JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue
JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)
An International Scholarly Open Access, Peer-reviewed, Refereed Journal

ADVANCEMENTS IN PLAGIARISM DETECTION: A COMPREHENSIVE REVIEW AND PROPOSAL

[1]Dr. Anita Dixit , [2]P N Unnathi, [3]Rahul J Guttedar and [4]Suraj More
[t] Assistant Professor, [2] B.E Student, [3]B.E Student, [4] B.E Student
Information Science and Engineering
SDM College of Engineering and Technology, Dharwad, India

Abstract : Plagiarism detection is crucial for upholding academic integrity and ensuring originality in scholarly works. This research paper surveys traditional and contemporary approaches to plagiarism detection, encompassing text-based, document-based, and image-based methods. Additionally, we propose a comprehensive plagiarism checker system that integrates various detection modules, including text, document, and image plagiarism detection, along with a paraphrasing module, web scraping, API integration, and report generation capabilities. By reviewing existing methodologies and introducing our proposed system, this paper contributes to the advancement of plagiarism detection tools, aiming to provide a robust solution for detecting and preventing plagiarism in academic and professional contexts.

***IndexTerms*- Plagiarism Detection, Literature review, Machine learning, Deep learning, Academic Integrity.**

I. INTRODUCTION

Plagiarism stands as a pervasive challenge confronting academic and professional realms, casting a shadow over the authenticity and integrity of scholarly pursuits. At its core, plagiarism entails the misappropriation of intellectual property, blurring the lines between original thought and derivative imitation. This insidious practice not only erodes the foundation of knowledge dissemination but also compromises the trust upon which academic and professional communities thrive. In response to this existential threat, the field of plagiarism detection has emerged as a bastion of vigilance, leveraging a myriad of methodologies and technologies to safeguard the sanctity of intellectual discourse.

This paper embarks on a comprehensive exploration of existing methods for detecting plagiarism, encompassing text-based, document-based, and image-based techniques. Through a meticulous examination of both classical and contemporary methodologies, we aim to illuminate the multifaceted nature of plagiarism detection and unravel the intricate challenges that beset this critical endeavour. From rule-based algorithms to sophisticated machine learning models, each approach offers unique insights into the detection and prevention of academic misconduct, highlighting the dynamic interplay between technology, ethics, and academic integrity.

Moreover, this paper presents a pragmatic proposal - a plagiarism checker system designed to offer a user-friendly alternative to existing methods. By integrating diverse detection modules, algorithms, and user-centric functionalities, our system aims to provide a practical solution for plagiarism detection. While acknowledging the existence of well-established plagiarism checkers with highly accurate methods, our intention is to contribute to the field by offering an accessible and intuitive tool.

II. LITERATURE REVIEW

Plagiarism detection methods have evolved significantly over the years, ranging from traditional techniques to modern approaches leveraging advanced technologies. In this section, we delve into existing literature on plagiarism detection methodologies, offering insights into both classical and contemporary methods.

Traditional techniques, such as manual inspection and rule-based algorithms, have long been the cornerstone of plagiarism detection. Manual inspection involves human reviewers meticulously scrutinising documents for instances of textual similarity. While this method can be effective in detecting blatant cases of plagiarism, it is labour-intensive and time-consuming, making it impractical for large-scale analysis. Rule-based algorithms, on the other hand, automate the detection process by applying predefined rules to identify plagiarised content. While these algorithms offer scalability, they often lack the sophistication to detect nuanced forms of plagiarism.

In recent years, modern approaches to plagiarism detection have gained prominence, fueled by advancements in machine learning and artificial intelligence. Deep learning-based models, such as convolutional neural networks (CNNs) and recurrent neural networks (RNNs), have shown promise in automatically extracting features from textual data and identifying patterns indicative of plagiarism. These models offer improved accuracy and efficiency compared to traditional techniques, particularly in handling large volumes of data. However, they may require substantial computational resources and training data to achieve optimal performance.

Optimization algorithms represent another frontier in plagiarism detection, aiming to enhance the effectiveness and efficiency of detection models. Techniques such as genetic algorithms and differential evolution optimise model parameters to improve detection accuracy and reduce false positives. While these algorithms can yield significant performance gains, they may also introduce complexity and computational overhead.

By examining key research studies in the field, we aim to provide a comprehensive overview of the strengths and limitations of traditional and modern plagiarism detection methods. Through this review, we seek to identify areas for further research and development, with the ultimate goal of advancing the state-of-the-art in plagiarism detection and promoting academic integrity.

III. Analysis and Discussion of Plagiarism Detection Studies

[1] This study presents a deep learning-based technique for plagiarism detection, focusing on high-quality vector representations of words. It compares various neural network-based models and document representation methods. Strengths include a comprehensive comparison of different deep learning architectures and representation methods, offering insights into their effectiveness. However, limitations may include a potential lack of diversity in the datasets used for evaluation and the computational complexity of some deep learning models.

[2] This study proposes an improved differential evolution (DE) algorithm to optimise the learning process of a BERT-based plagiarism detection model. It aims to enhance the performance of the model by optimising the initialization of model weights. Strengths include introducing a novel optimization algorithm tailored specifically for BERT-based plagiarism detection models, potentially improving accuracy and efficiency. However, limitations may include the evaluation being limited to specific datasets or scenarios, requiring further validation across diverse datasets.

[3] This paper provides an overview of plagiarism detection methods and tools, categorising them into verbatim/literal plagiarism and intelligent plagiarism detection approaches. Strengths include offering a comprehensive overview of different types of plagiarism detection techniques, providing valuable insights into their categorization and characteristics. However, limitations may include a lack of in-depth analysis or evaluation of specific methods, focusing more on categorization rather than detailed comparison or assessment of effectiveness.

[4] This study presents a novel approach to plagiarism detection in student programming assignments, treating it as an information retrieval problem and using a query-by-example approach. Strengths include introducing a unique approach to plagiarism detection in a specific domain (student programming assignments), offering potential efficiency advantages over traditional methods. Limitations may include the evaluation being limited to specific programming languages or assignment types, requiring further investigation of effectiveness across diverse datasets or programming paradigms.

[5] This paper discusses academic plagiarism detection using machine learning techniques, providing insights into challenges and opportunities in the field. Strengths include offering a comprehensive review of academic plagiarism detection methods, highlighting the importance of machine learning, and discussing challenges and future directions. Limitations may include a lack of detailed technical analysis or empirical evaluation of specific methods, focusing more on the broader landscape of academic plagiarism detection research.

[6] This study presents plagiarism detection in computer programming using feature extraction from ultra-fine-grained repositories, leveraging machine learning techniques. Strengths include introducing a novel approach to plagiarism detection in computer programming, focusing on fine-grained coding activity logs and machine learning-based feature extraction. Limitations may include the effectiveness of the approach depending on the availability and quality of coding activity logs, and its generalizability to other programming languages or contexts requiring further investigation.

IV. SYSTEM OVERVIEW

The proposed plagiarism checker system is designed to fulfil the demand for a reliable and user-friendly tool capable of identifying plagiarism across diverse content formats. The envisioned system architecture comprises several pivotal components, each contributing to the system's effectiveness and usability.

1. User Interface: At the forefront of the system is its user interface, engineered to provide an intuitive and seamless experience for users. Through clear design and navigation elements, users will access various functionalities such as inputting text, uploading documents or images, and generating reports.
2. Authentication and Authorization: Ensuring the security and integrity of the system, robust authentication and authorization mechanisms will be implemented. User authentication during login and stringent access controls will guarantee that only authorised users can interact with the system's features.
3. Detection Modules: Central to the system's functionality are its detection modules, which will employ advanced algorithms to identify instances of plagiarism. These modules will encompass text, document, and image analysis techniques, utilising methodologies such as TF-IDF, cosine similarity, and integration with external resources like the Google search API.
4. Paraphrasing Functionality: A distinctive feature of the system will be its paraphrasing functionality, allowing users to generate alternative content variants. This feature will promote originality while discouraging direct content replication.
5. Web Scraping and API Integration: To enhance its plagiarism detection capabilities, the system will integrate with web scraping tools and external APIs. By leveraging resources such as the Google Custom Search API, the system will retrieve relevant web content for comparison and analysis.
6. Report Generation: The system will facilitate comprehensive reporting by generating downloadable PDF reports. These reports will encapsulate input data, search results, and similarity scores, empowering users with detailed insights for further analysis and reference.

In essence, the proposed plagiarism checker system embodies a holistic approach to plagiarism detection, combining robust algorithms with user-friendly design elements to deliver a powerful yet accessible tool for maintaining academic integrity and originality.

V. METHODOLOGY

The methodology for the proposed plagiarism checker system revolves around a systematic approach from initial concept to proposed deployment. Commencing with a thorough analysis of user requirements through surveys, key features were prioritised based on their significance in plagiarism detection and content paraphrasing. Careful consideration was given to selecting a suitable technology stack, including Django for rapid development, NLTK and WordNet for natural language processing, and the Google Custom Search API for web search integration. The design of the system's database employed a relational schema using Django's ORM for optimal performance. Security measures were ensured through robust implementation of user authentication and authorization protocols. Plagiarism detection algorithms were constructed utilising TF-IDF, cosine similarity, and external APIs. Integration of paraphrasing functionality with NLTK and WordNet facilitated creative content generation. Comprehensive documentation was maintained throughout the process to capture all aspects of development, from codebase annotations to user instructions. This methodology underscores our commitment to delivering a robust plagiarism detection and content enhancement solution.

VI. CONCLUSION

In conclusion, this survey paper has provided an overview of various methods for plagiarism detection, highlighting the strengths, limitations, and advancements in the field. Through the literature review, we identified different approaches, from traditional techniques to modern deep learning-based models. The proposed system overview outlines a methodology for developing a comprehensive plagiarism checker system, leveraging insights gained from the literature survey. By addressing existing gaps and challenges, such as accuracy and scalability, this proposed system aims to contribute to the ongoing improvement of plagiarism detection tools. Moving forward, further research should focus on exploring novel algorithms and specialised tools to enhance the effectiveness of plagiarism detection systems, ensuring the preservation of academic integrity and intellectual honesty.

REFERENCES

[1] F. El Mostafa Hambi and F. Benabbou, "A deep learning based technique for plagiarism detection: a comparative study," **Int. J. Artif. Intell.**, vol. 9, no. 1, pp. 1-12, Mar. 2020,doi: 10.11591/ijai.v9.i1.pp81-90

[2] S. V. Moravvej, S. J. Mousavirad, D. Oliva, G. Schaefer and Z. Sobhaninia, "An Improved DE Algorithm to Optimise the Learning Process of a BERT-based Plagiarism Detection Model," 2022 IEEE Congress on Evolutionary Computation (CEC), Padua, Italy, 2022, pp. 1-7, doi: 10.1109/CEC55065.2022.9870280.

[3] Khaled, Farah, and Mohammed Sabbih H. Al-Tamimi. "Plagiarism detection methods and tools: An overview." Iraqi Journal of

Science (2021): 2771-2783,DOI: 10.24996/ijs.2021.62.8.30.
[4] Pajić, Enil, and Vedran Ljubović. "Improving plagiarism detection using genetic algorithm." 2019 42nd International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO). IEEE, 2019.
[5] Bohra, Anjali, and N. C. Barwar. "A deep learning approach for plagiarism detection system using BERT." Congress on Intelligent Systems: Proceedings of CIS 2021, Volume 2. Singapore: Springer Nature Singapore, 2022.
[6] V. Ljubovic and E. Pajic, "Plagiarism Detection in Computer Programming Using Feature Extraction From Ultra-Fine-Grained Repositories," in IEEE Access, vol. 8, pp. 96505-96514, 2020, doi: 10.1109/ACCESS.2020.2996146.
[7] Chavan, Hiten, et al. "Plagiarism detector using machine learning." International Journal of Research in Engineering, Science and Management 4.4 (2021): 152-154.
[8] Foltýnek, Tomáš, Norman Meuschke, and Bela Gipp. "Academic plagiarism detection: a systematic literature review." ACM Computing Surveys (CSUR) 52.6 (2019): 1-42.
[9] Wahle, Jan Philip, et al. "Identifying machine-paraphrase plagiarism." International Conference on Information. Cham: Springer International Publishing, 2022.
[10] El-Rashidy, Mohamed A., et al. "Reliable plagiarism detection system based on deep learning approaches." Neural Computing and Applications 34.21 (2022): 18837-18858.