# UPERF:Urdu Proximity Enhanced Retrieval Framework

Samreen Kazi, Shakeel Khoja

Proceedings of the 38th Pacific Asia Conference on
Language, Information and Computation (PACLIC 38)

Nathaniel Oco, Shirley N. Dita, Ariane Macalinga Borlongan, Jong-Bok Kim (eds.)

2024

# UPERF:Urdu Proximity Enhanced Retrieval Framework

**Samreen Kazi** and **Shakeel Khoja**
School of Mathematics and Computer Science
Institute of Business Administration (IBA)
Karachi, Pakistan
{sakazi, skhoja}@iba.edu.pk

## Abstract

Traditional document retrieval for Urdu faces challenges due to the language's complex morphological structure and limited resources. While existing approaches rely heavily on term-matching techniques, they often fail to capture semantic relationships effectively. This paper introduces the Urdu Proximity Enhanced Retrieval Framework (UPERF), which combines traditional retrieval models with modern embedding techniques through an optimized weighting scheme. Using the UND corpus of 2.8M documents, we evaluate various configurations of Word2Vec, FastText, Doc2Vec, and mBERT models alongside traditional approaches. Our framework employs grid search to determine optimal weights for combining TF-IDF, BM25, and embedding-based proximity measures. The results show that Word2Vec with stemmed text preprocessing and cosine similarity achieves a Recall@5 of 0.85, significantly outperforming baseline methods. Analysis of document rankings demonstrates that our weighted approach better aligns with human relevance judgments compared to individual methods.

## 1 Introduction

Document retrieval is a fundamental task in information retrieval that involves fetching relevant documents from a large corpus based on a user's query. This task is particularly challenging in low-resource languages like Urdu, the official language of Pakistan, spoken by millions globally[1]. The scarcity of annotated data and linguistic resources further complicates document retrieval in Urdu (Iqbal et al., 2021). Traditional vector space models, such as TF-IDF, are commonly used for document retrieval. However, their reliance on term frequency limit their effectiveness in capturing semantic nuances (Kazi and Khoja, 2021) (Kazi and Khoja, 2024) (Rasolofo and Savoy, 2003) (Beigbeder and Mercier, 2005). The need for effective document retrieval in Urdu has become increasingly critical, especially with the surge in online educational materials and digital content in Urdu following the COVID-19 pandemic (Kazi et al., 2023). Previous efforts in document retrieval for low-resource languages have primarily focused on traditional approaches, such as boolean retrieval and TF-IDF (Magueresse et al., 2020)(Novak et al., 2022). While these methods are effective to some extent, they often fall short in capturing the deeper semantic relationships within the text. Research efforts in Urdu information retrieval have recognized the critical need to build specialized test collections to build and evaluate effectiveness of IR models, ranking algorithms, and various natural language processing techniques (Shaukat et al., 2022). However, inherent linguistic differences between Urdu and English, including different syntactic and morphological structures, script variations, and a scarcity of resources, pose significant obstacles to the direct application of English-based algorithms in Urdu language processing (Nasim and Haider, 2022). This study investigates enhancing Urdu document retrieval by incorporating proximity measures with established models like BM25 (Robertson et al., 2009) and embedding-based techniques. Initially, documents were retrieved using traditional models such as TF-IDF and BM25. The relevance of these documents was then refined by integrating proximity-based scores, enabling accurate ranking. A grid search was employed to optimize the weighting of proximity measures during the re-ranking process, re-

[1]https://www.ethnologue.com/language/urd/

sulting in a more effective document retrieval approach for Urdu. By incorporating proximity measures, the system addresses the limitations of traditional term-matching models and improving the ranking of relevant documents.

The remainder of this paper is structured as follows: Section 2 provides a related work, Section 3 outlines the methodology, Section 4 presents the results, and Section 5 concludes the paper.

## 2 Related Work

This section presents a brief description of the previous research on Urdu document retrieval and the impact of various algorithms on retrieval performance. Traditional approaches to document retrieval, such as Boolean retrieval and vector space models , while effective in specific contexts, often fail to capture deeper semantic relationships within text (Aronson et al., 1994) (Dang et al., 2024). Several studies have addressed these limitations by introducing more advanced techniques such as semantic distance measures, and query expansion techniques (Jiang et al., 2019). However, as evident from the literature review, the impact of distance measures on Urdu document retrieval remains largely unexplored (Daud et al., 2017). Although distance measures are fundamental in determining how documents are compared and ranked in response to user queries, directly influencing the accuracy and relevance of retrieved results. (Riaz, 2008) aimed to establish a baseline for Urdu IR by creating a test reference collection for Urdu. The study followed the TREC methodology (Harman, 1993) and evaluated models such as Boolean retrieval and the Vector Space Model (VSM). This work highlighted the need for specialized test collections for Urdu IR to improve evaluation performance. (Rasheed and Banka, 2018)investigated the impact of query expansion techniques for improving information retrieval (IR) in the Urdu language. The study emphasized that the inherent morphological complexity of Urdu and its scarcity of linguistic resources make traditional IR methods less effective. To address this, the authors explored different query expansion techniques to enhance the retrieval of relevant documents. (Rasheed et al., 2021b) evaluated different models for query expansion in Urdu IR, such as Pseudo-Relevance Feedback (PRF) and Automatic Query Expansion. They showed significant improvements in retrieval precision using models like KL, Bo1, and Bo2, but also emphasized the challenges posed by Urdu's linguistic complexities. (Rasheed et al., 2021a) discussed the development of an Urdu test collection based on TREC guidelines. They emphasized the importance of proximity-based methods, especially when combined with BM25, in enhancing retrieval effectiveness in low-resource languages. (Shaukat et al., 2022) developed a comprehensive benchmark for evaluating information retrieval systems in Urdu using TREC guidelines. The study introduced proximity-based models to improve retrieval performance by incorporating non-binary relevance judgments across a large collection of Urdu news documents. This work underscored the need for robust test collections that go beyond binary relevance measures, which are essential for addressing the challenges posed by Urdu's complex linguistic structure. (Shoaib et al., 2023)presented a Context-Aware Urdu Information Retrieval System aimed at improving the precision and recall of search results in Urdu. This system addresses challenges unique to the Urdu language, such as word sense ambiguity (WSA), stemming, and complex morphology, by leveraging Web Semantic Search Engine (WSSE) technologies. The authors developed an ontology-based retrieval system that uses quad formats rather than triplets, incorporating subject, object, predicate, and context to better handle ambiguity in queries. While these studies represent significant advancements in Urdu document retrieval, a notable gap remains in evaluating the impact of distance measures on document retrieval performance. Although techniques like query expansion and semantic distance have been explored, a comprehensive analysis of how various distance metrics enhance document retrieval for Urdu has yet to be conducted (Asim et al., 2019). This research aims to address that gap by optimizing distance measures within established models such as BM25, TFIDF and embedding-based techniques to improve retrieval effectiveness for Urdu.

## 3 Methdology

This section outlines the approach undertaken in developing the Urdu Proximity Enhanced Retrieval Framework (UPERF), which incorporates proximity measures into traditional and embedding-based models for Urdu document retrieval. The methodology is divided into several stages:

- Data preprocessing
- Traditional retrieval
- Embedding generation
- Enhanced score calculation
- Document re-ranking
- Evaluation

### 3.1 Data preprocessing

The input data includes both Urdu documents and a user query. Before proceeding with the retrieval, the data undergoes a preprocessing stage where URLs, non-Urdu alphabets, punctuation marks, and diacritics are removed to ensure clean text. We used the Stanza[2] library from Stanford NLP for tokenization, stopword removal, stemming, and lemmatization to normalize the text. This process reduces words to their base forms and ensures uniformity in document representation.

### 3.2 Traditional Retrieval

After preprocessing, we constructed feature matrices using unigrams, bigrams, and trigrams to capture various levels of n-gram information, essential for handling multi-word queries effectively. The documents were then transformed into vectors using the Term Frequency-Inverse Document Frequency (TF-IDF) metric. The TF-IDF metric is computed as:

$$\text{TF-IDF}(t, d) = \text{TF}(t, d) \cdot \log\left(\frac{N}{\text{DF}(t)}\right) \quad (1)$$

where:

- t is the term(unigram, bigram, or trigram),
- d is the document,
- N is the total number of documents, and
- DF(t) is the number of documents containing the term

Additionally, we calculated BM25 scores, which are based on a probabilistic model by considering document length and term saturation. The BM25 score is calculated as:

$$\text{BM25}(q, d) = \sum_{i=1}^{n} \log\left(\frac{N - \text{DF}(t_i) + 0.5}{\text{DF}(t_i) + 0.5}\right) \cdot \frac{(k_1 + 1) \cdot \text{TF}(t_i, d)}{k_1 \cdot (1 - b + b \cdot \frac{|d|}{\text{avgdl}}) + \text{TF}(t_i, d)} \quad (2)$$

where:

- $q$ is the query,
- $d$ is the document,
- $k_1$ and $b$ are BM25 parameters, and
- avgdl is the average document length.

These two scores TF-IDF and BM25;are used to identify relevant documents in the first stage of retrieval.

### 3.3 Embedding Generation

To enhance retrieval beyond traditional scoring, we generated embeddings for the documents and queries using Word2Vec, FastText, and doc2vec models trained on a large Urdu corpus. Additionally, we used a pre-trained mBERT model[3] to generate contextual embeddings. These embeddings capture the semantic meaning of the words. For generating document embeddings, we applied TF-IDF Weighted Averaging to the word embeddings within each document, giving more importance to words with higher TF-IDF scores. We trained these embeddings on the UND collection (Shaukat et al., 2022) to further finetune them for Urdu document retrieval.

### 3.4 Proximity Score Calculation

Proximity measures are calculated to determine the semantic similarity between the query embeddings and document embeddings. The following proximity measures were used

[2]https://stanfordnlp.github.io/stanza/

[3]https://huggingface.co/google-bert/bert-base-multilingual-uncased

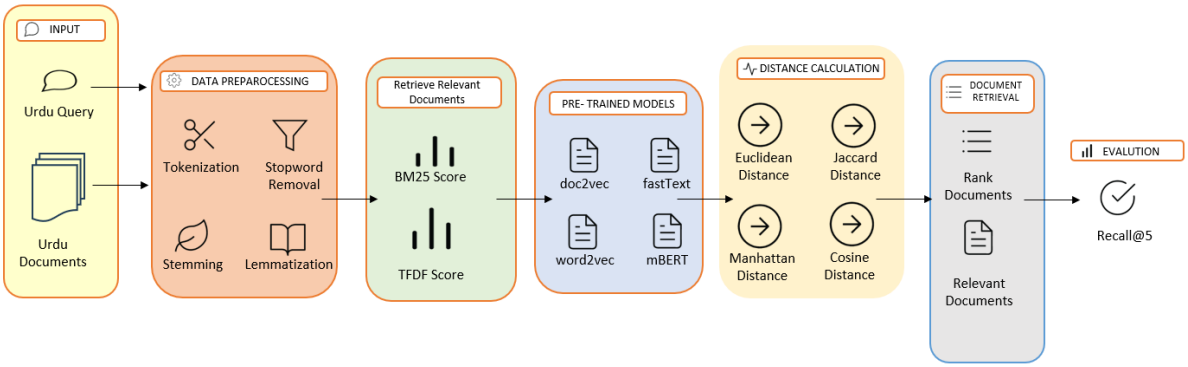


Figure 1: Urdu Proximity Enhanced Retrieval Framework (UPERF)

- Euclidean Distance

$$d_{\text{Euclidean}}(A, B) = \sqrt{\sum_{i=1}^{n}(A_i - B_i)^2} \quad (2)$$

where 'A' and 'B' are the embedding vectors of the query and document, respectively.

- Manhattan Distance

$$d_{\text{Manhattan}}(A, B) = \sum_{i=1}^{n}|A_i - B_i| \quad (3)$$

- Cosine Distance (based on Cosine Similarity)

$$\text{Cosine Similarity}(A, B) = \frac{A \cdot B}{||A|| \cdot ||B||} \quad (4)$$

$$\text{Cosine Distance} = 1 - \text{Cosine Similarity} \quad (5)$$

### 3.5 Weighted Combination of Scores

The final relevance score for each document is calculated by combining the traditional retrieval scores (TF-IDF and BM25) with the proximity-based scores. The combination is done using a weighted formula:

$$\text{Final Score} = \alpha \cdot \text{TF-IDF} + \beta \cdot \text{BM25} + \gamma \cdot \text{Proximity Score} \quad (6)$$

$\alpha$, $\beta$, and $\gamma$ the weights assigned to the TF-IDF, BM25, and Proximity Scores, respectively. These weights are optimized using grid Search to find the best combination that maximizes retrieval performance. The grid search tests multiple values of$\alpha$, $\beta$, and $\gamma$ and selects the combination that yields the highest performance metric.

### 3.6 Document Re-Ranking

Once the final score is computed, the documents are re-ranked based on their final relevance score. The higher the final score, the more relevant the document is considered to the query, and it is placed higher in the resultant ranking.

## 4 Experiments and Results

This section first introduces the dataset used in this study, followed by a detailed description of the series of experiments conducted.

### 4.1 Dataset

The dataset used in this research is the Urdu News Document (UND) corpus (Shaukat et al., 2022), consisting of 2,887,169 news articles collected from 11 newspapers, covering topics such as law, government, sports, and international relations as shown in Table 1. The documents were scraped, cleaned, and converted to TREC-Standard SGML format, including fields like document ID, title, publication date, and full text. The corpus was further processed for relevance judgments using IR techniques such as BM25, TF-IDF, and Boolean similarity, making it suitable for evaluating retrieval techniques.

### 4.2 Evaluation Metrics

For the evaluation of our retrieval framework, we utilize Recall@5 to measure performance.

**Recall@5** is a metric used to evaluate how well the system retrieves relevant documents within the top 5 results. It is defined as

Table 1: Urdu News Document (UND) Corpus Overview

| Aspect | Description |
|---|---|
| Dataset Name | UND Corpus |
| Documents | 2,887,169 |
| Source | 11 newspapers |
| Topics | Law, govt, sports, etc. |
| Format | TREC-standard SGML |
| Fields | ID, Title, Date, Text |
| Queries | 105 queries (35 base, 3 variants each) |
| Judgments | Highly Relevant<br>Fairly Relevant<br>Marginally Relevant<br>Irrelevant |
| IR Methods | BM25, TF-IDF, Boolean |
| Purpose | Retrieval, proximity, embeddings |

the fraction of relevant documents that are retrieved among the top 5 documents returned by the system. The formula for Recall@5 is as follows:

$$\text{Recall@5} = \frac{\text{Relevant documents in top 5}}{\text{Total relevant documents}} \tag{7}$$

### 4.3 Experimental Setup

Our study followed a systematic approach to evaluate various retrieval models and their configurations. To ensure comprehensive assessment, we designed multiple experimental combinations, testing different aspects of the retrieval process ranging from embedding dimensions to similarity measures. Table 2 presents the complete experimental framework, where each component (A through F) represents a different aspect of our evaluation setup. From component A through F, we performed multiple experimental combinations by systematically varying each parameter. These combinations were derived from: 3 embedding models with varying dimensions (100, 150, 200) and window sizes (3, 5, 7), 2 traditional models, 1 contextual model, 3 preprocessing variants, 2 query types, and 4 similarity measures. All experiments were conducted on the UND corpus and evaluated using Recall@5.

Table 2: Experimental Configurations and Parameters

| Exp ID | Component | Parameters |
|---|---|---|
| A | Embedding Models | |
| | A.1 Word2Vec | Vector size: {100, 150, 200} |
| | A.2 FastText | Window size: {3, 5, 7} |
| | A.3 Doc2Vec | |
| B | Traditional Models | |
| | B.1 TF-IDF | N-grams: {bigram, trigram} |
| | B.2 BM25 | Default parameters |
| C | Contextual Model | |
| | C.1 mBERT | Pre-trained weights |
| D | Preprocessing | |
| | D.1 Raw Text | No modification |
| | D.2 Stemmed | Root form reduction |
| | D.3 Lemmatized | Canonical form |
| E | Query Types | |
| | E.1 Single-word | One word per query |
| | E.2 Multiple-word | Multiple words per query |
| F | Similarity Measures | |
| | F.1 Cosine | Angular similarity |
| | F.2 Euclidean | L2 distance |
| | F.3 Manhattan | L1 distance |
| | F.4 Jaccard | Set overlap |

### 4.4 Performance Analysis

The experimental results, presented in Table 3, demonstrate varying performance across different models, preprocessing techniques, and query types. Word2Vec (vs150_ws3) emerged as the best performing model, achieving a peak Recall@5 of 0.85 with stemmed text and multiple-word queries. This performance can be attributed to its ability to effectively capture semantic relationships in the Urdu text. The mBERT model showed strong performance with multiple-word queries (0.79 with stemmed text) but struggled with single-word queries (0.33), indicating its dependence on contextual information. Among traditional approaches, TF-IDF(Trigram) demonstrated moderate performance (0.76 for stemmed text, multiple-word queries), while BM25 achieved lower scores (0.36). FastText, despite its subword-level processing capability, peaked at 0.74, not surpassing Word2Vec's performance. Doc2Vec consistently underperformed, reaching only 0.23 at its highest, suggesting limitations in document-level embedding for fine-grained retrieval tasks. Across all models, two consistent patterns emerged: stemmed text outperformed both raw and lemmatized preprocessing, and multiple-word queries consistently yielded better results than single-word queries. This suggests that reducing morphological variations while maintaining adequate context is crucial for effective Urdu document

retrieval.

Table 4: Comprehensive Weight Distribution Analysis using Word2Vec (vs150_ws3) with Cosine similarity for proximity score. All scores normalized to [0,1] range before combining.

| Config. Type | Weights ($\alpha$,$\beta$,$\gamma$) | Recall@5 |
|---|---|---|
| *Individual Baselines:* | | |
| TF-IDF only | (1.0, 0.0, 0.0) | 0.45 |
| BM25 only | (0.0, 1.0, 0.0) | 0.48 |
| Proximity only | (0.0, 0.0, 1.0) | 0.85 |
| *TF-IDF Enhanced:* | | |
| Heavy TF-IDF | (0.8, 0.1, 0.1) | 0.50 |
| Moderate TF-IDF | (0.6, 0.2, 0.2) | 0.55 |
| Light TF-IDF | (0.4, 0.3, 0.3) | 0.75 |
| *BM25 Enhanced:* | | |
| Heavy BM25 | (0.1, 0.8, 0.1) | 0.52 |
| Moderate BM25 | (0.2, 0.6, 0.2) | 0.58 |
| Light BM25 | (0.3, 0.4, 0.3) | 0.73 |
| *Proximity Enhanced:* | | |
| Heavy Prox. | (0.1, 0.1, 0.8) | 0.82 |
| Moderate Prox. | (0.2, 0.2, 0.6) | 0.80 |
| Light Prox. | (0.3, 0.3, 0.4) | 0.78 |
| *Balanced:* | | |
| Equal weights | 0.33, 0.33, 0.34 | 0.72 |

### 4.5 Weighted Distribution Analysis

The comprehensive analysis of our weighted combination formula reveals interesting patterns across different weight distributions. As shown in Table 4, we first established baselines with individual components: TF-IDF (0.45), BM25 (0.48), and Word2Vec proximity with cosine similarity (0.85). The weight variations demonstrate that heavily emphasizing a single component (0.8 weight) generally underperforms balanced approaches. TF-IDF emphasis shows gradual improvement as weights become more balanced, from 0.50 (heavy) to 0.75 (light emphasis). Similar patterns emerge with BM25 emphasis, improving from 0.52 to 0.73. Notably, proximity-based configurations consistently outperform pure lexical approaches. Even with heavy proximity emphasis (0.8 weight), the system maintains strong performance (0.82), though slightly below the pure proximity baseline (0.85). This suggests that while embedding-based similarity is crucial, some contribution from traditional retrieval methods helps maintain robust performance. The balanced configuration (0.33, 0.33, 0.34) achieves 0.72, indicating that equal weighting of components may not be optimal. The best performing combination maintains a slight emphasis on proximity while balancing traditional approaches.

### 4.6 Document Re-Ranking Analysis

To evaluate the practical effectiveness of different ranking methods, we analyzed two representative queries from distinct domains in the UND corpus. Table 5 presents a comparison of rankings across different approaches against human-judged ground truth.

For the sports domain query "پاکستان اور بھارت کا میچ" (Pakistan-India Match), we observe varying ranking behaviors. Traditional methods (TF-IDF+BM25) prioritized term matching, placing document 2362784 (fairly relevant) first, while relegating the highly relevant document 2368653 to fourth position - likely due to exact matches of terms "بھارت" and "میچ". The Word2Vec approach demonstrated better semantic understanding by ranking the highly relevant document first, though with some inconsistencies in subsequent rankings. The combined weighted approach ($\alpha$=0.3, $\beta$=0.3, $\gamma$=0.4) shows interesting trade-offs. While it ranked a fairly relevant document (2378593) ahead of the highly relevant one, it maintained better overall relevance distribution in subsequent positions. This suggests the weighting scheme helps balance lexical and semantic signals, though not perfectly replicating human judgment patterns. A similar pattern emerges for the medical domain query "ڈاکٹروں کی ہڑتال" (Doctors Strike). Each method shows distinct ranking behaviors, with the combined approach demonstrating improved but imperfect ranking. The placement of document 2392190 (fairly relevant) before 2817668 (highly relevant) indicates that even weighted combinations of different retrieval signals may prioritize documents differently than human assessors.

These results demonstrate that while our weighted framework improves upon individual approaches, but future work could explore more sophisticated techniques such as dynamic weighting schemes, contextual relevance

Table 3: Document Retrieval Results on UND Dataset for Various Query Types and Preprocessing Techniques

| Model | Preprocessing | Query Type | Cosine | Euclidean | Jaccard | Manhattan |
|---|---|---|---|---|---|---|
| **Word2Vec (vs150_ws3)** | Raw Text | Multiple Word | 0.82 | 0.80 | 0.75 | 0.81 |
| | | Single Word | 0.40 | 0.39 | 0.35 | 0.40 |
| | **Stemmed Text** | **Multiple Word** | **0.85** | 0.83 | 0.79 | 0.84 |
| | | Single Word | 0.42 | 0.41 | 0.37 | 0.42 |
| | Lemmatized Text | Multiple Word | 0.83 | 0.81 | 0.76 | 0.82 |
| | | Single Word | 0.41 | 0.40 | 0.36 | 0.41 |
| **mBERT** | Raw Text | Multiple Word | 0.78 | 0.72 | 0.74 | 0.76 |
| | | Single Word | 0.33 | 0.30 | 0.31 | 0.32 |
| | Stemmed Text | Multiple Word | 0.79 | 0.73 | 0.75 | 0.77 |
| | | Single Word | 0.34 | 0.31 | 0.32 | 0.33 |
| | Lemmatized Text | Multiple Word | 0.78 | 0.72 | 0.74 | 0.76 |
| | | Single Word | 0.33 | 0.30 | 0.31 | 0.32 |
| **TF-IDF (trigram)** | Raw Text | Multiple Word | 0.75 | 0.75 | 0.56 | 0.75 |
| | | Single Word | 0.36 | 0.36 | 0.26 | 0.36 |
| | Stemmed Text | Multiple Word | 0.76 | 0.76 | 0.57 | 0.76 |
| | | Single Word | 0.37 | 0.37 | 0.27 | 0.37 |
| | Lemmatized Text | Multiple Word | 0.75 | 0.75 | 0.56 | 0.75 |
| | | Single Word | 0.36 | 0.36 | 0.26 | 0.36 |
| **FastText (vs200_ws7)** | Raw Text | Multiple Word | 0.73 | 0.72 | 0.68 | 0.73 |
| | | Single Word | 0.35 | 0.34 | 0.32 | 0.35 |
| | Stemmed Text | Multiple Word | 0.74 | 0.73 | 0.69 | 0.74 |
| | | Single Word | 0.36 | 0.35 | 0.33 | 0.36 |
| | Lemmatized Text | Multiple Word | 0.73 | 0.72 | 0.68 | 0.73 |
| | | Single Word | 0.35 | 0.34 | 0.32 | 0.35 |
| **Doc2Vec (vs150_ws3)** | Raw Text | Multiple Word | 0.22 | 0.21 | 0.19 | 0.20 |
| | | Single Word | 0.11 | 0.10 | 0.09 | 0.10 |
| | Stemmed Text | Multiple Word | 0.23 | 0.22 | 0.20 | 0.21 |
| | | Single Word | 0.12 | 0.11 | 0.10 | 0.11 |
| | Lemmatized Text | Multiple Word | 0.22 | 0.21 | 0.19 | 0.20 |
| | | Single Word | 0.11 | 0.10 | 0.09 | 0.10 |
| **BM25** | Raw Text | Multiple Word | 0.35 | 0.33 | 0.28 | 0.34 |
| | | Single Word | 0.16 | 0.14 | 0.12 | 0.15 |
| | Stemmed Text | Multiple Word | 0.36 | 0.34 | 0.29 | 0.35 |
| | | Single Word | 0.17 | 0.15 | 0.13 | 0.16 |
| | Lemmatized Text | Multiple Word | 0.35 | 0.33 | 0.28 | 0.34 |
| | | Single Word | 0.16 | 0.14 | 0.12 | 0.15 |

modeling, or learning-to-rank approaches to better align automated rankings with human relevance assessments. The current framework establishes a foundation for developing such advanced retrieval mechanisms for the Urdu language.

## 5 Conclusion

This study presents UPERF, a comprehensive framework for Urdu document retrieval that effectively bridges traditional and modern approaches. Our experimental results across multiple models and configurations demonstrate

Table 5: Ranking Analysis Across Different Methods with Ground Truth Comparison

| **Query** | **Ground Truth (Top 5)** | **TF-IDF+BM25 (Top 5)** | **Word2Vec (Top 5)** | **Combined (Top 5)** |
|---|---|---|---|---|
| پاکستان اور بھارت کا میچ (Pakistan-India Match) | 2368653[HR], 2362784[FR], 2378593[FR], 2008560[MR], 556316[IR] | 2362784[FR], 2008560[MR], 556316[IR], 2368653[HR], 2378593[FR] | 2368653[HR], 2378593[FR], 2362784[FR], 2008560[MR], 556316[IR] | 2378593[FR], 2368653[HR], 2362784[FR], 2008560[MR], 2367037[MR] |
| ڈاکٹروں کی ہڑتال (Doctors Strike) | 2817668[HR], 2392190[FR], 2373033[FR], 2367037[MR], 2367590[MR] | 2392190[FR], 2817668[HR], 2367037[MR], 2373033[FR], 2367590[MR] | 2817668[HR], 2373033[FR], 2392190[FR], 2367590[MR], 2367037[MR] | 2392190[FR], 2817668[HR], 2367590[MR], 2373033[FR], 2367037[MR] |

HR: Highly Relevant, FR: Fairly Relevant, MR: Marginally Relevant, IR: Irrelevant

several key findings: (1) embedding-based proximity measures, particularly Word2Vec with cosine similarity, significantly outperform traditional term-matching approaches, (2) stemmed text preprocessing consistently yields better results across all models, and (3) our weighted combination approach achieves better alignment with human relevance judgments compared to individual methods.The framework's effectiveness is particularly evident in the re-ranking analysis, where it successfully maintains the proper ordering of documents based on relevance levels while balancing both lexical and semantic matching. This is crucial for practical applications where retrieval accuracy directly impacts user experience. Future work could explore integration with newer transformer architectures like BERT and RoBERTa, fine-tuned specifically for Urdu. Additionally, incorporating query expansion techniques and pseudo-relevance feedback could further enhance retrieval performance for complex and ambiguous queries. These developments, combined with UPERF's strong foundation, hold promise for advancing information retrieval capabilities in low-resource languages.

## References


Alan R Aronson, Thomas C Rindflesch, and Allen C Browne. 1994. Exploiting a large thesaurus for information retrieval. In *RIAO*, volume 94, pages 197–216.

Muhammad Nabeel Asim, Muhammad Wasim, Muhammad Usman Ghani Khan, Nasir Mahmood, and Waqar Mahmood. 2019. The use of ontology in retrieval: a study on textual, multilingual, and multimedia retrieval. *IEEE Access*, 7:21662–21686.

Michel Beigbeder and Annabelle Mercier. 2005. An information retrieval model using the fuzzy proximity degree of term occurences. In *Proceedings of the 2005 ACM symposium on Applied computing*, pages 1018–1022.

Edward Kai Fung Dang, Robert Wing Pong Luk, and Qing Li. 2024. A study of word bigrams for pseudo-relevance feedback in information retrieval. *Journal of Universal Computer Science*, 30(11):1511.

Ali Daud, Wahab Khan, and Dunren Che. 2017. Urdu language processing: a survey. *Artificial Intelligence Review*, 47:279–311.

Donna Harman. 1993. Overview of the first trec conference. In *Proceedings of the 16th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 36–47.

Muntaha Iqbal, Bilal Tahir, and Muhammad Amir Mehmood. 2021. Cure: Collection for urdu information retrieval evaluation and ranking. In *2021 International Conference on Digital Futures and Transformative Technologies (ICoDT2)*, pages 1–6. IEEE.

Jyun-Yu Jiang, Mingyang Zhang, Cheng Li, Michael Bendersky, Nadav Golbandi, and Marc Najork. 2019. Semantic text matching for long-form documents. In *The world wide web conference*, pages 795–806.

Samreen Kazi and Shakeel Khoja. 2021. Uquad1. 0: development of an urdu question answering training data for machine reading comprehension. *arXiv preprint arXiv:2111.01543*.

Samreen Kazi, Shakeel Khoja, and Ali Daud. 2023. A survey of deep learning techniques for machine reading comprehension. *Artificial Intelligence Review*, 56(Suppl 2):2509–2569.

Samreen Kazi and Shakeel Ahmed Khoja. 2024. Context-aware question answering in urdu. In *Proceedings of the 7th International Conference on Natural Language and Speech Processing (ICNLSP 2024)*, pages 233–242.

Alexandre Magueresse, Vincent Carles, and Evan Heetderks. 2020. Low-resource languages: A review of past work and future challenges. *arXiv preprint arXiv:2006.07264*.

Zarmeen Nasim and Sajjad Haider. 2022. Impact of distance measures on urdu document clustering. In *Proceedings of the 36th Pacific Asia Conference on Language, Information and Computation*, pages 183–189.

Erik Novak, Luka Bizjak, Dunja Mladenić, and Marko Grobelnik. 2022. Why is a document relevant? understanding the relevance scores in cross-lingual document retrieval. *Knowledge-Based Systems*, 244:108545.

Imran Rasheed and Haider Banka. 2018. Query expansion in information retrieval for urdu language. In *2018 Fourth International Conference on Information Retrieval and Knowledge Management (CAMP)*, pages 1–6. IEEE.

Imran Rasheed, Haider Banka, and Hamaid M Khan. 2021a. Building a text collection for urdu information retrieval. *ETRI Journal*, 43(5):856–868.

Imran Rasheed, Haider Banka, and Hamaid Mahmood Khan. 2021b. Pseudo-relevance feedback based query expansion using boosting algorithm. *Artificial Intelligence Review*, 54(8):6101–6124.

Yves Rasolofo and Jacques Savoy. 2003. Term proximity scoring for keyword-based retrieval systems. In *European Conference on Information Retrieval*, pages 207–218. Springer.

Kashif Riaz. 2008. Concept search in urdu. In *Proceedings of the 2nd PhD workshop on Information and Knowledge Management*, pages 33–40.

Stephen Robertson, Hugo Zaragoza, et al. 2009. The probabilistic relevance framework: Bm25 and beyond. *Foundations and Trends® in Information Retrieval*, 3(4):333–389.

Saba Shaukat, Asma Shaukat, Khurram Shahzad, and Ali Daud. 2022. Using trec for developing semantic information retrieval benchmark for urdu. *Information Processing & Management*, 59(3):102939.

Umar Shoaib, Laiba Fiaz, Chinmay Chakraborty, and Hafiz Tayyab Rauf. 2023. Context-aware urdu information retrieval system. *Transactions on Asian and Low-Resource Language Information Processing*, 22(3):1–19.