International Journal of Trend in Scientific Research and Development (IJTSRD)
Volume 9 Issue 1, Jan-Feb 2025 Available Online: www.ijtsrd.com e-ISSN: 2456 – 6470

Doc-Sheild Plagiarism Detection Improving Accuracy and Efficiency Enhancement in Text and Image Similarity Detection

Mrs. Pethota Swaroopa[1], T. Sri Harsha[2], P. Shivani[3], K. Aashritha[4], P. Akshay[5]

[1]Assistant Professor, [2,3,4,5]Student,
[1,2,3,4,5]CSE Department, ACE Engineering College, Hyderabad, Telangana, India

ABSTRACT

Plagiarism in both the academic and professional world attacks the integrity of the intellect and stifles innovation. Traditional plagiarism detection systems can adequately identify text duplication but may lack the efficiency of detecting paraphrased, translated, or contextual variations of text content as well as plagiarized graphical elements, including flowcharts. This paper will outline a complete system that uses a combination of Artificial Neural Networks (ANNs), Natural Language Processing (NLP), and deep learning techniques for both textual and visual plagiarism. It has proved to be much more accurate in terms of plagiarism detection, especially with an 81.91% success rate for flowchart plagiarism, providing an efficiency gain for semantic similarity detection, establishing a new standard for academic integrity tools.

How to cite this paper: Mrs. Pethota Swaroopa | T. Sri Harsha | P. Shivani | K. Aashritha | P. Akshay "Doc-Sheild Plagiarism Detection Improving Accuracy and Efficiency Enhancement in Text and Image Similarity Detection" Published in International Journal of Trend in Scientific Research and Development (ijtsrd), ISSN: 2456-6470, Volume-9 | Issue-1, February 2025, pp.651-652, URL: www.ijtsrd.com/papers/ijtsrd73834.pdf

IJTSRD73834

1. INTRODUCTION

Academic dishonesty, especially plagiarism, undermines the integrity of intellectual contributions. Conventional plagiarism detection tools rely heavily on string-matching algorithms to identify exact or partial matches in text. However, these systems struggle with paraphrased, contextually altered, or translated content and entirely overlook visual data like flowcharts and diagrams, which often carry critical information in research.

To bridge these gaps, this paper brings together semantic similarity detection, NLP, and ANN-based image analysis for developing a robust system capable of detecting both textual and visual plagiarism. Advanced linguistic analysis and deep learning enhance the accuracy of detection, especially for paraphrased and multilingual content while furthering its reach to image-based elements like flowcharts.

2. Research Approaches

2.1. Machine Learning and NLP-Based Approaches

It exploits advanced NLP techniques and the state-of-the-art models using machine learning; such models pre-trained with transformer like BERT and GPT that will analyse the semantic similarity between them. It understands the subtlety in contextual information with better effectiveness while detecting paraphrasing or translated text.

2.2. Limitation of the Previous Models

The most striking disadvantages are false positives produced by state-of-the-art machine learning-based models with their computationally demanding process, along with a challenge toward detecting visual plagiarism especially with increasing numbers of technical research studies based on structured diagrams, for instance, flowcharts.

3. Proposed System

3.1. System Architecture

The core modules proposed include

1. Text Preprocessing: Tokenization, lemmatization, and removal of stop-words in order to eliminate noise, while improving concentration of the actual information.
2. Semantic Similarity Module: Deep learning models are used to determine similarity between text portions, identifying paraphrasing and contextually altered content.
3. Image Plagiarism Detection Module: ANNs are used to analyse the flowchart structure, nodes, and edges to detect similarities. In this module, an accuracy of 81.91% is obtained in detecting plagiarized flowcharts.

3.2. Methodology

The system was trained on a comprehensive dataset of academic papers containing instances of textual and image plagiarism. It uses cosine similarity metrics and ANN-based structural analysis for accurate detection. The dual-layered semantic similarity check ensures robustness against paraphrased and cross-lingual plagiarism.

3.3. Innovations and Expected Outcomes

The system introduces a novel approach to integrated plagiarism detection by combining NLP and ANN techniques. Its ability to handle multilingual and image-based content significantly broadens its applicability, promising improved accuracy and reduced false positives in academic and professional settings.

4. Implementation and Testing

4.1. Data Collection

The corpus consists of articles, essays, and web content in several languages, marked for textual and image-based plagiarism. The images used in the flowcharts are obtained from the CLEF-IP 2012 database.

4.2. Training and Evaluation

The ANN model was trained with supervised learning algorithms, and its performance was evaluated using metrics such as precision, recall, and F1-score. Preliminary results show that the detection of complex plagiarism forms is highly improved compared to the traditional models.

4.3. Results

The system was able to achieve 81.91% accuracy in the detection of plagiarized flowchart images compared to other methods. The NLP module also showed high precision in detecting paraphrased and contextually altered text.

5. Discussion

The integration of ANN and NLP techniques addresses the major limitations of traditional plagiarism detection systems, particularly in handling paraphrased, image-based, and cross-lingual content. The incorporation of transformer-based models and context-aware approaches makes the system effective in detecting similarities across multiple languages and formats. The cross-lingual capability significantly broadens its usability in diverse academic and professional settings, reducing the reliance on language-specific datasets. While the system exhibits high accuracy, further optimization is required to enhance computational efficiency and scalability for large-scale applications. Future work may focus on expanding the range of supported languages and refining the detection algorithms to address nuanced forms of plagiarism.

6. Conclusion

The proposed system represents a huge step forward in plagiarism detection. The textual and image-based challenges it provides solutions for make the system quite robust. By focusing on semantic similarity and contextual analysis, the system establishes a new standard for academic integrity tools. Future developments aim to enhance its applicability across diverse languages and real-world scenarios, fostering a culture of originality and authenticity in academia and beyond.

References:

[1] Dong, Y., et al. "A Semantic-Based Plagiarism Detection Approach Using Word Embeddings." Journal of Educational Computing Research, vol. 57, no. 1, 2019.

[2] Soni, S., and Roberts, R. "Deep Learning for Paraphrase Detection." Proceedings of the 12th ACM Conference on Text Mining, 2021.

[3] Leacock, C., Chodorow, M., and Gamon, M. "Contextual Similarity in Plagiarism Detection: Improving Accuracy through Deep Learning." Natural Language Engineering, vol. 28, no. 2, 2022.

[4] Zhang, H., et al. "Cross-Lingual Plagiarism Detection Using Transformer-Based Models." ACM Transactions on Asian and Low-Resource Language Information Processing, vol. 20, no. 5, 2023.

[5] Gupta, R., and Singh, P. "Enhancing Plagiarism Detection with Multimodal Data Fusion." IEEE Transactions on Knowledge and Data Engineering, vol. 35, no. 3, 2023.

[6] Brown, T., et al. "Language Models are Few-Shot Learners." Advances in Neural Information Processing Systems (NeurIPS), 2020.