
Hybrid Search Engine for Maharashtra Government Schemes

Aishwarya Gunjal¹, Kanchan Khemnari², Pawan Poojary³, Pranav Kulkarni⁴, Prof.
Manish Khodaskar⁵, Mr. Raviraj Joshi⁶

¹⁻⁵ IT Dept., Pune Institute of Computer Technology, Pune, India

⁶ L3Cube-Labs, Pune

Emails: gunjalaishwarya625@gmail.com, kanchankhemnar170@gmail.com,
pawanpoojary989@gmail.com, pranavkulkarni1115@gmail.com, mrkhodaskar@pict.edu,
raviraj.j1991@gmail.com

Abstract.

Language incompatibilities and restricted semantic search make it difficult for Marathi-speaking citizens to access government programs. A hybrid multilingual retrieval system that combines transformer-based embeddings with BM25 is presented. We compare MahaSBERT and IndicSBERT on a carefully selected bilingual dataset of 64 schemes. With MahaSBERT, F1-score of 0.85 and MRR 0.91, hybrid retrieval performs 7% better than lexical retrieval and much better than semantic retrieval alone. This illustrates the effectiveness of hybrid searches for administrative domains with limited resources.

Keywords. Information Retrieval, Hybrid Search, Marathi, Government Schemes, BM25, MahaSBERT

1. INTRODUCTION

Government welfare schemes in Maharashtra remain underutilized due to limited awareness and language barriers among Marathi-speaking citizens. Traditional lexical search such as BM25 struggles with semantic differences between queries and scheme descriptions, while transformer-based models may miss exact scheme names. This study proposes a hybrid multilingual search engine combining MahaSBERT, IndicSBERT, and translation-based retrieval for Marathi and English schemes. The proposed approach improves retrieval accuracy by up to 8.7% for Marathi queries compared with standalone methods.

2. RELATED WORK

BM25 [3] supports keyword matching but struggles with vocabulary mismatch, while BERT-based models [7,8] capture semantics but may miss exact terms. MahaSBERT [5] and IndicBERT [6] improve Marathi and multilingual retrieval. Existing portals such as Aaple Sarkar [9] have limited language support. This study proposes a hybrid multilingual retrieval system for government schemes..

3. DATASET AND TEST SET

We scraped official portals (india.gov.in, Aaple Sarkar) to create a bilingual corpus of 64 government schemes. The scheme name, description, department, category, and eligibility are all included in each entry. Google Translate was used to translate the English data to Marathi, with 20% manual checking. The corpus is summarized in Table 1.

Table 1: Bilingual Government Scheme & Public Service Corpus Statistics

Metric	English	Marathi	Metric
Number of Records	64	64	Number of Records
Avg. Title Length (words)	3.8 ± 2.1	4.6 ± 2.4	Avg. Title Length (words)
Avg. Description Length (words)	178.3 ± 96.4	171.5 ± 92.8	Avg. Description Length (words)
Description Range	5–620	6–590	Description Range
Domain Categories	14	14	Domain Categories
Language Type	Formal administrative	Formal administrative	Language Type
Corpus Type	Structured descriptive	Structured descriptive	Corpus Type
Content	Title + Description + URL	Title + Description + URL	Content

To verify the effectiveness of the search algorithm, we created a test dataset with over 150 actual questions (in both Marathi and English). Each query was manually matched to the government program that best suited its objectives. Each query was compared to the existing scheme dataset for evaluation, and the top results were examined to determine whether the correct scheme was present. Basic relevance testing was done by hand to verify the outputs.

3.1. SYSTEM ARCHITECTURE

The system is made up of UI, processing, and results layers (Fig. 1). Users enter queries with optional filters in either English or Marathi. Three searches are carried out concurrently by the processing layer: (1) lexical (BM25 on Elasticsearch), (2) semantic (embedding similarity via Elastic Search), and (3) hybrid. The hybrid score combines normalized lexical and semantic scores:

$$S_{\text{hybrid}}(q,d) = \alpha \cdot \text{norm}(S_{\text{lexical}}) + (1 - \alpha) \cdot \text{norm}(S_{\text{semantic}}) \quad (1)$$

with $\alpha = 0.5$ (empirically set). Results are ranked by hybrid score, deduplicated, and presented with explanations (Scheme links, Similarity breakdown).

4. MODELS AND METHODS

5. EXPERIMENTAL SETUP

6. RESULTS AND DISCUSSION

Table 2: Performance Comparison Across Retrieval Methods

Model	Search Type	Precision	Recall	MRR	F1-score
IndicSBERT	Hybrid	0.795	0.950	0.872	0.779
IndicSBERT	Lexical	0.781	0.946	0.864	0.766
IndicSBERT	Semantic	0.575	0.804	0.663	0.572
MahaSBERT	Hybrid	0.808	0.946	0.879	0.793

Model	Search Type	Precision	Recall	MRR	F1-score
IndicSBERT	Hybrid	0.795	0.950	0.872	0.779
IndicSBERT	Lexical	0.781	0.946	0.864	0.766
MahaSBERT	Lexical	0.781	0.946	0.864	0.766
MahaSBERT	Semantic	0.562	0.775	0.655	0.562

7. LIMITATIONS AND FUTURE WORK

The 64 schemes in the current dataset are modest; there are translation errors and vocabulary particular to the domain. Future developments will include voice-based search, adaptive hybrid weighting, personalized suggestions, and language expansion.

8. CONCLUSION

We built a hybrid marathi search engine that combines transformer embeddings with BM25 for Maharashtra government initiatives. For Marathi, hybrid retrieval with MahaSBERT improves Accuracy@5 by 8.7%, outperforming lexical and semantic approaches alone. This illustrates the usefulness of language-specific embeddings and the efficacy of hybrid techniques in administrative domains with limited resources.

9. REFERENCES

- [1] A. Joshi, A. Kajale, J. Gadre, S. Deode, and R. Joshi, ‘L3Cube-MahaSBERT and HindSBERT: Sentence BERT models and benchmarking BERT sentence representations for Hindi and Marathi’, arXiv preprint arXiv:2211.11187, 2021.
- [2] S. Deode, J. Gadre, A. Kajale, A. Joshi, and R. Joshi, ‘L3Cube-IndicSBERT: A simple approach for learning cross-lingual sentence representations using multilingual BERT’, arXiv preprint arXiv:2304.11434, 2023.
- [3] N. Reimers and I. Gurevych, ‘Sentence-BERT: Sentence embeddings using Siamese BERT networks’, Proc. 2019 Conf. on Empirical Methods in Natural Language Processing and the 9th Int. Joint Conf. on Natural Language Processing (EMNLP-IJCNLP), pp. 3980-3990, Hong Kong, Nov., 2019.
- [4] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, ‘BERT: Pre-training of deep bidirectional transformers for language understanding’, Proc. 2019 Conf. of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT), pp. 4171-4186, Minneapolis, Jun., 2019.
- [5] S. Robertson and H. Zaragoza, ‘The probabilistic relevance framework: BM25 and beyond’, Foundations and Trends in Information Retrieval, vol. 3, no. 4, pp. 333-389, 2009.

Biographies



Aishwarya Gunjal received her Bachelor's degree in Information Technology Engineering from Pune Institute of Computer Technology, Pune, in 2026.



Kanchan Khemnagar received her Bachelor's degree in Information Technology Engineering from Pune Institute of Computer Technology, Pune, in 2026.



Pranav Kulkarni received his Bachelor's degree in Information Technology Engineering from Pune Institute of Computer Technology, Pune, in 2026.



Pawan Poojary received his Bachelor's degree in Information Technology Engineering from Pune Institute of Computer Technology, Pune, in 2026.



Prof. Manish Khodaskar, Assistant Professor at Pune Institute of Computer Technology.



Mr. Raviraj Joshi Senior Data Scientist at NVIDIA, Affiliations: L3Cube Labs Pune, Indian Institute of Technology, Madras.