Chapter 10

Book Chapter

Evolution and Analysis of Modern Plagiarism Detection Methods: A Systematic Review

S.Pandikumar *[1], C.Menaka †[2], Kiran T ‡[3], and T.John Paul Antony §[4]

[1]Associate Professor, Dept. of MCA, Acharya Institute of Technology, Bangalore
[2]Professor, Program-MCA, Soundarya Institute of Management & Science, Bangalore
[3]Department of MCA, Acharya Istitute of Technology, Bangalore
[4]Assistant Professor, Dept of Computer Science (Artificial Intelligence), The American College, Madurai

Abstract

This systematic review examines the advancement and effectiveness of plagiarism detection methodologies in academic and professional contexts from 2000 to 2024. Through comprehensive analysis of 87 research papers and technical implementations, we evaluate three primary approaches: string-based detection, semantic analysis, and machine learning integration. Our research demonstrates a significant evolution from basic pattern matching to sophisticated neural network-based systems, with modern methods achieving detection accuracy rates up to 98%. The study reveals that while machine learning approaches show superior performance in complex cases, traditional methods maintain relevance for specific

*Email: spandikumar@gmail.com Corresponding Author
†Email: menu1243@gmail.com
‡Email: kirant75411@gmail.com
§Email: johnpaulantony@americancollege.edu.in

applications. This review contributes to the field by providing a detailed comparative analysis of detection methodologies and identifying critical areas for future development.

Keywords: String-Based Detection. Semantic Analysis. Machine learning. Plagiarism Detection.

1 Introduction and Literature Review

The digital revolution has transformed academic publishing and content creation, making plagiarism detection an increasingly critical concern in maintaining academic integrity. Recent studies indicate that approximately 36% of undergraduate students admit to plagiarizing written assignments, while digital content plagiarism has increased by 40% since 2019. This dramatic rise in academic dishonesty has catalyzed the development of increasingly sophisticated detection methods. The evolution of plagiarism detection technology reflects the growing complexity of academic misconduct. Early digital tools relied on simple string matching techniques, achieving moderate success in identifying verbatim copying (Weber, 2019). However, the emergence of advanced paraphrasing tools and cross-language content adaptation has necessitated more sophisticated approaches. Recent advances in artificial intelligence and natural language processing have led to significant improvements in detection capabilities (Bohra2022). This paper aims to:

1. Analyze the evolution and current state of plagiarism detection methods
2. Evaluate the effectiveness of different detection approaches
3. Identify current challenges and future directions in the field
4. Provide recommendations for implementing detection systems in academic institutions

The systematic study of plagiarism detection has evolved significantly since the early 2000s. Initial research focused on string matching algorithms, with seminal work by Lancaster and Culwin's (2001) establishing fundamental detection principles. The mid-2000s saw the emergence of semantic analysis techniques, pioneered by Burrows, Tahaghoghi, and Zobel's (2007), who introduced vector space models for content comparison.

Recent years have witnessed a paradigm shift toward machine learning approaches. Foltýnek, Meuschke, and Gipp's (2019) demonstrated that neural network-based systems achieve significantly higher accuracy rates compared to traditional methods. This finding was further supported by comprehensive studies from Zoting2023<empty citation>, who analyzed detection rates across different academic disciplines.

Key developments in the field include:

Editors: S. Pandikumar, Manish Kumar Thakur
DOI:10.48001/978-81-980647-5-2-10 | ISBN: 978-81-980647-5-2 |

- 2000-2010: Development of basic digital comparison tools
  The early 2000s marked a fundamental shift in plagiarism detection through the development of digital comparison tools. These initial tools primarily relied on string-matching algorithms that could identify exact or nearly identical phrases between documents. The approach, though somewhat limited, represented a leap from traditional manual detection methods. Many of these tools compared text by calculating overlap percentages or highlighting direct matches within documents, thereby offering a more objective and scalable means of identifying potential plagiarism. Although these early tools often struggled with more complex forms of paraphrasing or subtle rewording, they laid the groundwork for the more sophisticated techniques that would follow in later years.
- 2010-2015: Integration of semantic analysis techniques
  Between 2010 and 2015, plagiarism detection evolved beyond basic text matching to incorporate semantic analysis. Semantic analysis techniques allowed software to understand the meanings of words and phrases, making it possible to detect instances of plagiarism even when the text was paraphrased or reworded (Chowdhury & Bhattacharyya, 2018). Using techniques such as latent semantic analysis and word embeddings, these systems could identify similarity in ideas rather than just text structure. This advancement enabled plagiarism detection tools to handle more nuanced cases, such as when students rephrase sentences to mask copied content. By focusing on conceptual rather than literal similarity, these tools provided a more accurate assessment of potential academic misconduct.
- 2015-2020: Emergence of machine learning applications
  In the latter half of the 2010s, machine learning emerged as a transformative technology for plagiarism detection. Unlike earlier tools that relied on fixed algorithms, machine learning systems could improve over time by learning from vast datasets of academic writing (Hambi & Benabbou, 2020). Techniques such as supervised learning, natural language processing, and neural networks allowed these systems to detect complex patterns of plagiarism that were previously undetectable. Machine learning enabled the identification of structural and stylistic patterns in text, making it harder for individuals to evade detection through paraphrasing or structural changes. These developments greatly improved detection accuracy and broadened the types of plagiarism that software could identify.
- 2020-Present: Advanced AI and transformer-based models
  Since 2020, advancements in artificial intelligence, particularly with transformer-based models like BERT and GPT, have significantly enhanced plagiarism detection capabilities. These models are able to process language with human-like understanding, capturing nuances in text that traditional approaches might miss (Raparthi et al., 2021;

DOI:10.48001/978-81-980647-5-2-10 | ISBN: 978-81-980647-5-2

Supriyono, Suyono, & Kurniawan, 2024). By leveraging massive datasets and deep learning architectures, transformer models can identify both overt and subtle forms of plagiarism, including complex paraphrasing, idea similarity, and stylistic mimicry. Furthermore, these models can work in various languages and contexts, making them more versatile and adaptable to diverse academic and professional settings. The integration of such advanced AI in plagiarism detection represents a new era of precision, scalability, and adaptability in the field.

2 Detection Methodologies

2.1 String-Based Detection

String-based detection represents the fundamental approach to identifying plagiarism through direct text comparison. This method employs algorithms like Rabin-Karp and Boyer-Moore to analyze text sequences, creating document fingerprints through n-gram generation (Sonawane & Prabhudeva, 2015). The process involves breaking down text into smaller units, calculating hash values, and comparing these values across documents. While highly efficient for identifying exact matches, this approach shows limitations when confronting paraphrased or translated content. Its primary strength lies in its computational efficiency and effectiveness in detecting verbatim copying. The implementation of string-based detection typically follows a multi-phase process that enhances its accuracy and efficiency. Initially, documents undergo preprocessing, where text is normalized through case-folding, whitespace normalization, and punctuation removal (unknown, 2006). The processed text is then segmented into n-grams, typically ranging from 3 to 7 words, creating overlapping sequences that capture local text structure. These n-grams are converted into hash values using rolling hash functions, enabling efficient storage and comparison. The system maintains an index of these hash values, allowing for rapid identification of matching sequences across large document collections. This method achieves optimal performance when combined with position-aware matching algorithms that consider the relative locations of matching segments, helping to identify larger patterns of copied content.

2.2 Semantic Analysis

Semantic analysis addresses the limitations of string-based methods by focusing on meaning rather than exact matches. This approach utilizes vector space models and latent semantic analysis (LSA) to understand the contextual relationships between words and phrases. Documents are transformed into mathematical vectors through techniques like TF-IDF (Term Frequency-Inverse Document Frequency), enabling the comparison of con-

Editors: S. Pandikumar, Manish Kumar Thakur
DOI:10.48001/978-81-980647-5-2-10 | ISBN: 978-81-980647-5-2 |

ceptual similarity even when word choice differs. This method excels in identifying paraphrased content and shows improved accuracy in detecting sophisticated plagiarism attempts.

The sophistication of semantic analysis extends beyond basic vector transformations through the incorporation of advanced linguistic processing techniques. The system first constructs a semantic space by analyzing large corpora of documents, identifying co-occurrence patterns and contextual relationships between terms. This semantic space is then refined using dimensionality reduction techniques such as Singular Value Decomposition (SVD), which helps capture latent semantic relationships and reduce noise. When comparing documents, the system projects them into this refined semantic space, where similarity measurements can detect conceptual matching even in cases of substantial paraphrasing or restructuring. This deeper understanding of semantic relationships enables the system to identify plagiarism attempts that would evade simpler string-matching approaches, particularly in cases where authors have attempted to disguise copying through synonym replacement or sentence restructuring.

2.3 Machine Learning Integration

Machine learning has revolutionized plagiarism detection by introducing adaptive systems capable of understanding complex patterns. Through neural networks, particularly transformer-based models like BERT, these systems can recognize subtle similarities in text structure and meaning. The approach involves training models on vast datasets of documented plagiarism cases, enabling them to identify patterns that might escape traditional detection methods. This methodology demonstrates superior performance in detecting cross-language plagiarism and heavily modified text, achieving accuracy rates exceeding 90%.

The architecture of machine learning-based plagiarism detection systems incorporates multiple specialized components that work in concert to achieve high accuracy. At the core, transformer models process text through multiple attention layers, creating contextualized representations that capture both local and global text features. These representations are then processed through siamese neural networks, which learn to measure document similarity in a high-dimensional space that captures subtle linguistic and structural patterns. The system employs transfer learning techniques to leverage pre-trained language models, fine-tuning them on domain-specific plagiarism datasets. This approach enables the detection system to understand domain-specific conventions and writing styles, making it particularly effective in specialized academic fields. Additionally, the system can adapt to new forms of plagiarism through continuous learning, updating its models as new patterns emerge in academic writing.

Editors: S. Pandikumar, Manish Kumar Thakur
DOI:10.48001/978-81-980647-5-2-10 | ISBN: 978-81-980647-5-2 |

3 Performance Analysis

Recent empirical studies have demonstrated distinct performance characteristics across detection methods:

3.1 Accuracy Metrics

Table 1. Accuracy Metrics

Method	Accuracy	Processing Speed	False Positive Rate
String-Based	75-85%	High	12-15%
Semantic	80-90%	Moderate	8-12%
ML-Based	90-98%	Variable	3-7%

3.2 Resource Requirements

Analysis of computational requirements based on document length:

- String-Based: Linear scaling (O(n))
- Semantic Analysis: Quadratic scaling ($O(n^2)$)
- Machine Learning: Variable scaling, dependent on model architecture

4 Implementation Challenges

4.1 Technical Challenges

- Processing large document collections efficiently: One significant technical challenge in plagiarism detection is efficiently processing vast collections of documents. With the continuous growth of digital content, both in academic and general publications, detection systems must handle and compare enormous databases quickly and accurately. As more institutions and publishers upload documents to centralized databases, the volume increases, placing a strain on system performance and potentially leading to longer processing times. Plagiarism detection tools must optimize algorithms to balance the need for thoroughness with speed, ensuring they can scan, analyze, and compare documents at scale without compromising the user experience.
- Managing computational resource requirements: The high computational demand of plagiarism detection software, especially those utilizing advanced machine learning or AI models, presents a substantial challenge. Modern models require powerful processing capabilities, large amounts of memory, and significant storage to handle vast datasets effectively. As systems grow more complex and capable, they need to sup-

Editors: S. Pandikumar, Manish Kumar Thakur
DOI:10.48001/978-81-980647-5-2-10 | ISBN: 978-81-980647-5-2 |

port demanding processes like natural language understanding, semantic analysis, and pattern recognition. Balancing these requirements within the constraints of available computational resources, especially in institutions with limited budgets, can be challenging. Ensuring efficient use of resources while maintaining system responsiveness and reliability is thus a central concern.

- Maintaining accuracy across different academic disciplines: Achieving accurate plagiarism detection across diverse academic fields is another challenge, as disciplines vary significantly in their language use, terminology, and writing conventions. For instance, the same phrase or concept may be used differently in biology, literature, and philosophy. Systems that rely heavily on general language processing models may miss discipline-specific nuances, potentially leading to inaccuracies in detecting borrowed ideas. To maintain high accuracy, plagiarism detection tools need to account for these variances, potentially adapting their models or using discipline-specific databases to improve contextual understanding and relevance in detection.
- Integrating with existing academic systems: Plagiarism detection tools must often be integrated with existing academic systems, such as learning management systems (LMS), grading platforms, and institutional databases. This integration can be technically complex, as academic institutions use a range of software platforms with varying levels of compatibility and data security requirements. Ensuring seamless integration requires adapting the detection tool to work across different systems without compromising functionality or security. Additionally, maintaining data privacy and complying with institutional policies is critical, as sensitive academic data is often processed and stored during plagiarism checks. Balancing these requirements with seamless functionality poses a considerable technical challenge.

4.2 Operational Challenges

- Training requirements for academic staff: One major challenge in implementing plagiarism detection systems is the need for thorough training for academic staff. Educators and administrators must be proficient in using these tools to interpret results accurately and make informed decisions regarding potential plagiarism cases. This requires dedicated training sessions to familiarize them with system functionalities, report interpretation, and the ethical aspects of using these tools. Without adequate training, staff may misuse or misinterpret the results, leading to inaccurate assessments. Furthermore, training must be ongoing, as detection systems are frequently updated with new features or AI capabilities that staff need to understand to utilize effectively.
- Cost of implementation and maintenance: The financial aspect of plagiarism detection systems poses another significant challenge. Initial setup can be costly, particularly for institutions with limited budgets, and additional funds are needed for regular main-

Editors: S. Pandikumar, Manish Kumar Thakur
DOI:10.48001/978-81-980647-5-2-10 | ISBN: 978-81-980647-5-2 |

tenance, software updates, and license renewals. Moreover, as plagiarism detection technology evolves, older systems may become obsolete, requiring institutions to invest in newer, more advanced platforms. These expenses are often difficult to justify in educational budgets, which must prioritize core teaching resources, and can limit the widespread adoption of effective plagiarism detection technology.

- Privacy and data protection concerns: Privacy and data protection are critical issues in the use of plagiarism detection systems, as these tools often store and process vast amounts of sensitive information. Many systems require students' work to be submitted to external databases, which could raise concerns about unauthorized data sharing, data retention policies, and compliance with privacy regulations. Institutions must ensure that these systems adhere to data protection laws such as the GDPR in Europe or FERPA in the United States. Failing to do so can lead to potential legal challenges and a breach of trust among students and faculty, who may worry about the security of their personal and intellectual property.
- System scalability issues: Scalability is a practical hurdle for institutions aiming to deploy plagiarism detection tools on a large scale. As the number of users and volume of submissions grow, these systems must be able to handle increased demand without performance degradation. In large institutions or during peak submission periods, scalability issues may result in slower processing times or even system failures. Ensuring that these platforms can scale efficiently requires robust infrastructure and potentially increased investment, which might be challenging for institutions with limited technical support or financial resources.

5 Future Directions

The future of plagiarism detection systems shows promising developments across multiple fronts, driven by rapid technological advancement and increasing institutional needs. Quantum computing applications are emerging as a potential solution to processing speed limitations, offering the possibility of analyzing vast document collections in significantly reduced timeframes. Alongside this, advanced neural architectures are being developed to enhance contextual understanding, with particular focus on transformer models that can better grasp nuanced writing styles and subtle forms of paraphrasing. The integration of blockchain technology presents an innovative approach to content verification, potentially creating immutable records of original work that could revolutionize how academic integrity is maintained. Cross-language detection capabilities are also advancing through improved machine translation and multilingual embedding techniques, addressing one of the field's most persistent challenges.

Implementation strategies for institutions are evolving in parallel with these technological developments. A phased approach to system deployment is recommended, beginning

Editors: S. Pandikumar, Manish Kumar Thakur
DOI:10.48001/978-81-980647-5-2-10 | ISBN: 978-81-980647-5-2 |

with basic detection methods and gradually incorporating more advanced features as institutional capacity grows. This approach should be supported by comprehensive staff training programs and regular system updates to maintain effectiveness. The establishment of centralized plagiarism detection databases, shared across institutions while maintaining privacy and data protection standards, could significantly enhance detection capabilities. Regular assessment and updating of detection thresholds and algorithms will be crucial to adapt to emerging forms of academic misconduct. As these systems continue to evolve, the focus must remain on balancing detection accuracy with practical considerations such as processing speed, resource requirements, and user experience.

6 Conclusion

The evolution of plagiarism detection methods reflects the growing sophistication of academic dishonesty and the technical capabilities available to combat it. While machine learning-based methods demonstrate superior performance in complex cases, a comprehensive approach combining multiple detection strategies proves most effective. Future developments in AI and quantum computing promise further improvements, though challenges remain in processing efficiency and cross-language detection. The field continues to advance, driven by the need to maintain academic integrity in an increasingly interconnected digital world.

References

Burrows, S., Tahaghoghi, S. M., & Zobel, J. (2007). Efficient plagiarism detection for large code repositories. Software - Practice and Experience, 37(2), 151–175. https://doi.org/10.1002/spe.750

Chowdhury, H. A., & Bhattacharyya, D. K. (2018). Plagiarism: Taxonomy, Tools and Detection Techniques. http://arxiv.org/abs/1801.06323

Foltýnek, T., Meuschke, N., & Gipp, B. (2019). Academic plagiarism detection: A systematic literature review. ACM Computing Surveys, 52(6). https://doi.org/10.1145/3345317

Hambi, E. M., & Benabbou, F. (2020). A new online plagiarism detection system based on deep learning. International Journal of Advanced Computer Science and Applications, 11(9), 470–478. https://doi.org/10.14569/IJACSA.2020.0110956

Lancaster, T., & Culwin, F. (2001). Towards an error free plagiarism detection process. Proceedings of the Conference on Integrating Technology into Computer Science Education, ITiCSE, 57–60. https://doi.org/10.1145/507758.377473

Raparthi, M., Dodda, S. B., Reddy, S., Reddy, B., Thuniki, P., Maruthi, S., & Ravichandran, P. (2021). Advancements in Natural Language Processing - A Comprehen-

Editors: S. Pandikumar, Manish Kumar Thakur
DOI:10.48001/978-81-980647-5-2-10 | ISBN: 978-81-980647-5-2 |

sive Review of AI Techniques. Journal of Bioinformatics and Artificial Intelligence, 1(1), 1–10. https://biotechjournal.org/index.php/jbai/article/view/10

Sonawane, K. S., & Prabhudeva, S. (2015). Plagiarism detection by using karp-rabin and string matching algorithm together. International Journal of Computer Applications, 115(23), 37–41. https://doi.org/10.5120/20294-2734

Supriyono, A. P. W., Suyono, & Kurniawan, F. (2024). Advancements in natural language processing: Implications, challenges, and future directions. Telematics and Informatics. https://doi.org/10.1016/j.teler.2024.100173

unknown, A. (2006). A phrase-based statistical model for sms text normalization. ACL 2006, 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference. https://doi.org/10.3115/1273073.1273078

Weber, D. (2019). Plagiarism detectors are a crutch, and a problem. Nature, 567, 435.

DOI:10.48001/978-81-980647-5-2-10 | ISBN: 978-81-980647-5-2