Reciprocal Rank Fusion Based Hybrid Dense–Sparse Information Retrieval on Code-Mixed Banglish Social Media Text

Burhanuddin Merchant[1,*,†], Ashwaq Khazi[1,†] and Sheetal S. Sonawane[1]

1SCTR's Pune Institute of Computer Technology, Pune, India

Abstract

Social media platforms generate vast amounts of code-mixed text, such as Banglish (Bengali-English), which poses unique challenges for information retrieval due to spelling variations, transliterations, and informal usage. Traditional sparse retrieval methods like BM25 fail to fully capture semantic meaning, while dense embedding models such as Sentence Transformers may overlook lexical matches. In this work, we propose a hybrid retrieval framework that integrates BM25 and a triplet-tuned Sentence Transformer model using Reciprocal Rank Fusion (RRF). Our approach leverages the complementary strengths of sparse and dense retrieval, ensuring robust performance on noisy Banglish social media data. We evaluate our system on the FIRE 2025 code-mixed information retrieval shared task, achieving 6th place with a MAP score of 0.123, NDCG score of 0.376, P@5 of 0.293, and P@10 of 0.21. The results demonstrate that RRF fusion significantly improves retrieval effectiveness compared to standalone methods, making it a promising strategy for code-mixed information retrieval.

Keywords

Information Retrieval, Code-Mixed Text, Banglish, Reciprocal Rank Fusion, Dense Retrieval, Sparse Retrieval, Social Media Text

1. Introduction

With the exponential growth of textual data from diverse digital sources, the need for robust and efficient information retrieval (IR) mechanisms has become increasingly critical. Social media platforms, in particular, serve as a major source of multilingual and code-mixed content, especially in linguistically diverse regions.

Banglish, a code-mixed variety combining Bengali and English, is widely used on social media platforms among Bengali speakers [1, 2, 3]. This linguistic phenomenon introduces several challenges for IR systems due to its unique characteristics—such as inconsistent spelling, varied transliteration schemes, informal grammar, and the seamless integration of two distinct languages within the same text.

Bengali, an Indo-Aryan language spoken primarily in the eastern regions of the Indian subcontinent, ranks as the seventh most spoken language globally, with nearly 300 million speakers. Despite this, most IR research has focused predominantly on English, with limited exploration into languages like Bengali and their code-mixed variants.

IR for code-mixed languages poses several challenges, including:

- Limited availability of large-scale datasets
- Complex linguistic and structural variations
- Inconsistent transliteration patterns
- Informal and noisy social media language

The Forum for Information Retrieval Evaluation (FIRE) 2025 shared task on code-mixed IR from social media data provides a standardized framework to tackle these challenges. However, conventional IR methods face significant limitations in this domain. Sparse retrieval models such as BM25 [4] excel at exact lexical matching but fail to capture semantic relationships—particularly when concepts appear in different languages or transliteration forms. Conversely, dense retrieval models using neural embeddings [5] capture semantic similarity effectively but may overlook important lexical cues crucial for precision.

To address these limitations, we propose a hybrid retrieval framework that combines the strengths of both sparse and dense retrieval methods through Reciprocal Rank Fusion (RRF) [6]. Our approach integrates BM25 for lexical matching with a fine-tuned Sentence Transformer model [7] optimized for code-mixed Banglish text. The fusion mechanism ensures a balance between lexical precision and semantic recall, thereby enhancing overall retrieval effectiveness.

The main contributions of this work are as follows:

- A hybrid retrieval framework tailored for code-mixed Banglish social media data
- A triplet-tuned Sentence Transformer model for enhanced cross-lingual semantic understanding
- Comprehensive evaluation on the FIRE 2025 shared task, achieving competitive performance (6th place)
- Analysis of the complementary roles of sparse and dense retrieval in code-mixed scenarios
- Empirical demonstration of the effectiveness of RRF-based fusion for code-mixed IR

2. Related Work

Research in information retrieval has seen significant progress over the decades, yet most advancements have been centered around English-language data [8]. In recent years, increasing attention has been directed toward IR for code-mixed text, driven by the growing prevalence of multilingual social media communication.

Chanda and Pal [1] investigated the impact of stopword removal on code-mixed IR, highlighting the unique linguistic and orthographic challenges present in such data. Traditional IR approaches typically fall into two categories: lexical matching and semantic embedding methods—each offering distinct strengths and limitations.

Historically, IR systems have focused on three fundamental tasks [8, 9]:

- Capturing and representing key textual information
- Scoring documents based on relevance
- Selecting the most relevant documents for retrieval

Early IR systems were predominantly rule-based, followed by production-rule models. The introduction of Term Frequency (TF) and Inverse Document Frequency (IDF) weighting [9] marked a significant milestone, representing text in vector space and assigning term weights based on occurrence frequency. Subsequently, language modeling approaches [10] further advanced the field by introducing probabilistic frameworks for document ranking, paving the way for modern neural and hybrid retrieval methods.

2.1. Code-Mixed Information Retrieval

The FIRE 2025 shared task on code-mixed information retrieval from social media data [11] established a comprehensive evaluation framework for this challenging domain. The task focuses on Banglish text, which combines Bengali and English in various forms, creating significant challenges for traditional retrieval systems due to inconsistent transliteration, informal language use, and semantic ambiguity across languages. Code-switching research has gained momentum with comprehensive surveys [12] and evaluation benchmarks like GLUECoS [13] and LinCE [14]. Social media platforms have become primary sources for code-mixed content [15, 16], making this domain particularly relevant for information retrieval research.

2.2. Sparse Retrieval Methods

Sparse retrieval methods, particularly BM25 [4], have been the backbone of information retrieval systems for decades. These methods excel at exact lexical matching but struggle with vocabulary mismatch problems, especially in multilingual and code-mixed scenarios where the same concept can be expressed in multiple languages or transliteration schemes. The vector space model [9] laid the foundation for modern sparse retrieval, while probabilistic models [10] provided theoretical frameworks for ranking documents based on query likelihood.

2.3. Dense Retrieval Methods

Significant developments started happening in the dense retrieval space after the introduction of neural networks for natural language processing [17]. The transformer architecture [18] revolutionized the field, leading to powerful pre-trained models like BERT [19] and RoBERTa [20]. Recent advances in neural information retrieval have introduced dense retrieval methods using pre-trained language models [5]. Sentence Transformers [7] and similar architectures can capture semantic relationships between queries and documents, but may miss important lexical signals that are crucial for precision in specialized domains. Advanced models like ColBERT [21] and neural ranking approaches [22] have further improved dense retrieval effectiveness.

2.4. Multilingual and Cross-lingual Retrieval

Cross-lingual information retrieval has been extensively studied [23, 24], with recent advances in multilingual representations [25] enabling better cross-language understanding. For Indic languages specifically, significant progress has been made with resources like IndicNLP [26] and models like IndicBART [27], which provide strong foundations for Bengali and other Indian language processing.

2.5. Hybrid Retrieval Approaches

Hybrid retrieval systems that combine sparse and dense methods have shown promising results in various domains [28, 29]. The complementary nature of lexical and semantic matching makes hybrid approaches particularly suitable for challenging scenarios like code-mixed text retrieval. Reciprocal Rank Fusion [6] has emerged as an effective method for combining different ranking systems without requiring score normalization. Recent work on hybrid approaches includes RankT5 [30] and SPLADE [31], which demonstrate the effectiveness of combining different retrieval paradigms.

3. Methodology

Our hybrid retrieval framework combines sparse and dense retrieval methods through Reciprocal Rank Fusion to leverage the complementary strengths of both approaches.

3.1. Sparse Retrieval Component

We employ BM25 as our sparse retrieval baseline, which provides strong lexical matching capabilities. BM25 is particularly effective for capturing exact term matches and handling proper nouns, technical terms, and transliterated words that are common in Banglish text.

3.2. Dense Retrieval Component

For dense retrieval, we fine-tune a Sentence Transformer model using triplet loss on code-mixed Banglish data. The model learns to encode semantic relationships between queries and documents in a shared embedding space, enabling retrieval based on semantic similarity rather than just lexical overlap.

3.3. Reciprocal Rank Fusion

Reciprocal Rank Fusion combines the ranked lists from both sparse and dense retrieval methods. For each query, we obtain two ranked lists: one from BM25 and another from the fine-tuned Sentence Transformer. The RRF score for each document is calculated as:

$$RRF(d) = \sum_{r \in R} \frac{1}{k + rank_r(d)} \tag{1}$$

where R is the set of rankers (BM25 and Sentence Transformer), $rank_r(d)$ is the rank of document d in ranker r, and k is a constant (typically 60) that controls the impact of lower-ranked documents.

3.4. Model Fine-tuning

We fine-tune the Sentence Transformer model using triplet loss on a dataset of Banglish queries and relevant documents. Our implementation uses the paraphrase-multilingual-MiniLM-L12-v2 model as the base, which provides strong multilingual capabilities essential for code-mixed text understanding. The triplet loss encourages the model to place relevant query-document pairs closer in the embedding space while pushing irrelevant pairs apart:

$$L_{triplet} = \max(0, \delta + d(q, d^-) - d(q, d^+)) \tag{2}$$

where q is the query, d^+ is a relevant document, d^- is an irrelevant document, $d(\cdot, \cdot)$ is the cosine distance, and δ is the margin.

3.5. Bengali to Banglish Conversion

To handle the mixed-language nature of the queries, we implement a Bengali to Banglish converter that transliterates Bengali Unicode text to romanized form. This preprocessing step ensures compatibility with the multilingual Sentence Transformer model and improves retrieval performance by normalizing the representation of Bengali terms.

4. Experimental Setup

4.1. Dataset

We evaluate our approach on the FIRE 2025 shared task dataset for code-mixed information retrieval from social media data [32]. The dataset contains Banglish social media posts and queries with varying degrees of English and Bengali content, representing realistic social media usage patterns. The training set consists of 20 queries with relevance judgments, while the test set contains 30 queries for evaluation.

The dataset presents several challenges typical of social media text:

- Inconsistent spelling and transliteration schemes
- Informal language and abbreviations
- Code-switching between Bengali and English
- Noisy text with grammatical errors
- Varying levels of language mixing within documents

4.2. Implementation Details

Our system is implemented using Python with the sentence-transformers library for dense retrieval and the rank-bm25 library for sparse retrieval. The implementation follows these key specifications:

Dense Retrieval Component:

- Base model: paraphrase-multilingual-MiniLM-L12-v2

- Fine-tuning: 2 epochs with triplet loss
- Batch size: 8
- Learning rate: Default with 50 warmup steps
- Maximum sequence length: 512 tokens

Sparse Retrieval Component:

- Algorithm: BM25 with Okapi normalization
- Parameters: k1=1.2, b=0.75
- Preprocessing: Basic tokenization and cleaning
- Language handling: Bengali to Banglish conversion

Fusion Parameters:

- RRF constant: k=60
- Top-k documents: 1000 per query
- Score normalization: Min-max for weighted fusion

4.3. Evaluation Metrics

The effectiveness of the generated retrievals is evaluated using standard information retrieval metrics established in the literature [8, 33]:

- Mean Average Precision at 10 (MAP@10) - measures the average precision across all relevant documents
- Normalized Discounted Cumulative Gain at 10 (NDCG@10) [34] - accounts for the graded relevance and position of documents
- Precision at 5 (P@5) - measures the fraction of relevant documents in the top 5 results
- Precision at 10 (P@10) - measures the fraction of relevant documents in the top 10 results

These metrics are widely used in information retrieval evaluation [35] and provide comprehensive assessment of retrieval effectiveness across different aspects of performance.

4.4. Baseline Methods

We compare our hybrid RRF approach against several baseline methods:

- BM25 (sparse retrieval only)
- Sentence Transformer (dense retrieval only)
- Weighted fusion (linear combination of normalized BM25 and dense scores)

5. Results and Discussion

Our experimental evaluation demonstrates the effectiveness of the hybrid RRF approach for code-mixed Banglish information retrieval on the FIRE 2025 shared task.

5.1. FIRE 2025 Shared Task Results

Our hybrid RRF system achieved competitive performance on the FIRE 2025 shared task, ranking 6th among participating teams. Table 1 shows our official results compared to the task baseline and our internal component analysis.

The results demonstrate that RRF effectively combines the strengths of both sparse and dense retrieval methods, consistent with findings in other hybrid retrieval studies [28, 29]. Our system achieved a MAP@10 score of 0.123, NDCG@10 of 0.376, P@5 of 0.293, and P@10 of 0.210, securing 6th place in the

Table 1
FIRE 2025 Shared Task Results and Component Analysis

Method	MAP@10	NDCG@10	P@5	P@10
BM25 Only	0.089	0.298	0.200	0.167
Sentence Transformer Only	0.105	0.332	0.233	0.183
Weighted Fusion	0.118	0.361	0.267	0.200
RRF (Our Submission)	**0.123**	**0.376**	**0.293**	**0.210**
Ranking	*6th place*	*6th place*	*6th place*	*6th place*

competition. While BM25 excels at capturing exact lexical matches [4], the Sentence Transformer model provides better semantic understanding [7]. The RRF fusion approach achieves superior performance compared to individual methods and simple weighted fusion, aligning with theoretical expectations about rank-based fusion methods [6]. This performance is competitive within the context of code-mixed retrieval challenges [36] and demonstrates the effectiveness of hybrid approaches for multilingual scenarios [25].

5.2. Analysis of Code-Mixed Challenges

Our analysis reveals that the hybrid approach is particularly effective for handling the unique challenges of code-mixed text:

- **Spelling variations and transliteration inconsistencies**: The Bengali to Banglish converter helps normalize different transliteration schemes, while BM25 captures exact matches for consistent spellings.
- **Semantic relationships across languages**: The fine-tuned Sentence Transformer model effectively captures semantic similarity between Bengali and English expressions of the same concept.
- **Informal social media language patterns**: The combination of lexical and semantic matching handles both formal terms and informal social media expressions.
- **Query-document language mismatch**: RRF fusion ensures that documents are retrieved even when queries and documents use different languages for the same concept.

5.3. Component Analysis

The component analysis in Table 1 shows that each retrieval method contributes differently to the overall performance:

- BM25 provides a solid baseline with strong precision for exact matches
- Sentence Transformer improves recall by capturing semantic relationships
- Weighted fusion shows improvement but is sensitive to score normalization
- RRF fusion achieves the best performance by effectively combining rankings without requiring score calibration

6. Conclusion

We presented a hybrid information retrieval framework that effectively addresses the challenges of code-mixed Banglish social media text. By combining BM25 sparse retrieval with a fine-tuned Sentence Transformer model through Reciprocal Rank Fusion, our approach achieves competitive performance on the FIRE 2025 shared task, ranking 6th with a MAP@10 score of 0.123.

The key findings of our work include:

- RRF fusion significantly improves retrieval effectiveness over standalone sparse or dense methods, achieving 38% improvement in MAP@10 over BM25 alone
- The hybrid approach is particularly robust for handling code-mixed text challenges, including transliteration inconsistencies and cross-language semantic relationships
- Fine-tuning dense models on code-mixed data with triplet loss is crucial for optimal performance
- Bengali to Banglish conversion preprocessing enhances compatibility with multilingual models
- RRF fusion outperforms weighted fusion by avoiding score normalization issues

6.1. Future Work

Future research directions include exploring more sophisticated fusion techniques beyond RRF [31, 30], investigating the impact of different pre-trained multilingual models [25, 27], and extending the approach to other code-mixed language pairs documented in recent surveys [12]. Additionally, incorporating user feedback mechanisms [37], query expansion techniques for multilingual scenarios [24], and advanced preprocessing methods for code-mixed text [13] could further improve retrieval performance. The development of specialized evaluation metrics for code-mixed retrieval tasks, similar to existing benchmarks [35, 14], would also benefit the research community. Integration with recent advances in neural ranking [29] and cross-lingual representations [26] presents additional opportunities for improvement.

Acknowledgments

Generative AI Declaration

During the preparation of this work, the author(s) used Claude, Gemini, and Grammarly for grammar and spelling checks, as well as for paraphrasing and rewording. After using these tools/services, the author(s) reviewed and edited the content as needed and take full responsibility for the publication's content.

References

[1] S. Chanda, S. Pal, The effect of stopword removal on information retrieval for code-mixed data obtained via social media, SN Comput. Sci. 4 (2023). URL: https://doi.org/10.1007/s42979-023-01942-7. doi:10.1007/s42979-023-01942-7.

[2] S. Chanda, S. Pal, Overview of the shared task on code-mixed information retrieval from social media data, in: Proceedings of the 16th Annual Meeting of the Forum for Information Retrieval Evaluation, FIRE '24, Association for Computing Machinery, New York, NY, USA, 2025, p. 29–31. URL: https://doi.org/10.1145/3734947.3735670. doi:10.1145/3734947.3735670.

[3] S. Chanda, S. Pal, Overview of the shared task on code-mixed information retrieval from social media data, in: FIRE 2024 Working Notesl, CEUR Workshop Proceedings, 2024, p. 124–128. URL: https://ceur-ws.org/Vol-4054/T2-1.pdf.

[4] S. Robertson, H. Zaragoza, The probabilistic relevance framework: Bm25 and beyond, Foundations and Trends in Information Retrieval 3 (2009) 333–389.

[5] V. Karpukhin, B. Oğuz, S. Min, P. Lewis, L. Wu, S. Edunov, D. Chen, W.-t. Yih, Dense passage retrieval for open-domain question answering, in: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, 2020, pp. 6769–6781.

[6] G. V. Cormack, C. L. Clarke, S. Buettcher, Reciprocal rank fusion outperforms condorcet and individual rank learning methods, in: Proceedings of the 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval, 2009, pp. 758–759.

[7] N. Reimers, I. Gurevych, Sentence-bert: Sentence embeddings using siamese bert-networks, in: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, 2019, pp. 3982–3992.

[8] C. D. Manning, P. Raghavan, H. Schütze, Introduction to information retrieval, Cambridge University Press, 2008.

[9] G. Salton, A. Wong, C.-S. Yang, A vector space model for automatic indexing, Communications of the ACM 18 (1975) 613–620. doi:10.1145/361219.361220.

[10] J. M. Ponte, W. B. Croft, A language modeling approach to information retrieval, in: Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, 1998, pp. 275–281. doi:10.1145/290941.291008.

[11] S. Chanda, K. Tewari, S. Pal, Overview of the cmir track at fire 2025: Code-mixed information retrieval from social media data, in: FIRE '25: Proceedings of the 17th Annual Meeting of the Forum for Information Retrieval Evaluation. December 17-20, Varanasi, India, Association for Computing Machinery (ACM), New York, NY, USA, 2025.

[12] S. Sitaram, K. R. Chandu, S. K. Rallabandi, A. W. Black, A survey of code-switched speech and language processing, arXiv preprint arXiv:1904.00784 (2019). doi:10.48550/arXiv.1904.00784.

[13] S. Khanuja, S. Dandapat, A. Srinivasan, S. Sitaram, M. Choudhury, Gluecos: An evaluation benchmark for code-switched nlp, in: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, 2020, pp. 3575–3585. doi:10.18653/v1/2020.acl-main.329.

[14] G. Aguilar, S. Kar, T. Solorio, Lince: A centralized benchmark for linguistic code-switching evaluation, in: Proceedings of the 12th Language Resources and Evaluation Conference, 2020, pp. 1803–1813.

[15] H. Kwak, C. Lee, H. Park, S. Moon, What is twitter, a social network or a news media?, in: Proceedings of the 19th International Conference on World Wide Web, 2010, pp. 591–600. doi:10.1145/1772690.1772751.

[16] M. Efron, Information search and retrieval in microblogs, Journal of the American Society for Information Science and Technology 62 (2011) 996–1008. doi:10.1002/asi.21512.

[17] A. Rogers, O. Kovaleva, A. Rumshisky, A primer on neural network models for natural language processing, Journal of Artificial Intelligence Research 61 (2020) 65–95. doi:10.1613/jair.1.11640.

[18] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, I. Polosukhin, Attention is all you need, Advances in Neural Information Processing Systems 30 (2017).

[19] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding, in: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2019, pp. 4171–4186. doi:10.18653/v1/N19-1423.

[20] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, V. Stoyanov, Roberta: A robustly optimized bert pretraining approach, arXiv preprint arXiv:1907.11692 (2019). doi:10.48550/arXiv.1907.11692.

[21] O. Khattab, M. Zaharia, Colbert: Efficient and effective passage search via contextualized late interaction over bert, in: Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval, 2020, pp. 39–48. doi:10.1145/3397271.3401075.

[22] C. Xiong, Z. Dai, J. Callan, Z. Liu, R. Power, End-to-end neural ad-hoc ranking with kernel pooling, in: Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval, 2017, pp. 55–64. doi:10.1145/3077136.3080809.

[23] D. W. Oard, B. J. Dorr, A comparative study of query and document translation for cross-language information retrieval, in: Conference of the Association for Machine Translation in the Americas, 1997, pp. 472–483. doi:10.1007/3-540-49478-2_42.

[24] J.-Y. Nie, Cross-language information retrieval based on parallel texts and automatic mining of parallel texts from the web, in: Proceedings of the 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval, 2010, pp. 819–826. doi:10.1145/1835449.1835602.

[25] A. Conneau, K. Khandelwal, N. Goyal, V. Chaudhary, G. Wenzek, F. Guzmán, E. Grave, M. Ott, L. Zettlemoyer, V. Stoyanov, Unsupervised cross-lingual representation learning at scale, in: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, 2020, pp. 8440–8451. doi:10.18653/v1/2020.acl-main.747.

[26] A. Kunchukuttan, D. Kakwani, S. Golla, G. NC, A. Bhattacharyya, M. M. Khapra, P. Kumar, Ai4bharat-indicnlp corpus: Monolingual corpora and word embeddings for indic languages, in: Findings of the Association for Computational Linguistics: EMNLP 2020, 2020, pp. 4948–4961. doi:10.18653/v1/2020.findings-emnlp.445.

[27] R. Dabre, S. Doddapaneni, H. Mhaske, M. Chitale, M. M. Khapra, P. Goyal, R. Aralikatte, Indicbart: A pre-trained model for indic natural language generation, in: Findings of the Association for Computational Linguistics: ACL 2022, 2022, pp. 1849–1863. doi:10.18653/v1/2022.findings-acl.145.

[28] P. Shaw, P. Pasupat, K. Toutanova, Multi-stage hybrid retrieval for biomedical question answering, in: Proceedings of the 2nd Clinical Natural Language Processing Workshop, 2019, pp. 109–118. doi:10.18653/v1/W19-1914.

[29] J. Lin, R. Nogueira, A. Yates, Pretrained transformers for text ranking: BERT and beyond, Morgan & Claypool Publishers, 2021. doi:10.2200/S01123ED1V01Y202108HLT053.

[30] H. Zhuang, Z. Qin, R. Jagerman, K. Hui, J. Ma, J. Lu, J. Ni, X. Wang, M. Bendersky, Rankt5: Fine-tuning t5 for text ranking with ranking losses, in: Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval, 2021, pp. 2045–2049. doi:10.1145/3404835.3463098.

[31] T. Formal, B. Piwowarski, S. Clinchant, Splade: Sparse lexical and expansion model for first stage ranking, in: Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval, 2021, pp. 2288–2292. doi:10.1145/3404835.3463098.

[32] S. Chanda, K. Tewari, S. Pal, Findings of the code-mixed information retrieval from social media data (cmir) shared task at fire 2025, in: K. Ghosh, T. Mandl, S. Pal, S. Majumdar, A. Chakraborty (Eds.), Forum for Information Retrieval Evaluation (Working Notes) (FIRE 2025) December 17-20, Varanasi, India, CEUR-WS.org, 2025.

[33] E. M. Voorhees, The trec-8 question answering track report, Proceedings of TREC 99 (1999) 77–82.

[34] K. Järvelin, J. Kekäläinen, Cumulated gain-based evaluation of ir techniques, ACM Transactions on Information Systems 20 (2002) 422–446. doi:10.1145/582415.582418.

[35] N. Thakur, N. Reimers, A. Rücklé, A. Srivastava, I. Gurevych, Beir: A heterogeneous benchmark for zero-shot evaluation of information retrieval models, in: Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track, 2021.

[36] P. Majumder, M. Mitra, T. Ghosal, A. Ekbal, S. Sett, J. K. Singh, Overview of the fire 2020 track: Information retrieval from microblogs during disasters, in: Proceedings of the 12th Annual Meeting of the Forum for Information Retrieval Evaluation, 2020, pp. 1–6. doi:10.1145/3441501.3441540.

[37] I. Ruthven, M. Lalmas, A survey on the use of relevance feedback for information access systems, The Knowledge Engineering Review 18 (2003) 95–145.