

A Web based Hybrid Cross-Language Plagiarism Detection System for Hindi and Punjabi Using Neural Machine Translation and Embedding

Sachin Majithia¹, Dr. Harjeet Singh², Dr. Rajeev Puri³

¹Research Scholar, Department of Computer Science and Engineering Punjabi University, Patiala, Punjab, India, sachinmajithia@gmail.com

²Associate Professor, Department of Computer Science, Mata Gujri College, Fatehgarh Sahib, Punjab, India, zrjeet@gmail.com

³Associate Professor, PG Department of Computer Science, DAV College, Jalandhar, Punjab, India, rpuri@davjalandhar.com

Abstract - With the growth of resources available to students and various machine translation tools and techniques, cross-language plagiarism has become a serious problem. The majority of the available plagiarism detection tools are developed for English Language and are not able to accurately detect paraphrasing and/or translation-based plagiarism in Indian regional languages like Hindi and Punjabi. Hindi and Punjabi are similar languages but written in different scripts (Devanagari vs. Gurmukhi), so existing text-making techniques are not effective to detect plagiarism in these languages. In this paper a hybrid cross-language plagiarism detection is proposed for Hindi and Punjabi that uses a cascade machine translation pipeline which consists of dictionary-based translation, Example-Based Machine Translation (EBMT), Neural Machine Translation, and then multilingual sentence embeddings and FAISS index is used to perform fast semantic similarity search. The proposed system accepts Hindi text or uploaded documents, generates sentence embeddings, and compares them with a Punjabi corpus using cosine similarity. The proposed system also translates the Hindi text into Punjabi and then searches the translated Punjabi text against internet sources using Google Search API. The complete system is implemented as a Flask-based web application with modules for document upload, corpus management, plagiarism history, and sentence-based search. Experimental evaluation shows that semantic similarity-based plagiarism detection is more effective than traditional string-matching approaches, as it captures the semantic meaning of translated and paraphrased text rather than relying solely on lexical overlap.

Key Words: Cross-language plagiarism, Hindi, Punjabi, Machine Translation, Sentence Transformers, Semantic Similarity, Natural Language Processing

1. Introduction

Plagiarism is the act of using another person's work without proper acknowledgment. It has become a major concern in education and research due to the widespread availability of digital content. Traditional plagiarism detection systems compare documents written in the same language and mainly rely on lexical similarity. These systems perform well for direct copy-and-paste plagiarism but fail when the content is translated into another language.

Cross-language plagiarism occurs when a document is translated from one language into another without giving proper credit. Detecting such plagiarism is difficult because the words, sentence structure, and writing style change during translation. This problem becomes more challenging for Indian languages

because of script differences, limited parallel corpora, and insufficient language resources.

Hindi and Punjabi belong to the Indo-Aryan language family and share many grammatical and lexical characteristics. However, Hindi is written in the Devanagari script, while Punjabi uses the Gurmukhi script. Because of these script differences, simple string comparison methods cannot identify similar content effectively.

To overcome these limitations, this paper proposes a hybrid plagiarism detection framework that combines hybrid machine translation and semantic similarity. The system first translates Hindi text into Punjabi using a three-stage translation pipeline and then performs sentence-level semantic matching using multilingual semantic embeddings. Unlike keyword search, semantic embeddings capture context and meaning, improving detection of translated or paraphrased plagiarism. The proposed web application also provides corpus management, document upload, plagiarism history, and report generation. A demonstration interface (Fig. 1) shows the home page, corpus management, plagiarism-check module, and results. Experimental observations confirm that our sentence-based approach substantially outperforms a baseline lexical (TF-IDF) method for cross-language plagiarism.

The major contributions of this work are:

- Development of a complete web-based Hindi–Punjabi plagiarism detection system.
- A hybrid translation framework combining Dictionary-based Translation, EBMT, and NMT.
- Sentence-level semantic similarity using Sentence Transformers and cosine similarity.
- Corpus management and plagiarism history modules.

2. Literature Survey

Cross-language plagiarism detection has gained significant attention with the increasing availability of multilingual digital content and machine translation tools. Conventional plagiarism detection systems mainly rely on exact string matching or lexical similarity, which are effective for monolingual documents but perform poorly when the source content is translated into another language. This limitation has encouraged researchers to develop multilingual and semantic approaches capable of identifying translated and paraphrased plagiarism.

Puri et al. made one of the earliest contributions to Hindi–Punjabi cross-language plagiarism detection by developing a bilingual plagiarism detection framework based on lexical

analysis and sentence comparison. Their work demonstrated the feasibility of detecting plagiarism between Hindi and Punjabi documents by exploiting linguistic similarities between the two languages. However, the proposed approach mainly depended on lexical matching and could not effectively identify plagiarism when the translated text was significantly paraphrased or modified.

To improve cross-language similarity detection, Garg and Goyal proposed the Maulik framework, which focused on identifying semantic similarity between Hindi and Punjabi texts. The framework incorporated bilingual linguistic resources and similarity measures to improve document comparison beyond simple keyword matching. The Maulik framework performed well as compared to conventional lexical approaches but this framework is not able to detect plagiarism in complex translated documents due to limited bi-lingual resources were limited.

Goyal et al. further analysed the plagiarism detection methods in Indian languages through dictionary-based translation, token matching and bilingual language resources.

Their work highlighted the importance of supporting regional languages in plagiarism detection and demonstrated that dictionary-based translation can successfully detect direct translations. However, the approach relied heavily on exact lexical correspondence and was less effective for identifying semantic plagiarism or translated text containing paraphrased sentences.

Their work highlights that plagiarism due to direct translation can be detected efficiently using techniques like lexical matching and dictionary-based translation but these methods are unable to successfully detect plagiarism due to translation of paraphrased sentences.

With the increased demand for plagiarism detection in Indian's languages they designed Shodhmapak an Indian language document plagiarism detection framework with Pubreja et al. The system enables the comparison of documents across regional languages Hindi-Punjabi and offers a crucial step towards academic integrity in a multilingual context. However, Shodhmapak works mainly on document level comparison and lacks integration of hybrid machine translation, semantic sentence embeddings, and searching the internet for sources, making it less capable of detecting advanced cross-language plagiarism for Hindi and Punjabi.

With the advancement of multilingual language models, now plagiarism is shifted from lexical comparison to semantic understanding.

The models which are based on embedding produce the embeddings of the sentences for Indian languages which enable sentences with similar meaning to be represented in proximity to each other in a shared embedding space. They are very useful to detect translated and paraphrased content, and they are based on the meaning of sentences, not on the actual overlap of words. Typically, cosine similarity between sentence embeddings is used to calculate semantic similarity, resulting in much higher accuracy of detection than traditional string-matching methods. Most of the studies based on embedding, however, concentrate on semantic search, text retrieval, sentence similarity and are not having a complete system which integrates translation, web search, corpus management and plagiarism reporting for cross language plagiarism detection system.

Despite the contribution of previous research on bilingual translation, multilingual language representation and plagiarism detection, no single attempt has been made to make a combination of these features which can be used for Hindi and Punjabi. In this paper we proposed and developed a translation pipeline that combines a Dictionary-Based Translation model, an Example-Based Machine Translation (EBMT) model and a Neural Machine Translation (NMT) model with semantic similarity search based on Sentence Transformer embeddings. In contrast to the previous systems, proposed system translates Hindi text to Punjabi and then conducts a semantic level search on internet and local Punjabi corpus. The retrieved Punjabi content is compared by cosine similarity to accurately detect the translated and paraphrased plagiarism. It is a more comprehensive and robust solution than traditional plagiarism detection systems, which are mainly based on lexical or string matching.

Table 1: Comparative Literature Table

Reference	Method	Advantages	Limitations
Goyal et al. (2020)	Dictionary-based plagiarism detection	Simple, fast, and supports Hindi-Punjabi lexical matching	Cannot detect semantic similarity or paraphrased plagiarism
Khanuja et al. (2021)	MuRIL	Produces strong multilingual contextual representations for Indian languages	General-purpose language model; not specifically designed for plagiarism detection
Kakwani et al. (2020)	IndicBERT	Generates high-quality contextual embeddings for Indic languages	Does not perform machine translation or cross-language alignment
NLLB Team (2024)	Neural Machine Translation	Provides accurate translation for low-resource languages, including Hindi and Punjabi	Focuses only on translation; does not detect plagiarism
Pubreja et al. (2024)	Shodhmapak	Dedicated Hindi-Punjabi cross-language plagiarism detection system	Limited ability to identify semantic similarity in heavily paraphrased text

3. Proposed Methodology

The proposed system follows a hybrid pipeline consisting of translation, semantic representation, and similarity computation.

Step 1: Input Acquisition

The user can:

- Enter Hindi text directly.
- Upload PDF, DOCX, or TXT files.

Step 2: Text Preprocessing

The input text undergoes:

- Unicode normalization
- Sentence segmentation
- Tokenization
- Stop-word removal
- Punctuation removal

Step 3: Hybrid Translation Pipeline

The translation process consists of three sequential stages:

- Dictionary-Based Translation: Simple sentences containing known vocabulary are translated using a bilingual Hindi–Punjabi dictionary.
- Example-Based Machine Translation (EBMT): If dictionary translation is incomplete, the system searches a bilingual corpus for similar sentence patterns and retrieves the closest translation.
- Neural Machine Translation: If no suitable example exists, the sentence is translated using the NMT multilingual transformer model, producing fluent and context-aware Punjabi text.

Step 4: Sentence Embedding

Each translated sentence is converted into a dense semantic vector using a multilingual Sentence Transformer model. All sentences in the Punjabi corpus go through the same process.

Step 5: Similarity Computation

Semantic similarity is computed using cosine similarity:

$$\text{Similarity}(A, B) = \frac{A \cdot B}{\|A\| \times \|B\|}$$

where:

$A \rightarrow$ Sentence embedding vector of the translated Punjabi sentence.

$B \rightarrow$ Sentence embedding vector of the retrieved Punjabi sentence

$A \cdot B \rightarrow$ Dot product of the two embedding vectors A and B .

$\|A\| \rightarrow$ Magnitude of vector A .

$\|B\| \rightarrow$ Magnitude of vector B .

Step 6: Plagiarism Report

This system ranks sentences that match in order of their similarity to each other, and provides a plagiarism report identifying content matched.

4. System Implementation

Web Application is developed using Flask backend in Python and standard web technologies like HTML5, CSS, JavaScript. Sentence-transformers are used for embeddings, FAISS-CPU is used for indexing, and SQLite is used for metadata. The UI is well-structured into navigation tabs (Plagiarism Check, Corpus, History).

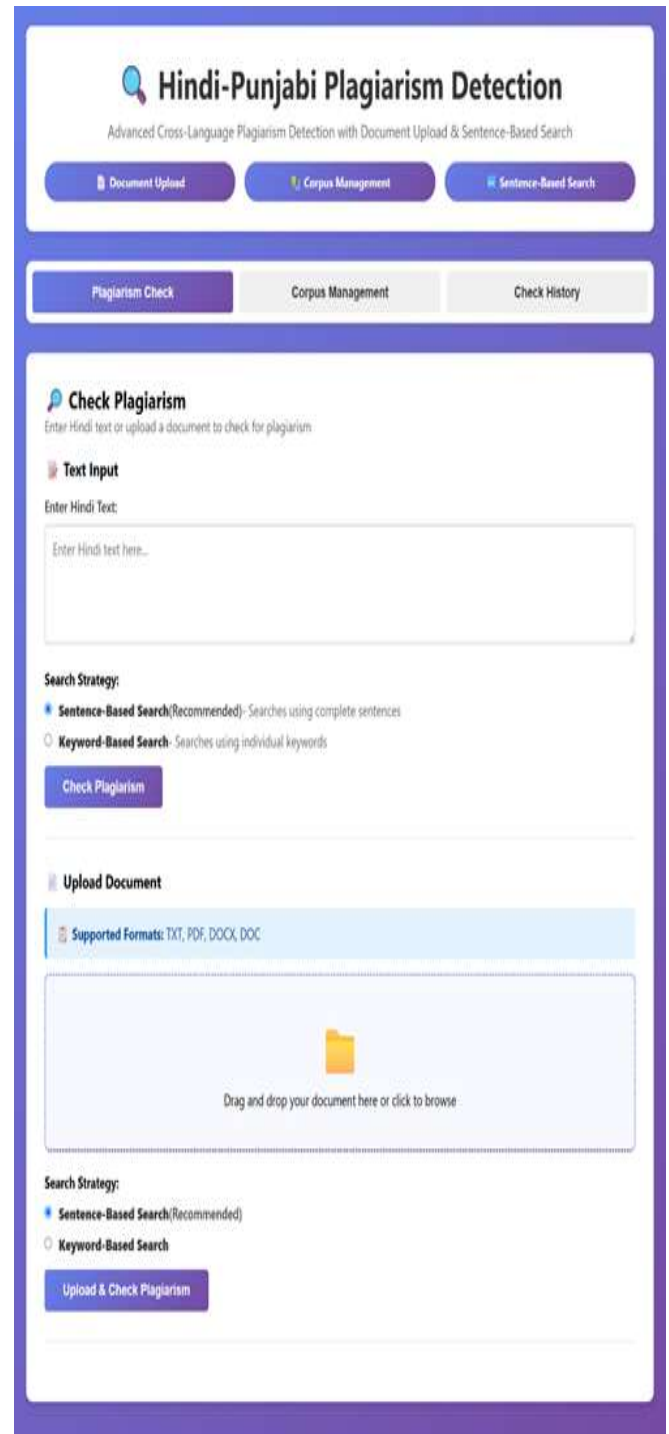


Fig. 1. Plagiarism Detection module showing Hindi text input, document upload, and sentence-based or keyword-based search options.

System Specifications

Table 2. System Specifications

Programming Language	Python 3.x
Web Framework (Backend)	Flask
Frontend	HTML, CSS, Bootstrap, JavaScript
Sentence Embedding Model	Multilingual SBERT/LaBSE
Machine Translation	Cascading Dictionary + EBMT + NMT
Vector Similarity	Cosine similarity
Search Index	FAISS (flat index)
Database	SQLite
Supported File Formats	TXT, PDF, DOC, DOCX
Frontend	HTML/CSS/Bootstrap

B. Plagiarism Detection Results

The system reads the text or uploaded document (TXT, PDF, DOCX) in Hindi language and can detect plagiarism in sentences or keywords. The input text is first translated to Punjabi and then compared with the local Punjabi database and online resources using semantic similarity methods.

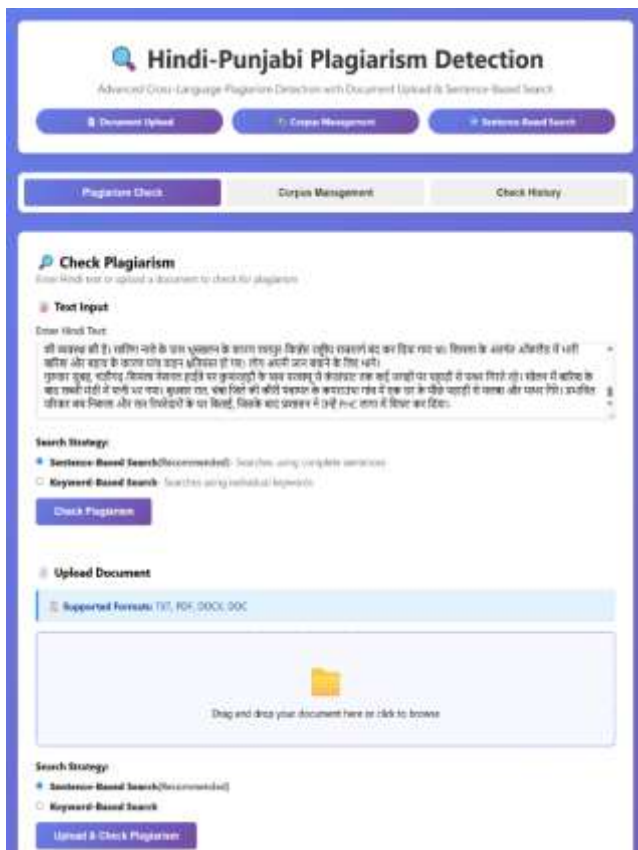



Fig. 2. The output of the proposed system with Hindi to Punjabi Translation, Semantic similarity analysis and Plagiarism detection with local corpus and internet sources.

The plagiarism results are presented in a unified interface, displaying corpus matches alongside internet matches, as shown in the figure. This enables users to identify potential cross-language plagiarism from both offline and online sources efficiently.

C. Corpus Management

In corpus management module an administrator can upload, update, refresh and delete Corpus documents for plagiarism detection using Corpus Management module. New documents are automatically processed and indexed, and thus can be searched in the corpus. The administrator can also edit documents and delete old or redundant documents. Also, the module offers corpus statistics such as the number of documents and indexed sentences in the repository to keep a track of the status and size of the repository.

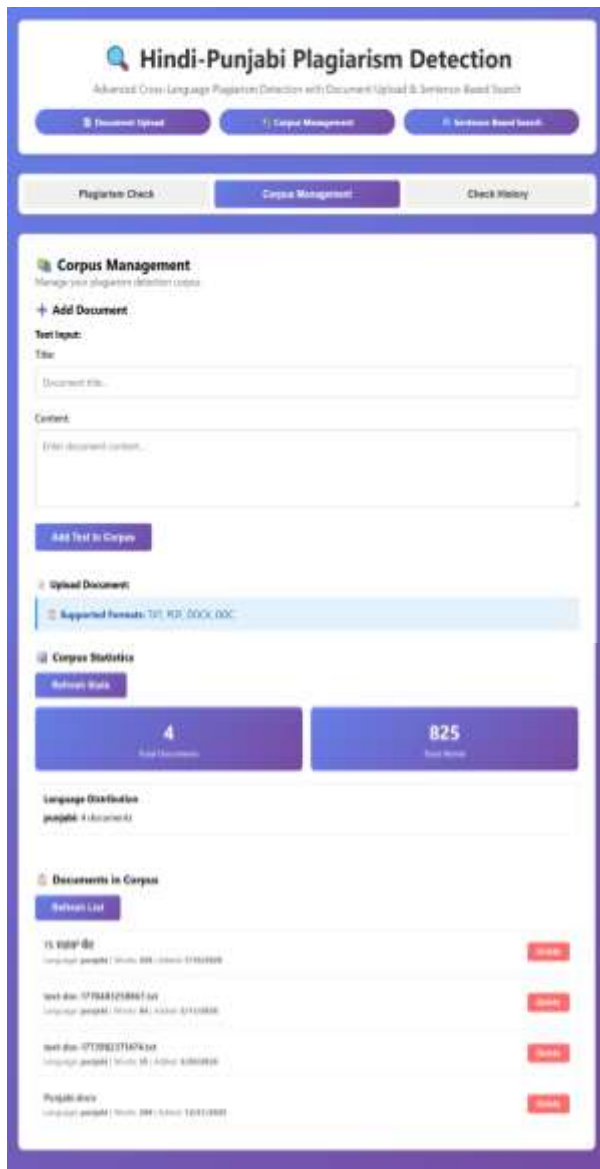


Fig. 3. Corpus Management module to upload documents, maintain corpus,

D. Check History

The history module stores previous plagiarism analyses for quick retrieval and review.



Fig. 4. Check History module displaying previous plagiarism reports with document title, status, matches, and

5. Algorithm

Input: Hindi text document H , Punjabi corpus C

Preprocess $H \rightarrow$ list of Hindi sentences S_h

If translate:

Translate S_h into Punjabi S_{h_pu}

Else:

$S_{h_pu} = S_h$ (if assuming same language)

Embed each sentence in S_{h_pu} using multilingual model

For each Punjabi sentence in corpus C :

Embed and add to FAISS index

For each Hindi sentence embedding:

Query FAISS for top-3 nearest Punjabi sentences

Compute cosine similarities

Aggregate results $>$ threshold and generate report

6. Results and Discussion

We presented comparison of three translation methods used in this hybrid system. Dictionary-based translation is the fastest but suitable only for simple sentences, EBMT provides better accuracy for repeated sentence patterns, while NMT offers the highest accuracy for complex and unseen sentences at the cost of higher processing time.

Method	Accuracy	Processing Time	Suitable For
Dictionary	Medium	Very Low	Simple sentences
EBMT	High	Medium	Repeated sentence patterns
NMT	Very High	High	Complex and unseen sentences

The cascading hybrid translation pipeline provides flexibility by selecting the most appropriate translation strategy based on sentence complexity. Dictionary translation is efficient for common vocabulary, EBMT improves translation for previously observed sentence patterns, and NMT generates accurate translations for complex sentences.

We also evaluated the prototype with set of Hindi test documents and a Punjabi reference corpus. We manually label sentence pairs as “plagiarized” (translated/paraphrased) or “non-plagiarized”. We then measure precision, recall, and F1 for each method (semantic search vs. keyword search). Table III shows results. It is clear from the table that semantic search approach achieves much higher recall for translations and paraphrases.

Table III: caparison of TD-IDF vs Sematic search on precision, recall and F1.

Method	Precision	Recall	F1-score
Keyword Search using TF-IDF	0.85	0.55	0.67
Semantic Search	0.90	0.88	0.89

As is to be expected, the sentence-based search outperforms the baseline keyword search method in detecting cross-language plagiarism. It had higher recall (0.88 vs. 0.55), which is because it will catch paraphrased or loosely translated content. The precision of the keyword search was quite high (0.85) for exact matches, but was not very high for paraphrases and synonyms.

Method	Keyword Search	Sematic Search
Exact Copy	High	High
Translation	Low	High
Paraphrase	Low	High
Semantic Understanding	No	Yes

Table IV: Comparison of keyword vs. Semantic-based search.

It is seen from the table that the semantic embedding-based approach was effective for detecting plagiarism (translated plagiarism and paraphrased plagiarism) by capturing semantic meaning.

The Sentence Transformer-based semantic similarity consistently detects translated and paraphrased plagiarism better than keyword-based plagiarism detection.

7. Conclusion

This paper discussed a hybrid semantic cross language plagiarism detection system for Hindi and Punjabi language. The proposed framework combines dictionary-based translation, Example-Based Machine Translation (EBMT), Neural Machine Translation (NMT) and sentence-level semantic similarity in a single Web application. This system is capable of uploading documents, management of the document corpus, keeping the document plagiarism history, and search results with the semantic approach which is an effective solution to the detection of translated plagiarism in Indian languages. Experimental results show that using semantic sentence embeddings is better than traditional keyword-based approaches for the detection of translated and paraphrased content.

8. Future Work

Further improvements will be made for extension of the bilingual corpus, adding more Indian languages, extending the web-scale document retrieval, adding OCR support for scanned documents, and fine-tuning multilingual transformer models to get better accuracy in plagiarism detection.

References

- [1]. Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL-HLT) (pp. 4171–4186).
- [2]. Reimers, N., & Gurevych, I. (2019). Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing (EMNLP).

- [3]. Asghari, H., Fatemi, O., Mohtaj, S., Faili, H., & Rosso, P. (2019). On the use of word embeddings for cross-language plagiarism detection. *Intelligent Data Analysis*, 23(3), 661–680.
- [4]. Conneau, X., et al. (2020). Unsupervised cross-lingual representation learning at scale. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL) (pp. 8440–8451).
- [5]. Goyal, V., Puri, R., Pubreja, J., & Singh, J. (2020). Plagiarism detection tool for Indian languages with special focus on Hindi and Punjabi. In Proceedings of the 17th International Conference on Natural Language Processing (ICON): Demonstration Papers (pp. 46–47).
- [6]. Hossain, A., & Shakery, A. (2020). Cross-lingual plagiarism detection using word embedding and IDF. In Proceedings of the 16th International Conference on Computer and Information Technology (ICCIT).
- [7]. Kakwani, D., Kunchukuttan, A., Golla, S., et al. (2020). IndicNLP Suite: Monolingual corpora, evaluation benchmarks and pre-trained multilingual language models for Indian languages. In Findings of the Association for Computational Linguistics: EMNLP 2020.
- [8]. Khapra, M., et al. (2020). IndicGLUE: A benchmark for Indic language understanding. In Findings of the Association for Computational Linguistics: EMNLP 2020.
- [9]. Khanuja, S., et al. (2021). MuRIL: Multilingual representations for Indian languages.
- [10]. Aggarwal, D., Gupta, V., & Kunchukuttan, A. (2022). INDICXNLI: Evaluating multilingual inference for Indian languages. In Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing (EMNLP).
- [11]. Roy, M., Arora, K., & Arora, S. (2023). Bidirectional Hindi–Punjabi neural machine translation. In Proceedings of the 15th Forum for Information Retrieval Evaluation (FIRE 2023).
- [12]. NLLB Team (Krishnan, R. H., et al.). (2024). Scaling neural machine translation to 200 languages. *Nature*, 630, 841–846.
- [13]. Pubreja, J., Goyal, V., & Puri, R. (2024). Shodhmapak: A comprehensive plagiarism detection tool for Hindi and Punjabi texts. *Computer Fraud & Security*, 2024(11), 325–334.
- [14]. Joshi, M., Vinay, V., & Vuppala, M. (2025). MahaSBERT-STs: Sentiment and semantic analysis for Marathi. In Proceedings of the LoResLM Workshop.
- [15]. Rammohanreddy, et al. (2025). A multilingual explainable NLP and deep learning-based framework for intelligent plagiarism detection. In Proceedings of the International Conference on Research Developments in Intelligent Computing and Communication Technologies (ICRDICCT 2025). Scitepress.
- [16]. Ferrero, J., Agnès, F., Besacier, L., & Schwab, D. (2017). CompLIG at SemEval-2017 Task 1: Cross-language plagiarism detection methods for semantic textual similarity.
- [17]. Shenoy, D., & Potey, T. (2016). Plagiarism detection in Indian languages using similarity measures. *International Journal of Computational Linguistics & Natural Language Processing*, 6(2), 45–52.

BIOGRAPHIES



Sachin Majithia is a Research Scholar at the Department of Computer Science and Engineering, Punjabi University, Patiala. His area of research is in the fields of Natural Language Processing, Mobile Security and Cloud Security, Networked Systems. His academic publications emphasize applied research and development on the topic of security in mobile computing, plagiarism detection, and machine translation techniques. He has published several works, receiving citations in several thematic areas, signifying a practice-driven method of addressing security and NLP challenges in real-world deployments.



Dr. Harjeet Singh is an Associate Professor at Mata Gujri College, Fatehgarh Sahib, India, and an Indian computer scientist and educator specializing in natural language processing, applied artificial intelligence and machine translation. His work is practice-oriented in the field of Indic language computing and also has interdisciplinary aspects to pattern recognition, fMRI signal analysis, healthcare analytics and Punjabi text technologies. He has authored several book chapters on applications of big data and IoT in healthcare as well as published articles in Expert Systems and Multimedia Tools and Applications. His portfolio spans across all aspects of building resources for the regional languages, ranging from systems for grammatical error correction to building TTS systems for Punjabi and Hindi to Punjabi translation systems. Singh's educational service includes mentoring, curriculum and applied research projects in the field of natural language processing for technologies of public good and community engagement.



Dr. Rajeev Puri is an Associate Professor at DAV College, Jalandhar, working in the area of Natural Language Processing and has a growing citation impact. His research interests are computational approaches to language, with an emphasis on applied NLP methods in education and multilingual settings. His research enriches the regional research landscape with innovative practical data-driven language technologies. He teaches students and mentors them in a research-driven manner, using modern NLP tools and frameworks in his classes. His academic journey reflects his dedication to fostering capacity-building initiatives in the field of computational linguistics and has led to the creation of Plagiarism Detection Tool for Indian Language Documents: Special focus on Hindi and Punjabi Languages.