Received: 16 October 2019 | Revised: 16 September 2020 | Accepted: 5 October 2020

DOI: 10.4218/etrij.2019-0458

ORIGINAL ARTICLE

ETRI Journal WILEY

Building a text collection for Urdu information retrieval

Imran Rasheed[1] | **Haider Banka**[1] | **Hamaid M. Khan**[2]

[1]Department of Computer Science and Engineering, Indian Institute of Technology (ISM), Dhanbad, India

[2]Aluteam, Fatih Sultan Mehmet Vakif University, Istanbul, Turkey

Correspondence
Imran Rasheed, Department of Computer Science and Engineering, Indian Institute of Technology (ISM), Dhanbad, India.
Email: imranrasheed@cse.ism.ac.in

Abstract

Urdu is a widely spoken language in the Indian subcontinent with over 300 million speakers worldwide. However, linguistic advancements in Urdu are rare compared to those in other European and Asian languages. Therefore, by following Text Retrieval Conference standards, we attempted to construct an extensive text collection of 85 304 documents from diverse categories covering over 52 topics with relevance judgment sets at 100 pool depth. We also present several applications to demonstrate the effectiveness of our collection. Although this collection is primarily intended for text retrieval, it can also be used for named entity recognition, text summarization, and other linguistic applications with suitable modifications. Ours is the most extensive existing collection for the Urdu language, and it will be freely available for future research and academic education.

KEYWORDS

Assessors agreement, relevance judgment, text collection construction and evaluation, Urdu corpus, Urdu information retrieval

1 | INTRODUCTION

Urdu belongs to the Perso-Arabic cluster of languages [1] and is mainly composed of words from Arabic, Persian, and Sanskrit. It is the national language of Pakistan and has over 300 million speakers spread worldwide, with a large portion of this population residing in the Indian subcontinent [2,3]. Urdu was initially derived from the Perso-Arabic script of Iran, is written from right to left like Arabic or Persian, and is characterized by the Nasta`liq format [4,5]. The family tree of Urdu traces back to a mixture of Indo-European, Indo-Iranian, and Indo-Aryan lingo evolution [6]. Urdu is known to have a rich and complex morphology [7,8] and its syntax structure is composed of a combination of Persian, Sanskrit, English, Turkish, and Arabic structures.

Research on information retrieval (IR) prior to the 1990s was relatively limited and immature. This is because only limited resources and data collections were available for evaluation. Experimentation based on new algorithms and techniques for various IR and natural language processing (NLP) tasks, as well as the development of language tools, requires benchmark collections. Worldwide, most text processing related research occurs through evaluation-based consortiums such as the Text Retrieval Conference (TREC),[1] which is co-sponsored by the National Institute of Standards and Technology[2] and the US Department of Defense. TREC was started in 1992 as part of the TIPSTER Text Program. Its goal was to provide a basis for research within the IR community by providing the infrastructure necessary for the large-scale evaluation of text retrieval methodologies. The TREC text

[1]https://trec.nist.gov/ Last visited: 28-01-2020.

[2]https://www.nist.gov/ Last visited: 28-01-2020.

collection mainly consists of a set of news documents. The TREC ideology was adopted in other initiatives such as the Conference and Labs of the Evaluation Forum, which was formerly known as the Cross-Language Evaluation Forum.[3] The NII Testbeds and Community for Information Access Research[4] and the Forum for Information Retrieval Evaluation (FIRE)[5] provide benchmark test collections and offer platforms for participating and engaging in various text processing tasks. They also provide service for languages from different geographic regions. Together, these forums have developed text collections for English and several other European and Asian languages. Western languages have many resources from the IR perspective, whereas Urdu lags significantly in terms of available resources. TREC emphasized test collections in English during its initial phases, but eventually included monolingual and cross-lingual retrieval activities for other European and Asian languages. However, the availability of several advanced benchmark collections for these languages through evaluation-based consortiums was critical for their development. Such collections are not available for the Urdu language. Based on the lack of resources and dedicated consortiums for Urdu, research on Urdu has largely been limited to domain-specific or task-based research. For example, some interesting works have considered Urdu for basic operations such as stemming, lemmatization, chunking, information extraction, and NER [2,9,10]. Data have been collected from email, tweets, or written and spoken words from various news agencies by different organizations and research groups [5,11–17]. However, for standard text searching and various retrieval operations, a large collection of data covering a wide range of topics of general interest are required. The corpus presented here is intended to fill the gap in IR research on Urdu for experimentation and evaluation on a scale that has never been attempted before.

The main contributions of this study can be summarized as follows.

- We construct a text collection for the Urdu language consisting of 85 000 documents covering 52 topics of interest.
- Various retrieval models supported by Terrier are empirically compared and the best model is identified for the Urdu language.
- The performance of well-known classifiers is evaluated using the Urdu language.

The remainder of this paper is organized as follows. We discuss the methodology used for the construction of our corpus in Section 2. Section 3 provides a detailed description of the segments of our collection. Experimental results and discussion are presented in Section 4. Section 5 presents an example application of IR for Urdu. Our conclusions are provided in Section 6.

2 | METHODOLOGY

A standard benchmark collection is essential for any type of language research. We attempted to construct an Urdu benchmark for various linguistic studies and text processing tasks, such as ad hoc IR, text summarization, text clustering, categorization, NER, and question answering. A process flow diagram is presented in Figure 1, which shows the four stages of the development of our text collection for Urdu.

2.1 | Data conversion

Our initial corpus was 640 GB in size and contained a mixture of data in various file formats, including InPage files (*.inp*), as well as Photoshop, Word, Excel, PDF, and image files. These files were collected over a span of four years. From this corpus, 553.5 MB of files were extracted in the *.inp* format. A non-Unicode InPage file[6] consists of multiple news items from different domains. To transform news items into readable text format, they are converted into a standard TREC format using the following steps. First, all individual news items are extracted from the InPage file and saved into another InPage file without images. This file is then transformed into the UTF-8 format so it can be read by any text editor. Next, the obtained Unicode file is split into a number of text files according to the number of news items. A unique number <DOCNO> is assigned to each file as a combination of the date of publishing, category type, and sequence number, allowing files to be differentiated. Finally, each file is converted into a standard TREC format by assigning an opening <DOC> tag at the beginning and a closing </DOC>tag at the end. The <TEXT> tag is used to enclose the news content. A sample document in the standard TREC format is presented in Figure 2.

2.2 | Preprocessing

In this module, various methods are applied to reform the collected text. For example, operations such as tokenization

[3]http://clef2018.clef-initiative.eu// Last visited: 28-01-2020.

[4]www.research.nii.ac.jp/ntcir/index-en.html Last visited: 28-01-2020.

[5]http://fire.irsi.res.in/fire/2019/home Last visited: 28-01-2020.

[6]InPage is a standard text editor for creating pages in Urdu newspapers, books, and magazines using the power of the Nasta'liq style of Arabic script.

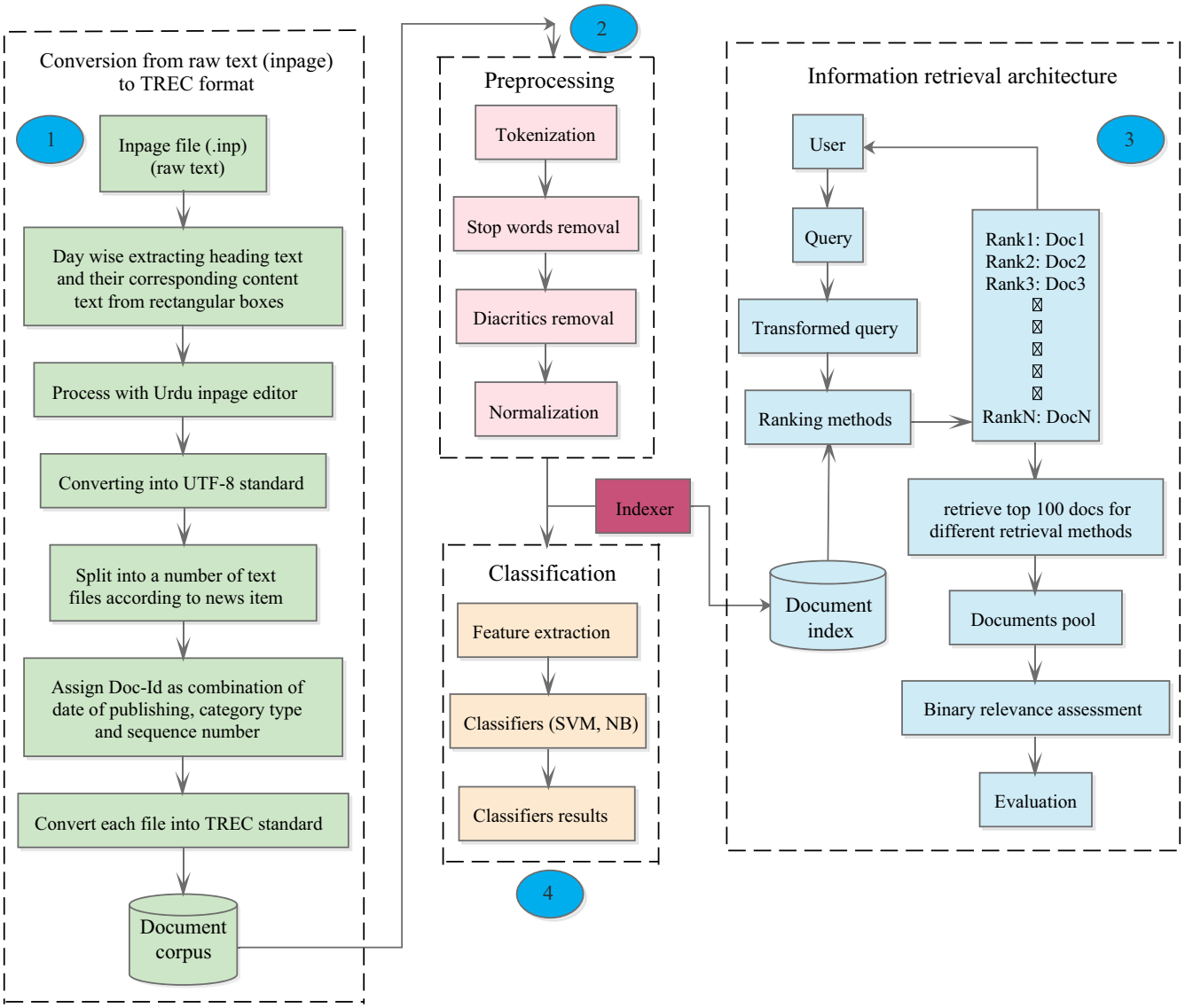


FIGURE 1 Process flow diagram showing the stages of the development of our Urdu text collection

[18], stopword removal [10,19], normalization, and diacritics removal [20,21] are applied to remove extraneous text.

2.2.1 | Tokenization

Tokenization is the process of splitting strings in a given document into words known as tokens by using a tokenizer to read delimiters such as /-"[]():?<>! Characters [22].

2.2.2 | Normalization

Some characters in the Urdu alphabet have more than one Unicode because they belong to two languages, such as Urdu and Arabic. Such characters should be replaced using the alternative Urdu alphabet to prevent the creation of multiple copies of a word.

2.2.3 | Diacritic removal

Diacritics are non-functional terms that are used to ease text reading. In principle, they are used to add significance to text to make it more meaningful to readers. However, diacritics can add too much significance to text, particularly when they appear in isolation, because many beginners end up making mistakes when reading Urdu text. During the preprocessing stage, all diacritics are removed to homogenize the text [23].

2.2.4 | Stopword removal

Stopwords are the most common terms in any language. These words are used to complete the structure of a sentence, but are not individually meaningful. They are

<DOC>
<DOCNO>26_July_2012_Sportz4</DOCNO>
<TITLE>
ٹیسٹ رینکنگ :ہاشم آملہ تیسرے نمبر پر آگئے
Test rankings: Hashim Amla reached third place
</TITLE>
<TEXT>
جنوبی افریقا کے جیک کیلس اور ہاشم آملہ آئی سی سی ٹیسٹ رینکنگ میں دوسرے اور تیسرے نمبر پر آگئے ہیں۔ عالمی رینکنگ میں چھ میں سے چار ٹاپ پوزیشن جنوبی افریقا کے بیٹسمینوں کے پاس ہے۔ سری لنکا کے کمار سنگاکارا بدستور ٹاپ پر ہیں۔ جیک کیلس دوسرے، ہاشم آملہ ٹرپل سنچری اسکور کرنے کے بعد تیسرے، چندر پال چوتھے، ای بی ڈی ویلیئرز پانچویں اور گریم اسمتھ چھٹے نمبر پر ہیں۔ آسٹریلوی مائیکل کلارک کا نمبر ساتواں ہے۔ پاکستان کے یونس خان نویں، اظہر علی گیارہویں اور مصباح الحق تیرہویں نمبر پر ہیں۔ وہ آل راؤنڈرز میں بنگلہ دیشی شکیب الحسن کی جگہ ٹاپ پر آگئے ہیں۔ بولنگ میں ڈیل اسٹین 896 پوائنٹس کے ساتھ ٹاپ پر ہیں۔ پاکستانی آف اسپنر سعید اجمل کا نمبر دوسرا ہے۔ لیفٹ آرم اسپنر عبدالرحمن دسویں نمبر پر ہیں۔

South Africa's Jacques Kallis and Hashim Amla have come second and third in the ICC Test rankings. South African batsmen have four top positions in the World Ranking. Sri Lanka's Kumar Sangakkara is on top. Jacques Kallis is second, Hashim Amla occupy the third position after scored tripple century, Chandra Paul fourth, AB de Villiers fifth and Graeme Smith sixth position. Australian Michael Clarke is the seventh position. Pakistan's Younis Khan is the ninth, Azhar Ali eleven and Misbah-ul-Haq are in the thirteenth position. He is on the top of the Bangladeshi Shakib-ul-Hasan in All-rounders list. In the Bowling, Dale Steyn is on top with 896 points. Pakistan's spinner Saeed Ajmal is the second. Left Arm Spinner Abdul Rahman is in the tenth number.
</TEXT>
</DOC>

FIGURE 2 Sample file in the TREC format

removed from the text target before applying any algorithm to reduce the size of the vocabulary. Stopwords are considered as a negative list, and they do not contribute to the indexing process.

2.3 | Classification

Text classification refers to the process of dividing documents into various categories based on their content and titles. Text classification is used extensively in many application domains, such as document categorization, medical diagnosis, and image processing, as well as by many different communities such as the database, data mining, machine learning, and IR communities. We discuss this aspect in greater detail in Section 5.

2.4 | Retrieval architecture

Our architecture can be broadly classified into three basic IR processes, namely the representation of user information needs, document content, and comparisons, as shown in Figure 1. Text is further indexed for the document representation of addressed queries. Indexing is the process of document representation [24]. A retrieved document is ranked based on its relevance, and then the top-*k* ranked documents are selected for pooling and significance assessment.

3 | CORPUS CONSTRUCTION

Because the availability of a standard benchmark collection is a prerequisite for research in the domain of IR, it is important to establish a standard benchmark collection for any language that provides great value to research in the domain of IR. However, to the best of our knowledge, such a collection does not exist for the Urdu language. In this work, we followed the TREC specifications for developing a standard text collection [25,26]. Our collection consists of data from various news sources in India. Daily Roshni,[7] which is a widely published news source, contributed raw data from over four years in proprietary file formats for text and images, as shown in Figure 3.

[7]http://www.dailyroshni.net/ Last visited: 25-04-2019.

FIGURE 3 Sample format of an InPage document

TABLE 1 Document analysis at different levels of preprocessing

Category	#Docs	#Tokens	#Types
Articles	2385	2 833 680	65 794
Opinion	194	284 449	18 093
Sports	7710	1 783 486	39 523
MuslimWorld	10 261	2 764 127	48 379
Health	3233	1 491 193	32 906
UPNews	2053	445 930	20 949
International	6710	1 922 930	43 863
Culture	1916	909 927	22 880
Economical	436	181 991	10 401
Social	3279	1 458 940	40 981
Political	8316	4 490 603	54 147
Entertainment	3778	736 442	25 888
National	10 477	2 766 881	50 479
LocalNews	21 481	5 900 216	65 067
ScienceTechnology	568	228 431	13 115
Miscellaneous	2507	347 881	20 884

Note: "Token" refers to the total number of words in a text.
"Type" refers to the number of distinct words in a text.

TABLE 2 Length distribution of the collected documents

Document Length (words)	Number of documents
9 to 508	67 782
509 to 1008	13 904
1009 to 1508	2520
1509 to 2008	623
2009 to 2508	263
2509 to 3008	125
3009 to 3508	34
3509 to 4008	28
4009 to 4508	11
4509 to 5008	7
5009 to 5508	4
5509 to 6008	3

3.1 | Document collection

The encoding schemes in original articles are non-standard and font based. Therefore, documents were converted into UTF-8 encoding to homogenize the corpus. The final collection of 85 304 articles was prepared using a wide range of categories with varying sizes. A summary of the documents in each category and the length distributions of the documents are presented in Tables 1 and 2, respectively, while Table 3 outlines some attributes of the collection.

3.2 | Comparative study of datasets

This section presents a comparative study of the proposed collection and existing datasets. The following reports have

TABLE 3 Attributes of the Urdu collection

Attributes	Values
Source Name	Daily Roshni
Source URL	www.dailyroshni.net
Time-Period	15 April 2010 to 26 December 2014
Collection size (MB)	255.20
Document format	Text
No. of documents in the corpus	85 304
Total no. of terms	28 547 107
No. of unique terms[a]	198 365
Average length of documents (words)	345.37
No. of categories	16
No. of topics	52
Encoding Type	UTF-8

[a]Unique terms refer to the number of distinct words in a text or collection.

provided information regarding constructing corpora in Urdu and related languages. There are only a few collections available, which have only limited access permissions. The corpus developed by Becker-Riaz only contains 7000 Web news documents [9], but it can be used for making relevance judgments and queries. Ijaz and Husain [15] reported an Urdu text corpus containing 18 million words, including 104 341 unique words. Their corpus was extracted from two news websites and categorized into several domains, including finance, culture, and science. However, their corpora are inaccessible for public use based on license constraints. Additionally, Urooj et al. [11] developed a corpus of 100K Urdu words known as the CLE Urdu Digest corpus, which was extracted from Urdu Digest magazine. The data were divided into 13 known sections for data categorization and annotated with part-of-speech (POS) tags. This Urdu corpus is available for public access, but is insufficient in size to represent a language model for various types of data testing. Therefore, a large Urdu text collection consisting of several domains is required to represent a standard language model ideally. Table 4 provides a brief summary of existing Urdu collections.

3.3 | Topics

First, a set of 148 topics was extracted from our Urdu corpus based on the manual observation of published news items. Two topics were taken from FIRE ad hoc queries for Hindi because we had data from an overlapping period. Each topic consists of natural language statements based on user requirements [27]. A typical topic contains four sections of *title*, *description*, *narrative*, and unique identification number <topic-id>for documentation.

For the formulation of a topic, queries were developed and narratives were observed. Topics were further modulated based on the number of documents such that topics with more than 80 or less than 10 retrieved documents were discarded for being too simple (that is number of relevant documents appearing in pooling is greater than 80 out of 1000) or too complex (that is number of relevant documents appearing in pooling is less than 10 out of 1000), respectively, for analysis.

As a result, only 52 topics were retained for final evaluation. The formed queries were also delineated as a mixture of short (title field), medium (title and description fields), and long queries (composed of title, description, and narration fields). The average length of the 52 topic titles is 6.61 terms, with minimum and maximum lengths of 3 and 14 terms, respectively. An example developed topic is presented in Figure 4 and the length distributions of the queries are presented in Figure 5.

3.4 | Pooling and judgment

The significance of any IR process is dependent on user satisfaction, which can be measured using either precision[8] or recall.[9] Recall is computationally expensive for large collections, and precision requires nominal feedback from experts. Therefore, relevance judgment was accomplished using a technique called pooling. Here, a pool based on the TREC specifications [25] was prepared for the set of 52 topics with eight different models, including language modeling, vector-space modeling [28], and divergence from randomness (DFR) [28] (for example, BM25, In_expB2, PL2, InL2, DFR_BM25, In_expC2, Dirichlet, and term frequency inverse document retrieval (TF-IDF)).

4 | RESULTS AND DISCUSSION

4.1 | Corpus evaluation

The top 100 ranked documents in each topic (out of 1000 documents) were collected for relevance assessment using the Terrier[10] IR platform [29]. A total of 41 600 relevance judgments were performed for 52 topics using all eight models, including 11 035 unique documents.[11] Finally, 1223 relevant documents were retrieved based on this process. Their

[8]$\text{Precision} = \frac{\text{Number of relevant items retrieved}}{\text{Number of retrieved items}}$

[9]$\text{Recall} = \frac{\text{Number of relevant items retrieved}}{\text{Number of relevant items in colection}}$

[10]http://www.terrier.org/ Last visited: 28-01-2020.

[11]The contents of these documents may be exactly the same, but exist with different names in a collection.

TABLE 4 List of previously developed collections for the Urdu language

Corpus name	Year	#Docs	#Topics	#Categories	#words	Encoding	Task
Our proposed corpus (ROSHNI)	2019	85 304	52	Sixteen	29 461 732	Unicode	Urdu text retrieval, stemming, stopword removal, text categorization, text summarization, and other NLP tasks
CLPD-UE-19 Corpus [41]	2019	2398 (Total) 1588 (Plagiarized) 810 (Non-Plagiarized)	-	Ten	-	Unicode	Cross-lingual plagiarism detection for Urdu-English language pairs
Urdu Language Corpus [42]	2019	-	-	Nine	3 084 039	Unicode	Computational linguistic and statistical analysis
RUCD [12]	2016	-	-	-	103 566 (Total) 71 437 (Training) 32 129 (Testing)	Unicode	Development of text editors for Romanized Urdu and other high-level computational resources
Naive collection [40]	2015	5000	-	Four	64 563	Unicode	Text categorization
CLE (erstwhile CRULP) [11]	2012	348	-	Six	100 000	Unicode	Computational linguistic, POS tagging, NER
IJCNLP [43]	2008	-	-	-	40 000	Unicode	5 documents (training) 1 document (testing)
EMILLE [2]	2003	300	-	Six	1 640 000 (Urdu written) 512 000 (Urdu spoken)	Unicode	NER, translation, anaphoric annotation, POS tagging, language engineering analysis
Becker-Riaz [9]	2002	7000	-	-	20 000–50 000	Unicode	NER and IR

<topic>
<topid>10</topid>
<title>
آدرش ہاؤسنگ سوسائٹی گھوٹالے استعفی
Adarsh housing society scam resigns
</title>
<desc>
آدرش ہاؤسنگ سوسائٹی گھوٹالے کے بعد وزیر اعلی استعفی
Chief Minister resigns after Adarsh housing society scam
</desc>
<narr>
مہاراشٹر کے وزیر اعلی اشوک چاوان کا استعفی سے متعلقہ دستاویزات ہونے چاہئے۔ دستاویزات کو آدرش ہاؤسنگ سوسائٹی گھوٹالے کے بعد وزیر اعلی استعفی میں ملوث ہونے سے متعلق معلومات پر مشتمل ہونا چاہئے۔ گھوٹالے کے دیگر پہلوؤں کے بارے میں خبروں کے مضامین (دیگر گرفتاریوں / استعفی وغیرہ وغیرہ) غیر متعلقہ ہیں

Relevant documents should contain information related to Maharashtra Chief Minister Ashok Chavan's resignation due to his involvement in the Adarsh Housing Society scam. News articles about other aspects of the scam (other arrests / resignations etc.) are irrelevant.
</narr>
</topic>

FIGURE 4 Sample topic (Topic 2, ``Adarsh Housing Scams'')

statistics are presented in Table 5. For each of the 52 queries, the numbers of relevant documents retrieved by all models in the pool are presented in Figure 6.

The pool generated in this manner was then exhaustively judged by experts for binary relevance [30]. Experts with at least a Master's degree in Urdu were asked to distinguish relevant documents. To obtain an agreement between experts, a Kappa[12] (k) statistic was applied with a coefficient of 0.6 considered as acceptable and a coefficient of 0.8 considered as good on the relevance scale [31]. Table 6 lists the resulting Kappa values, which indicate satisfactory agreement among the judged pairs. The query-wise distributions of documents for different retrieval models are presented in Figure 7. The observed minimum and maximum numbers of retrieved documents for all retrieval models are similar, and there are only small differences in their quartile deviations. The lower and upper parts of the bar slices represent the values of the first (Q1) and third (Q3) quartiles, respectively, and the midlines of the bar slices represent the median of the document distribution for each model.

4.2 | Experimental results

To analyze the effectiveness of the retrieval models, all experiments were performed within the Terrier platform. Terrier is a modular platform for the rapid development of wide-ranging IR applications with indexing, retrieval, and evaluation of English and non-English documents. For all of our experiments, we used the Terrier tool for document retrieval. During indexing, only single terms from documents and queries, excluding phrase terms, were pre-indexed. To make the terms consistent and normalized, diacritics were removed and transcoded uniformly, as suggested byAkram and others [21]. The Assas-Band stemmer was used for stemming, while tokenizing was performed using a tool developed by Humayoun and others [20]. For all of our experiments, the parameters were set to Terrier's default values. Terrier supports a large variety of weighting models. We adopted the following weighting models:

1. Probabilistic Model [32]
   a. BM25: One of the best-known term-weighting schemes. BM stands for best match. This method accounts for three components: the term frequency, IDF, and the length of the document.
2. DFR Models [33]
   a. InL2: An IDF model with Laplace after-effects and normalization 2. This model can be used for tasks that require initial precision.
   b. PL2: A Posisson model with Laplace after-effects and normalization 2. This model can be used for tasks that require initial precision. PL2 is one of the DFR weighting models.

[12] $\kappa = \frac{P(A) - P(E)}{1 - P(E)}$

FIGURE 5 Length distributions of the queries

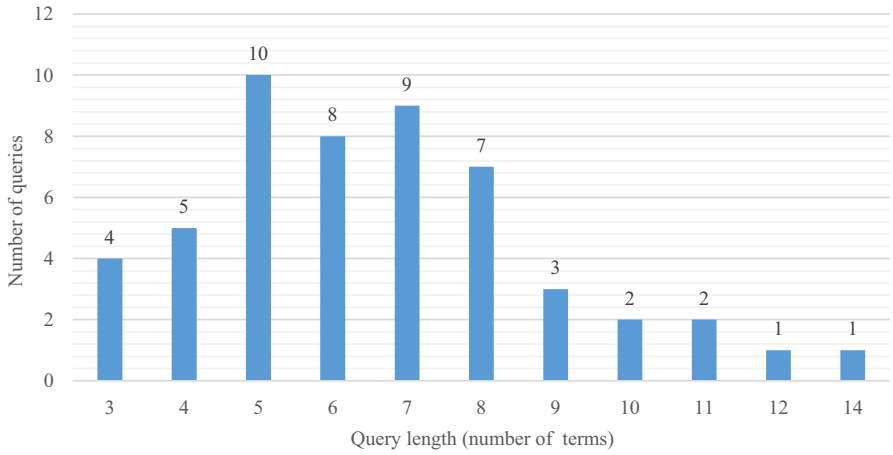


TABLE 5 Statistics of pooled documents

Attributes	Values
Number of topics	52
Pool-depth	100
Number of pooled docs	41 600
Number of relevant docs	1223
Average pooled docs per query	800
Average relevant docs per query	23.51
Minimum reldocs per query	11
Maximum reldocs per query	43

c. DFR_BM25 (DFR): The DFR version of BM25.
d. In-expB2: Inverse expected document frequency model with Bernoulli after-effects and normalization 2. The logarithms are base 2. This model can be used for classic ad hoc tasks.
e. In-expC2: Inverse expected document frequency model with Bernoulli after-effects and normalization 2. The logarithms are base e. This model can be used for classic ad hoc tasks.

3. Language Model
   a. Dirichlet: Language modeling with Bayesian smoothing and a Dirichlet prior.
4. Vector space model [34]
   a. TF-IDF: This weight is a statistical measure for evaluating the importance of a word in a selected document within a corpus. Importance increases proportionally to the number of times a word appears in a document, but is offset by the frequency of the word in the corpus.

It is difficult to judge that the performance of retrieval model A is better than that of retrieval model B based on a single query. The information process for any text collection must be quantified in terms of overall effectiveness [35–37]. Here, the performances of all eight retrieval systems on 52 queries based on binary relevance were measured using the TREC_eval[13] program, as shown in Table 7 and Figure 8.

[13]www.trec.nist.gov/trec_eval/ Last visited: 28-01-2020.

TABLE 6 Pairwise agreement among three experts for the evaluation of relevance

Judges pair	Agreement k (%)
Experts A, B	79.04%
Experts B, C	84.81%
Experts A, C	92.42%
Mean	85.42%

To measure the performance of retrieval models, calculations were performed for 11 recall levels (0.0, 0.1, 0.2, ..., 1.0) and the precision values were interpolated[14] at each point for all queries in the evaluation benchmark. The average scores for all 52 queries were mapped to the mean average precision (MAP) values of the retrieval models.

According to Table 7, at a recall of 0, TF-IDF yields the highest precision score among all models, buts its precision is reduced when the recall value is greater than 0.3, falling behind BM25. The MAP value of TF-IDF is the highest among all of the models. This means that the TF-IDF model provides good overall performance on the Urdu collection.

Precision values P@k were computed at different document threshold values k (5, 10, 15, 20, 30, 100) to assess the entire system. One can see that there are significant differences between all models at P@5 in Figure 8, whereas the precision values are approximately equal at other values of k (10, 15, 20, 30 and 100). However, BM25 exhibits greater efficiency at every value of k.

5 | APPLICATIONS OF CLASSIFICATION

A massive rise in digital information on the Internet and demand for search engine optimization has put increased pressure on developers to improve user experiences. We prepared a

[14]$P_{\text{inter}}(r) = max_{r' \geq r} P(r')$

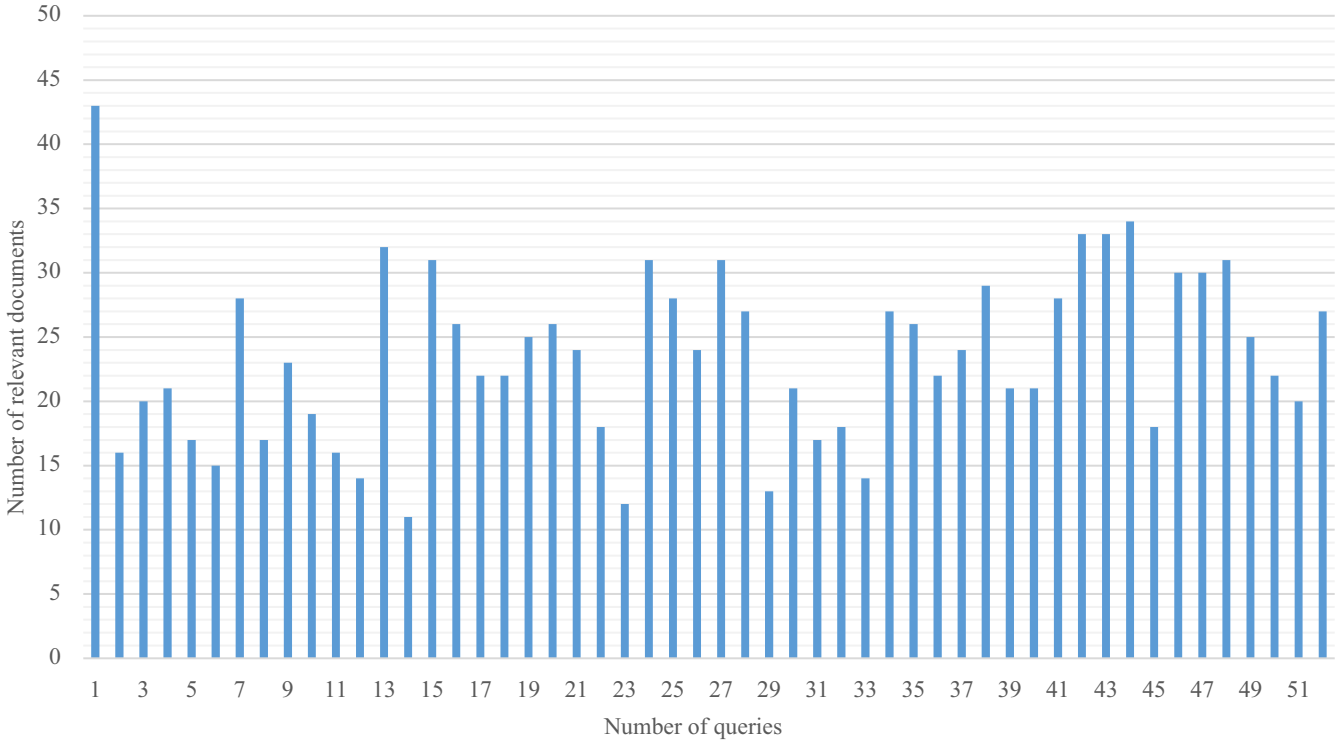


FIGURE 6 Numbers of retrieved relevant documents per query in the pool

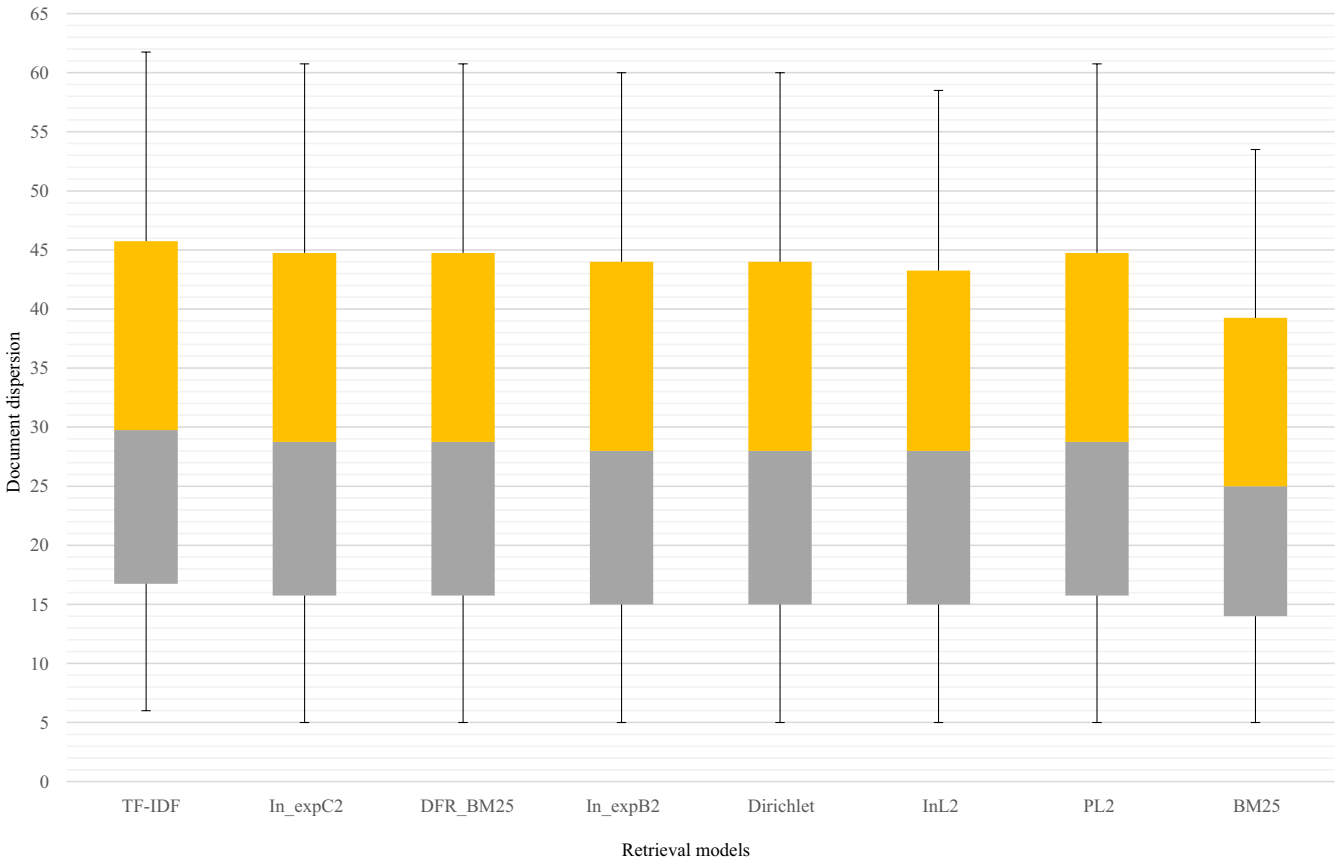


FIGURE 7 Query-based document dispersion for the eight retrieval models

TABLE 7 Precision and recall values for the eight retrieval systems

Recall	InL2	DFR_BM25	Dirichlet	In_expB2	PL2	TF_IDF	BM25	In_expC2
0.00	0.5212	0.6152	0.5719	0.5149	0.6081	0.6810	0.6187	0.6011
0.10	0.4692	0.5122	0.4865	0.4720	0.5020	0.5613	0.5401	0.5179
0.20	0.3860	0.4145	0.3952	0.3899	0.4134	0.4745	0.4312	0.4242
0.30	0.3678	0.3875	0.3755	0.3688	0.3830	0.3998	0.4007	0.3864
0.40	0.3415	0.3549	0.3384	0.3432	0.3531	0.3636	0.3652	0.3571
0.50	0.3096	0.3187	0.3109	0.3118	0.3149	0.3143	0.3315	0.3205
0.60	0.2725	0.2799	0.2682	0.2728	0.2768	0.2674	0.2876	0.2794
0.70	0.2321	0.2341	0.2229	0.2308	0.2305	0.2288	0.2358	0.2312
0.80	0.1979	0.2034	0.1949	0.1971	0.2019	0.2004	0.2058	0.2013
0.90	0.1705	0.1760	0.1695	0.1714	0.1734	0.1712	0.1811	0.1737
1.00	0.1496	0.1538	0.1480	0.1499	0.1514	0.1602	0.1593	0.1519
MAP	0.2859	0.3064	0.2920	0.2852	0.3033	0.3203	0.3162	0.3052

FIGURE 8 Precisions at document thresholds for the eight retrieval systems

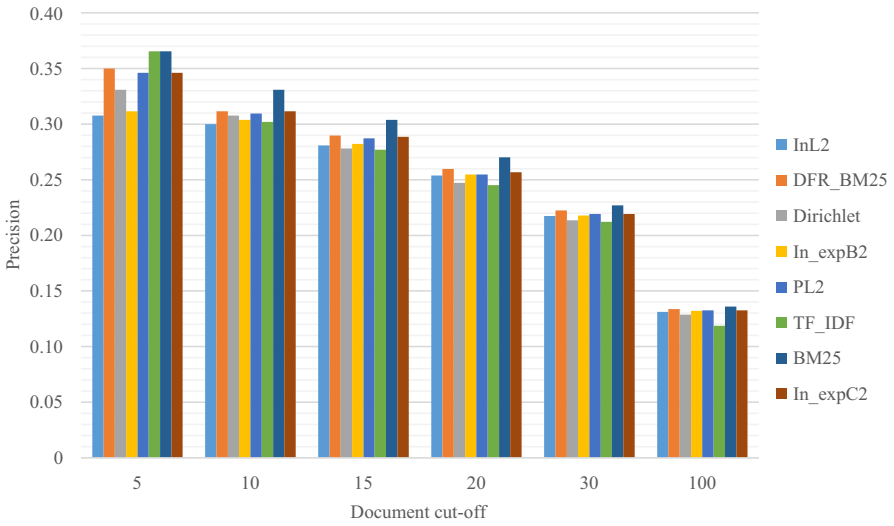


TABLE 8 Performances of different classifiers on our Urdu collection

10 Cross fold	SVM(%)	NB(%)	DT(%)	KNN-1(%)	KNN-3(%)	KNN-5(%)
100	67.90	48.70	61.00	46.30	48.00	50.00
200	71.30	52.60	63.00	47.20	48.60	49.90
300	74.30	53.90	65.70	48.60	50.10	51.10

standard benchmark collection for Urdu to simplify various essential operations such as IR, data mining, and NLP to streamline the text categorization of various data types for medical diagnosis, image processing, text filtering, and many other applications. In contrast to Latin script, text categorization for Urdu is a more difficult task based on its complex morphology and challenges related to space insertions or omissions.

Several machine learning algorithms are available based on supervised or unsupervised techniques. Several classification methods have been successfully tested using different software systems and are freely available, such as Mallet [38] and WEKA [39]. Additionally, WEKA also supports built-in methods such as tokenization, stopword removal, feature selection, and feature weighting. To assign weights to each term, we adopted TF-IDF weighting schemes, but the baseline feature selection results were not satisfactory. Therefore, we reduced the weights of less important features by applying the information gain (IG) method to select the best features and improve the efficiency of our classifiers. Four different supervised learning classifiers were tested for text classification, while the IG method was used for feature selection, as suggested by the authors of [40]. We performed k-fold cross-validation to divide the given data into k equal groups. We utilized a value of k = 10. This means that we mixed the data and then split the data into 10 groups

to evaluate the efficiency of each machine learning classifier. Accuracy was measured for the top-100, top-200, and top-300 selected features, and the results were compared for final analysis. The results in Table 8 reveal that the best performance is obtained by the support vector machine (SVM) classifier, followed by the decision tree (DT), regardless of the number of selected features. The other classifiers (that is, naive Bayes (NB) and k-nearest neighbors (KNN)) perform relatively poorly for different numbers of selected features.

6 | CONCLUSIONS AND FUTURE ENHANCEMENTS

Experimental research on any language is strongly dependent on the availability of linguistic resources, primarily large text collections, which have been largely unavailable for the Urdu language, significantly slowing its progress compared to other advanced languages. In this article, we described the construction of an Urdu text collection that can be used for evaluating various Urdu text processing activities. The proposed text collection for Urdu is expected to facilitate much-needed progress. Our text collection includes 16 different types of categories covering politics, sociology, sports, etc. This collection was constructed according to the TREC standard and tested using different standard retrieval models and techniques for relevance assessment. The results were very promising. Our collection can be used for many machine learning applications and various NLP tasks. The entire collection is freely available for academic research. We hope that the availability of this resource will help make Urdu information retrieval an exciting and productive field.

ACKNOWLEDGMENTS

We sincerely thank Mr Zahoor Ahmad Shora, who is the Chief Editor of "Daily Roshni," for his generous contribution of freely sharing raw data for our collection. We are also thankful to the students and scholars of the Department of Urdu and Linguistics, Aligarh Muslim University, Aligarh, who helped generate topics and evaluate relevance for our text collection.

ORCID

Imran Rasheed https://orcid.org/0000-0001-9550-1294

REFERENCES

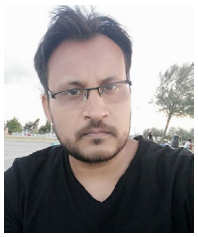
1. A. Hardie, *Developing a tagset for automated part-of-speech tagging in Urdu*, in Proc. Corpus Linguistics (Lancaster, UK), Mar. 2003.
2. P. Baker et al., *Corpus data for south asian language processing*, in Proc. Workshop South Asian Lang. Process. (EACL), (Budapest, Hungary), Apr. 2003, pp. 1–8.
3. K. Riaz, *Baseline for Urdu IR evaluation*, in Proc. ACM workshop Improving non english web searching (Napa Valley, CA, USA), Oct. 2008, pp. 97–100.
4. A. Daud, W. Khan, and D. Che, *Urdu language processing: A survey*, Artif Intell Rev **47** (2017), 279–311.
5. M. Sharjeel, R. M. A. Nawab, and P. Rayson, *Counter: Corpus of urdu news text reuse*, Lang. Res. Eval. **51** (2017), 777–803.
6. M. Humayoun, H. Hammarström, and A. Ranta, *Urdu morphology, orthography and lexicon extraction*, M.S. thesis, Department of Computer Science and Engineering, Chalmers tekniska högskola, Göteborg, Sweden, 2006.
7. V. Gupta, N. Joshi, and I. Mathur, *Design & development of rule based inflectional and derivational Urdu stemmer,* in Proc. Int. Conf. Futuristic Trends Comput. Anal. Knowl. Manag. (ABLAZE), (Greater Noida, India), Feb. 2015, pp. 7–12.
8. I. Rasheed, H. Banka, and H. M. Khan, *Pseudo-relevance feedback based query expansion using boosting algorithm*, Artif. Intell. Rev. (2021), https://doi.org/10.1007/s10462-021-09972-4.
9. D. Becker, and K. Riaz, *A study in Urdu corpus construction*, in Proc. Workshop Asian Lang. Resour. Int. Stand. vol. 12, (Stroudsburg, PA, USA), Aug. 2002, pp. 1–5.
10. K. Riaz, *Concept search in Urdu*, in Proc. PhD workshop Inf. Knowl. Manag. (Napa Valley, CA, USA), Oct. 2008, pp. 33–40.
11. S. Urooj et al., *Cle Urdu digest corpus*, in Proc. Conf. Lang. Technol. (SNLP), (Lahore, Pakistan), (2012), pp. 53–59.
12. F. Baseer, A. Habib, and J. Ashraf, *Romanized Urdu corpus development (rucd) model: Edit-distance based most frequent unique unigram extraction approach using real-time interactive dataset*, in Proc. Int. Conf. Innov. Comput. Technol. (INTECH), (Dublin, Ireland), Aug. 2016, pp. 513–518.
13. S. A. Ali et al., *Salience analysis of news corpus using heuristic approach in Urdu language*, Int. J. Comput. Sci. Netw. Secur. (IJCSNS), **16** (2016), no. 4, 28–36.
14. Q. Abbas, *Building a hierarchical annotated corpus of Urdu: The Urdu. kon-tb treebank*, in International Conference on Intelligent Text Processing and Computational Linguistics, Springer, Berlin, Germany, 2012, pp. 66–79.
15. M. Ijaz, and S. Hussain, *Corpus based Urdu lexicon development*, in Proc. Conf. Lang. Technol. (CLT07), vol. 73, (Peshawar, Pakistan), Aug. 2007.
16. I. Hanif et al., *Cross-language Urduenglish (clue) text alignment corpus*, in Proc. Working notes CLEF (Toulouse, France), Sept. 2015.
17. R. Rahimi, A. Shakery, and I. King, *Extracting translations from comparable corpora for cross-language information retrieval using the language modeling framework*, Inf Process Manage **52** (2016), no. 2, 299-318.
18. M. Karthikeyan, and P. Aruna, *Probability based document clustering and image clustering using content-based image retrieval*, Appl. Soft Comp. **13** (2013), no. 2, 959-966.
19. Z. Ahmad et al., *Urdu nastaleeq optical character recognition*, World Acad. Sci., Eng. Technol. **26** (2007), pp. 249–252.
20. M. Humayoun et al., *Urdu summary corpus*, in Proc. Int. Conf. Lang. Resour. Eval. (Reykjavik, Iceland), May 2014, pp. 796–800. https://github.com/humsha/USCorpus
21. Q. A. Akram, A. Naseer, and S. Hussain, Assasband, an affix-exception-list based Urdu stemmer, in Proc. Workshop Asian Lang. Resour. (Suntec, Singapore), Aug. 2009, pp. 40–47.
22. I. Rasheed and H. Banka, *Query expansion in information retrieval for Urdu language*, in Proc. Int. Conf. Inf. Retr. Knowl. Manag. (CAMP), (Kota Kinabalu, Malaysia), Mar. 2018, pp. 171–176.
23. I. Rasheed et al., *Urdu text classification: A comparative study using machine learning techniques*, in Proc. Int. Conf. Digit. Inf. Manag. (ICDIM) (Berlin, Germany), Sept. 2018, pp. 274–278.

24. K. Batri, S. Lakshmi, and B. Sathiyabhama, *Trade-off between the number of index-terms and the information retrieval system`s performance*, Kuwait J. Sci. **44** (2017), no. 4, 49–56.
25. N. Craswell et al., *Overview of the trec-2003 web track*, in Proc. Text Retr. Conf. (TREC), vol. 3, (Gaithersburg, MD, USA), 2002.
26. A. AleAhmad et al., *Hamshahri: A standard persian text collection*, Knowl. Based Syst. **22** (2009), no. 5, 382-387.
27. A. Kanapala and S. Pal, *Test collection for legal ir from online discussion forums*, in Proc. Forum Inf. Retr. Eval. (Bangalore, India), Dec. 2014, pp. 126–129.
28. J. M. Ponte and W. B. Croft, *A language modeling approach to information retrieval*, in Proc Int. ACM SIGIR Conf. Res. Dev. Inf Retr. (Melbourne, Australia), Aug. 1998, pp. 275–281.
29. I. Ounis et al., *Terrier information retrieval platform*, in Advances in Information Retrieval, vol. 3408, Springer, Berlin, Germany, 2005, pp. 517–519.
30. E. M. Voorhees, *Overview of trec 2003*, in Proc. Text Retr. Conf. (TREC), (Gaithersburg, MD, USA), Nov. 2003, pp. 1–13, https://tsapps.nist.gov/publi cation/get_pdf.cfm?pub_id=150467
31. L. Cohen, L. Manion, and K. Morrison, *The ethics of educational and social research*, in Research Methods in Education, 8th ed. Routledge, London, UK, 2013, https://doi.org/10.4324/9780203720967
32. S. E. Robertson et al., *Okapi at trec-4*, in Proc. Text REtrieval Conf. (London, UK), Oct. 1996, pp. 73–96, http://cites eerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.33.3342.
33. G. Amati and C. J. Van Rijsbergen, *Probabilistic models of information retrieval based on measuring the divergence from randomness*, ACM Trans. Inf. Syst. (TOIS), **20** (2002), no. 4, 357–389.
34. G. Salton, A. Wong, and C. S. Yang, *A vector space model for automatic indexing*, Commun. ACM **18** (1975), no. 11, 613–620.
35. C. D. Manning, and H. Schütze, Foundations of Statistical Natural Language Processing, vol. **999**, MIT Press, Cambridge, MA, USA, 1999, https://nlp.stanford.edu/fsnlp/.
36. P. Clough and M. Sanderson, *Evaluating the performance of information retrieval systems using test collections*, Inf. Res. **18** (2013), no. 2.
37. W. B. Croft, D. Metzler, and T. Strohmann, Search Engines: Information retrieval in practice, Pearson Education, Boston, MA, USA, 2010.
38. A. K. McCallum, Mallet: A machine learning for language toolkit, 2002, http://mallet.cs.umass.edu/.
39. E. Frank et al., Weka-a machine learning workbench for data mining, in Data mining and knowledge discovery handbook, Springer, Boston, MA, USA, 2009, pp. 1269–1277.
40. T. Zia, M. P. Akhter, and Q. Abbas, *Comparative study of feature selection approaches for Urdu text categorization*, Malaysian J. Comput. Sci. **28** (2015), no. 2, 93–109.
41. I. Haneef et al., *Design and development of a large cross-lingual plagiarism corpus for urdu-english language pair*, Sci. Program. **2019** (2019), 1–11.
42. N. Khan, M. P. Bakht, and R. A. Wagan, *Corpus construction and structure study of Urdu language using empirical laws*, in Proc. Int. Conf. Data Sci. (Karachi, Pakistan), Feb. 2019, pp. 9–14.
43. S. Hussain, *Resources for Urdu language processing*, in Proc. Workshop Asian Lang. Resour. IJCNLP, (Hyderabad, India), Jan. 2008, pp. 99–100, https://www.aclweb.org/anthology/I08-7017.pdf.

AUTHOR BIOGRAPHIES

Imran Rasheed received his MS degree in Computer Applications from the Department of Computer Science, Aligarh Muslim University. He is a senior research fellow at the Department of Computer Science and Engineering of the Indian Institute of Technology (ISM), Dhanbad, India. His research interests include information retrieval, natural language processing, and machine learning.

Haider Banka received his MS degree in Engineering from the University of Calcutta in 2003 and his PhD from Jadavpur University in 2008. He has contributed more than 50 research papers and authored three books. He has acted as a referee for many reputable journals, including IEEE Transactions on Evolutionary Computation, Pattern Recognition, and Information Sciences. He has served as an editorial board member, chairs, and PC member of several national and international conferences. He was awarded an EU-INDIA fellowship from April 2006 to July 2006 from the Department of Computer and Information Science (DISI) of the University of Genova, Italy. He works as an associate professor and HOD in the Department of Computer Science and Engineering of the Indian Institute of Technology (ISM), Dhanbad, India. His current research interests include soft computing, bioinformatics, algorithms, data mining, pattern recognition, combinatorial optimization, and AI.



Hamaid M. Khan received his PhD at Istanbul University, Cerrahpasa, last year. He was appointed as an Assistant Professor at Fatih Sultan Mehmet Vakif University in March, 2019. He received his MS degree in Engineering and Business Management from the King's College in London, UK in 2009 and his BS degree in Mechanical Engineering from Aligarh Muslim University, India in 2006. He is an Assistant Professor at Aluteam, Fatih Sultan Mehmet Vakif University, Istanbul, Turkey. His current research interests include interdisciplinary work in text retrieval and machine learning.