International Journal for Modern Trends in Science and Technology
Volume 12, Issue 07, pages 18-25.
ISSN: 2455-3778 online
Available online at: http://www.ijmtst.com/vol12issue06.html
DOI: https://doi.org/10.5281/zenodo.20739587

Plagiarism Detection and Document Similarity Analysis System

RVVSV Prasad, Gubbala Pavani Kumari, Gadireddy Dharani, Kota Tarun Kumar

Department of Information Technology , Swarnandhra College of Engineering & Technology, Seetharamapuram, Narsapur, Andhra Pradesh, INDIA.

To Cite this Article
RVVSV Prasad, Gubbala Pavani Kumari, Gadireddy Dharani& Kota Tarun Kumar (2026). Plagiarism Detection and Document Similarity Analysis System. International Journal for Modern Trends in Science and Technology, 12(07), 18-25. https://doi.org/10.5281/zenodo.20739587

Article Info
Received: 06 June 2026; Revised: 29 June 2026; Accepted: 01 July 2026.

KEYWORDS

Plagiarism Detection, Document Similarity Analysis, Machine Learning, Natural Language Processing (NLP), Text Pre-processing, TF-IDF, Cosine Similarity, Spring Boot, MySQL

ABSTRACT

The rapid growth of digital content and the widespread availability of online resources have significantly increased the risk of plagiarism in academic, research, and professional environments. Ensuring the originality of documents has become a critical challenge due to the vast amount of textual data available on the internet. Traditional plagiarism detection methods rely on manual verification, which is time-consuming, inefficient, and prone to human error. To overcome these limitations, this project proposes a Plagiarism Detection and Document Similarity Analysis System that automates the process of identifying similarities between documents using machine learning and natural language processing techniques. The proposed system performs text pre-processing steps such as tokenization, stop-word removal, and normalization to prepare the input data for analysis. It utilizes Term Frequency–Inverse Document Frequency (TF-IDF) for feature extraction and cosine similarity to compute the similarity between document vectors. The system is developed using frontend technologies such as HTML, CSS, and JavaScript, along with backend technologies including Java and Spring Boot. A MySQL database is used to store user data, uploaded documents, and similarity results. The system provides a user-friendly interface that allows users to upload documents and receive detailed similarity reports, including similarity percentage and highlighted matching content. The modular architecture ensures scalability, efficiency, and ease of maintenance. The proposed system improves accuracy, reduces manual effort, and enhances the reliability of plagiarism detection processes. This solution can be effectively applied in educational institutions, research organizations, and content development platforms to ensure originality and maintain academic integrity. Future enhancements may include integration with cloud-based storage, multi-language support, and advanced deep learning techniques for improved semantic analysis.

1. INTRODUCTION

In the modern digital era, the rapid growth of the internet and the widespread availability of online resources have significantly transformed the way information is accessed, shared, and utilized. This advancement has greatly contributed to education, research, and content creation by enabling quick and easy access to vast amounts of information. However, along with these benefits, it has also introduced serious challenges, particularly in maintaining the originality and authenticity of content. One of the major concerns is plagiarism, which refers to the act of copying or reproducing someone else's work, ideas, or expressions without proper acknowledgment. Academic institutions, research organizations, and content industries are increasingly facing difficulties in ensuring originality due to the abundance of digital content. Traditional plagiarism detection methods rely on manual checking, which is inefficient, time-consuming, and prone to human error, creating a strong need for an automated and intelligent solution.

1.1 Problem Statement and Objectives

The primary problem addressed in this project is the lack of an efficient and automated system capable of detecting plagiarism and measuring similarity between documents. With the increasing number of document submissions in academic and professional environments, manual verification has become impractical and unreliable. Existing basic systems often focus only on exact text matching and fail to identify semantic similarities where the meaning of the content remains the same despite changes in wording. This limitation makes plagiarism detection less effective and allows many cases to go unnoticed.

To overcome these challenges, the objective of this project is to develop a Plagiarism Detection and Document Similarity Analysis System using machine learning and natural language processing techniques. The system aims to preprocess textual data efficiently through techniques such as tokenization and normalization, and then compute similarity between documents using mathematical models like TF-IDF and cosine similarity. It also focuses on generating detailed similarity reports that clearly highlight matching content. Additionally, the system is designed with a user-friendly interface to ensure ease of use, high accuracy, fast processing, and scalability for handling large datasets.

1.2 Significance and Applications of the Project

The proposed system plays a crucial role in maintaining academic integrity and promoting ethical practices in content creation. By automating the plagiarism detection process, it significantly reduces manual effort and saves time for educators, evaluators, and organizations. The system encourages originality among students, researchers, and content creators by providing accurate and fast analysis of document similarity. Its ability to handle large datasets and detect both exact and semantic similarities makes it highly reliable and effective.

The system has wide-ranging real-world applications across multiple domains. It can be used in educational institutions for verifying assignments and projects, in research organizations for validating research papers, in publishing industries and content writing platforms for ensuring originality, and in corporate environments for checking reports and documents. The target users of this system include students, teachers, researchers, content creators, and organizations that require document validation. For example, a student can use the system to check the originality of an assignment before submission, a teacher can evaluate multiple assignments efficiently, and a researcher can ensure the uniqueness of their work before publication.

2 LITERATURE SURVEY

2.1 Overview of Plagiarism and Academic Integrity

Plagiarism has emerged as a significant issue in academic, research, and professional environments due to the rapid growth of digital content and easy access to online information. It involves the unauthorized use or reproduction of another individual's work without proper acknowledgment. According to Akbari (2021), plagiarism is not only limited to direct copying but also includes practices such as paraphrasing, translation, and content manipulation to disguise the original source. Bretag and Mahmud (2009) discussed the concept of self-plagiarism, emphasizing that reusing one's own previously published work without citation raises ethical concerns. Similarly, Devlin (2006) highlighted the importance of institutional policies and preventive measures in minimizing plagiarism among students. These studies collectively emphasize that plagiarism is

not just a technical issue but also an ethical and educational concern.

Furthermore, Gunnarsson et al. (2014) emphasized the role of educators and librarians in teaching students how to avoid plagiarism. Their work highlights that awareness and proper guidance can significantly reduce instances of academic misconduct. Milovanović et al. (2023) also pointed out the challenges faced by students in maintaining originality, especially when writing academic content in non-native languages. The increasing availability of digital resources has made plagiarism detection more complex, requiring advanced tools and techniques. Therefore, automated plagiarism detection systems have become essential in maintaining academic integrity and ensuring originality in scholarly work

2.2 Techniques for Plagiarism Detection

Plagiarism detection techniques have evolved significantly over the years, moving from simple text matching methods to advanced machine learning and natural language processing approaches. Traditional methods primarily relied on string matching and keyword comparison, which were limited in detecting paraphrased or semantically similar content. Alvi et al. (2021) proposed a method for paraphrase type identification using contextual information and word embeddings. Their approach demonstrates that understanding the context of words is crucial for detecting advanced forms of plagiarism. Similarly, Gangadharan et al. (2020) explored the use of deep neural networks and word embedding techniques to improve paraphrase detection, showing promising results in identifying semantic similarities.

Mahian et al. (2017) discussed similarity measurement techniques in academic contexts, emphasizing the importance of mathematical models such as cosine similarity for comparing document vectors. Memon (2020) further clarified the distinction between similarity and plagiarism, stating that similarity scores alone do not always indicate unethical practices. Recent studies have also focused on identifying non-traditional forms of plagiarism, such as algorithmically generated content. Cabanac and Labbé (2021) highlighted the prevalence of nonsensical, machine-generated research papers in scientific literature, indicating the need for more robust detection mechanisms.

Overall, modern plagiarism detection systems integrate multiple techniques, including text pre-processing, feature extraction, and similarity measurement, to achieve higher accuracy and reliability.

2.3 Challenges and Issues in Plagiarism Detection

Despite advancements in detection techniques, several challenges remain in accurately identifying plagiarism. One of the major challenges is detecting paraphrased content, where the original text is reworded while retaining its meaning. Ansorge et al. (2021) demonstrated how paraphrasing tools can be used to manipulate text, making it difficult for traditional detection systems to identify copied content. Jones (2009) introduced the concept of back-translation, where text is translated into another language and then translated back to the original language to disguise plagiarism. This method poses a significant challenge for detection systems that rely solely on surface-level text comparison. Jones and Sheridan (2015) further explored this technique, describing it as an advanced cyber strategy to bypass plagiarism detection tools.

Another emerging issue is the use of "tortured phrases," which are artificially generated or modified phrases that replace commonly used scientific terms. Cabanac et al. (2021) identified this phenomenon as a growing concern in scientific publications. Else (2021) also reported that such phrases are often used in fabricated research papers, making detection more complex. Additionally, Kannangara (2017) discussed the ethical concerns associated with the use of word-spinning software, which automatically generates variations of text to avoid detection. These tools can produce content that appears original but lacks meaningful contribution. The limitations of current systems highlight the need for advanced approaches that can analyse semantic meaning and context rather than relying solely on textual similarity.

3 PROPOSED METHODOLOGY

The proposed system, Plagiarism Detection and Document Similarity Analysis System, is designed to overcome the limitations of existing systems by using advanced machine learning and natural language processing techniques.

This system automates the process of detecting plagiarism by analysing textual data and calculating similarity between documents. It uses pre-processing

techniques such as tokenization and stop-word removal to clean the data, followed by feature extraction using TF-IDF. Cosine similarity is then applied to measure the similarity between document vectors.The system is developed using a three-tier architecture, including frontend (HTML, CSS, JavaScript), backend (Java, Spring Boot), and database (MySQL). It provides a user-friendly interface where users can upload documents and receive detailed similarity reports.

3.1 SYSTEM ARCHITECTURE

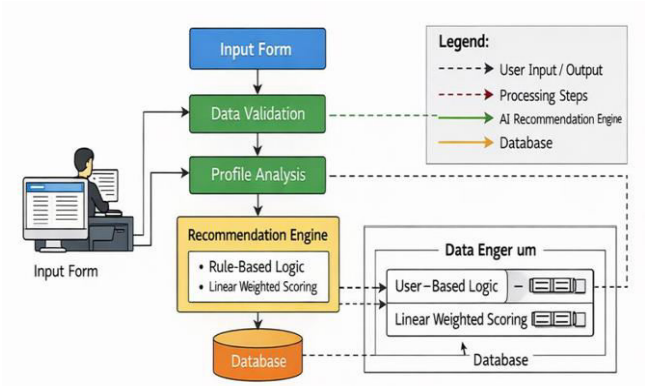

Fig 3.1 Proposed System Architecture

This diagram represents the overall architecture of the Plagiarism Detection and Document Similarity Analysis System. It consists of three main layers: frontend, backend, and database. The frontend provides user functionalities such as login, document upload, comparison, and report viewing. The backend, developed using Spring Boot, handles processing through controller, service, and repository layers. It performs text extraction, pre-processing, and similarity calculation using TF-IDF and cosine similarity algorithms. The database stores user details, documents, and similarity reports. The system efficiently processes user requests and generates accurate similarity results.

3.2 USE CASE DIAGRAM

The Use Case Diagram represents the interaction between the user and the Plagiarism Detection and Document Similarity Analysis System. It shows the various functionalities that the system provides to the user, such as registration, login, document upload, similarity checking, and report viewing. The user acts as the primary actor who interacts with the system to perform these operations. This diagram helps in understanding the overall system behaviour and how different functionalities are accessed by the user.

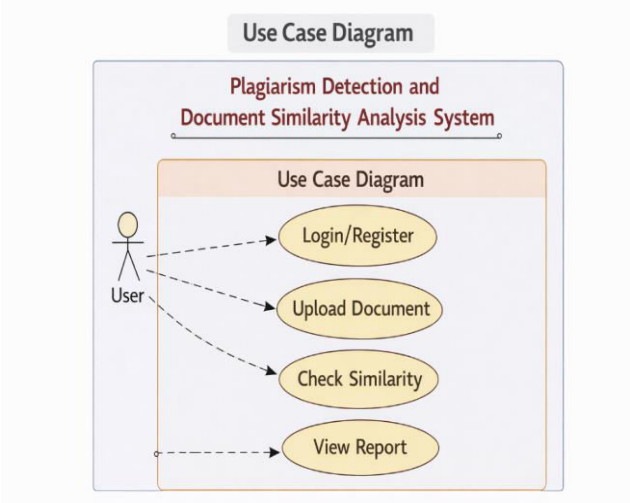

Fig 3.2 Use Case Diagram

3.3 SEQUENCE DIAGRAM

The sequence diagram shows how the user interacts with the plagiarism detection system step by step. The user uploads a document, and the request is sent to the controller. The controller forwards it to the service layer for processing, where text pre-processing and similarity calculation are performed. The results are stored in the database and then returned back through the controller to the user interface, where the similarity result is displayed.

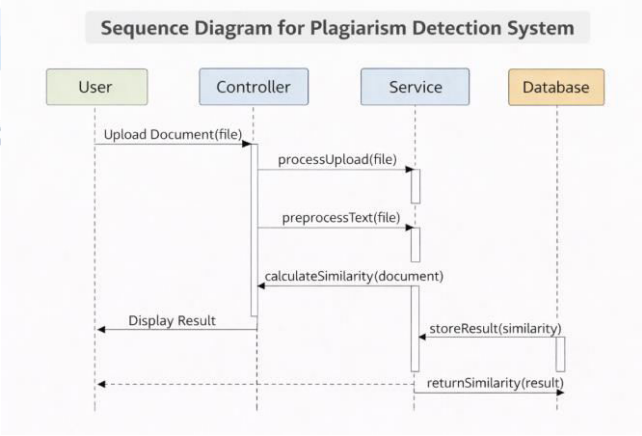

Fig 3.4 Sequence Diagram

3.4. EVALUATION METRICS

The performance of the proposed Plagiarism Detection and Document Similarity Analysis System is evaluated using a combination of quantitative metrics. Since the system utilizes techniques such as TF-IDF, cosine similarity, and natural language processing (NLP), multiple evaluation parameters are considered to

measure accuracy, relevance, and effectiveness of similarity detection.

A. Accuracy

Accuracy measures the proportion of correctly identified plagiarism cases (both plagiarized and non-plagiarized documents) out of all evaluated document pairs. It indicates how effectively the system classifies document similarity.

$$Accuracy = \frac{TP + TN}{TP + TNFP + FN}$$

Where:

- TPTPTP = True Positives (correctly detected plagiarism)
- TNTNTN = True Negatives (correctly identified non-plagiarism)
- FPFPFP = False Positives
- FNFNFN = False Negatives

A higher accuracy indicates that the system performs well in distinguishing between plagiarized and original content.

B. Precision

Precision evaluates how many of the documents identified as plagiarized are actually plagiarized. It focuses on reducing false positives (incorrect plagiarism detection).

$$Precision = \frac{TP}{TP + FP}$$

High precision ensures that the system provides reliable plagiarism detection results without wrongly flagging original content.

C. Recall

Recall measures the system's ability to detect all actual plagiarism cases present in the dataset. It focuses on minimizing missed plagiarism instances.

$$Recall = \frac{TP}{TP + FN}$$

A higher recall indicates that the system effectively identifies most of the plagiarized documents.

4 RESULTS

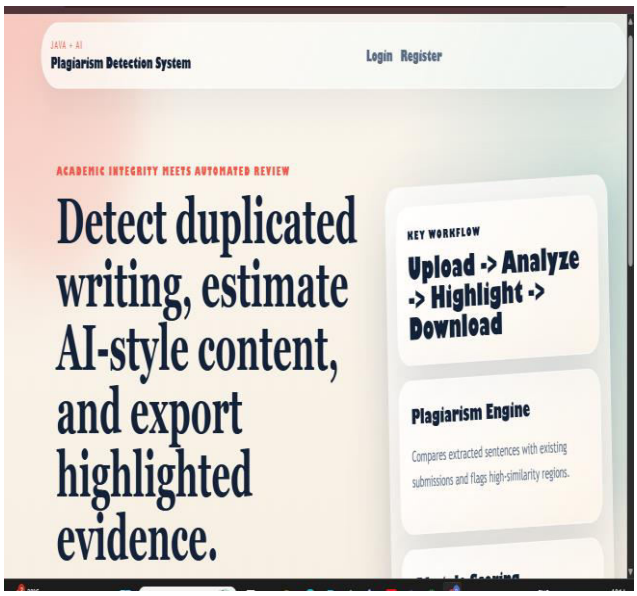

Fig 4.1: Login Page Interface

This diagram represents the overall workflow of a plagiarism detection system. It illustrates how users upload documents, which are then analysed by the system to identify duplicated or AI-generated content. The system compares the text with existing data, highlights similar sections, and finally allows users to download a detailed report with marked evidence.

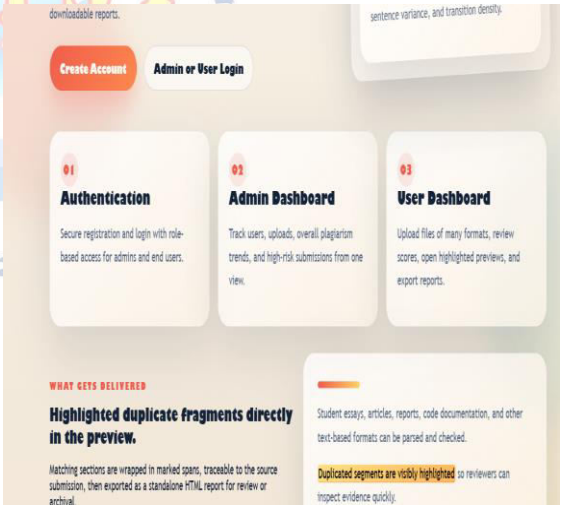

Fig-4.2 Plagiarism Detection System Functional Modules Diagram

This diagram illustrates the key functional modules of the plagiarism detection system, including authentication, admin dashboard, and user dashboard. It shows how users securely log in, upload and analyse documents, while administrators monitor system activity, user uploads, and plagiarism trends. The system highlights duplicated content and provides downloadable reports for review.

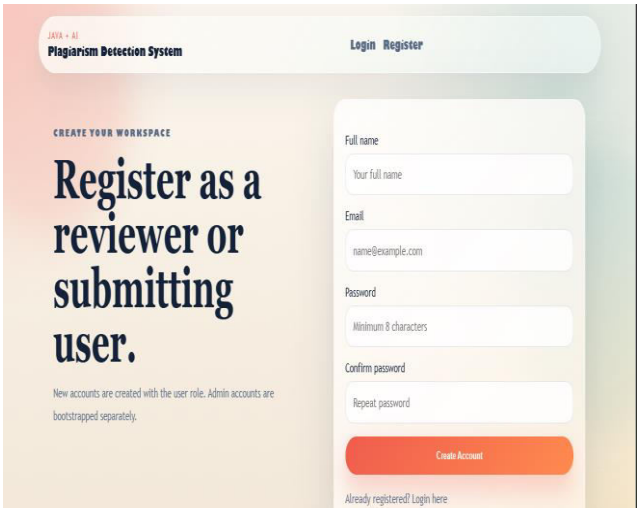

Figure 4.3 User Registration Interface Diagram

This diagram represents the user registration interface of the plagiarism detection system. It allows new users to create an account by entering details such as full name, email, and password.

The system supports role-based access, enabling users to register as reviewers or submitting users, ensuring secure and organized access to system features.

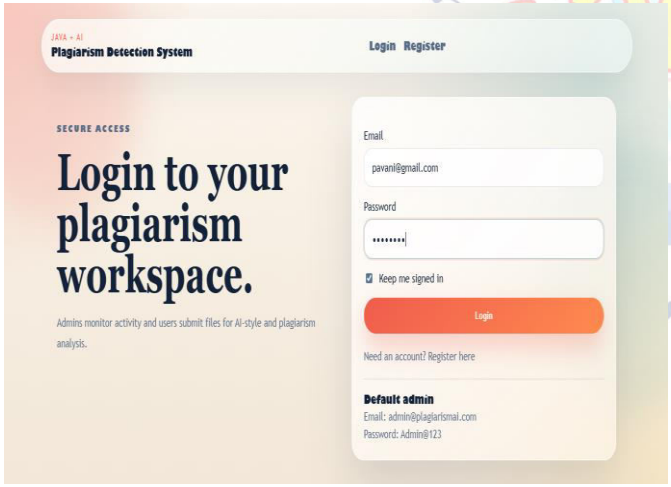

Fig 4.4 User Login Diagram

This diagram represents the login interface of the plagiarism detection system. It allows registered users and administrators to securely access their accounts by entering their email and password.

The interface includes features like “keep me signed in” and provides access to the system workspace for uploading, analysing, and reviewing plagiarism reports.

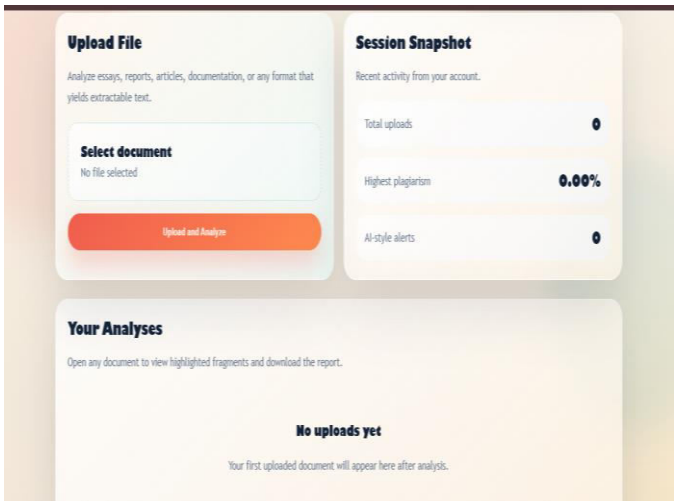

Fig 4.5 User Dashboard Workflow Diagram

This diagram represents the user dashboard of the plagiarism detection system. It allows users to upload documents for analysis, view session statistics such as total uploads and plagiarism scores, and access their analysis history. The dashboard provides an organized interface to manage files, review highlighted results, and download detailed plagiarism reports.

5 CONCLUSION

The Plagiarism Detection and Document Similarity Analysis System has been successfully designed and implemented to address the growing need for automated plagiarism detection in academic and professional environments. With the increasing availability of digital content, ensuring originality has become a major challenge. This system provides an effective solution by automating the process of detecting similarities between documents using computational techniques.

The system integrates frontend, backend, and database components to provide a complete and user-friendly application. The frontend interface allows users to upload documents and view results easily, while the backend, developed using Spring Boot, processes the data efficiently. The MySQL database ensures proper storage and retrieval of user data and similarity results.

The implementation of text pre-processing techniques such as tokenization and normalization improves the quality of data before analysis. Feature extraction using TF-IDF enables the system to identify important terms within documents, while cosine similarity provides an efficient method to measure similarity between

document vectors. These techniques ensure accurate and reliable plagiarism detection.

The system successfully reduces manual effort, improves accuracy, and saves time for users. It is scalable, easy to use, and suitable for applications in educational institutions, research organizations, and content development platforms. The modular design allows easy maintenance and future enhancements.

Overall, the project demonstrates how machine learning and natural language processing techniques can be effectively used to solve real-world problems related to plagiarism detection and document similarity analysis.

6 FUTURE ENHANCEMENT

Although the proposed system provides an efficient solution for plagiarism detection, there is scope for further improvements and enhancements to increase its functionality and performance.

One possible enhancement is the integration of the system with online databases and search engines. This would allow the system to compare documents not only with locally stored data but also with a vast range of online content, improving detection accuracy.

Another area for improvement is the incorporation of advanced machine learning and deep learning techniques. Methods such as word embeddings, neural networks, and semantic analysis can be used to detect deeper levels of similarity, including contextual and conceptual similarities.

The system can also be extended to support multiple languages. Currently, it focuses on text in a single language, but adding multilingual support would make it more versatile and useful for global users.

Improving the user interface and user experience is another potential enhancement. Adding features such as real-time analysis, graphical representation of results, and downloadable reports can make the system more interactive and user-friendly.

Security can also be enhanced by implementing advanced authentication mechanisms and data encryption techniques to protect user data and documents.Additionally, the system can be deployed on cloud platforms to improve scalability and accessibility. Cloud integration would allow users to access the system from anywhere and enable efficient handling of large datasets.

In conclusion, the proposed system provides a strong foundation for plagiarism detection, and with further enhancements, it can be developed into a more advanced and comprehensive solution capable of handling complex real-world scenarios.

Conflict of interest statement

Authors declare that they do not have any conflict of interest.

REFERENCES

[1] A. Akbari, "Spinning-translation and the act of plagiarising: How to avoid and resist," Journal of Further and Higher Education, vol. 45, pp. 49–64, 2021.

[2] F. Alvi, M. Stevenson, and P. Clough, "Paraphrase type identification for plagiarism detection using contexts and word embeddings," International Journal of Educational Technology in Higher Education, vol. 18, p. 42, 2021.

[3] L. Ansorge, "Tortured phrases are not automatically unethical," European Science Editing, vol. 50, e135388, 2024.

[4] L. Ansorge, K. Ansorgeová, and M. Sixsmith, "Plagiarism through paraphrasing tools—The story of one plagiarized text," Publications, vol. 9, p. 48, 2021.

[5] J. Aronson, "When I use a word … artificial translation," Aug. 6, 2021. [Online]. Available: https://blogs.bmj.com/bmj/2021/08/06/jeffrey-aronson-when-i-use-a-word-artificial-translation

[6] S. Baskaran et al., "Is there plagiarism in the most influential publications in the field of andrology?" Andrologia, vol. 51, e13405, 2019.

[7] T. Bretag and S. Mahmud, "Self-plagiarism or appropriate textual re-use?" Journal of Academic Ethics, vol. 7, pp. 193–205, 2009.

[8] [8] G. Cabanac and C. Labbé, "Prevalence of nonsensical algorithmically generated papers in the scientific literature," Journal of the Association for Information Science and Technology, vol. 72, pp. 1461–1476, 2021.

[9] G. Cabanac, C. Labbé, and A. Magazinov, "Tortured phrases: A dubious writing style emerging in science," arXiv preprint arXiv:2107.06751, 2021.

[10] M. Devlin, "Policy, preparation, and prevention: Proactive minimization of student plagiarism," Journal of Higher Education Policy and Management, vol. 28, pp. 45–58, 2006.

[11] H. Else, "'Tortured Phrases' give away fabricated research papers," Nature, vol. 596, pp. 328–329, 2021.

[12] V. Gangadharan et al., "Paraphrase detection using deep neural network based word embedding techniques," in Proc. 4th Int. Conf. Trends in Electronics and Informatics (ICOEI), 2020, pp. 517–521.

[13] J. Gunnarsson, W. J. Kulesza, and A. Pettersson, "Teaching international students how to avoid plagiarism," The Journal of Academic Librarianship, vol. 40, pp. 413–417, 2014.

[14] J. R. Higgins, F.-C. Lin, and J. P. Evans, "Plagiarism in submitted manuscripts," Research Integrity and Peer Review, vol. 1, p. 13, 2016.

[15] M. Jones, "Back-translation: The latest form of plagiarism," in Proc. Asia Pacific Conf. Educational Integrity, 2009, pp. 1–7.

[16] M. Jones and L. Sheridan, "Back translation: An emerging sophisticated cyber strategy," Assessment & Evaluation in Higher Education, vol. 40, pp. 712–724, 2015.
[17] D. N. Kannangara, "Quality, ethics and plagiarism issues," MIER Journal of Educational Studies Trends and Practices, vol. 7, pp. 24–32, 2017.
[18] P. Lay, M. Lentschat, and C. Labbé, "Investigating the detection of tortured phrases," in Workshop on Scholarly Document Processing, 2022, pp. 32–36.
[19] O. Mahian et al., "Measurement of similarity in academic contexts," Publications, vol. 5, p. 18, 2017.
[20] S. Manley, "The use of text-matching software's similarity scores," Accountability in Research, vol. 30, pp. 219–245, 2023.