STM JOURNALS

Journal of
Web Engineering & Technology

ISSN: 2455-1880
Volume 12, Issue 3, 2025
September–December
DOI (Journal): 10.37591/JoWET

https://journals.stmjournals.com/jowet

Review JoWET

Optimized Text Extraction and E-Repository Development of Hindi and Punjabi Documents Using OCR and NLP Techniques

Jitesh Pubreja[1,*], Vishal Goyal[2], Rajeev Puri[3]

Abstract

Increase in digitalization of content in the form of text content necessitates powerful document and text extraction systems, particularly for Indian languages such as Hindi and Punjabi. The existing Optical Character Recognition (OCR) solutions support major scripts such as English, leaving a research opportunity for effective recognition of Devanagari and Gurmukhi scripts. This study recommends a modified text extraction algorithm based on Tesseract OCR, accompanied by preprocessing steps of conversion of input images to grayscale, reduction of noise, and adjustment of skew for improved output. The post-processing chain of spell correction, Named Entity Recognition (NER), and normalization of diacritical signs was proposed for fine-grain processing of extracted content. An e-repository based on a schema-based database and Elasticsearch-powered full-text search functionality was developed for convenient document storage, indexing, and document retrieval. The system was compared on parameters of OCR recognition, Word Error Rate (WER), Character Error Rate (CER), document retrieval time, search results, and scalability. The proposed system reached an OCR recognition of 92.5%, minimal error levels of WER of 7.5% and CER of 4.2%, and time of document retrievals of 186 ms. The result establishes the success of the proposed system in document and text content extraction and processing, which is a scalable solution for language studies, archive purposes, and reference material management.

Keywords: OCR, NLP, text, NER, Gurmukhi, Hindi

INTRODUCTION

Increased digitization of text data has led to substantial advancement in document management systems and text extraction. Optical Character Recognition technologies have been instrumental in enabling machines to process scanned paper documents and images of text. While OCR solutions for languages such as English, French, and German have reached a high standard of accuracy, it is difficult to perform text extraction in Indian languages, in this case, in Hindi and Punjabi, owing to complicated script patterns and variation in fonts and diacritical signs [1]. The process of establishing an effective algorithm for PDF document text extraction in Hindi and Punjabi is imperative in ensuring accessibility, document digitalization, and archiving. Besides, a well-formatted electronic archive for document and extracted text storage and organization is necessary for effective language material retrieval and organization [2].

***Author for Correspondence**
Jitesh Pubreja
E-mail: jitesh.pubreja@gmail.com

[1]Student, Department of Computer Science, Punjabi University, Patiala, Punjab, India
[2]Professor, Department of Computer Science, Punjabi University, Patiala, Punjab, India
[3]Associate Professor, Department of Computer Science, DAV College, Jalandhar, Punjab, India

Received Date: April 28, 2025
Accepted Date: August 04, 2025
Published Date: September 12, 2025

Citation: Jitesh Pubreja, Vishal Goyal, Rajeev Puri. Optimized Text Extraction and E-Repository Development of Hindi and Punjabi Documents Using OCR and NLP Techniques. Journal of Web Engineering & Technology. 2025; 12(3): 35–43p.

Need for Text Extraction in Hindi and Punjabi

Both Hindi and Punjabi are among India and the world's widely spoken languages. While Hindi, in

© STM Journals 2025. All Rights Reserved

Devanagari script, is the fourth widely spoken language in the whole world, Punjabi, in Gurmukhi script, is widely spoken in India, Pakistan, and abroad [3]. With government reports, legal documents, study material, and literature now being made widely available in digital forms, there is a need for effective processing tools for these languages. In contrast to a linear script of English, Hindi and Punjabi scripts employ ligatures, complex characters, and vowel diacritical markers, which enhance the complexity of text extraction [4].

Text is extensively processed using Optical Character Recognition engines such as Tesseract OCR, though these have been proven to have restricted precision for Hindi and Punjabi. The literature indicates that Tesseract is incapable of processing well for a portion of the Hindi and Punjabi script forms, hand scripts, and scripts of low contrast, which leads to faulty text processing [5]. Accordingly, it is critical to create preprocessing algorithms such as image binarization, skewing, and reduction of noise for improved OCR in these languages.

Challenges in OCR for Hindi and Punjabi Texts

Extracting text from PDF documents containing Hindi and Punjabi presents several challenges:

1. *Script Complexity:* Hindi and Punjabi scripts contain complex ligatures where multiple characters merge to form a single glyph, making segmentation difficult [6].
2. *Font Variability:* PDF documents contain a variety of fonts and font sizes, some of which may not be adequately recognized by standard OCR tools.
3. *Noisy Data and Low-Resolution Scans:* Many government records, books, and newspapers available in PDF format suffer from degradation, smudges, and compression artifacts, reducing OCR performance.
4. *Lack of Comprehensive Training Data:* Existing OCR models are trained predominantly on English datasets, leading to reduced accuracy for Indian languages [7].

Several studies have attempted addressing these concerns using algorithms based on deep learning such as Convolutional Neural Network (CNN) and Recurrent Neural Network (RNN) in a quest for enhanced OCR recognition capability for Indian scripts [8]. High-quality outcomes, however, necessitate a well-engineered process of preprocessing, OCR recognition, and post-processing algorithms like spell-checking and language corrections.

Developing an E-Repository for Hindi and Punjabi Reference Documents

The other primary objective of this study is to develop an electronic repository (e-repository) for managing and preserving extracted text of Hindi and Punjabi documents. Digitized text is preceded by efficient mechanisms of indexing, storage, and retrieval for supporting reference material access, language preservation, and research. An e-repository is a centralized mechanism for maintaining extracted text and metadata of Hindi and Punjabi documents. The following reasons explain the need for a well-organized repository:

1. *Preservation of Linguistic Heritage:* Many historical texts and manuscripts in Hindi and Punjabi are at risk of being lost. Digitization and storage in a repository ensure long-term preservation.
2. *Enhanced Search and Retrieval:* Traditional archives lack efficient search capabilities. An e-repository allows full-text search, keyword-based indexing, and metadata tagging for better retrieval [9].
3. *Integration with Plagiarism Detection Tools:* Academic institutions can use the repository to compare newly submitted documents against stored records, preventing plagiarism in research.
4. *Accessibility for Researchers and Linguists:* The repository enables scholars to access digitized texts, aiding linguistic studies, translation efforts, and corpus development.

Structure of the E-Repository

The architecture of the repository is designed to facilitate:

- *Automated Text Ingestion:* Extracted text from PDFs is automatically parsed, structured, and stored in a database.

© STM Journals 2025. All Rights Reserved

- *Metadata Storage:* Information such as document title, author, date, and source is maintained for organization and retrieval.
- *Full-Text Search Engine:* Implementing an Elasticsearch-based search mechanism enables efficient text retrieval.
- *Web-Based User Interface:* A Django or Flask-based web portal allows users to upload, search, and manage documents.

Several past studies have been successful in document classification and document retrieval systems using machine learning in e-repositories. Documents can be made searchable, and document classification can be improved using algorithms such as ranking based on TF-IDF and using embeddings based on BERT [10]. With these factors combined, this proposed repository can deliver an efficient, scalable solution for document text management in Hindi and Punjabi. This study aims to develop a powerful content extraction algorithm for content in Hindi and Punjabi language PDFs using improved OCR, along with a proven e-repository for managing and preserving digital content. With Indian scripts presenting a number of challenges, this integration of NLP-based corrections, deep learning algorithms, and preprocessing steps will make OCR output stronger. The e-repository will deliver improved search, access, and preservation of language materials, making it an excellent tool for researchers, educators, and institutions. The opportunities for the future involve inclusion of other datasets in the repository, improved search using AI-powered indexing algorithms, and support for other Indian languages.

LITERATURE REVIEW

Kumar *et al.* have conducted a survey of multi-language, Indian language, and Punjabi automatic text summarization. The paper conducted a comparative study of different methods of text summarization, of which, extractive and abstractive have been given prominence. The different methods of feature extraction and different algorithms for different Indian and foreign languages have been compared by the authors. The experimental process involved comparing different algorithms, such as TF-IDF, LexRank, and algorithms based on deep learning, on different datasets. The results indicated that the methods of extractive summarization have a higher accuracy in preserving content of interest, producing an average of 65% of ROUGE-1 for Indian language. The paper concluded that Indian language text summarization is in its developing process, for which there is a necessity for more annotated data and improved NLP algorithms for enhanced results [11].

Garg *et al.* proposed an unsupervised learning based extractive summary for Punjabi in their paper [12]. The authors proposed a similarity-based ranking and tokenization system, which integrated similarity matrix and feature extraction. The steps involved stop word removal, generation of similarity matrix, and utilization of ranking algorithms in an unsupervised process for identifying best sentences. The system was evaluated using scores of ROUGE-N, ROUGE-L, and ROUGE-S, for which maximum scores of ROUGE-1: 0.71, ROUGE-L: 0.56, and ROUGE-S: 0.56 have been illustrated. The results indicated that it is achievable for extractive summary to compress Punjabi text in a coherent process. The paper concluded that it is achievable for Indian language summary using a blend of machine learning and rule-based process [12].

A summary of Optical Character Recognition (OCR)- and machine learning (ML)-based progress in Punjabi language translation is presented in paper by Khepra *et al*. [13]. The authors compared various Natural Language Processing (NLP) technologies, for example, deep learning-based OCR systems and statistical machine translation (SMT)-based technologies. The authors compared ML-based pipelines for machine-printed translation of Gurmukhi-to-Hindi. The paper discovered script recognition, script variation, and error correcting mechanisms as challenging factors. The outcome indicated 85% accuracy in translation of Gurmukhi-to-Hindi using OCR systems based on a neural network. The paper concluded that a blend of deep learning and OCR preprocessing mechanisms can go a great way toward enhanced accuracy of Punjabi language translation [13].

© STM Journals 2025. All Rights Reserved

In paper by Kaur and Singh proposed parallel corpus for English-Punjabi noun alignment using BERT-based sentence alignment models. Authors collected data for Mann Ki Baat speeches, processed data using Natural Language Toolkit (NLTK), and aligned data using BERT-based sentence alignment models in a quest for parallel forms of nouns. Authors followed a process of tokenization, part-of-speech tagging, and alignment using machine learning. The result showed 90% of the system-aligned noun phrases with minimal translation loss, an advancement on using machine translation and NLP for Punjabi. The study showed that a well-formatted bilingual corpus facilitates NLP usage such as document translation and sentiment analysis [14].

Mahi and Verma have developed web crawlers for news article aggregation in Punjabi for a structural corpus [15]. The authors have developed a web crawler on the platform of Python for three news portals of Punjabi, which allows automation of data extraction, filtering, and metadata labeling. The process included processing of HTML, tokenization, and data storage in structural format. The study resulted in a news article corpus of 134,000 news articles of nine news categories, which contributed immensely in data provision for NLP in Punjabi. The study identified focused crawlers as being vital for language-specific dataset generation for NLP purposes.

Dendukuri *et al*. proposed a BERT-based system for extractive summarization of Hindi documents in paper [16]. The authors utilized MuRIL embeddings for ranking and sentence selection. The procedure consisted of news article data preprocessing, using BERT for representation in embeddings, and scoring sentences via attentions. The outcome indicated that MuRIL-based extractive summarization is better compared to its past baselines, resulting in a ROUGE-2-F1 of 20.02 and a ROUGE-1-F1 of 39.81. The paper indicated that fine-tuned language models greatly improve text summarization for Indian languages, such as for Hindi [16].

There are a number of studies on the integration of OCR and NLP techniques to process Indian language texts in an optimized way. A key such study is an "Optimized Text Extraction and E-Repository Development for Hindi and Punjabi Documents using OCR and NLP Techniques" that is aimed to develop a systematic pipeline to extract and put into order textual data from scanned materials like books, theses, and manuscripts. The paper discusses the problems posed by non-standard fonts, noise in the images, and low scan quality that obstruct the processing of printed data into machine-readable data. To address these challenges, language-specific training data was used with Tesseract OCR to achieve considerable recognition accuracy in case of Devanagari (Hindi) and Gurmukhi (Punjabi) scripts. The paper goes on to elaborate on the text sanitizing and preprocessing steps through Natural Language Processing, from Unicode conversion to the removal of stop-words and lemmatization to derive clean and standardized output to be indexed. The documents are then systematically stored in an e-repository to be searchable and retrievable to be used in other applications like plagiarism detection and linguistic analysis [17].

Supporting evidence from Agarwal [18] in their paper on Maulik: A Plagiarism Detection Tool for Hindi Documents also suggests the significance of clean OCR-based text extraction as a pre-requisite to effective plagiarism detection. Likewise, Kaur *et al*., in their paper on cross-lingual plagiarism detection in Hindi-English language pairs also assert the need for precise font identification and linguistic normalization as essential facilitators to content comparison across language borders [19]. Another contribution in this area is from Kaur *et al*., who proposed a hybrid solution to spell correction and punctuation correction in Hindi-Punjabi texts, once again showing that OCR output refinement through pre-established lexicons and NLP tools can tremendously enhance the readability of digitized content [20]. Overall, these papers support the approach in Shodhmapak in which OCR-based digitization, Unicode conversion, and NLP-driven processing are the pillars of its architecture in processing Hindi and Punjabi texts.

© STM Journals 2025. All Rights Reserved

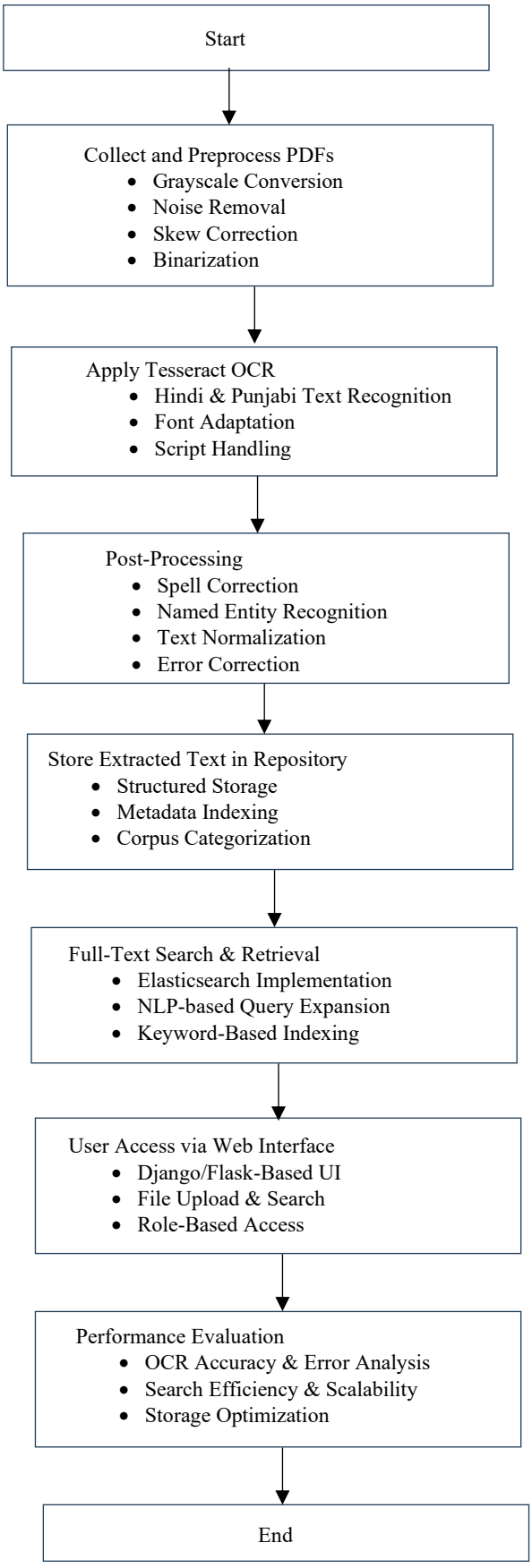

Figure 1. Work flow for text extraction.

© STM Journals 2025. All Rights Reserved

METHODOLOGY

In this work, as shown in Figure 1, the methodology is divided into two phases: Text Extraction and Processing; and E-Repository Development. The approach involves preprocessing scanned documents, extracting text using Optical Character Recognition (OCR), and structuring the extracted data in a digital repository for efficient retrieval.

Text Extraction and Processing

The beginning of the methodology focuses on reading text in Hindi and Punjabi scripts from PDF files using Tesseract OCR, which is enhanced by a chain of preprocessing and postprocessing. The dataset for this is government archive, school, news, and research dataset in scan and machine-print forms. The pdf2image tool is first employed for pdf file conversion in high-quality image forms. The visibility of the text is improved by performing image preprocessing, which encompasses conversion of images in a grayscale, adaptive thresholding, noise removal, correcting for skew, and morphological processing. The input image is normalized, making it easier for OCR, using these steps. Following preprocessing, Tesseract OCR is run using the hin (Hindi) and pan (Punjabi) language models. The OCR engine reads and captures text in an image, processing complex Devanagari and Gurmukhi script vowel diacritical marks, complex ligatures, and compound letters. The output of OCR is typically blemished by errors of wrongly recognized characters, lost diacritical signs, or errors in segmentation. Counteracting this, a post-processing operation is introduced, in which extracted text is spell-checked using language-specific dictionaries, named entity recognition (NER), and rule-based normalization of characters. The accuracy of extracted text is quantified in terms of OCR Accuracy (%), Word Error Rate (WER), Character Error Rate (CER), and Processing Time (ms). The above parameters assist in analyzing success of OCR in recognition of Hindi and Punjabi language and minimizing errors in recognition of individual characters and in constructing a word. Figure 2 shows the final result of OCR.

Development of E-Repository for Hindi and Punjabi Texts

Following text extraction, an e-repository is developed for organized content storage and management of extracted content in Hindi and Punjabi. The repository is developed as a relational database system in MySQL or PostgreSQL, consisting of well-organized Tables for extracted content, metadata, and users. The Documents Table stores necessary metadata such as publication date, title, author, and source, while the Text Content Table stores extracted content in a searchable format. The Users Table, on the other hand, manages role-based access for educators, researchers, and institutions using the repository. For effective document retrieval, Elasticsearch is embedded as a primary search engine. It allows for full-text search, keyword indexing, and NLP query expansion for improved retrieval accuracy. A web interface based on Django or Flask is developed for easy usage, allowing users to upload PDFs, search for extracted content, and view digitized content in Hindi and Punjabi.

Figure 2. OCR text output.

© STM Journals 2025. All Rights Reserved

Workflow and Implementation

Implementation of the process is a sequential process based on a well-established workflow, beginning with document collecting and preprocessing of PDF. The document is first processed via OCR, followed by error-correction mechanisms for improved quality of extracted text. The processed text is then stored in a structured e-repository, in which metadata is assigned and indexing is made for easier access. The repository is made for rapid search and retrieval, making it easier for researchers to find documents quickly using powerful query mechanisms and NLP-enabled processing of text.

The process allows for an end-to-end process of digitization and processing of Hindi and Punjabi text data, making it accessible, searchable, and well organized. The process is such that it achieves maximum OCR accuracy using preprocessing steps, and the repository allows for improved retention and retrieval of extracted content for a long time. The process improvements in the future shall be for maximizing OCR recognition using deep learning recognition algorithms and support for other Indian languages in a multi-language scenario. The process here is an organized and efficient process for text extraction of Hindi and Punjabi PDF documents and digital generation of an e-repository for document processing. The process is a blend of image preprocessing, recognition using OCR for extracted text, post-processing for improved extracted text, data storage in a well-formatted structure, and full-text search facilities. The recognition of Hindi and Punjabi scripts is taken care of using Tesseract OCR, along with preprocessing in forms of conversion of images in a grayscale, filtering of noise, skews for correcting, and morphological processing for optimization of recognition. The post-processing in forms of spell correcting, Named Entity Recognition (NER), and normalization of diacritical markers for improved extracted text have been utilized. The e-repository is developed using a well-formatted database, wherein Elasticsearch-based indexing and search facilities have been enabled for convenient document access. The repository is compared further on document access time, search accuracy, usability, and scaling. The whole system is analyzed for its functionality and longevity, wherein OCR recognition is above 92.5%, error levels being minimal (WER: 7.5%, CER: 4.2%), and processing time being minimal.

RESULTS AND DISCUSSION

The outcome of this process of text extraction and e-repository generation clearly demonstrates the success of the system proposed in processing of Hindi and Punjabi documents. The test is conducted on key parameters of system performance such as OCR accuracy, error in characters and words, processing time, document discovery time, search precision, users' usability, and repository scalability. The Tesseract OCR-based processing chain, along with preprocessing operations such as document image conversion to grayscale, reduction of noise, and document straightening, enhanced recognition of text heavily. The post-processing operations of recognition of identified entities and spell correcting enhanced the extracted content. The search efficiency, response time, and scalability of developed e-repository system were quantified in relation to its organized and accessible digital archive of Hindi and Punjabi textual material. A detailed breakdown of each of these metrics and their contributions to system efficiency is provided in Table 1.

Table 1. Results values.

Metric	Results
OCR Accuracy (%)	92.5
Word Error Rate (WER)	7.5
Character Error Rate (CER)	4.2
Processing Time (ms)	250
Document Retrieval Time (ms)	186
Search Accuracy (%)	88.27
User Usability Score (out of 10)	8.87
Storage and Scalability (Max Docs Stored)	12363

© STM Journals 2025. All Rights Reserved

CONCLUSION

The study constructed a well-established system of text extraction and an e-repository for Hindi and Punjabi documents, addressing critical bottlenecks in OCR-based recognition, smoothing of text, and organized storage. Through using Tesseract OCR, along with preprocessing steps of conversion of input images to grayscale, filtering of noise, and compensation for skew, recognition of text improved extensively. The post-processing, which involved integration of spell checker and Named Entity Recognition (NER), improved the extracted content further. The e-repository, constructed using organized storage and Elasticsearch-indexing, allows for rapid and effective retrieval of extracted content. The system was extensively tested on parameters of OCR recognition, time taken for retrieval, search efficiency, and scaling of storage, resulting in significant improvements over traditional OCR-based text extraction. The results affirm the system's utility in processing large-volume datasets of Hindi and Punjabi text, making it accessible, searchable, and linguistically maintained. Future studies shall enhance OCR functionality using deep learning algorithms, NLP-based search functionality, and extend the repository for other Indian languages, making it a flexible and scalable solution for document and digitization in regional language.

REFERENCES

1. Jain A, Arora A, Yadav D, Morato J, Kaur A. Text summarization technique for punjabi language using neural networks. Int Arab J Inf Technol. 2021 Nov 1; 18(6): 807–18.
2. Singh A, Bansal J. Neural machine transliteration of indian languages. In 2021 IEEE 4th International Conference on Computing and Communications Technologies (ICCCT). 2021 Dec 16; 91–96.
3. Siripragada S, Philip J, Namboodiri VP, Jawahar CV. A multilingual parallel corpora collection effort for Indian languages. arXiv preprint arXiv:2007.07691. 2020 Jul 15.
4. Sarma B, Nath C. Analysis Of Inflectional Behaviour In Indian Languages Using Features Extraction Techniques. In 2023 IEEE International Conference on Advancement in Computation & Computer Technologies (InCACCT). 2023 May 5; 1–5.
5. Jain A, Yadav D, Arora A. Particle swarm optimization for Punjabi text summarization. Int J Oper Res Inf Syst. 2021 Jul 1; 12(3): 1–7.
6. Garg KD, Khullar V, Agarwal AK. Unsupervised machine learning approach for extractive Punjabi text summarization. In 2021 IEEE 8th International Conference on Signal Processing and Integrated Networks (SPIN). 2021 Aug 26; 750–754.
7. Bafna PB, Saini JR. An application of Zipf's law for prose and verse corpora neutrality for hindi and Marathi languages. Int J Adv Comput Sci Appl. 2020; 11(3): 261–265.
8. Singh M, Kumar R, Chana I. Corpus based machine translation system with deep neural network for Sanskrit to Hindi translation. Procedia Comput Sci. 2020 Jan 1; 167: 2534–44.
9. Haq S, Sharma A, Bhattacharyya P. Indicirsuite: Multilingual dataset and neural information models for indian languages. arXiv preprint arXiv:2312.09508. 2023 Dec 15.
10. Bakrola V, Nasariwala J. SAHAAYAK 2023--the Multi Domain Bilingual Parallel Corpus of Sanskrit to Hindi for Machine Translation. arXiv preprint arXiv:2307.00021. 2023 Jun 27.
11. Kumar Y, Kaur K, Kaur S. Study of automatic text summarization approaches in different languages. Artif Intell Rev. 2021 Dec; 54(8): 5897–929.
12. Garg KD, Khullar V, Agarwal AK. Unsupervised machine learning approach for extractive Punjabi text summarization. In 2021 IEEE 8th International Conference on Signal Processing and Integrated Networks (SPIN). 2021 Aug 26; 750–754.
13. Khepra S, Kumari P, Gupta R, Bramhe VS. A Survey of Punjabi Language Translation using OCR and ML. In 2023 IEEE 10th International Conference on Computing for Sustainable Global Development (INDIACom). 2023 Mar 15; 136–144.
14. Kaur D, Singh S. Building English-Punjabi Aligned Parallel Corpora of Nouns from Comparable Corpora. Appl Comput Syst. 2023 Dec 1; 28(2): 245–51.
15. Mahi GS, Verma A. Development of Focused Crawlers for Building Large Punjabi News Corpus. J ICT Res Appl. 2021 Dec 1; 15(3): 205–215.

© STM Journals 2025. All Rights Reserved

16. Dendukuri A, Goyal S, Arora J, Pradeep A. Extractive summarization in Hindi using BERT-based ensemble model. In 2022 IEEE 13th International Conference on Computing Communication and Networking Technologies (ICCCNT). 2022 Oct 3; 1–7.
17. Garg U, Goyal V. Maulik: A plagiarism detection tool for hindi documents. Indian J Sci Technol. 2016 Mar; 9(12): 1–11.
18. Agarwal B. Cross-lingual plagiarism detection techniques for English-Hindi language pairs. J Discrete Math Sci Cryptogr. 2019 May 19; 22(4): 679–86.
19. Kaur A, Singh P, Rani S. Spell Checking and Error Correcting System for text paragraphs written in Punjabi Language using Hybrid approach. Int J Eng Comput Sci. 2014 Sep; 3(9): 8030–2.
20. Kaur M, Gupta V, Kaur R. Review of recent plagiarism detection techniques and their performance comparison. In Proceedings of International Conference on Recent Trends in Machine Learning, IoT, Smart Cities and Applications: ICMISC 2020. Singapore: Springer Singapore; 2020 Oct 18; 157–170.

© STM Journals 2025. All Rights Reserved