iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 11 **Issue:** IX **Month of publication:** September 2023

DOI: https://doi.org/10.22214/ijraset.2023.55658

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 11 Issue IX Sep 2023- Available at www.ijraset.com

Cross Language Information Retrieval based on Automatic Query Translation for Marathi Documents

Suhas D. Pachpande[1], Parag U. Bhalchandra[2]
[1]*Department of Computer Science, Sant Gadge Baba Amravati University, Amravati*
[2]*School of Computational Sciences, S.R.T.M. University, Nanded*

Abstract: ***The current research article explores the realm of Cross Language Information Retrieval (CLIR) and its significance in the digital age. It addresses the challenges faced in CLIR, including lexical and semantic disparities, the scarcity of parallel corpora, cultural nuances, and more. The article discusses innovative solutions encompassing Machine Translation, Query Expansion, Cross-Lingual Word Embeddings, and Multilingual Information Retrieval Models to enhance CLIR's effectiveness. Furthermore, it sheds light on Information Retrieval Models, such as the Boolean Model, Vector Space Model (VSM), and Probabilistic Models, explaining their principles and applications. The study also presents experimental results highlighting the limitations of monolingual IR models and the effectiveness of crosslingual techniques, such as translation and query expansion, in improving CLIR, making it a valuable tool for accessing information across languages.***
Keywords: Information Retrieval (IR), Cross Language Information Retrieval (CLIR), Cross-Lingual Evaluation, Information Retrieval Challenges.

I. INTRODUCTION

Information Retrieval (IR) stands at the core of our digital age, serving as the backbone of search engines, digital libraries, and countless applications that help users access and navigate the vast ocean of digital information (Manning et al., 2008). IR encompasses the systematic process of retrieving pertinent information from extensive collections, often referred to as document corpora, based on user queries or search terms (Safdar et al., 2020). This retrieval process plays a pivotal role in ensuring that users can access relevant, reliable, and timely information. IR techniques have evolved significantly over the years, ranging from early keyword-based searches to sophisticated ranking algorithms that consider factors such as relevance, context, and user intent. Central to this evolution is the recognition that effective IR goes beyond simple keyword matching; it involves understanding the nuances of human language and the multifaceted nature of information needs (Salton, G. 1989).

In parallel with the advancement of IR, the importance of Cross Language Information Retrieval (CLIR) has grown in an increasingly interconnected and multilingual world. CLIR is a specialized domain within IR, dedicated to overcoming language barriers in the retrieval of information (Voorhees E. M, 2009). It allows users to seek information in languages different from their own, addressing the global demand for multilingual access to knowledge (Suhas et al., 2020).

1) *Challenges in CLIR:* The challenges faced in Cross Language Information Retrieval (CLIR). CLIR is a complex field that encounters various linguistic, semantic, and practical obstacles (Chen B and Abedjan Z 2021). Understanding these challenges is crucial for developing effective CLIR systems and improving information access in multilingual contexts (Jian-Yun Nie 2003, Monika et al., 2015).

Here's a detailed description of some key challenges in CLIR:

a) Lexical Variability: Different languages often have distinct vocabularies, synonyms, and homonyms, making it challenging to match query terms to document terms accurately. In English, "car" may refer to an automobile, while in British English, "car" is often used to describe a railway carriage. In French, "actuellement" means "currently," but in Canadian French, it means "nowadays."

b) Semantic Divergence: Languages can express concepts differently, making it difficult to ensure precise query-document matching. Translating "apple" from English to French results in "pomme," but "pomme" can also mean "potato" in some contexts, leading to potential misinterpretation.

c) Lack of Parallel Corpora: Parallel corpora, which consist of aligned texts in two or more languages, are essential for training machine translation models and improving CLIR. Many languages lack sufficient parallel corpora, hindering the development of accurate translation systems (Monika et al., 2015). Without parallel corpora, machine translation systems may produce less accurate translations, affecting the quality of CLIR (Suhas et al., 2022).

d) Cultural Nuances: Documents often contain cultural references, idioms, or metaphors that may not be well-understood by users from different cultural backgrounds. These cultural nuances can lead to misunderstandings, particularly in humorous or metaphorical content.

e) Low-Resource Languages: Some languages have limited digital resources, such as online texts, which makes it challenging to build effective CLIR systems for these languages. Users of low-resource languages may experience poorer search results due to the scarcity of digital content.

f) Multilingual Query Ambiguity: Ambiguous queries, where a single query term can have multiple meanings in different languages, can lead to retrieval of irrelevant documents. The English query "bank" can refer to a financial institution or the side of a river. In French, "banque" and "rive" are the respective translations, but a CLIR system may struggle to disambiguate without context (Pu-Jen et al., 2004).

g) Cross-Lingual Polysemy: Polysemy occurs when a single word has multiple related meanings. In CLIR, polysemy across languages can lead to mismatches in query-document relevance. The word "bat" can mean a flying mammal or a piece of sporting equipment. Translating "bat" into another language may introduce ambiguity (Sanderson, M., & Clough 2010).

h) Data Sparsity: In some languages, there may be limited digital content available for indexing and retrieval, resulting in data sparsity. Sparse data can affect the statistical reliability of IR models, making it harder to accurately rank documents.

2) Solutions to CLIR: Addressing the challenges in Cross Language Information Retrieval (CLIR) requires innovative solutions that draw from various fields such as machine translation, information retrieval, and natural language processing (Ujjwal et al., 2016 and Brown et al., 2020).

Here, we discuss some of the key solutions and approaches to mitigating the challenges in CLIR:

a) Machine Translation (MT): Utilize advanced machine translation systems, including neural machine translation, statistical machine translation, and rule-based translation, to bridge the language gap between queries and documents. High-quality machine translation can significantly improve CLIR by ensuring that queries are accurately translated and that retrieved documents are comprehensible to users.

b) Query Expansion: Incorporate query expansion techniques to refine user queries and improve retrieval accuracy. Pseudo-relevance feedback, relevance modeling, and query expansion based on external resources (e.g., WordNet) can be employed. Expanding queries with synonyms, related terms, or translations can enhance the recall and precision of CLIR systems (Taan et al., 2016).

c) Cross-Lingual Word Embeddings: Employ cross-lingual word embeddings to map words or phrases across languages, enabling the transfer of semantic information. Cross-lingual embeddings help capture semantic relationships between languages, allowing for more accurate matching of query terms to document terms.

d) Multilingual Information Retrieval Models: Develop specialized information retrieval models that are designed to handle multiple languages simultaneously. These models can incorporate language-specific features and relevance signals. Multilingual models enhance the effectiveness of CLIR by considering language-specific characteristics while still accommodating cross-lingual retrieval (Brown et al., 2020).

e) Cross-Lingual Query Expansion: Extend query expansion techniques to include translations of query terms. This approach augments the query with terms in the target language, improving cross-lingual retrieval. Cross-lingual query expansion enhances the retrieval of documents in the desired language, even when users submit queries in their native language (Dan et al., 2008).

f) Parallel Corpora Creation: Collaborate on building parallel corpora, especially for low-resource languages. Crowdsourcing, machine translation, and cross-lingual information retrieval competitions can aid in corpus development. Access to high-quality parallel corpora facilitates the training of better machine translation models, benefiting CLIR accuracy.

g) Contextual and Cross-Lingual Information Extraction: Incorporate advanced techniques in natural language processing (NLP) for contextual understanding and information extraction from multilingual documents. Leveraging contextual information can help improve the accuracy of matching user intent with relevant documents, even when dealing with complex sentence structures.

h) Cross-Lingual Evaluation and Benchmarking: Establish robust evaluation metrics, datasets, and benchmarks specifically designed for CLIR to assess the performance of different techniques and systems. Effective evaluation ensures that CLIR systems are tested rigorously and can guide further advancements in the field.

i) Hybrid Approaches: Combine multiple techniques, such as machine translation, cross-lingual embeddings, and query expansion, to build hybrid CLIR systems that harness the strengths of each approach. Hybrid approaches can achieve superior retrieval performance by addressing multiple challenges simultaneously.

3) Information Retrieval (IR) models:

a) Boolean Model: The Boolean Model is a foundational IR model based on principles of binary logic. In this model, documents and queries are represented as sets of terms, and retrieval is performed through strict term matching using Boolean operators. The Boolean Model operates on the premise of precise term matching. It retrieves documents that exactly match the query terms, based on Boolean operators such as AND, OR, and NOT. In the Boolean Model, documents and queries are converted into binary vectors. Each dimension in the vector corresponds to a unique term from the vocabulary. A value of 1 indicates the presence of the term, and 0 indicates its absence. Documents are retrieved if they satisfy the query's Boolean expression. For instance, using "AND" retrieves documents containing all query terms, while "OR" retrieves documents containing any of the terms(Manning et al., 2008).

b) Vector Space Model (VSM): The Vector Space Model is a versatile and widely used IR model that represents documents and queries as vectors in a multidimensional space, facilitating ranking by relevance. VSM captures the essence of documents and queries by treating them as vectors in a high-dimensional space. Similarity between these vectors is used to gauge relevance. In VSM, documents and queries are transformed into numerical vectors, with each dimension corresponding to a term in the vocabulary. Term weighting schemes like TF-IDF are applied to signify term importance. Documents are ranked based on the cosine similarity between their vectors and the query vector. Higher cosine similarity scores indicate greater relevance. VSM enables partial matching and ranking (Manning et al., 2008).

c) Probabilistic Models: Probabilistic Information Retrieval Models treat relevance as a probabilistic event, estimating the likelihood that a document is relevant to a query based on term frequencies and document statistics. Probabilistic models view relevance as a probabilistic event. They calculate the probability that a document is relevant to a query by considering term frequencies, document lengths, and other factors. Documents and queries are represented as probabilistic models, typically incorporating term frequencies and document length normalization. Documents are ranked based on their estimated probability of relevance (Manning et al., 2008). These models take into account factors like term frequency, document length, and the likelihood of term occurrences.

II. MATERIALS AND METHODS

Working of Vector Space Model (VSM): The Vector Space Model (VSM) is a widely used Information Retrieval (IR) model that represents documents and queries as vectors in a multidimensional space, allowing for the measurement of similarity and ranking of documents by relevance. Here, we provide a detailed description of how the VSM works (Fig. 01).

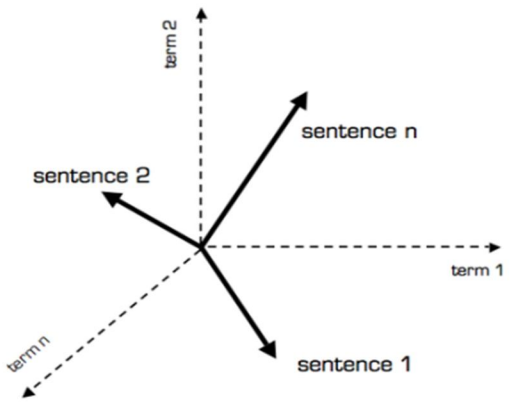


Fig. 01: Vector Space Model

1) Term-Document Matrix: VSM begins by constructing a term-document matrix. In this matrix, each row represents a term from the vocabulary, and each column represents a document from the collection. The values in the matrix typically represent the term frequency (TF) or other weighted measures like TF-IDF (Term Frequency-Inverse Document Frequency).The dimensions of the matrix are determined by the size of the vocabulary (number of unique terms) and the number of documents in the collection.
2) Term Weighting: TF-IDF: The most common term weighting scheme used in VSM is TF-IDF, which stands for Term Frequency-Inverse Document Frequency. This scheme assigns a weight to each term in a document based on its frequency in the document (TF) and its rarity across the entire document collection (IDF). Term Frequency represents how often a term occurs in a document. It is a measure of the importance of the term within the document.
3) Inverse Document Frequency (IDF): IDF measures the rarity of a term in the entire collection. It assigns higher weights to terms that are rare across the collection and lower weights to common terms.
4) Document and Query Vectorization: Vector representation is a term once the term-document matrix is prepared, documents and queries are transformed into numerical vectors. Each vector corresponds to a document or query. The dimensions of these vectors are determined by the vocabulary terms. Each dimension in the vector represents a unique term, and the value in each dimension represents the TF-IDF weight of that term in the document or query.
5) Cosine Similarity: Similarity calculation for the core of VSM's retrieval process is measuring the cosine similarity between the query vector and each document vector. Cosine similarity calculates the cosine of the angle between two vectors and provides a value between -1 and 1, where a higher value indicates greater similarity.
6) Ranking and Retrieval: Ranking by cosine similarity is ranked in descending order of cosine similarity scores. Documents with higher cosine similarity to the query are considered more relevant and are placed higher in the ranking. The top-ranked documents are retrieved and presented to the user as search results.

III. RESULT AND DISCUSSION

Experimental work:

Experimental work performed on English and Marathi documents both the monolingual and crosslingual IR Models implemented using VSM. The monolingual IR model which doesn't use any specific technique like translation for handling multiple languages is applied on English dataset first and then on Marathi dataset. Then the crosslingual IR model that incorporates techniques like translation, query expansion is applied on English dataset first and then on Marathi dataset. It has been observed that monolingual IR model fails to retrieve documents from Marathi dataset while the cross lingual IR model can successfully retrieve relevant documents from English and Marathi datasets.

Results obtained are demonstrated Fig. 02.

Keywords extracted from Query

Query ID	Keywords without CLIR	Keywords using CLIR
1	5	12
2	8	20
3	3	8
4	6	15
5	5	13
6	7	19
7	9	25
8	4	12
9	6	17
10	3	7

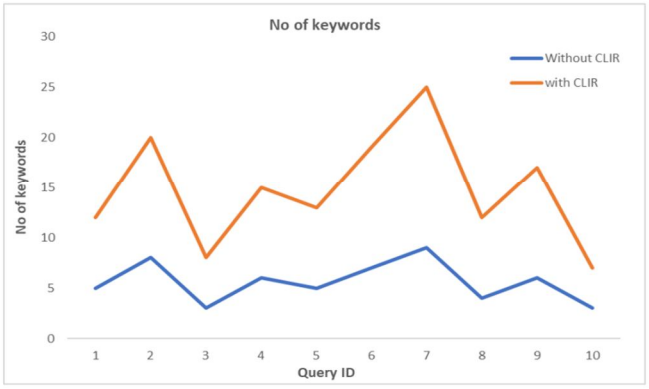


Fig. 02: Keywords for Query

During the preprocessing stage, keywords obtained from the English query are translated into Marathi and their synonyms in English and Marathi also are obtained. These additional words are appended to the list of keywords resulting into query having expanded coverage over both English and Marathi datasets.

Document weights computed using Monolingual IR (for English dataset)

Document No	Document Weight
2	0.64585
11	0.17397
41	0.17586
65	0
…	…
…	…
141	0.13771
158	0.17599
196	0.14530
202	0.12023
209	0.19183
211	0
…	…

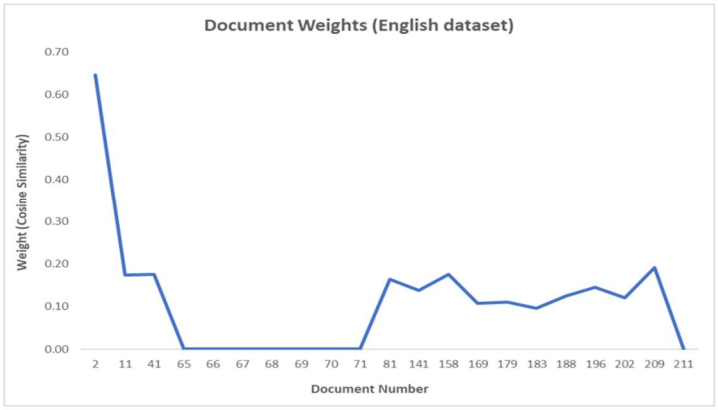


Fig. 03: Document weights computed using Monolingual IR for English.

Document weights computed using Monolingual IR (for Marathi dataset)

Document No	Document Weight
1	0
2	0
3	0
4	0
…	…
…	…
981	0
982	0
983	0
984	0
985	0
…	…

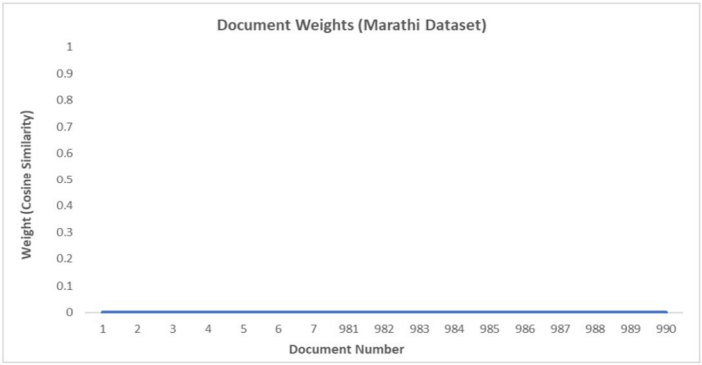


Fig. 04: Document weights computed using Monolingual IR for Marathi.

As can be observed in above fig. 04 and chart, the monolingual IR model fails to compute cosine similarity measures for Marathi documents since it doesn't incorporate any CLIR measure.

Experimental work conducted involved the implementation and evaluation of monolingual and IR models using VSM on both English and Marathi documents. The key findings from this investigation can be summarized as follows:

Document weights computed using Cross lingual IR (for Marathi dataset)

Document No	Document Weight
502	0.49584
511	0.32396
541	0.22585
565	0
…	…
…	…
658	0.32599
669	0
679	0.16125
683	0
688	0.07550
…	…

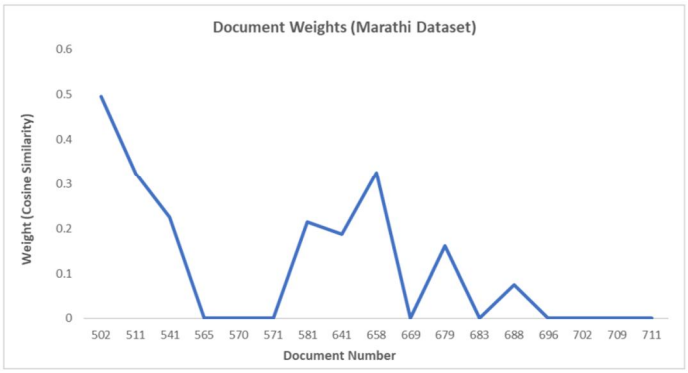


Fig. 05: Document weights computed using CLIR for Marathi.

As represented above (Fig. 05), it can be observed that the IR model which incorporates the CLIR techniques can successfully process Marathi documents and compute document weights as per the VSM.

Retrieval results of CLIR model for documents in English and Marathi datasets

Sr No	Weight (English)	Weight (Marathi)
1	0.64585	0.89585
2	0.19183	0.62599
3	0.17599	0.62397
4	0.17586	0.52586
…	…	…
…	…	…
12	0.10839	0.21963
13	0.09588	0.18839
14	0	0.16386
15	0	0.10839
16	0	0.12844
17	0	0.00000
…	…	…

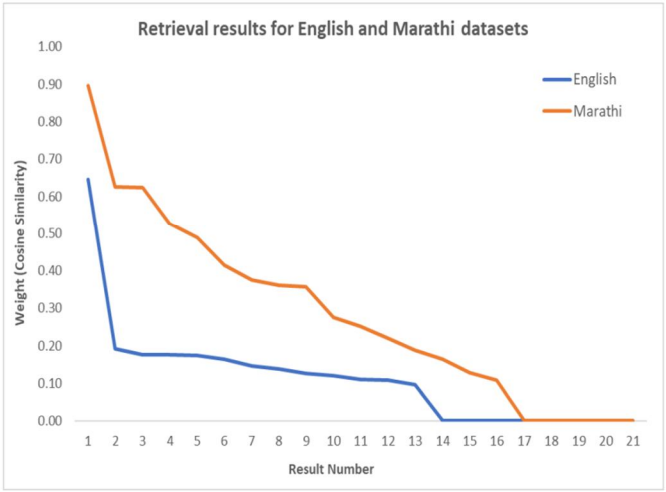


Fig. 06: Retrieval results of CLIR model for English and Marathi dataset.

The retrieval results in terms of relevant documents retrieved after applying the CLIR model on both English and Marathi datasets are demonstrated in Fig. 06. It is clearly observed that the CLIR model can successfully retrieve documents from datasets in both languages using a single query presented in English language.

IV. CONCLUSION

The monolingual IR model, which did not employ any specific techniques for handling multiple languages, was applied to both English and Marathi datasets. The results showed that this model failed to retrieve relevant documents from the Marathi dataset, indicating its limitations in crosslingual retrieval. In contrast, the crosslingual IR model, which incorporated techniques such as translation and query expansion, was applied to the same datasets. It was observed that the crosslingual IR model successfully retrieved relevant documents from both English and Marathi datasets. This demonstrated the effectiveness of crosslingual techniques in improving document retrieval across languages.

During the preprocessing stage, keywords from the English query were translated into Marathi, and their synonyms in both languages were obtained. These additional words were added to the list of keywords, resulting in a query with expanded coverage over both English and Marathi datasets. This approach contributed to the success of the crosslingual IR model.

Cosine similarity measures were computed for English documents using the VSM. The monolingual IR model failed to compute cosine similarity measures for Marathi documents since it lacked crosslingual capabilities. In contrast, the crosslingual IR model successfully processed Marathi documents and computed document weights based on the VSM. The retrieval results, in terms of relevant documents retrieved after applying the crosslingual IR model to both English and Marathi datasets, demonstrated the model's capability to retrieve documents in both languages using a single query presented in English.

REFERENCES

[1] Brown, A., et al. (2020). Advances in Cross-Language Information Retrieval: Third Workshop of the Cross-Language Evaluation Forum, CLEF 2002. Springer.
[2] Chen B and Abedjan Z (2021). "Interactive Cross-language Code Retrieval with Auto-Encoders," 36th IEEE/ACM International Conference on Automated Software Engineering (ASE), Melbourne, Australia, 2021, pp. 167-178, doi: 10.1109/ASE51524.2021.9678929.
[3] Dan Wu, Daqing He (2008), “ICE-TEA: an Interactive Cross-language Search Engine with Translation Enhancement”, SIGIR’08 (ACM), 20–24, 882-84.
[4] Jian-Yun Nie (2003), “Cross-Language Information Retrieval”, IEEE Computational Intelligence Bulletin, 6 (4) 19-23.
[5] Manning, C. D., Raghavan, P., & Schütze, H. (2008). Introduction to Information Retrieval. Cambridge University Press.
[6] Monika Sharma, Sudha Morwal (2015), “A Survey on Cross Language Information Retrieval”, International Journal of Advanced Research in Computer and Communication Engineering 4 (2), 384-387.

[7] Pu-Jen Cheng, Jei-Wen Teng, Ruei-Cheng Chen, Jenq-Haur Wang, Wen-Hsiang Lu, Lee-Feng Chien (2004), "Translating Unknown Queries with Web Corpora for CrossLanguage Information Retrieval", SIGIR'04 ACM 25 (9) 146-153.

[8] Safdar, Z., Bajwa, R.S., Hussain, S., Abdullah, H.B., Safdar, K., Draz, U., (2020). The role of Roman Urdu in multilingual information retrieval: A regional study. The Journal of Academic Librarianship 46, 102258. https://doi.org/10.1016/j.acalib.2020.102258

[9] Salton, G. (1989). Automatic Text Processing: The Transformation, Analysis, and Retrieval of Information by Computer. Addison-Wesley.

[10] Sanderson, M., & Clough, P. (2010). Evaluation of cross-language spoken document retrieval using the HLT-NAACL 2007 and TREC spoken document retrieval tracks. Information Retrieval, 13(2), 125-148.

[11] Suhas D. Pachpande, Parag U. Bhalchandra (2020). Cross Language Information Retrieval (CLIR): A Survey of Approaches for Exploring Web Across Languages. International Journal of Innovative Technology and Exploring Engineering (IJITEE), Volume-10 Issue-1, November 2020, 326-332.

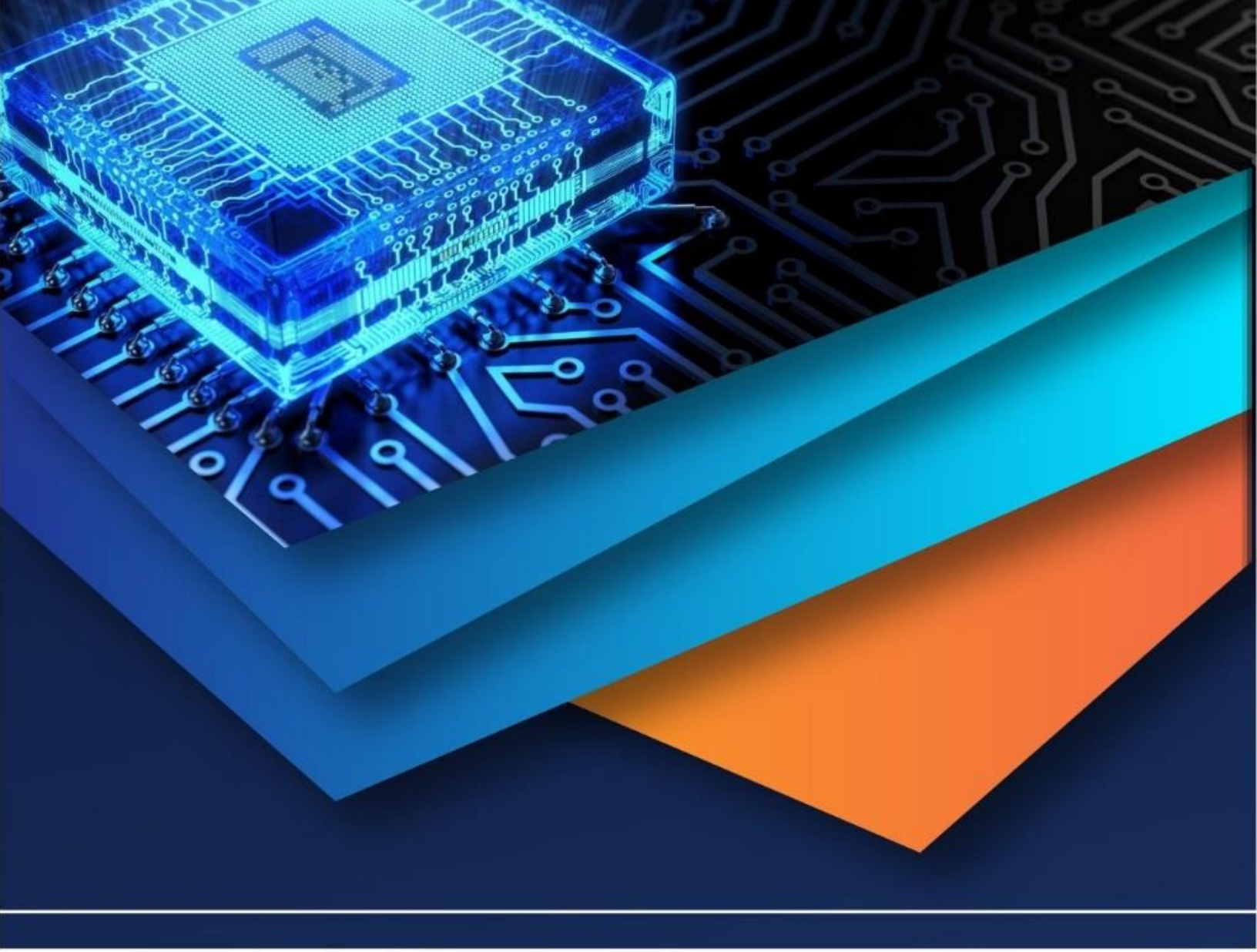
[12] Suhas D. Pachpande, Parag U. Bhalchandra and Ashok Gingine (2022). Framework of an Expert System for Intelligent Information Retrieval Across Languages using CLIR Techniques. AJOMC – Special issue on Research in Applied Science, Management and Technology, Vol. 7 No. 1. pp. 1662-1667

[13] Taan A A, Khan S. U. R., Raza A, Hanif A. M. and Anwar H (2021). "Comparative Analysis of Information Retrieval Models on Quran Dataset in Cross-Language Information Retrieval Systems," in IEEE Access, vol. 9, pp. 169056-169067, 2021, doi: 10.1109/ACCESS.2021.3126168.

[14] U. Archana, A. Khan, A. Sudarshanam, C. Sathya, A. K. Koshariya and R. Krishnamoorthy, "Plant Disease Detection using ResNet," 2023 International Conference on Inventive Computation Technologies (ICICT), Lalitpur, Nepal, 2023, pp. 614-618, doi: 10.1109/ICICT57646.2023.10133938.

[15] Ujjwal D, Rastogi P. and Siddhartha S (2016). "Analysis of retrieval models for cross language information retrieval," 10th International Conference on Intelligent Systems and Control (ISCO), Coimbatore, India, 2016, pp. 1-4, doi: 10.1109/ISCO.2016.7727028.

[16] Voorhees, E. M. (2009). Overview of the TREC 2009 Cross-Language Information Retrieval (CLIR) Track. In Proceedings of the Eighteenth Text Retrieval Conference (TREC 2009).



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY