ISSN: 3139-0862

AI-Based Plagiarism Detection System Using Deep Learning

Prof. Shital Patil, Komal Sonawane, Sakshi Jaiswal, Sanskar Sonawane, Karishma Verma
Department of Computer Engineering, Watumull Institute of Engineering and Technology, Mumbai University, India
komalsonawane816@gmail.com, sakshijaiswal037@gmail.com

Abstract
Plagiarism is a concern in educational institutions because students may use existing work without giving proper credit. Detecting rewritten text and processing scanned or handwritten documents can be challenging for plagiarism detection methods. This paper proposes an AI-based plagiarism detection system that combines Natural Language Processing (NLP) deep learning and document text extraction. The system provides portals for students and teachers. Students can submit documents scanned files and handwritten assignments in supported formats. Text extraction and Optical Character Recognition (OCR) are used to convert document content into text for analysis. The extracted text is examined using matching and semantic similarity techniques to identify possible similarities with reference documents and previous student submissions. The proposed system also provides teacher dashboards for reviewing student records submitted assignments and plagiarism reports. These reports are intended to help teachers examine matching passages and assess the submitted content. The systems performance will be evaluated using accuracy, precision, recall and F1-score. The main objective is to support integrity by improving the analysis of textual similarities including cases where the wording differs but the meaning remains similar.

Keywords: Artificial Intelligence, Deep Learning, Plagiarism Detection, Natural Language Processing, Optical Character Recognition, Handwritten Document Recognition, Sentence-BERT, Semantic Similarity.

I. INTRODUCTION

The availability of books, research articles, websites and educational resources has made it easier for students and researchers to access information. However, using existing material without giving credit can lead to plagiarism and raise concerns about academic integrity [9].

Conventional plagiarism detection methods mainly examine similarities between a submitted document and existing text. They can identify copied words, phrases and passages. Their effectiveness may be limited when the original content is rewritten or its sentence structure is changed [3] [8]. As a result, comparing words may not be sufficient to identify every case of possible plagiarism.

Document format is another consideration. Scanned pages and handwritten assignments cannot always be compared directly with text. Optical Character Recognition (OCR) and other text extraction techniques can help convert document content into a machine- form before further analysis [4]. However, recognition errors may affect the quality of the extracted text.

Natural Language Processing (NLP) and deep learning provide methods for analyzing textual content. Transformer-based models can represent information while Sentence-BERT generates sentence embeddings that can be compared using cosine similarity to estimate semantic similarity [1] [2]. These techniques can help identify passages that express ideas despite differences in wording. However, a high similarity score alone does not establish that plagiarism has occurred.

This paper proposes an AI-based plagiarism detection system that combines document text extraction, textual matching and semantic similarity analysis. The application is designed with portals for students and teachers. Students can submit supported document formats while teachers can access submitted assignments and review generated similarity reports. The proposed system aims to make document comparison more informative and assist teachers in examining similar content.

The system will be evaluated using accuracy, precision, recall and F1-score. Further evaluation will be needed to assess how effectively the approach handles rewritten passages and how document extraction quality influences the results.

II. LITERATURE SURVEY

Plagiarism detection has evolved from word matching to machine learning and deep learning methods. Past research provides a foundation for building systems that compare documents and identify passages. [3] [8]

2.1 Traditional Text-Matching Techniques

Traditional methods identify similarities by comparing words, phrases and sentences across documents. String matching and word-overlap techniques are particularly useful for detecting copying when the wording remains similar [3].

TF-IDF represents documents according to the importance of the terms they contain. Cosine similarity can then be applied to compare the resulting document vectors and estimate their similarity [9].

When a paragraph is rewritten with words but retains its original meaning word-based techniques may report a lower similarity score. This makes it difficult to identify some forms of paraphrased content using comparison alone [9].

2.2 Machine Learning Approaches

Support Vector Machines (SVM) and Logistic Regression can be used to classify pairs of text by examining selected features that indicate similarities or differences between them. These features might include word overlap, sentence structure or other text characteristics.

ISSN: 3139-0862

The effectiveness of machine learning models depends on the availability of labelled training examples the relevance of the selected features, the size of the dataset and the model training procedure.

2.3 Deep Learning for Plagiarism Detection

Deep learning techniques can learn language patterns. Recurrent Neural Networks (RNNs) and Long Short-Term Memory (LSTM) networks are used to process data and analyze language.

Transformer-based models can capture relationships in text. BERT considers information from both the right sides of a word to create its representation supporting tasks such as text classification and semantic analysis [1].

2.4 Sentence-BERT and Semantic Similarity

Reimers and Gurevych proposed Sentence-BERT, which converts sentences into representations called embeddings. Cosine similarity can be used to compare these representations and estimate how closely the sentences are related in meaning [2].

Sentence-BERT can identify passages that express ideas even when they contain different words. This makes it useful for examining paraphrased text that conventional word-matching techniques may overlook [2].

A high semantic similarity score does not necessarily indicate plagiarism. Two passages may be similar because they discuss the subject, present established facts or explain commonly accepted definitions.

2.5 Paragraph-Level Paraphrase Detection

Research on learning and paragraph-level paraphrase generation highlights the importance of examining similarities between complete passages rather than relying only on individual word matches [3] [8].

Paragraph-level analysis is useful when the wording or sentence order changes but the main idea remains similar. Deep learning approaches can be investigated to identify paraphrased content although their effectiveness needs to be evaluated.

The proposed project applies these concepts in a system designed to provide student and teacher dashboards, document submission, OCR-assisted text extraction, comparison with reference materials and similarity report generation.

2.6 OCR and Handwritten Document Processing

Optical Character Recognition (OCR) converts text in scanned documents and images into a format that computers can process. Recognizing content is more challenging because writing styles, letter shapes, spacing and image quality can vary [4].

Errors during text recognition may affect plagiarism detection because incorrectly extracted words can reduce the accuracy of comparisons with reference documents. Therefore the influence of OCR quality on detection performance should be evaluated [4].

III. RESEARCH GAP

The findings from the reviewed studies identify limitations that inform the design of the proposed system.

3.1 Limited support for document types:

Certain plagiarism detection approaches require document content to be available as digital text before comparison. Consequently, scanned pages and handwritten assignments require a text recognition step.

3.2 Difficulty detecting paraphrased content

Word-based comparison may fail to identify passages whose wording has changed while their central meaning remains similar. Semantic similarity analysis can help examine these cases.

3.3 OCR-related errors

Text extracted from scanned or handwritten documents may contain recognized characters and spelling errors. Such inaccuracies can affect the reliability of text comparisons.

3.4 Separation of student and teacher activities

The proposed application brings together student document submission and teacher functions, including record management, document access and review of plagiarism reports.

3.5 Limited explanation of similarity results

A single similarity score provides information about the reasons for flagging a passage. The proposed system is intended to display matching passages and their source details when the similarity score meets the selected threshold.

IV. METHODOLOGY

A) User Authentication and Document Upload: The application provides login portals for students and teachers. Students can submit supported files scanned PDFs and images containing handwritten assignments. Teachers can use their dashboard to view student records, access submitted documents and examine plagiarism reports.

B) Text Preprocessing: The system checks the file type. Digital documents are processed with text-extraction tools. Scanned and handwritten documents need OCR or handwriting recognition. After extraction the text is cleaned to remove formatting. The processed content is then divided into sentences or paragraphs for analysis.

C) Direct Text Similarity Analysis: The first step finds matches. The submitted document is compared with reference materials and previous submissions using TF-IDF, cosine similarity and phrase-matching techniques. These methods can identify overlapping text. May not detect every instance of paraphrasing [9].

D) Deep Learning-Based Semantic Analysis: In this stage Sentence-BERT converts text into embeddings. Cosine similarity is then used to compare these embeddings and identify passages that may convey meanings despite differences in wording [2].

E) Similarity Integration and Report Generation: In this stage Sentence-BERT converts text into embeddings. Cosine similarity is then used to compare these embeddings and identify passages that may convey meanings despite differences in wording [2].

F) Performance Evaluation: The systems performance will be assessed using accuracy, precision, recall and F1-score on labelled examples of copied, paraphrased and unrelated text. OCR quality and processing time will also be examined. Evaluating document types separately can help identify limitations, in individual processing stages. Performance

ISSN: 3139-0862

conclusions will be reported only after the system has been implemented and tested.

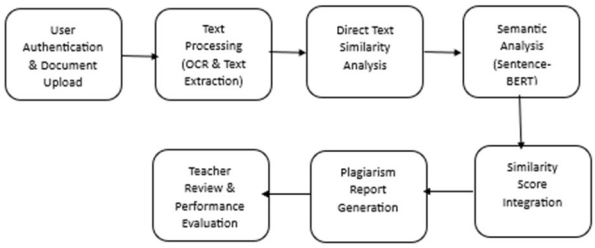


Figure 1: Proposed Methodology

FUTURE WORK

a. Better Handwriting Recognition: Future versions of the system might use handwriting recognition models to get text more easily from different handwriting styles, blurry pictures and low-quality scans.

b. Plagiarism Detection in Languages: The system can be made to work with different languages by using the right OCR tools, language models that work for many languages and steps that are specific to each language.

c. Better Ways to Find Ideas: Future studies might look at more advanced transformer models to find passages that mean the same thing but use different words or sentence structures.

d. Documents to Compare: A bigger collection of documents to compare could be created by adding more archives and other document databases. Good ways to organize and find these resources would make it easier to find the documents.

e. Reports That Explain Why Things Are Similar: Future reports could show parts next to each other and point out what is the same. Separate scores for word similarity and meaning similarity could help teachers see why certain parts were checked.

REFERENCES

[1] J. Devlin, M.-W. Chang, K. Lee and K. Toutanova "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding " in Proceedings of the 2019 Conference of the American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2019 pp. 4171-4186.
[2] N. Reimers and I. Gurevych "Sentence-BERT: Sentence Embeddings using BERT-Networks " in Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, 2019 pp. 3982-3992.
[3] M. Potthast, A. Eiselt, A. Barron-Cedeno, B. Stein and P. Rosso "Overview of the 3rd International Competition on Plagiarism Detection " in CLEF 2011 Labs and Workshop Notebook papers, 2011.
[4] R. Smith, "An Overview of the Tesseract OCR Engine " in Proceedings of the Ninth International Conference on Document Analysis and Recognition (ICDAR) vol. 2 2007. 629-633.
[5] M. A. Hearst, "TextTiling: Segmenting Text into Multi-Paragraph Subtopic Passages, " Computational Linguistics, vol. 23 No. 1 Pp. 33-64 1997.
[6] T. Mikolov, K. Chen G. Corrado and J. Dean, "Efficient Estimation of Word Representations in Vector Space " arXiv:1301.3781, 2013.
[7] M. A. Karpukhin et al. "Dense Passage Retrieval for Open-Domain Question Answering " in Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, 2020 6769-6781.
[8] M. Potthast et al. "Overview of the International Competition on Plagiarism Detection " in CLEF 2013 Evaluation Labs and Workshop 2013.
[9] C. D. Manning, P. Raghavan and H. Schütze, Introduction, to Information Retrieval. Cambridge University Press, 2008.
[10] V. I. Levenshtein, "Binary Codes of Correcting Deletions, Insertions and Reversals " Soviet Physics Doklady, vol. 10 No. 8 Pp. 707-710, 1966.