Spectrum of Engineering Sciences
ISSN (e) 3007-3138 (p) 3007-312X
Volume 4, Issue 3, 2026

A TECHNIQUE FOR BRIDGING LANGUAGE BARRIERS TO DEVELOP AN URDU-ENGLISH INFORMATION RETRIEVAL SYSTEM

Irsa Manzoor[1], Ramisha Manzoor[2], Dr. Tauqeer Safdar Malik[3], Muhammad Sajid Maqbool[*4]

[1]*Department of Information Technology, Bahauddin Zakariya University, Multan*
[2]*Center for Advanced Studies in Pure and Applied Mathematics, Bahauddin Zakariya University, Multan*
[3]*Department of Information Technology, Bahauddin Zakariya University, Multan*
[*4]*Department of Computer Science, NFC Institute of Engineering and Technology, Multan*

[1]irsamanzoor6@gmail.com, [2]ramishamanzoor858@gmail.com, [3]drsafdar@bzu.edu.pk,
[*4]sajid.maqbool@nfciet.edu.pk

[*4]ORCID:0000-0001-5583-3550

DOI: https://doi.org/10.5281/zenodo.21789432

Keywords
Cross-Lingual Information Retrieval, Urdu–English Retrieval, Machine Translation, Query Expansion, Cross-Lingual Embeddings, Natural Language Processing, Multilingual Information Access, Semantic Search.

Article History
Received: 13 January 2026
Accepted: 26 February 2026
Published: 12 March 2026

Copyright @Author
Corresponding Author: *
Muhammad Sajid Maqbool

Abstract

Cross-Lingual Information Retrieval (CLIR) aims to enable users to access information written in a language different from that of their search query. In a CLIR environment, a query submitted in one language can retrieve relevant documents written in another language, thereby facilitating multilingual information access. This is achieved by developing retrieval systems capable of matching queries and documents across different languages. In recent years, CLIR has emerged as one of the most challenging research areas in Information Retrieval due to the linguistic diversity, semantic variations, and translation complexities involved. Urdu, the national language of Pakistan, is spoken by approximately 163 million people worldwide. Its rich morphological structure, complex syntax, and large speaker population make Urdu Information Retrieval an important and specialized area of research. Despite the growing demand for multilingual information access, efficient CLIR solutions for the Urdu–English language pair remain limited. To address this gap, this study proposes a hybrid query model designed to improve retrieval effectiveness and optimize performance in a cross-lingual retrieval environment. The CLIR framework proposed above allows retrieving relevant documents both in the cases when the search queries are posed either in Urdu or in English in both monolingual and cross-lingual environments such as Urdu to English and English to Urdu information retrieval tasks. The assessment of the system was carried out using the Text Retrieval Conference (TREC) approach, where standardized methodology exists for testing the performance of IR systems via use of document collections, query topics and relevance assessments. For experimental purposes, the UIR-21 dataset was chosen, which is a collection of Urdu news articles in the style of TREC. In total, 37,000 documents were used in the research. For the purpose of cross-lingual IR, Urdu documents were translated to English using the Google Translate API. The performance of the model was investigated by comparing the effectiveness of the process of retrieving the documents when the query-document matching takes place in both Urdu-URDU and translated English-English languages. The criteria for assessment included precision, recall, and

F1-score. It should be noted that, according to the experimental results, English IRS was characterized by the highest precision, while Urdu IRS showed low precision.

1 Introduction

The diversity of languages is one of the defining attributes of human civilization that reflects the cultural, historical and ethnic identity of peoples all around the world [1]. Although language diversity adds variety to human communication and exchange of information, it creates significant obstacles in the modern age when fast and effective access to information becomes necessary for studying, researching and communicating [2]. Among those obstacles can be singled out the development of information retrieval systems that would help users to overcome language barriers and gain smooth access to information regardless of its language. Urdu and English are two languages which differ substantially but which are widely spoken by many people all over the world. The language of Urdu originates from the sub-continent of India and has a very rich literary tradition and is the national language of Pakistan [3]. In contrast, English operates as a lingua franca and is widely used in academic circles, business, science and international communications. Due to common history, culture and linguistic exchanges, the vocabulary of both languages has much in common despite great differences in writing system, morphology and syntax. Therefore, Urdu-English language pair represents an interesting field of study for CLIR.

There are various issues related to developing an efficient Urdu-English information retrieval system. The problems emerge mainly due to dissimilarities in writing systems, linguistic structure, word usage and semantics representation of the two languages. Solving the problems involves applying computational technologies that will be able to model similarities and differences in the two languages in order to provide effective information retrieval across languages [5]. The present research is intended to examine challenges in relation to Urdu-English cross-lingual information retrieval and develop a new framework for overcoming the language barrier in multilingual information access. Building upon existing progress in CLIR and applying particular linguistic features of Urdu and English, the research will try to make its contribution to the advancement of multilingual retrieval technology. There are some approaches to solving the problems of Urdu-English information retrieval [6]. One of the approaches that has been identified as fundamental one in reducing linguistic difference between the two languages is machine translation. Modern machine translation techniques including Google Neural Machine Translation (GNMT) and transformers-based large language models have considerably enhanced translation accuracy and contextual comprehension [7]. These systems enable the translation of Urdu queries into English, allowing users to retrieve relevant English-language documents while interacting in their native language.

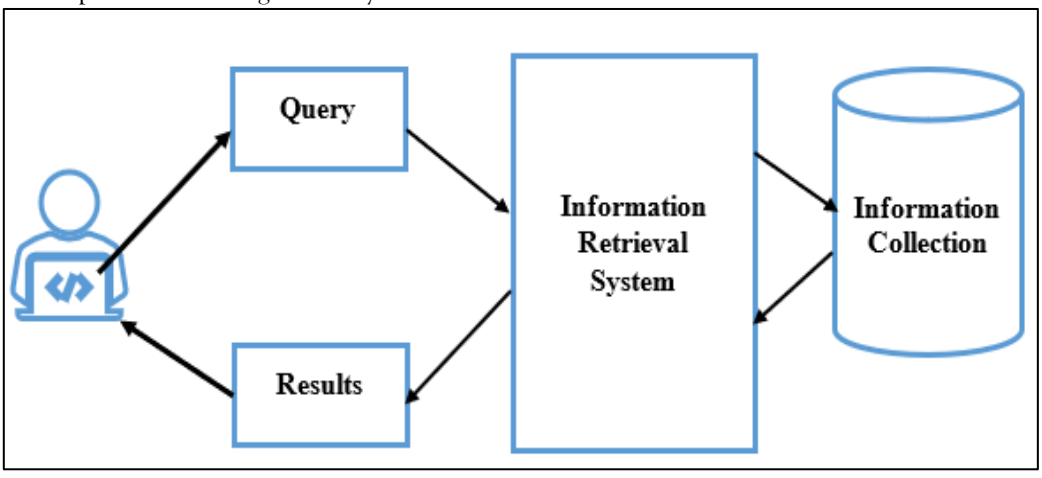

Figure 1 Generic process of IRS

A practical solution is the use of cross-lingual word embeddings and multilingual language models. mBERT and XLM are models that transform words from different languages into one semantic vector space. This common representation makes the process of semantic comparisons across languages possible and helps to retrieve contextually appropriate information despite the language in which the query and the document are written [8]. There are also query expansion methods, which help to enhance the performance of the search by expanding the user query with semantically similar words. In the case of Urdu-English retrieval system, the Urdu query may be expanded by English words and concepts. Thus, there are more chances to retrieve relevant English documents. This increases the recall and retrieval accuracy as well by overcoming the problems of vocabulary mismatch and translation ambiguity. The combination of machine translation, cross-lingual embeddings, and query expansion techniques allows modern Urdu-English information retrieval systems to cope with linguistic problems and provide effective cross-language information retrieval [11]. These systems have a great potential to increase knowledge discovery, multilingual communication, and information exchange. Therefore, the development of reliable Urdu-English CLIR systems is a critical step towards successful information retrieval in multilingual environment [12].

2 Literature Review

The following are the CLIRs and the monolingual IR systems which were studied in the subsequent studies as per the research findings. The well-known Salton's CLIRs formed the first CLIRs and English test collections which were established by their developers. The German version of English queries was transformed into German directly in the year 1970 by a German linguist named Walter Jüttler. Among the current research works, CLIRS is one of the hardest tasks to perform owing to the tremendous developments that have been achieved in information technology, that is, due to the many research papers published on the internet. Although English language is commonly used in the operation of nuclear power plants, there are instances where other languages prevail. As per the agenda of a Knowledge Base (or KB), the idea is to provide accurate and valuable answers to the user's queries [13]. However, the case could pose the questions like in case of the information needed in the user's mother tongue language not being there. For example, in many cases, when a user from the Internet is making searches on the topic "Machine Learning" using Hindi language, the results tend to come comparatively low as well as sparse in comparison to those results which one could get when making the search in the English language. Similarly, another example is when one tries to search for the information related to "Rule of the Hindus in India or Pakistan", then he may get very few results as compared to those results which one could get by translating the query into Hindi [14]. In such a case, the primary purpose of the CLIR process is to obtain relevant and accurate results. In essence, much of what CLIR has done in the past was with the aim of producing multi-lingual retrieval systems in which better translation was involved. Apart from this summary, some research works should also be studied which will be required for further discussion [15].

According to Chayapathi et al. [4] the language in which the question is asked does not have an effect on the precision of the information given as an answer. Some instances will be taken into account where the Background Linguistic Information Retrieval Framework is applied. For example, a patent attorney while carrying out a patent search goes after all the relevant documents no matter what language it may be in since it can also be seen when a journalist goes about seeking for international news, a company providing data analysis seeking business activities in other countries, or an individual searching for information, where he finds a combination of the words specified above. Info Retrieval (IR) refers to the extraction of information from the huge database employing diverse methodologies and structures such as Clumped Unstructured Data. The current status quo entails the presence of complicated networks of varied and disorganized data and did not facilitate any kind of social interaction that exists all over the world. Language is the

most vital barrier which a person faces while interacting with many people. Cross-information barrier problem is solved by CLIR, which means cross-language information retrieval. Retrieval system is being greatly talked about nowadays. The Query Extension (Q.E.) is one of the techniques for introducing a related query. The critical terms are introduced in the initial query in order to enhance the search process of the retrieved documents using letter L, i, R, and E. "What is the use of learning something if you are not going to put it into practice?" has been the theme of my experience with the exploration work. Translating into Hindi-English as well as Kannada-English language combinations, the participants performed the CLIR experiment by finding the English documents using queries in both Hindi and Kannada languages. In Q. E., simple searches and improved queries are used that have been improved by OKAPI scoring in the primary experiment. It has been observed from the experiment that using the appropriate search terms, strategically positioned can be an effective method in improving the precision of the retrieved files without or with the use of the exact positional information, provided that the OKapiBM25 has been used. The entity-based and passage-based methods further enhance the result in both Hindi-English and Kannada-English domains.

In the new method proposed by Bisma Ayaz in 2016, semantic engine has been used for Urdu_e-novels. Semantic engine has been designed in three stages: ontology modeling module, information retrieval, and dynamic ontology update. Domain modeling acts as 'brain' in this case and usually depicts the knowledge domain through domain ontology. Semantic queries and the information collected through ontologies are the next stage which is being used to collect all the required information. Furthermore, the engine is capable of handling changes in the domain ontology due to its adaptive nature. It becomes possible through the engine to derive various fresh quality parameters which might not be possible through deep analysis and generating wrong behavior. In addition to it, the engine provides a special platform for media professionals to exhibit their own work.

A research [5] by Daniel Bekele et al. CLIR is fundamentally based on the assumption that there are people who have the capability of accessing a language other than the language of the query. In translation, it may receive the query in one language and translate it into another depending on the capability. The process involves the use of an application that would solve the problem of the comparison between the query in one language and the documents in another. Afaan Oromo is characterized by these, among many other features, thus making it one of the best-known languages in Ethiopia. Ethiopia. Apart from the fact that Afaan Oromo has the highest number of speakers in the country, not to mention the efforts made to conduct research involving these speakers. This technique allows the learner to interact with the English books written by knowledgeable speakers of Afaan Oromo. This research is aimed at developing a retrieval system of information in the two languages simultaneously to enable Afaan Oromo speakers to have access to universal information databases through 11 queries in their language. Both bilingual and monolingual confirmations are equally significant in the evaluation of the system. Recall and Precision were used as the evaluation parameters of the system.

The method of solving the issue in IR systems has been suggested by Tianchi Bi Sa et al. [6] in 2020. It involves using deep learning approaches like Neural Machine Translation (NMT). First of all, it has to be noted that these kinds of solutions have been proven effective in multiple domains. In contrast, NMT requires sentence pairs from the query to be trained on the vast amounts of out-of-domain data. However, during inference, there is no way for the system to evaluate any conditions that may influence the result of the algorithm. The usage of the deep learning method enables the analysis of the context of the created search index. The search index usually consists of the visual media of the present day. Even in the case of flying machines, there will still be a lack of objects that could be used as downstream delivery. What the researchers propose is the innovative way of solving the problems mentioned above is the broadening of the scope of the general lexical search. Humanization: space becomes viewed as

something relative to the specific number of meaningful terms taken from the search system database. As a result, translation candidates are the translation units of a model that will be trained to learn how to query (Goldberg et al., 1997). Practically, in CLIR, the e-Commerce search engine approach "includes examined and implemented methods" that are functioning. 1 Our approach has shown to be efficient in the field of Neural Machine Translation, despite being applied to the challenging field of Neural Machine Translation. Our accuracy level is basically inexistent in terms of translation and searching. In the research conducted by Maryamah et. al. [8] an article was written describing how Google Translate was modified in accordance with their study. The dictionary of words and word embedding resulted in an innovative approach to Arabic-Indonesian cross-language information retrieval via machine translation. "We get a BLEU score of on average 0. 47 for this approach, which is more than Example Input: As such, it affects the mood and thinking of the characters, ultimately determining the ending of the story. Example Output: It ultimately modulates the outlook and thinking of these characters, which in a way affects the ending of the story. that 0. 43 received from Translation Software."

Thus, the support of the organization for its peers will make the cooperation and interdependence possible in the organization's industry. In this connection, it can be observed that this can serve as an excellent platform for the patients to receive their health records and for the medical professionals to exchange their information. The method is much better than Google Translate since the translation service offered by it is of a much higher quality. One of the other problems is that there is no strategy mentioned in a complete sentence. Knowing the exact saying is a perfect basis for this task since the repeated words are quite predictable. The translator should know the meaning of the terms used in the specific context, and this is the problem. The words' environment within a sentence is going to be one of the factors considered by the learning system when the brain develops. Better understanding of the future adaptability of the different species in the urban environment is the main goal of this research.

3 Methodology

The proposed framework comprised two major components, as illustrated in Figure 2. The system was designed to operate on static corpus collections containing documents in two languages, namely Urdu and English. The model leveraged pre-existing monolingual document collections for both languages to support efficient retrieval in a controlled experimental environment.

The first component of the system allowed monolingual information retrieval. Here, users could use either Urdu or English as query languages, and consequently, the relevant documents would be retrieved in the same language as used in the query. More specifically, for an Urdu query, relevant documents would be retrieved from the Urdu corpus while, for an English query, relevant documents would be retrieved from the English corpus. This first component of the system thus enabled evaluation of baseline performance for the two monolingual retrieval settings. The second component of the proposed framework included the use of Google Translate API to allow cross-lingual information retrieval capabilities. This component allowed translation-based query processing in order to overcome the language barrier in retrieving documents from the two different language corpora. For example, in a case where a user submitted a query in Urdu but sought retrieval of documents in English, the query would first be translated into English using the Google Translate API before it is processed in order to retrieve relevant documents from the English document collection.

Overall, the proposed model effectively supported both monolingual and cross-lingual retrieval scenarios within a unified framework. The integration of a translation module enabled seamless interaction between Urdu and English information spaces, thereby enhancing the accessibility of multilingual information and improving retrieval flexibility for end users.

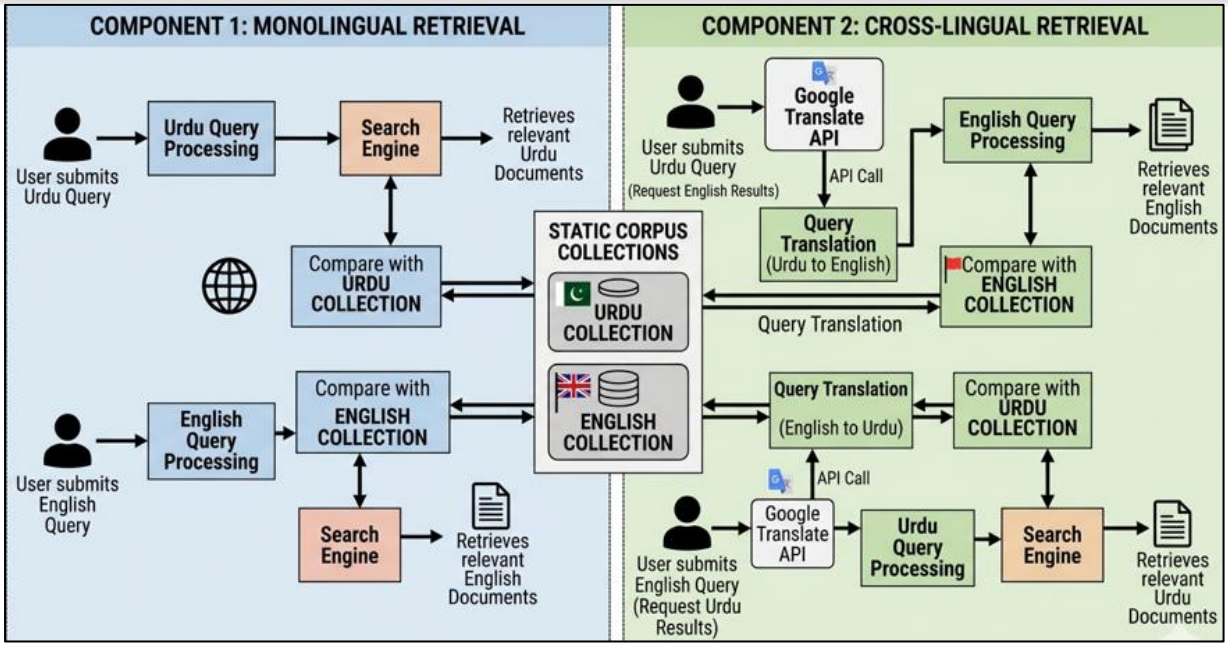

Figure 2 Proposed CLIR Model

3.1 Dataset

The test collection of the corpus UIR-21 consists of three parts: (a) the initial set of Urdu News Documents that acts as a base for searching the required information; (b) a set of Urdu Enquiries that is used to search the relevant material; (c) pairwise relevance assessment. The data used almost satisfies the requirements of TREC concerning the approach. According to the rules of TREC, the data set should contain three components. First, the data set should contain the collection of documents, followed by the query, and the annotation or the relevance assessment at last. The most recently available data set posted online on April 30, 2022, UIR-21, has been selected. It includes 2,887,169 Urdu documents taken from various newspapers. The ARY News data has been used, which consisted of approximately 37,000 Urdu documents. From them, 10,000 Urdu documents have been selected, translated into English through the Google API and 10,000 English documents have been obtained. The representation of the Urdu document before applying the translation API is given in Figure 3. We use only two relevancy levels in our study (Relevant and not relevant) from the Relevance judgment scale in the dataset.

- Irrelevant document: The document has no information related to the topic.
- Marginally relevant document: The document just hints at the topic. There is no additional or more information provided by the document except that related to the topic.
- Fairly relevant document: The document has more information regarding the topic than the topic statement, but the coverage is incomplete. In case of multi-faceted topics, it covers only few subtopics or perspectives.
- Highly relevant document: The document discusses the issues of the topic comprehensively. In case of multi-faceted topics, it covers all the subtopics or perspectives.

3.1.1 Preprocessing

Documents could be converted to English language by use of Google API Translator. Figure 3 illustrates the conversion from Urdu language to English of the document that had been referred to earlier in Urdu.

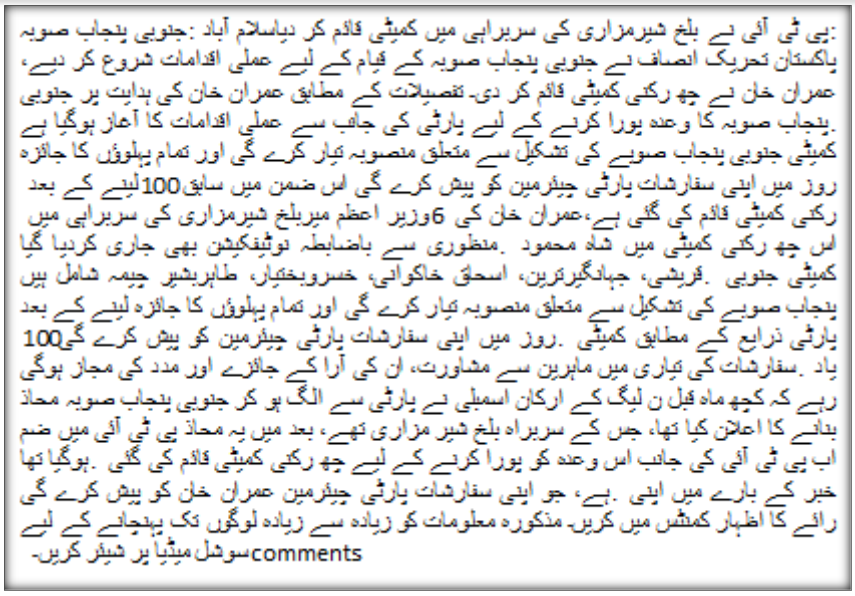
جنوبی پنجاب صوبہ: پی ٹی آئی نے بلخ شیرمزاری کی سربراہی میں کمیٹی قائم کر دیاسلام آباد
پاکستان تحریک انصاف نے جنوبی پنجاب صوبہ کے قیام کے لیے عملی اقدامات شروع کر دیے،
عمران خان نے چھ رکنی کمیٹی قائم کر دی۔ تفصیلات کے مطابق عمران خان کی ہدایت پر جنوبی
پنجاب صوبہ کا وعدہ پورا کرنے کے لیے پارٹی کی جانب سے عملی اقدامات کا آغاز ہوگیا ہے
کمیٹی جنوبی پنجاب صوبے کی تشکیل سے متعلق منصوبہ تیار کرے گی اور تمام پہلوؤں کا جائزہ
روز میں اپنی سفارشات پارٹی چیئرمین کو پیش کرے گی اس ضمن میں سابق100لینے کے بعد
رکنی کمیٹی قائم کی گئی ہے،عمران خان کی 6وزیر اعظم میربلخ شیرمزاری کی سربراہی میں
اس چھ رکنی کمیٹی میں شاہ محمود .منظوری سے باضابطہ نوٹیفکیشن بھی جاری کردیا گیا
کمیٹی جنوبی .قریشی، جہانگیرترین، اسحٰق خاکوانی، خسروبختیار، طاہربشیر چیمہ شامل ہیں
پنجاب صوبے کی تشکیل سے متعلق منصوبہ تیار کرے گی اور تمام پہلوؤں کا جائزہ لینے کے بعد
پارٹی ذرایع کے مطابق کمیٹی .روز میں اپنی سفارشات پارٹی چیئرمین کو پیش کرے گی100
یاد .سفارشات کی تیاری میں ماہرین سے مشاورت، ان کی آرا کے جائزے اور مدد کی مجاز ہوگی
رہے کہ کچھ ماہ قبل ن لیگ کے ارکان اسمبلی نے پارٹی سے الگ ہو کر جنوبی پنجاب صوبہ محاذ
بنانے کا اعلان کیا تھا، جس کے سربراہ بلخ شیر مزاری تھے، بعد میں یہ محاذ پی ٹی آئی میں ضم
اب پی ٹی آئی کی جانب اس وعدہ کو پورا کرنے کے لیے چھ رکنی کمیٹی قائم کی گئی .ہوگیا تھا
خبر کے بارے میں اپنی .ہے، جو اپنی سفارشات پارٹی چیئرمین عمران خان کو پیش کرے گی
رائے کا اظہار کمنٹس میں کریں۔ مذکورہ معلومات کو زیادہ سے زیادہ لوگوں تک پہنچانے کے لیے
سوشل میڈیا پر شیئر کریں۔comments

Figure 3 Used Corpus UIR-21

The translation of the Urdu documents is given below. At present, we have two languages for the document set: one being the original language (Urdu) and the other being the translated version (English). There are a total of 35 query sets in the document collection, and there are two variations for each query. The 35 query sets along with their variations have been meticulously translated into English. With two document sets and two query sets, our system is now ready to perform cross-lingual information retrieval.

South Punjab Province: PTI has constituted a committee headed by Balkh Shermazari Islamabad: Pakistan Tehreek -e -Insaf has initiated practical steps for the establishment of southern Punjab province, Imran Khan has set up a six -member committee. According to the details, the party has begun to fulfill the promise of South Punjab province on the direction of Imran Khan. The committee will prepare a plan to form a southern Punjab province and after reviewing all aspects, it will submit its recommendations to the party chairman in 100 days. Official notification was also issued with Khan's approval. The six -member committee includes Shah Mehmood Qureshi, Jahangir Tareen, Ishaq Khakwani, Khusroqat, Tahirbashir Cheema. The committee will prepare a plan to form a southern Punjab province and submit its recommendations to the party chairman within 100 days after reviewing all aspects. According to party sources, the committee will be able to consult experts in preparation of recommendations, review their views and assist. It is to be remembered that a few months ago, the members of the PML -N had announced to separate from the party and form a southern Punjab province, which was headed by Balkh Sher Mazari, later the front was merged with the PTI. Now a six -member committee has been set up to fulfill this promise by PTI, which will submit its recommendations to party chairman Imran Khan. Express your opinion about the news in the comments. Share on social media to convey the above information to more and more people. comments

Figure 4 English Text

Diagram 4 contains the translation of the Urdu document to English. We use the Google Translate API for translating the Urdu documents to English, which is available in Python.

3.1.2 Queries Keywords

All the keywords of each query that apply to the corpus and are mentioned in the relevance judgment of corpus UIR-21 [23] are mentioned below in the tables. The keywords exist in both Urdu and English languages. The keywords of the English query are listed in the first column,

whereas the keywords of the Urdu query are mentioned in the second column.

3.2 Evaluation Parameters

The evaluation criteria in IRS are applied for measuring the efficiency and performance of the device in finding statistics relevant to the queries raised by users. The common evaluation criteria include (Precision, Recall, F1-Score).

3.3 Evaluation Setup

Evaluation of the proposed models is done by performing the above-mentioned six queries and doing a search in the document set. The English query is compared to the English document set, while the Urdu query is compared to the Urdu document set.

3.3.1 Relevancy judgment

The following Table 1, obtained from the corpus, presents the relevance judgments for the six selected queries. Table 1 shows the predefined relevance judgments for all queries in the Information Retrieval System (IRS) test. All the queries (from 1 to 6) are linked to the list of Document IDs deemed relevant to the particular query. For example, query 1 corresponds to documents such as 5111.Txt, 8491.Txt and 21572.Txt. This means that these documents are relevant to query 1, according to the predefined relevancy judgments. In similar fashion, the other queries correspond to their respective relevant documents. The predefined relevancy judgments act as a standard for measuring the overall effectiveness of the IRS in retrieving relevant documents for each query, allowing for a quantitative assessment of the system's effectiveness in statistics retrieval for the Urdu and English languages.

Table 1 Predefined Relevancy Judgment of each query

Query	Documents
6	"7090,10469,11759,12097,13561,13856 & 25408"
5	3516.txt,6132.txt,16164.txt,23704.txt,27812.txt and 31972.txt
4	"2530,3525,3684,6083,10525,12036, 12087,12330,12373,13839,13904,15427, 15854,16100,16966,16986,17384,18411, 18459,18940,19782,25363,25398,25691 & 29433"
3	"11225,20464 & 20571"
2	"524,787 & 904"
1	"5111,8491 & 21572"

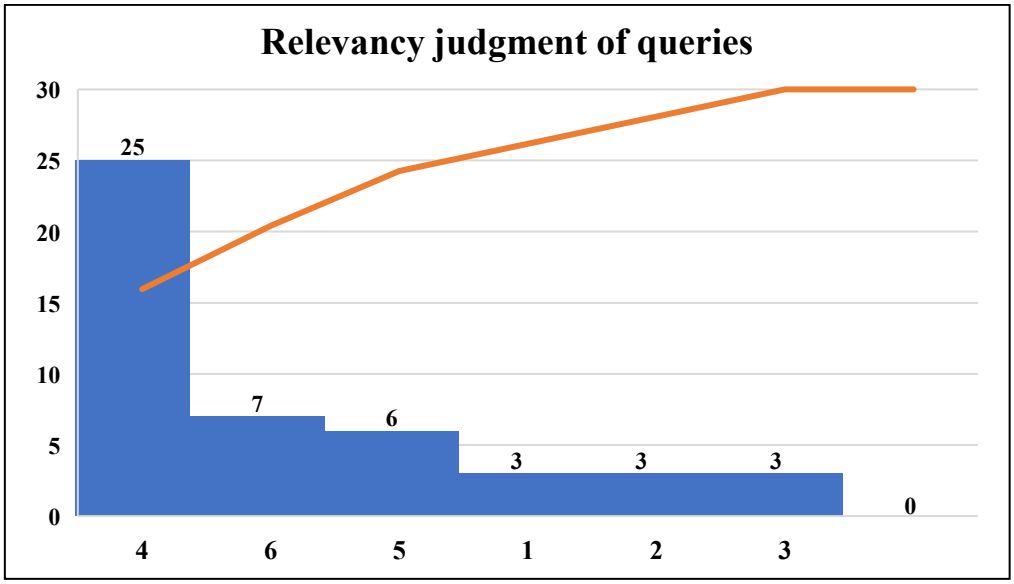

Figure 5 Relevance Judgment of each query

Figure 5 depicts the number of documents containing relevant judgments. Documents associated with Q1, Q2, Q3, and Q4 can be obtained (3,3,3,3). The highest number of documents (25) was obtained from Q5. Q6 obtained six and seven Docs.

3.3.2 Suggested Urdu IRS

Table 2 gives an illustration of the suggested output from the Urdu IRS for various queries. Here, each query (one through six) is linked with a set of report ID numbers that could be recommended as relevant via the system. For instance, query 1 will be associated with documents 5111.Txt, 8000.Txt, 8491.Txt, and 16525. Txt.

Table 2 Proposed Urdu IRS results

Query	Documents-ID
6	"3467,3481,3594,4699,7090,9642,10479,11759,12212,13377,13561, 13856,14927,17299,17873,18116,20768,21103,22243,22425,22440, 22872,23499,23518,23547,24187,25918,25966,26386,29470,29516, 30344,30394,30622,30997,32586,33030,33311,33830,35095,35878, 36324,36511, & 36804"
5	"3132 & 16164"
4	"488,575,2530,5720,6083,10528,12036,12087,12330,12373,13904, 15427,15854,16100,16966,16986,17384,17780,18441,18659,19782, 25363,25398, & 25691"
3	"133,1011,1546,1850,2526,3148,3442,3508,4446,4570,4894,5783, 6333,6559,7162,7602,7635,7760,7882,7888,8376,8773,8913, & 9605"
2	"524,22852,32016, &32025"
1	"5111,8000,8491 & 16525"

The above-mentioned penalties represent the output produced by IRS mainly relying on its rules and indexing technique. The table is meant to be used as a means to measure the efficiency of IRS and its capacity to produce applicable content in response to customer requests in Urdu.

3.3.3 Suggested English IRS

The table below shows the expected output from the English Information Retrieval System (IRS) with different queries. In a way similar to Table 3 with Urdu queries, each question number (from 1 to 6) is connected with file IDs recommended by the gadget as relevant documents. For example, question number 1 is related to files 5111.txt, 8591.txt, and 21572.txt.

Table 3 Proposed English IRS results

Query	Documents
6	7090,10479,11759,12097,13561,13856 & 25406"
5	3516,6133,16174,23704,27812 & 31972"
4	"2530,3525,3684,6083,10525,12036,12087, 12336,12373,13849,13904,15427,15854, 16200,16966,16986,17384,18411,18459, 18940,19782,25363,25398, & 25691"
3	"11225,20564 & 20571"
2	"524,787 & 904"
1	"5111,8591 & 21572"

The table shows a picture of the performance of the tool in retrieving possibly applicable files for each of the queries in English. This is done through the outcomes obtained from the indexing and retrieval capabilities of the tool, which are meant to help the clients find

relevant statistics. The desk forms the basis of assessing the effectiveness of IRS in handling English queries.

3.3.4 Urdu IRS, English IRS, and Relevance Judgment

The comparison between the relevance judgment of queries and the performance of the two information retrieval systems (IRS) is given in Table 4 below. In each question from 1 to 6, the table shows the number of relevant files according to the relevancy judgment, the number of files recommended using the Urdu IRS, and the number of files recommended using the English IRS.

Table 4 Comparison of all three IRS

Query	Relevancy Judgement	Urdu IRS	English IRS
6	7	44	7
5	6	2	6
4	25	24	24
3	3	24	3
2	3	4	3
1	3	4	3

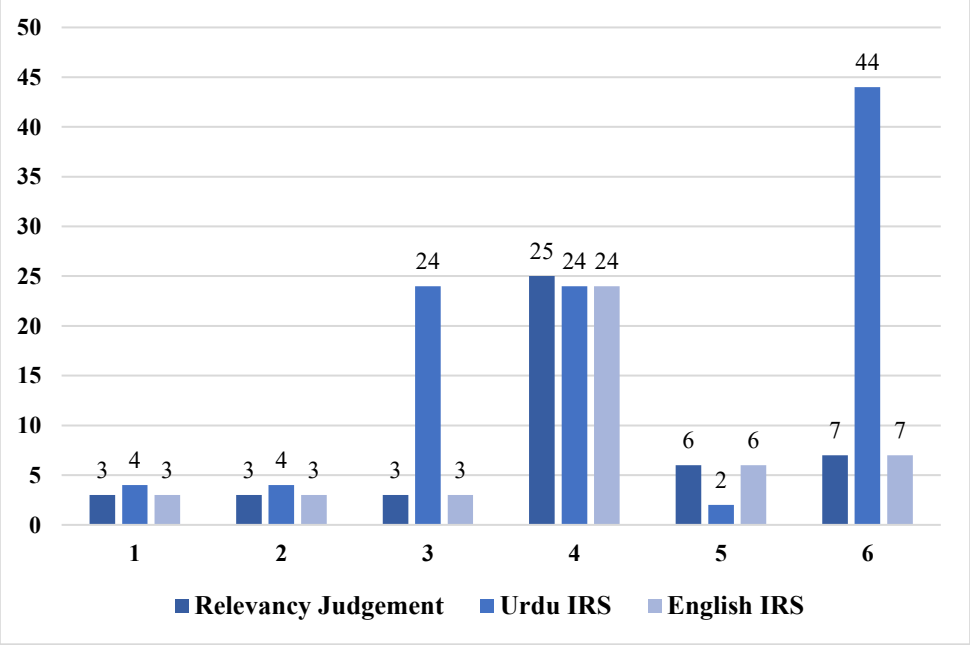

Figure 6 Comparison of all IRS

This comparison allows us to evaluate how well each machine did in terms of finding the relevant documents in comparison to ground truth relevance. A higher diversity in the Urdu IRS or English IRS column indicates a better performance in terms of finding relevant documents for that particular question. The following table is a quantified representation of the efficiency of the two IRSs in finding the relevant statistics for Urdu and English queries.

3.3.5 Urdu with Relevance Judgment

Table 5 shows the comparison of the overall performance of the Urdu Information Retrieval System (IRS) with the relevance judgment in terms of the measures like Precision, Recall, and F1-Score for each query (1 to 6). The measure "Precision" is calculated based on the ratio of the number of relevant files from the total number of retrieved files. The measure "Recall" is the percentage of relevant files that have been retrieved.

Table 5 Urdu VS Relevancy Judgment

Query	Precision	Recall	F1-Score
6	0.11	0.71	0.20
5	0.5	0.17	0.25
4	0.75	0.72	0.73
3	0.08	0.67	0.15
2	0.25	0.33	0.29
1	0.5	0.67	0.57
Average	**0.39**	**0.56**	**0.40**

For instance, the Precision for question 1 is 0.50 meaning that half of the documents retrieved through the Urdu IRS were pertinent. The Recall value for question 1 is 0.67 meaning that the Urdu IRS was able to retrieve two-thirds of the pertinent documents. The F1-Score that evaluates both Recall and Precision is 0.57 meaning that there is an equilibrium between Recall and Precision values.

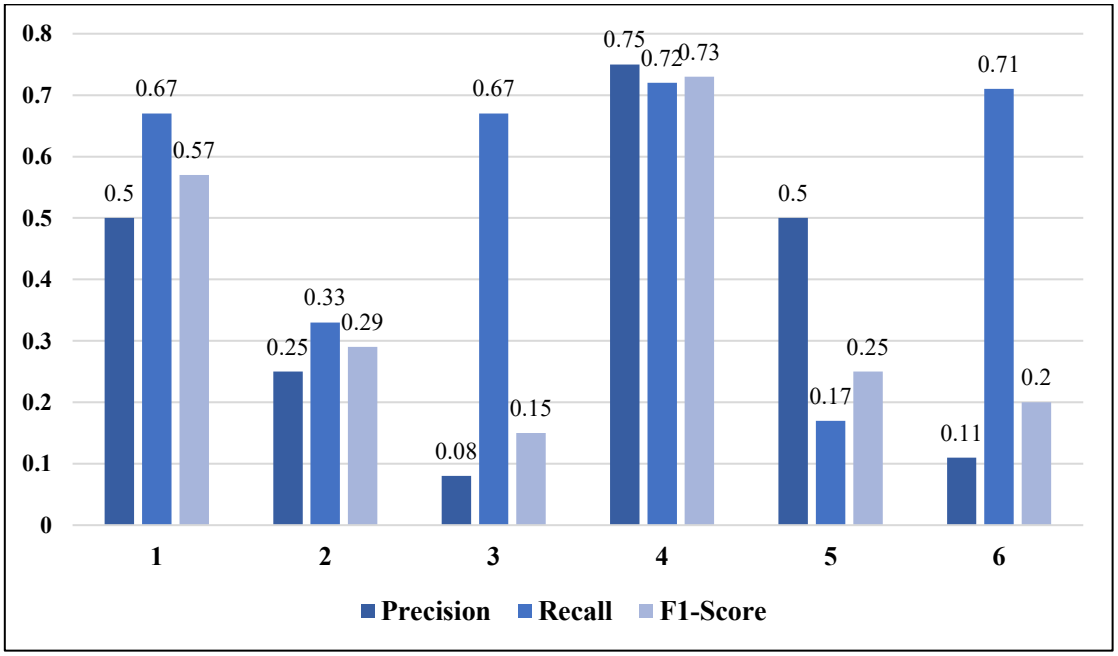

Figure 7 Urdu Vs Relevancy

In summary, Figure 7 provides an analysis of the performance of the Urdu IRS in comparison with the relevancy judgments, indicating areas where the tool succeeded and where it can be improved upon. The mean Precision, Recall, and F1 Score give us an indication of the performance of the tool as a whole.

3.3.6 Comparison of IRS

Comparison between the results obtained through the execution of the English IRS and the relevance judgement as well as the application of metrics like Precision, Recall and F1-Score for each question (Questions 1-6) can be seen in Table 6 below. The Precision is used to measure the ratio of the relevant documents to the total number of retrieved documents; Recall is the ratio of the relevant documents to the total number of retrieved documents while the F1-Score is the harmonic mean of the Precision and Recall. As an illustration, in Question 1, the Precision is 0.50 which means that half of the retrieved documents are relevant. The Recall is 0.67, implying that the English IRS can retrieve about three-quarters of the relevant documents.

Table 6 Relevance vs. English

Query	Precision	Recall	F1
6	0.11	0.71	0.20
5	2	0.67	1
4	0.88	0.84	0.86
3	0.13	1	0.22
2	0.75	1	0.86
1	0.5	0.67	0.57
Average	**0.70**	**0.79**	**0.61**

The F1-Score, where both Precision and Recall are taken into account, is 0.57, implying that there is a moderate balance between Precision and Recall. The averaged Precision, Recall, and F1-Score allow presenting an overall picture of the average system behavior across all queries, thus demonstrating the level of performance of the English IRS against the relevancy assessment.

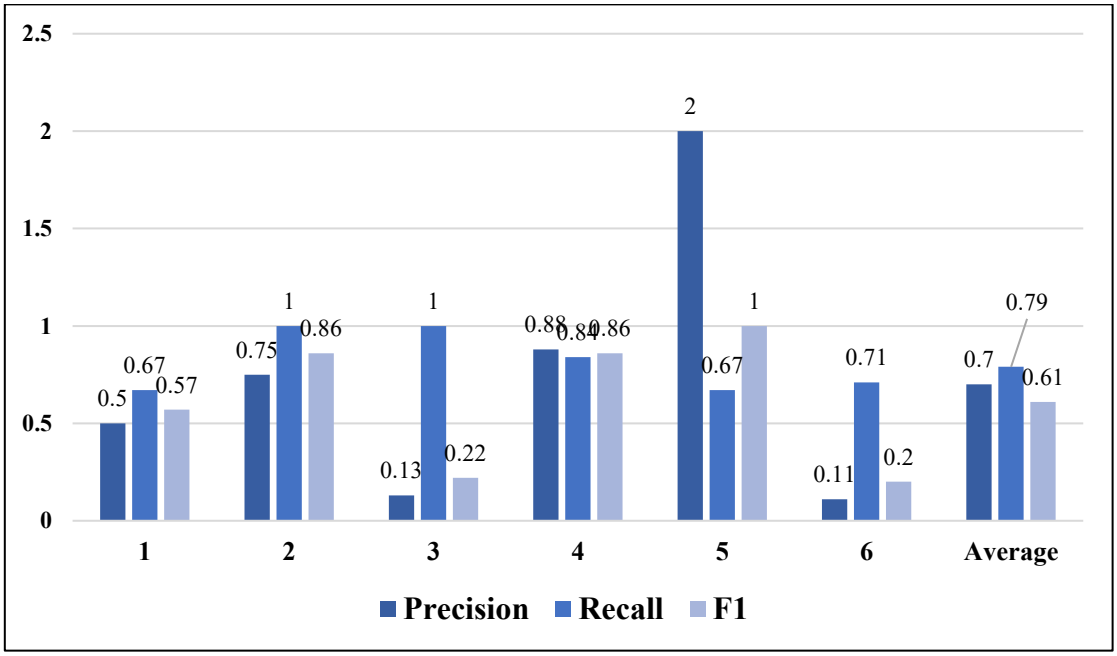

Figure 8 Relevance vs. English

4 Conclusion

Cross-Lingual Information Retrieval (CLIR) systems were developed to enable users to access information written in languages different from those used in their search queries. Such systems allowed queries submitted in one language to retrieve relevant documents written in another language by establishing a framework for comparing multilingual queries and documents. CLIR has remained one of the most challenging research areas in the field of Information Retrieval due to the linguistic complexities involved in multilingual environments. Urdu, the national language of Pakistan, is spoken by approximately 163 million people worldwide and exhibits rich morphological characteristics that require specialized attention in Information Retrieval (IR) research.

In this study, a hybrid query model was developed to address the lack of efficient CLIR systems for the Urdu–English language pair. The proposed framework supported both Urdu and English queries and facilitated document retrieval in monolingual as well as cross-lingual settings, including Urdu-to-English and English-to-Urdu retrieval scenarios. The proposed system was tested following the TREC paradigm that involves a corpus of documents, query

topics, and relevance judgments to evaluate the retrieval performance of the system. To test the performance of the proposed hybrid query model, standard measures were used including Precision, Recall, and F1-Score. Results indicated variability in the retrieval performance of the proposed model depending on the language settings. The Urdu Information Retrieval System (IRS) performed best in terms of precision while the English IRS obtained comparatively low precision. On the other hand, the English IRS obtained the highest recall value of 80% while the Urdu IRS had a minimum recall of 50%. These results revealed the relationship between the precision and recall in the cross-lingual retrieval setting and also demonstrated the efficiency of the proposed hybrid query model. Future research would focus on improving the accuracy of retrieval through the use of complete queries and using the whole UIR-21 document corpus. Additionally, the proposed framework could be expanded to include more languages.

REFERENCES

E. Ghanbari and A. Shakery, "A Learning to a rank framework based on cross-lingual loss function for cross-lingual information retrieval," 2021.

B. Ayaz, W. Altaf, F. Sadiq, H. Ahmed, and M. A. Ismail, "Novel Mania: A semantic search engine for Urdu," ICOSST 2016 - 2016 Int. Conf. Open Source Syst. Technol. Proc., pp. 42–47, 2017, DOI: 10.1109/ICOSST.2016.7838575.

E. Ghanbari and A. Shakery, "Query-dependent learning to rank for cross-lingual information retrieval," Knowl. Inf. Syst., 2018, DOI: 10.1007/s10115-018-1232-8.

Chayapathi A R, G Sunil Kumar, Manjunath Swamy B E , Thriveni J4, and Venugopal K R, "cloud-based multi-language indexing using cross lingual information retrieval approaches," vol. 9, no. 1, pp. 1283–1293, 2021.

D. Bekele, "A Cross-Lingual Information Retrieval (CLIR) System for Afaan Oromo-English using a Corpus-Based Approach," vol. 4, no. 05, pp. 1287–1293, 2015.

T. Bi, "Constraint Translation Candidates : A Bridge between Neural Query Translation and Cross-lingual Information Retrieval."

S. Saleh and S. Saleh, "Cross-lingual information retrieval systems: Methods and Challenges," 2017.

M. Maryamah, A. Agus Zainal, S. Riyanarto and H. Ahmad Makki, "Adapting Google Translate using Dictionary and Word Embedding for Arabic-Indonesian Cross- lingual Information Retrieval," pp. 205–209, 2020.

S. Shadi and P. Pavel, 2020. "Document Translation vs. Query Translation for Cross-Lingual Information Retrieval in the Medical Domain", In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 6849–6860, Online. Association for Computational Linguistics.

P. M. and S. P. Manaal Faruqui, "Soundex-based Translation Correction in Urdu–English Cross-Language Information Retrieval," Proc. Work. CLIA IJCNLP, pp. 25–29, 2011, [Online]. Available: http://www.cs.cmu.edu/~mfaruqui/Faruqui-

R. Zbib, Rabih, Hartmann, William, DeYoung, Jay Rivkin, Noah Olin, Franklin, Schwartz, Richard Makhoul, John, "Neural-Network Lexical Translation for Cross-lingual IR from Text and Speech," no. July, pp. 645–654, 2019.

E. Boschee, Barry, Joel Billa, Jayadev Freedman, Marjorie Gowda, Thamme,"SARAL : A Low-Resource Cross-Lingual Domain-Focused Information Retrieval System for Effective Rapid Document Triage," pp. 19–24, 2019.

Picchi, E. & Peters, C, (2000) "Cross-language information retrieval: a system for comparable corpus querying, in Cross-Language Information Retrieval," G. Grefenstette, Editor. 2000, Kluwer Academic Publishing: Massachusetts. pp. 81-90.

IEER
Institute for Excellence in Education & Research

Lawrie, D., Mayfield, J., Oard, D., & Yang, E. (2022). HC4: A New Suite of Test Collections for Ad Hoc CLIR. arXiv preprint arXiv:2201.09992.

Maqbool, M. S., Hanif, I., Iqbal, S., Basit, A., & Shabbir, A. (2023). Optimized feature extraction and cross-lingual text reuse detection using ensemble machine learning models. Journal of Computing & Biomedical Informatics, 5(01), 26-40.

Abid, K., Aslam, N., Fuzail, M., Maqbool, M. S., & Sajid, K. (2023). An efficient deep learning approach for prediction of student performance using neural network. VFAST Transactions on Software Engineering, 11(4), 67-79.

Kanwal, F., Abid, M. K., Maqbool, M. S., Aslam, N., & Fuzail, M. (2023). Optimized classification of cardiovascular disease using machine learning paradigms. VFAST Transactions on Software Engineering, 11(2), 140-148.

Basit, A., Hanif, I., Maqbool, M. S., Qayyum, W., Hasnain, M. A., & Nazeer, R. (2023). Cross-lingual information retrieval in a hybrid query model for optimality. Journal of Computing & Biomedical Informatics, 5(01), 130-141.

Fazal, U., Khan, M., Maqbool, M. S., Bibi, H., & Nazeer, R. (2023). Sentiment analysis of omicron tweets by using machine learning models. VFAST Transactions on Software Engineering, 11(1), 67-75.

Hasnain, M. A., Ali, S., Malik, H., Irfan, M., & Maqbool, M. S. (2023). Deep learning-based classification of dental disease using x-rays. Journal of Computing & Biomedical Informatics, 5(01), 82-95.

Hasnain, M. A., Ali, Z., Maqbool, M. S., & Aziz, M. (2024). X-ray image analysis for dental disease: A deep learning approach using efficientnets. VFAST Transactions on Software Engineering, 12(3), 147-165.

Maqbool, M. S., Zahra, S. R., Ismail, S., Nadeem, M., Fatima, N., & Ahmad, J. (2026). A CNN-BASED FRAMEWORK FOR EFFICIENT DETECTION OF EYE DISEASE IN FUNDUS IMAGES. Spectrum of Engineering Sciences, 4(4), 1157-1169.

Rafiqee, M. M., Qaiser, Z. H., Fuzail, M., Aslam, N., & Maqbool, M. S. (2023). Implementation of efficient deep fake detection technique on videos dataset using deep learning method. Journal of Computing & Biomedical Informatics, 5(01), 345-357.

Zainab, F., Nazim, F., Kashaf, M., Aslam, N., & Maqbool, M. S. (2026). Predictive analytics for customer churn in subscription-based businesses using machine learning. Spectrum of Engineering Sciences, 4(4), 596-618.

Tanveer, K., Aslam, N., Saleem, H., Maqbool, M. S., Akhter, M., & Rashid, M. H. (2026). INTEGRATING BLOCKCHAIN AND MACHINE LEARNING FOR ROBUST IOT DATA INTEGRITY IN CRITICAL INFRASTRUCTURES. Spectrum of Engineering Sciences, 4(3), 1875-1886.

Aslam, N., Meeran, M. T., Aslam, M., Maqbool, M. S., & Saeed, B. (2025). Understanding Urban Expansion Through Multi-Temporal Satellite Data Analysis. Kashf Journal of Multidisciplinary Research, 2(09), 252-273.

Malik, F., Fuzail, M., Aslam, N., Sarwar, R., Abid, K., Maqbool, M. S., & Yousaf, A. (2024). A hybrid machine learning model to predict sentiment analysis on X. Journal of Computing & Biomedical Informatics, 6(02), 64-79.

Maqbool, M. S., Nazeer, R., Basit, A., & Zahra, K. (2023). Automated detection and localization of fungal infections on cotton leaves using YOLO-based object detection model. Machines and Algorithms, 2(2), 121-136.

Hasnain, M. A., Ali, Z., Rehman, K. U., Ehtsham, M., & Maqbool, M. S. (2024). Idd-net: A deep learning approach for early detection of dental diseases using x-ray imaging. Journal of Computing & Biomedical Informatics, 7(02).

Aslam, N., Maqbool, M. S., Nadeem, M., & Saleem, H. (2026). DEEPFAKESHIELD: ENHANCED VIDEO AUTHENTICITY DETECTION VIA CONVOLUTIONAL VISION TRANSFORMER. Spectrum of Engineering Sciences, 4(3), 1650-1665.

Manzoor, I., Maqbool, M. S., Shahzad, F., Nadeem, M., Zulfiqar, A., & Naqvi, S. Q. (2026). A FRAMEWORK FOR HATE SPEECH IDENTIFICATION USING OPTIMIZED TEXT FEATURES AND NATURAL LANGUAGE PROCESSING ON TWITTER DATASET. Spectrum of Engineering Sciences, 4(6), 1389-1405.

Maqbool, M. S., Fatima, N., Nazeer, R., Aslam, N., Abbas, F., Sumra, U., & Nadeem, M. (2025). A HYBRID DATASET-BASED ENSEMBLE STRATEGY FOR EFFICIENT BREAST CANCER DETECTION. Kashf Journal of Multidisciplinary Research, 2(12), 39-57.

Aslam, M., Maqbool, M. S., Aoun, M., Aslam, N., Razzaq, A. M., & Ali, S. (2026). HIGH-PERFORMANCE AND EFFICIENT BRAIN TUMOR SEGMENTATION FOR ENHANCED CLINICAL ANALYSIS. Spectrum of Engineering Sciences, 4(3), 195-210.

Israr, S., Maqbool, M. S., Hanif, I., Nadeem, M., Basit, A., & Batool, A. A. (2026). A ROBUST DEEP LEARNING FRAMEWORK FOR PREDICTING ACADEMIC SUCCESS USING ADVANCED NEURAL ARCHITECTURES. Spectrum of Engineering Sciences, 4(5), 1006-1022.

Madankar, M., Chandak, M. B., & Chavhan, N. (2016). Information retrieval system and machine translation: a review. Procedia Computer Science, 78, 845-850.

S. Sia, "CLIReval: Evaluating Machine Translation as a Cross-Lingual Information Retrieval Task," pp. 134–141, 2020.

K. R. Shahapure, "Analysis of Cross Language Information Retrieval methods", Term paper, 2013.

Jungwirth, M., & Hanbury, A. (2018). Replicating an Experiment in Cross-lingual Information Retrieval with Explicit Semantic Analysis. In CLEF (Working Notes)."

J. Li et al., "Cross-lingual low-resource set-to-description retrieval for global e-commerce," in AAAI 2020 - 34th AAAI Conference on Artificial Intelligence, 2020, pp. 8212–8219.

L. Zhao, R. Zbib, Z. Jiang, D. Karakos, and Z. Huang, "Weakly Supervised Attentional Model for Low Resource Ad-hoc Cross-lingual Information Retrieval," pp. 259–264, 2019.

Ameer, F., Maqbool, M. S., Aslam, N., & Raza, A. (2026). DEEP LEARNING MALWARE DETECTION FOR ANDROID MOBILE PHONES UTILIZING CCN AND TRANSFER LEARNING MODELS. *Spectrum of Engineering Sciences*, 4(3), 4213-4229.

S. Shaukat, A. Shaukat, K. Shahzad, and A. Daud, "Using TREC for developing semantic information retrieval benchmark for Urdu," Inf. Process. Manag., vol. 59, no. 3, p. 102939, 2022, doi: 10.1016/j.ipm.2022.102939.

L. Zhang et al., "The 2019 {BBN} Cross-lingual Information Retrieval System," in Proceedings of the workshop on Cross-Language Search and Summarization of Text and Speech (CLSSTS2020), 2020, no. May, pp. 44–51, [Online]. Available: https://aclanthology.org/2020.clssts-1.8.

M. Yarmohammadi et al., "Robust Document Representations for Cross-Lingual Information Retrieval in Low-Resource Settings," vol. 1, 2019.

L. Yao, B. Yang, H. Zhang, W. Luo, and B. Chen, "Exploiting Neural Query Translation into Cross Lingual Information Retrieval," 2020, [Online]. Available: http://arxiv.org/abs/2010.13659.

A. Xain, "Multilinguistic approach towards Information Retrieval System for Big Data," pp. 159–164, 2020.

D. Thenmozhi and C. Aravindan, "Ontology-based Tamil – English cross-lingual information retrieval system," Sādhanā, vol. 0123456789, 2018, doi: 10.1007/s12046-018-0942-7.

P. Shi, R. Zhang, H. Bai, and J. Lin, "Cross-Lingual Training of Dense Retrievers for Document Retrieval," 2021, pp. 251–253, doi: 10.18653/v1/2021.mrl-1.24.

V. Sharma and N. Mittal, "Refined stop-words and morphological variants solutions applied to Hindi-English cross-lingual information retrieval," in Journal of Intelligent and Fuzzy Systems, 2019, vol. 36, no. 3, pp. 2219–2227, doi: 10.3233/JIFS-169933.

V. K. Sharma, N. Mittal, and A. Vidyarthi, "Context-based Translation for the Out of Vocabulary Words Applied to Hindi-English Cross- Lingual Information Retrieval Context-based Translation for the Out of Vocabulary Words Applied," IETE Tech. Rev., vol. 0, no. 0, pp. 1–10, 2020, doi: 10.1080/02564602.2020.1843553.

V. K. Sharma and N. Mittal, "Cross-lingual information retrieval: A dictionary-based query translation approach," in Advances in Intelligent Systems and Computing, 2018, vol. 554, pp. 611–618, doi: 10.1007/978-981-10-3773-3_59.

Manning, C. D., Raghavan, P., & Schütze, H. (2008). Introduction to information retrieval (Vol. 1, p. 496). Cambridge: Cambridge University Press.