Performance Evaluation Of AI-Enhanced Detection Methods for Image-Based and Alphanumeric Substitution Plagiarism Manipulations

Mr. Yatheendra K V[1] & Dr. Sudhakara Arabagatte[2]
[1]Research Scholar, College of Computer Science, Srinivas University, Mangalore, India
[2] Professor, College of Computer Science, Srinivas University, Mangalore, India

***Abstract*—The sophistication of plagiarism techniques continues to evolve, making detection increasingly challenging. Among the most evasive manipulations are embedding text as images and performing alphanumeric substitutions that evade conventional text-matching algorithms. This paper presents a multi-layered AI-enhanced detection system that combines OCR-based extraction, proportionating algorithms, advanced regular expressions, and language model-assisted validation. Evaluated on a real-world dataset of 5000 academic reports, the proposed system achieved an accuracy of 85%, significantly improving upon the baseline accuracy of 55%. The results demonstrate the system's robustness in handling advanced manipulation tactics while also highlighting the continuous need for adaptive model improvements.**

***Index Terms*—plagiarism detection, AI, OCR, alphanumeric substitution, proportionating algorithm, academic integrity, experimental results.**

1. INTRODUCTION

As plagiarism detection systems continue to evolve, individuals seeking to circumvent these technologies are simultaneously developing increasingly sophisticated evasion techniques. Among these, two manipulation methods have become particularly prevalent: (i) the transformation of textual content into embedded image formats, and (ii) the substitution of standard alphanumeric characters with visually similar alternatives, often leveraging Unicode variations or custom fonts — a strategy commonly referred to as *alphanumeric substitution*. These techniques are specifically designed to exploit the limitations of conventional text-matching algorithms, which predominantly depend on the ability to extract and compare raw textual data directly from documents. As a result, such manipulations can significantly undermine the accuracy and reliability of standard similarity assessment tools.

To address these emerging challenges, this work presents a comprehensive multi-layered detection framework specifically engineered to identify and mitigate these advanced forms of obfuscation. The proposed system combines several complementary detection modules:

- Optical Character Recognition (OCR): Utilized to convert embedded text images back into machine-readable text, thereby neutralizing image-based masking techniques.
- Proportionating Algorithm: Designed to analyze and compare expected word densities, aiding in the detection of unusual word count fluctuations that may suggest embedded or altered content.
- Advanced Regular Expressions: Employed for the recognition of character-level manipulations and pattern irregularities introduced by alphanumeric substitutions.
- Language Model-Assisted Validation: Integrates state-of-the-art language models to assess semantic consistency and coherence, offering an additional layer of validation that can flag unnatural language constructs or unusual substitution patterns.

Extensive experimental evaluations conducted on diverse datasets demonstrate the system's ability to achieve significantly higher detection accuracy compared to traditional approaches. The results highlight substantial improvements in both recall and precision metrics, validating the framework's practical applicability and robustness in real-world academic and professional plagiarism detection scenarios.

2. RELATED WORK

Plagiarism detection has traditionally relied on text-matching algorithms that compare submitted content against large databases of existing documents to identify overlapping or similar text segments. While effective for direct copying and simple paraphrasing, numerous studies have highlighted significant limitations of these methods when faced with non-textual manipulations and character-based obfuscation techniques specifically designed to evade detection.

One major challenge arises from image-based text embedding, where portions of the textual content are converted into images and embedded within documents. To counteract this, Optical Character Recognition (OCR) systems, such as Tesseract, have been employed to extract textual content from these images. However, the performance of OCR systems is often compromised by factors such as image noise, varying resolutions, compression artifacts, non-standard fonts, overlapping elements, and diverse background textures, all of which introduce inaccuracies in the extracted text. These errors propagate through subsequent detection stages, reducing the overall effectiveness of traditional plagiarism detection pipelines.

In parallel, another increasingly common evasion technique involves alphanumeric substitution attacks, where visually similar characters from alternative Unicode code points or specialized fonts are used to replace standard characters (e.g., substituting the Latin letter 'A' with the mathematical bold '**A**' (U+1D400) or the Greek capital letter 'A' (U+0391)). While these characters are visually indistinguishable to human readers, they fundamentally alter the underlying digital representation of the text, thereby breaking string-matching algorithms and token-based comparison models. Rule-based normalization techniques—which attempt to map such variants back to their canonical forms—have shown some promise in mitigating these substitutions, but they often fall short due to the vast and evolving range of Unicode characters and font variations available for manipulation.

Recent advancements in language model-based contextual analysis have opened new possibilities for addressing such challenges more reliably. By leveraging the semantic and syntactic coherence of the surrounding text, language models can identify unnatural substitutions that violate linguistic patterns, even when the substitutions individually pass visual inspection. This allows for a more intelligent validation layer capable of detecting subtle manipulation patterns that may elude deterministic rule-based approaches.

Nevertheless, despite these technological advancements, limited large-scale real-world evaluations and the continuous development of novel manipulation strategies by users seeking to evade detection continue to expose vulnerabilities in existing plagiarism detection frameworks. This persistent gap in robust, scalable, and adaptive detection capabilities motivates the current study, which seeks to address these evolving challenges through a comprehensive, multi-layered detection approach that integrates OCR, character normalization, proportion-based analysis, and language model-assisted semantic validation.

3. METHODOLOGY

The proposed system utilizes a layered and modular architecture, combining traditional pattern detection methods with AI-powered validation components to effectively detect advanced plagiarism evasion techniques. Each module is designed to handle specific manipulation vectors, and collectively they contribute to a unified composite scoring system that generates the final assessment for each document. The overall architecture is both scalable and adaptive, allowing new modules to be integrated as manipulation tactics evolve.

3.1 OCR Extraction

The first layer of processing targets image-based text embedding, where text is hidden within image formats to circumvent traditional text extraction algorithms. To retrieve the embedded text, the system incorporates the widely used open-source Tesseract OCR engine. Recognizing the inherent limitations and variability in OCR accuracy, a multi-stage post-processing pipeline is applied to improve the quality of OCR output before it is passed to downstream modules.

The post-processing stages include:

- Noise Removal: Preprocessing techniques such as Gaussian blurring, thresholding, and morphological operations are applied to suppress background noise, shadows, and scanning

artifacts that commonly interfere with OCR recognition accuracy.
- Line Segmentation Correction: Algorithms analyze spatial distribution of characters and words to correct skewed or broken line segmentations, ensuring better text reconstruction from poorly aligned documents.
- Character Confidence Thresholding: OCR confidence scores are used to discard or flag low-confidence character outputs, minimizing the introduction of random noise or misrecognized symbols into subsequent stages.

By enhancing the reliability of text extraction, this module ensures that text embedded as images is made available for further analysis.

3.2 Proportionating Algorithm

Following OCR extraction, the proportionating algorithm performs a statistical assessment of the expected word density for the document. This module is based on empirical analysis of academic writing structures, where typical documents exhibit a fairly consistent range of word counts per page.

- For academic reports, research papers, and dissertations, observed data suggests an average of 200–300 words per page under standard formatting conditions.
- The system calculates an expected word count range for each document by multiplying its total page count by this empirical word density.
- When the actual extracted word count significantly deviates—particularly when it falls below the lower threshold—it suggests the presence of large non-text regions, which may result from text-to-image conversions or deliberate content suppression.
- Such deviations are flagged, providing early indicators of potential embedded manipulations that warrant further investigation.

This layer helps detect sophisticated manipulations that may evade purely text-based analysis by exploiting document structure-level inconsistencies.

3.3 Alphanumeric Substitution Detection

The third layer addresses character-based obfuscation, where standard characters are replaced by visually identical but semantically distinct Unicode alternatives. This form of manipulation directly targets the vulnerabilities in conventional string-matching and tokenization algorithms.

3.3.1 Advanced Regular Expressions

- A comprehensive library of regular expressions (Regex) is utilized to scan text for known sets of Unicode confusables—characters that resemble standard Latin alphabet characters but differ at the code-point level.
- These patterns are regularly updated as new substitution variants emerge, ensuring coverage across various language scripts (e.g., Cyrillic, Greek, Mathematical Alphanumerics, Fullwidth forms).
- Detected substitutions are normalized where possible, converting non-standard code points back to their canonical equivalents for consistency in downstream similarity comparisons.

3.3.2 Language Model Validation

Beyond deterministic substitutions, AI-powered language models are leveraged to capture deeper contextual anomalies that may indicate unnatural substitutions or fabricated word formations:

- Normalization & Correction: Language models assist in dynamically normalizing inconsistent substitutions that follow non-standard patterns.
- Semantic Coherence Validation: By analyzing sentence structures, grammar, and contextual relevance, language models can detect unnatural text fragments resulting from partial substitutions or character mixing that violate linguistic fluency.
- Error Detection: Text flagged by the model for having low semantic coherence is further escalated for manual or automated review.

The language models are deployed within isolated Docker containers, ensuring scalability, reproducibility, and fault isolation across multiple concurrent document processing pipelines. These models are queried via internal REST APIs, allowing seamless integration with the core detection system.

3.4 Composite Scoring Engine

Finally, the outputs generated by the OCR extraction, proportionating algorithm, and alphanumeric substitution detection modules are aggregated and synthesized within a composite scoring engine:

- Each module contributes a weighted score based on its confidence level and the severity of the detected manipulation.
- A tunable weighting mechanism is employed, allowing domain-specific calibration based on document types, user-defined sensitivity, and historical false-positive rates.
- The cumulative score determines the final classification of the document into one of the following categories:
- Clean Document: No substantial manipulations detected; safe for standard similarity evaluation.
- Suspicious Document: Potential manipulation indicators present; recommended for enhanced review or secondary verification.
- Severe Manipulation (doc:error): High-confidence evidence of significant evasion tactics; document flagged for immediate investigation or rejection.

This multi-layered architecture ensures that even highly sophisticated plagiarism attempts, involving both visual and character-level obfuscation, are accurately detected while minimizing false positives on legitimate submissions.

4. EXPERIMENT SETUP

4.1 Dataset

A dataset of 5000 academic reports was used for evaluation, collected from real-world plagiarism reports. These reports included:

- Clean documents
- Reports with image-based text
- Reports containing alphanumeric substitutions
- Reports with combined manipulations

4.2 Evaluation Metrics

The system performance was measured using standard classification metrics:

- Accuracy
- Precision
- Recall
- F1-Score

A baseline system — lacking OCR, proportionating, and substitution detection modules — was used for comparison.

5. RESULT

The results are summarized below:

5.1 Confusion Matrix

Outcome	Baseline	Proposed
True Positive	1350	2100 (out of 2500 manipulated)
True Negative	1400	2150 (out of 2500 clean)
False Positive	1250	350
False Negative	1000	400

- Correctly detect manipulation (True Positive)
- Correctly detect clean (True Negative)
- Miss manipulation (False Negative)
- Incorrectly flag clean as manipulated (False Positive)

System	Accuracy	Precision	Recall	F1-score
Baseline	55%	52%	54%	53%
Proposed	85%	87%	84%	85%

5.2 Performance Metrics

Accuracy = (TP + TN) / Total = (2100 + 2150) / 5000 = 85%

Precision = TP / (TP + FP) = 2100 / (2100 + 350) ≈ 85.7%

Recall = TP / (TP + FN) = 2100 / (2100 + 400) = 84%

F1 Score ≈ 85%

6. CONCLUSION

This study presents a novel AI-powered, multi-layered detection framework designed to address some of the most challenging and rapidly evolving manipulation techniques in modern plagiarism detection — specifically, text embedded within images and character-level alphanumeric substitutions. Unlike traditional text-matching algorithms, which are largely

ineffective against such advanced obfuscation strategies, the proposed system integrates multiple complementary modules, including OCR extraction, proportionating analysis, Unicode-based substitution detection, and language model-based semantic validation.

The system's effectiveness was rigorously evaluated on a dataset of 5,000 real-world academic and professional documents, which included a diverse mix of naturally occurring and synthetically manipulated samples. The evaluation results demonstrate a substantial improvement, with the proposed system achieving an overall detection accuracy of 85%, significantly outperforming the baseline accuracy of 55% observed in conventional plagiarism detection solutions. These findings underscore the framework's ability to detect previously undetectable manipulation techniques, thereby greatly enhancing the robustness and reliability of plagiarism assessment processes.

Beyond controlled evaluations, the system has been successfully deployed in real-world production environments, where it processes high volumes of submissions across varied domains, including academic institutions, publishing houses, and corporate content validation platforms. Importantly, the framework is designed with adaptive learning capabilities, allowing it to evolve alongside emerging evasion techniques through continuous updates to its rule sets, language models, and substitution libraries. This ensures that the detection system remains effective against newly developed manipulation patterns that may arise over time.

In summary, the proposed approach addresses critical gaps in existing plagiarism detection technologies by providing a scalable, extensible, and future-proof solution capable of handling both current and emerging forms of document manipulation. Future work will explore further enhancements, including deeper integration of transformer-based large language models (LLMs), cross-lingual manipulation detection, and fully automated self-learning modules that can autonomously adapt to newly observed evasion strategies without human intervention.

REFERENCES:

[1] Potthast, M., et al. (2014). Evaluating plagiarism detection. ACM Transactions on Intelligent Systems and Technology (TIST), 7(4), 1-27. DOI: 10.1145/2742360.

[2] Alzahrani, S. M., Salim, N., & Abraham, A. (2012). Understanding plagiarism linguistic patterns, textual features, and detection methods. IEEE Transactions on Systems, Man, and Cybernetics, Part C, 42(2), 133-149. DOI: 10.1109/TSMCC.2011.2134847.

[3] Gupta, V., & Garg, S. (2020). Machine Learning Based Approaches for Text Plagiarism Detection: A Survey. International Journal of Information Retrieval Research (IJIRR), 10(3), 1–15. DOI: 10.4018/IJIRR.2020070101.

[4] Manzoor, M. F., Farooq, M. S., & Abid, A. (2025). Stylometry-Driven Framework for Urdu Intrinsic Plagiarism Detection. Neural Computing and Applications. DOI: 10.1007/s00521-024-10966-w

[5] Vrbanec, T., & Meštrović, A. (2023). Comparison Study of Unsupervised Paraphrase Detection: Deep Learning – The Key for Semantic Similarity Detection. Expert Systems. DOI: 10.1111/exsy.13386

[6] Sharjeel, M., Iqbal, H. R., & Shafi, J. (2025). Urdu Paraphrased Text Reuse and Plagiarism Detection Using Pre-trained LLMs and Deep Neural Networks. Multimedia Tools and Applications.

[7] Pudasaini, S., Miralles-Pechuán, L., & Lillis, D. (2024). Survey on AI-Generated Plagiarism Detection: The Impact of Large Language Models on Academic Integrity. Journal of Academic Ethics. DOI: 10.1007/s10805-024-09576-x

[8] Sajid, M., Sanaullah, M., Fuzail, M., & Malik, T. S. (2025). Comparative Analysis of Text-Based Plagiarism Detection Techniques. PLOS ONE. DOI: 10.1371/journal.pone.0319551

[9] Amirzhanov, A., Turan, C., & Makhmutova, A. (2025). Plagiarism Types and Detection Methods: A Systematic Survey of Algorithms in Text Analysis. Frontiers in Computer Science. DOI: 10.3389/fcomp.2025.1504725

[10] Lee, J., Le, T., Chen, J., & Lee, D. (2023). Do Language Models Plagiarize? Proceedings of the ACM Web Conference (WWW). DOI: 10.1145/3543507.3583199