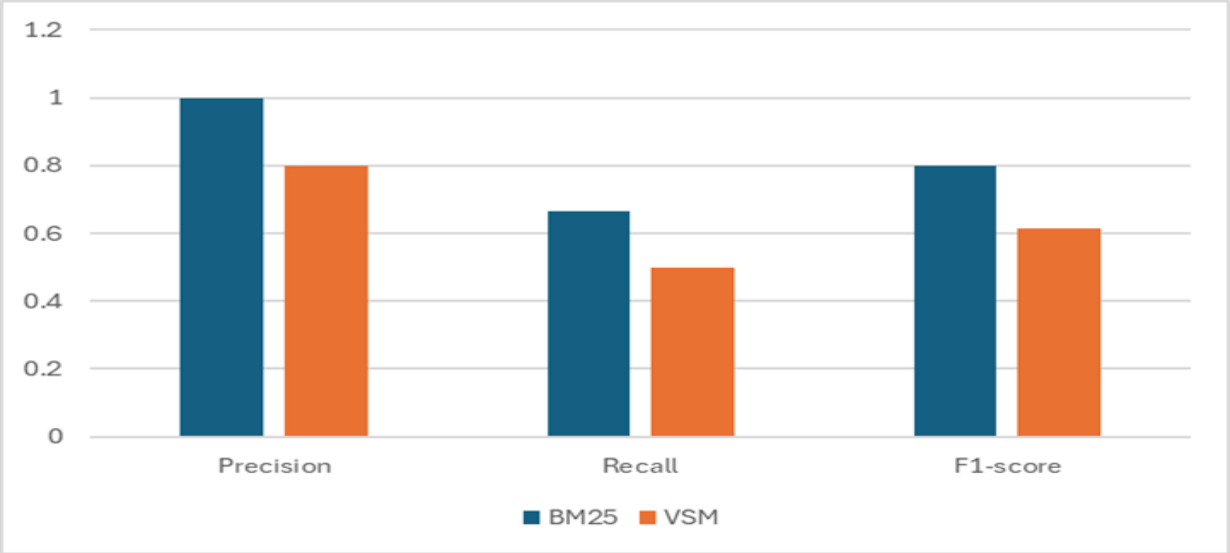

Fig 5: Comparison of BM25 and Vector Space Model (VSM) retrieval performance for the event-oriented query "ગુજરાતમાં ભારે વરસાદના કારણે અનેક જિલ્લાઓમાં પૂર જેવી સ્થિતિ"

As illustrated in Fig 5, BM25 demonstrates improved retrieval effectiveness for long and descriptive Gujarati queries because of its probabilistic ranking mechanism and document length normalization.

Based on the comparative experimental analysis, the following observations were derived:

• Retrieval effectiveness was observed to vary according to query length and contextual richness. VSM achieved better retrieval effectiveness for short Gujarati keyword-based queries.

• BM25 demonstrated improved retrieval performance for long and descriptive Gujarati queries.

• Retrieval effectiveness was significantly influenced by query structure and contextual richness.

• BM25 provided better contextual relevance through probabilistic ranking and document length normalization.

• VSM performed effectively for direct lexical matching and concise query retrieval.

• Both retrieval models successfully retrieved relevant Gujarati news documents under different query conditions.

• BM25 was observed to be more suitable for practical event-oriented Gujarati news retrieval tasks involving descriptive user queries.

• VSM served as an effective baseline retrieval model for short keyword-oriented searches in low-resource language environments.

5. CONCLUSION AND FUTURE WORK

Effective retrieval of Gujarati news documents from large-scale news corpora was achieved using classical Information Retrieval (IR) techniques. Comparative evaluation of BM25 and the Vector Space Model (VSM) was performed using short and descriptive Gujarati event-oriented queries on the GSF-2009 dataset. The experimental analysis demonstrated that retrieval effectiveness is significantly influenced by query structure and contextual richness.

For short keyword-based Gujarati queries, VSM achieved competitive retrieval performance because of its TF–IDF–

based vector similarity computation and direct lexical matching capability. In contrast, BM25 demonstrated improved retrieval effectiveness for long and context-rich queries through probabilistic relevance scoring and document length normalization. The obtained results confirm that BM25 provides more effective document ranking for descriptive Gujarati news retrieval tasks, whereas VSM serves as an efficient baseline retrieval model for concise keyword-oriented searches.

The study demonstrates that classical document retrieval approaches remain computationally efficient and practically effective for Gujarati information retrieval in low-resource language environments. The proposed framework establishes a baseline retrieval system for Gujarati news corpora and highlights the importance of appropriate retrieval models for different query formulations in event-oriented news retrieval applications.

Future enhancements may focus on multilingual Gujarati news retrieval, and temporal analysis of news documents. Retrieval effectiveness may also be improved through evaluation on larger Gujarati news corpora and optimization of ranking strategies for context-aware news retrieval applications in low-resource Indic languages.

6. REFERENCES

[1] Salton, G., Wong, A., Yang, C.S.: A vector space model for automatic indexing. Communications of the ACM 18(11), 613–620 (1975)

[2] Allan, J., Papka, R., Lavrenko, V.: On-line new event detection and tracking. In: Proc. 21st ACM SIGIR, pp. 37–45 (1998)

[3] Yang, Y., Pierce, T., Carbonell, J.: A study of retrospective and on-line event detection. In: Proc. 21st ACM SIGIR, pp. 28–36 (1998)

[4] Basaka, S.: Event detection from news in Indian languages using similarity-based pattern finding. In: Proc. FIRE 2020 (EDNIL) (2020)

[5] Singh, J., Goel, P., Debnath, A., Shrivastava, M.: A semantico-syntactic approach to event mention detection in Hindi. In: Proc. ISA Workshop (2021)

[6] Mirashi, A., Sonavane, S., Lingayat, P., Padhiyar, T., Joshi, R.: L3Cube-IndicNews: News-based datasets for Indic languages. arXiv:2401.02254 (2024)

[7] Haq, S., Sharma, A., Bhattacharyya, P.: IndicIRSuite: Multilingual datasets and retrieval models for Indian languages. arXiv:2312.09508 (2023)

[8] Xie, J., Zhang, Y., Kou, H., Zhao, X., Feng, Z., Song, L., Zhong, W.: A survey of the application of neural networks to event extraction. Tsinghua Science and Technology 30(2), 748–768 (2025)

[9] Nguyen, H., Shi, X., Li, J.: A survey on deep learning approaches for event extraction. IEEE Access 8, 16754–16769 (2020)

[10] Jiao, Y., Zhao, L.: Real-time extraction of news events based on BERT. International Journal of Advanced Networking and Monitoring and Controls 9(3), 24–34 (2024)

[11] Robertson, S., Zaragoza, H.: The probabilistic relevance framework: BM25 and beyond. Foundations and Trends in Information Retrieval 3(4), 333–389 (2009)

[12] Khoo, K.B., Ishizuka, M.: Topic extraction from news archive using TF*PDF algorithm. In: Proc. IEEE Int. Conf. on Web Intelligence, pp. 571–577 (2002)

[13] Balouchzahi, F., Shashirekha, H.L.: An approach for event detection from news in Indian languages using linear SVC. In: Proc. Forum for Information Retrieval Evaluation (FIRE 2020) – Event Detection from News in Indian Languages (EDNIL), CEUR Workshop Proceedings (2020)