LEX LOCALIS-JOURNAL OF LOCAL SELF-GOVERNMENT
ISSN:1581-5374 E-ISSN:1855-363X
VOL. 24, NO. 1(2026)

Scientific Text Fingerprinting in the Age of Artificial Intelligence

Dr: Ali Latreche, University Ahmed Draia, Adrar, Algeria
ali.latreche@univ-adrar.edu.dz
ORCID : 0009-0008-3536-4126
Dr: Mechid Mohammed, Ahmed Ben Ahmed University of Oran 2
mdmechid@gmail.com
ORCID : 0009-0003-4286-945X

Recieved: 09/12/2025 accepted: 16/03/2026 Published: 10/05/2026

Abstract
Plagiarism detection systems have become a critical component in safeguarding academic integrity, particularly in light of the rapid expansion of digital content and the emergence of text generation tools based on large language models. This study provides a systematic and comprehensive examination of plagiarism detection techniques grounded in scientific text fingerprinting within an artificial intelligence framework, highlighting the methodological shift from literal string matching to semantic detection of paraphrasing, translation, and obfuscated plagiarism. The article adopts a systematic literature review approach and proposes a multi-layered hybrid framework that integrates text fingerprinting and local algorithms, semantic text representations through embedding's, and language-specific processing layers addressing the morphological and syntactic challenges of Arabic. Furthermore, the study discusses standard evaluation protocols and benchmarking metrics, and offers a critical analysis of the strengths and limitations of existing approaches, alongside practical recommendations for academic institutions and scholarly publishers.

Keywords: Scientific Text, Text Fingerprinting, Semantic Analysis, Artificial Intelligence, Plagiarism Detection

1. Introduction

Academic integrity constitutes a foundational pillar of knowledge production and accumulation, as the entire research enterprise is grounded in trust in the originality of scholarly contributions and the accuracy of their bibliographic references. Early studies on textual similarity analysis in digital environments drew attention to the risks associated with the unauthorized reuse of textual material, particularly in light of the rapid expansion of electronic publishing and global academic databases (Broder, 1997; Broder et al., 1997). With the evolution of digital publishing ecosystems and the emergence of generative artificial intelligence tools, academic plagiarism is no longer confined to

literal copying; rather, it has evolved to encompass more sophisticated forms, including paraphrasing, translation, and the reconstruction of ideas and argumentative structures.

Within this context, artificial intelligence—especially natural language processing and machine learning techniques—has become a central driver in the development of plagiarism detection systems. These systems have progressively shifted from surface-level comparisons based on lexical matching toward deeper analytical approaches that account for linguistic structure, semantics, style, and contextual coherence (Potthast et al., 2010; Potthast et al., 2019). The concept of *text fingerprinting* has emerged as a core component of such systems, providing a compact representation of textual content that enables large-scale comparison while maintaining computational efficiency and scalability (Schleimer et al., 2003).

However, recent advances in artificial intelligence—particularly large language models—have exposed the limitations of relying solely on text fingerprinting techniques. These models are capable of generating ostensibly novel texts that preserve the underlying semantic content of multiple sources, thereby challenging traditional detection mechanisms. This development raises a critical scientific and methodological question concerning the adequacy of existing tools and highlights the need for hybrid approaches that integrate text fingerprinting with advanced semantic and stylometric analysis (Stamatatos et al., 2015).

Research Problem

This study is driven by the following central research question:
To what extent can artificial intelligence, through the integration of text fingerprinting techniques with semantic analysis, provide effective and accurate detection of diverse forms of plagiarism in academic texts, particularly in the presence of Arabic language complexities and increasingly sophisticated paraphrasing strategies?

Research Hypotheses

Based on this research problem, the study advances the following hypotheses:

1. Reliance on text fingerprinting alone achieves high effectiveness in detecting literal plagiarism but remains limited when confronted with paraphrased and semantically driven plagiarism.
2. Integrating text fingerprinting with AI-based semantic and stylometric analysis significantly improves detection accuracy and reduces false positive rates.
3. Due to its rich morphology and syntactic diversity, the Arabic language requires specialized processing strategies that cannot be directly transferred from models developed for Western languages.
4. Hybrid detection systems demonstrate greater adaptability to emerging forms of plagiarism, including those facilitated by generative language models.

Research Methodology

This study adopts a descriptive–analytical methodology through a systematic review of the scholarly literature on plagiarism detection and text fingerprinting, complemented by a comparative approach that examines differences among traditional, semantic, and hybrid techniques. In addition, the study draws on the analysis of established models and evaluation frameworks used in international research initiatives—such as PAN/CLEF—to identify best practices and prevailing research trends in the field.

Temporal and Spatial Scope of the Study

The temporal scope of this study spans approximately the period from 1997 to 2025, a phase marked by substantial developments in text similarity and plagiarism detection techniques. This period encompasses the evolution from early text fingerprinting algorithms to advanced artificial intelligence–based systems, including those relying on large language models. The spatial scope of the study is situated within the international academic environment, with particular emphasis on research and applications related to the Arabic language in universities and research centres, where linguistic complexity presents distinctive methodological challenges.

Research Objectives

This study aims to:

- Develop a comprehensive theoretical and methodological framework for contemporary plagiarism detection techniques.
- Analyse the theoretical foundations and technical mechanisms underlying text fingerprinting and its associated algorithms.
- Highlight the role of artificial intelligence in the transition from lexical matching to semantic-oriented analysis.
- Examine linguistic, ethical, and regulatory challenges, with particular attention to the Arabic academic context.
- Propose future research directions with practical applicability in higher education institutions.

2. Types of Plagiarism and Detection Challenges (with Illustrative Examples)

Classifying the various forms of plagiarism constitutes a crucial step toward understanding the nature of the problem and designing effective technical solutions. For explanatory and methodological purposes, this section presents simplified textual examples for each identified type, with the explicit clarification that these examples are purely illustrative and do not correspond to real published cases.

2.1 Verbatim Plagiarism

Description: Direct copying of the original text either in its entirety or with only minimal superficial modifications, such as changes in punctuation, spacing, or formatting.

Example Original text: "Artificial intelligence is a branch of computer science that aims to simulate human cognitive abilities such as learning and reasoning."

Copied text: "Artificial intelligence is a branch of computer science that aims to simulate human cognitive abilities such as learning and reasoning."

Analysis: The near-complete textual overlap makes this form of plagiarism straightforward to detect using text fingerprinting techniques and *n-gram*–based matching algorithms.

2.2 Near-Verbatim Plagiarism

Description: Copying a text while substituting a small number of words or introducing minor stylistic changes, without proper source attribution.

Example Original text: “Natural language processing techniques contribute to improving a computer’s understanding of human-written texts.”

Modified text: “Natural language processing methods help enhance a computer’s ability to understand texts produced by humans.”

Analysis: Lexical similarity remains high, allowing advanced text fingerprinting algorithms to detect this type of plagiarism with relatively high accuracy.

2.3 Paraphrasing Plagiarism

Description: Altering sentence structure and vocabulary while preserving the overall meaning of the original text.

Example Original text: “Modern plagiarism detection systems rely on artificial intelligence to analyse texts and compare them against large databases.”

Paraphrased text: “Contemporary systems for identifying plagiarism employ intelligent techniques that analyse textual content and compare it with extensive collections of sources.”

Analysis: This form is difficult to detect using surface-level matching alone and requires semantic analysis to identify underlying meaning similarity.

2.4 Translation Plagiarism

Description: Translating a text from one language into another without acknowledging the original source.

Example Original text (English): “Artificial intelligence systems are increasingly used to detect plagiarism in academic writing.”

Translated text (Arabic): “تُستخدم أنظمة الذكاء الاصطناعي بشكل متزايد لكشف الانتحال في الكتابة الأكاديمية.”

Analysis: No lexical overlap is present, making detection dependent on cross-lingual or multilingual plagiarism detection techniques.

2.5 Idea or Structural Plagiarism

Description: Appropriating the underlying idea, methodological framework, or argumentative structure of a work without copying its actual wording.

Example Original work: A study proposing a three-stage plagiarism detection process: preprocessing, feature extraction, and decision-making.

Plagiarised work: Another study presenting the same stages in the same logical order, with only terminological changes and without citing the original source.

Analysis: This type represents one of the most challenging forms of plagiarism to detect and typically requires conceptual and structural analysis rather than textual comparison.

2.6 Self-Plagiarism

Description: The reuse of substantial portions of an author's own previously published work without appropriate citation or disclosure.

Example: A researcher republishes an entire chapter from a master's thesis in a new journal article without referencing the original thesis.

Analysis: Despite involving the same author, this practice constitutes a breach of academic integrity and violates publication ethics.

2.7 AI-Assisted Plagiarism

Description: The use of text generation models to produce content that is semantically close to existing sources without proper attribution.

Example: Feeding a paragraph from a scholarly article into a language model and incorporating the generated output into an academic paper without citation.

Analysis: Such cases are difficult to detect using text fingerprinting alone and require advanced stylometric and semantic analysis techniques.

3. The Theoretical Foundations of Text Fingerprinting

Text fingerprinting constitutes one of the foundational technical pillars upon which plagiarism and academic misconduct detection systems have been built in digital environments. Its significance lies in its ability to provide an efficient mechanism for representing texts in a **compact and discriminative form**, enabling rapid and reliable comparison across large-scale document repositories. At its core, text fingerprinting transforms full textual content into a set of digital signatures or features that capture essential structural and linguistic properties, thereby facilitating the detection of similarity even within massive databases (Amirzhanov et al., 2025).

The conceptual importance of text fingerprinting stems from its capacity to reconcile two seemingly conflicting requirements: **data reduction** and **discriminative power**. By abstracting away from full-text representations while preserving similarity-relevant information, fingerprinting techniques enabled a paradigm shift from direct textual comparison to the comparison of compressed numerical representations. This shift laid the technical groundwork for scalable plagiarism detection systems capable of operating efficiently in near real-time conditions (Sajid et al., 2025).

This section aims to present the theoretical underpinnings of text fingerprinting by clarifying its conceptual definition, explicating its operational mechanisms, and reviewing the most influential algorithms associated with it. In doing so, it highlights text fingerprinting as the historical and methodological entry point that paved the way for modern plagiarism detection systems, including contemporary hybrid architectures that integrate fingerprinting with AI-driven semantic analysis.

3.1 Definition of Text Fingerprinting

Text fingerprinting is commonly defined as a **set of digital features or signatures extracted from a text**, designed to represent its distinctive characteristics in a compressed form without compromising the ability to detect similarity (Broder, 1997). In more recent literature, this definition has been reframed within a broader perspective that conceptualizes fingerprinting as a form of *task-oriented information reduction*, tailored specifically to support retrieval and comparison rather than deep semantic understanding (Amirzhanov et al., 2025).

Several key properties characterize effective text fingerprinting methods. These include **compactness**, whereby long texts are represented by a limited number of numerical values; **relative robustness**, which allows fingerprints to remain stable under minor textual variations; and **computational comparability**, enabling fast similarity computation at low computational cost. Together, these properties have rendered fingerprinting techniques indispensable for plagiarism detection in academic and institutional settings, particularly prior to the widespread adoption of deep semantic models. Even today, fingerprinting remains central as a **first-stage filtering mechanism** in most modern systems (Sajid et al., 2025).

3.2 Shingling and Similarity Measures

The **shingling** approach is based on segmenting a text into contiguous units—commonly referred to as *k-shingles* or *n-grams*—each representing a fixed-length sequence of words or characters. Within this framework, a document is modeled as a set of shingles, which can then be compared to those of other documents using set-based similarity metrics (Broder, 1997).

Among the most widely used measures in this context is the **Jaccard similarity coefficient**, which quantifies similarity as the ratio of the intersection to the union of two shingle sets. This measure enables the estimation of both global and partial similarity between texts, making it particularly effective for detecting verbatim copying and near-duplicate content (Amirzhanov et al., 2025).

Despite its conceptual simplicity and computational efficiency, shingling is subject to well-documented limitations. These include sensitivity to word order changes, the potentially large number of generated shingles for long documents, and a limited capacity to detect deeply paraphrased plagiarism. Consequently, recent research increasingly views shingling as a **surface-level representation strategy** that is highly effective in early detection stages but insufficient on its own for addressing complex forms of academic plagiarism (Wahle et al., 2021; Sajid et al., 2025).

3.3 The Winnowing Algorithm

The **Winnowing algorithm** represents one of the most influential methodological refinements in the domain of text fingerprinting, as it explicitly seeks to balance **detection accuracy** with **computational efficiency**. The algorithm operates by applying a sliding window over hashed shingles and selecting only the minimum hash value within each window. This process substantially reduces the number of retained fingerprints while preserving the ability to detect sufficiently long matching substrings (Schleimer et al., 2003).

A defining feature of Winnowing is its **guarantee property**, which ensures that any shared text fragment exceeding a predefined length threshold will be detected, regardless of its position within the document. This property has made Winnowing particularly well suited for large-scale plagiarism detection systems in academic and commercial contexts (Amirzhanov et al., 2025).

Contemporary studies indicate that Winnowing continues to be widely employed—not as a standalone solution, but as an **initial candidate retrieval stage** that precedes more computationally intensive analyses, such as embedding-based semantic similarity or deep neural language modeling (Tang et al., 2025). In this capacity, the algorithm retains its relevance within modern hybrid systems that integrate text fingerprinting with AI-driven semantic techniques.

4. Plagiarism Detection: From Surface Matching to Semantics

The evolution of plagiarism detection technologies reflects a profound methodological shift in how texts—and the relations among them—are conceptualized. Early approaches treated text primarily as a linear sequence of symbols amenable to literal comparison, whereas contemporary paradigms

increasingly view it as a complex semantic and contextual structure. This transition is not merely technical; it is also epistemological, insofar as it reconfigures the very notion of plagiarism from overt textual identity to a form of *knowledge reproduction* that may be concealed behind divergent linguistic realizations (Amirzhanov et al., 2025; Sajid et al., 2025). Accordingly, modern systems tend to move beyond what is *said* toward how meaning is constructed, preserved, and redistributed—even when lexical overlap is minimal (Wahle et al., 2021).

4.1 Surface Matching

Surface matching methods operate by directly comparing textual strings—at the level of words, spans, or sentences—using techniques such as **n-grams**, **text fingerprinting**, and exact (or near-exact) matching algorithms. Their principal advantages lie in computational simplicity and speed, which explains their historical centrality in early—and many still widely deployed—commercial plagiarism detection systems (Amirzhanov et al., 2025; Sajid et al., 2025).

Yet, the effectiveness of surface matching remains largely confined to *verbatim* or *near-verbatim* copying.It is substantially weakened by linguistic obfuscation strategies such as paraphrasing, syntactic reordering, and synonym substitution, especially when these transformations preserve meaning while reducing lexical overlap (Wahle et al., 2021). Moreover, because these methods privilege form over function, they often struggle to distinguish illegitimate reuse from legitimate similarity arising from terminological constraints and conventional phrasing in specialized scholarly domains (Amirzhanov et al., 2025). Consequently, surface matching is best understood as a necessary *front-end* stage—valuable for rapid screening, but insufficient as a standalone solution in complex academic settings (Sajid et al., 2025).

4.2 Stylometric Analysis

Stylometric analysis constitutes a qualitative leap in plagiarism detection by shifting attention from *what* is written to *how* it is written. It relies on extracting relatively stable stylistic signals—such as sentence-length distributions, syntactic patterning, function-word frequencies, and punctuation habits—in order to construct an authorial "style signature" (Manzoor et al., 2025).

This approach becomes particularly salient in **intrinsic plagiarism detection** and authorship-related misconduct, where no external source text is available for direct comparison. In such cases, abrupt stylistic discontinuities within a single document may function as probabilistic indicators of inserted material originating from a different authorial source (Manzoor et al., 2023; Manzoor et al., 2025). Nevertheless, stylometry remains inherently probabilistic and sensitive to confounding variables, including longitudinal style development, genre shifts, disciplinary conventions, and the effects of academic supervision or editorial intervention. For this reason, recent research increasingly emphasizes integrating stylometric evidence with complementary signals (e.g., semantic similarity and retrieval-based filtering) to strengthen robustness and reduce false positives (Sajid et al., 2025).

4.3 Semantic Analysis

Semantic analysis represents the most advanced stage in the trajectory of plagiarism detection, aiming to transcend surface form in order to model meaning and context. Contemporary semantic approaches rely on representation learning—particularly sentence- and document-level embeddings—and neural architectures that enable detection of conceptually equivalent passages even when lexical similarity is low (Tang et al., 2025). This is crucial for identifying *deep paraphrase plagiarism*, where the plagiarized text may appear linguistically dissimilar while retaining the same underlying conceptual structure (Wahle et al., 2021).

Recent work also highlights the strategic value of **hybrid semantic retrieval pipelines**, in which embedding-based filtering reduces the search space and approximate nearest-neighbor indexing (e.g., FAISS) enables scalable similarity search across large corpora (Tang et al., 2025). In parallel, proposals combining semantic modeling with higher-level relational reasoning—such as textual analogies and contextual reformulation—illustrate a broader movement toward semantic generalization beyond lexical overlap (Andriantsialo et al., 2024).

However, semantic methods raise persistent challenges related to computational cost, interpretability, and embedded model biases. As a result, current literature increasingly treats explainability not as an optional feature but as a prerequisite for institutional adoption, particularly in high-stakes academic contexts (Opitz et al., 2025). Hybrid frameworks that combine embedding-based similarity with knowledge-grounded explanation mechanisms have therefore gained traction as a means of improving both accuracy and interpretability (Saeed et al., 2024).

5. A Proposed Framework for a Hybrid Plagiarism Detection System (Fingerprinting + AI)

In light of the inherent limitations of traditional plagiarism detection systems on the one hand, and the computational and interpretability challenges associated with fully AI-driven approaches on the other, this paper proposes a **hybrid framework** that integrates the efficiency and speed of text fingerprinting techniques with the analytical depth afforded by semantic and stylistic AI-based methods (Potthast et al., 2010; Barrón-Cedeño et al., 2013).

The proposed framework is grounded in the principle of **functional separation between processing stages**. Surface-level techniques are deployed in the early phases to reduce the search space and alleviate computational burden, while intelligent models are activated in later stages to conduct fine-grained analysis and support final judgment. This integration seeks to achieve a pragmatic balance between **accuracy, efficiency, and interpretability**, rendering the framework particularly suitable for academic and institutional contexts (Clough, 2000).

5.1 Preprocessing

Preprocessing constitutes the cornerstone of any effective plagiarism detection system, as the quality of this stage directly influences the accuracy of subsequent analyses (Potthast et al., 2010). This phase encompasses a set of linguistic and structural operations, most notably the removal of non-semantic noise (such as unnecessary symbols, formatting artifacts, and extraneous markers), orthographic normalization, and the elimination of non-functional redundancy.

Additionally, preprocessing involves **text normalization** aimed at reducing formal variability that lacks semantic value, including the unification of character forms and the handling of morphological variation. This phase also includes segmenting the text into processable units (sentences, paragraphs, or semantically coherent segments). These steps are essential to ensure that any detected similarity in later stages reflects **genuine conceptual proximity rather than incidental formal variation**, thereby enhancing the epistemic validity of similarity assessments (Barrón-Cedeño et al., 2013).

5.2 Text Fingerprinting

At this stage, **text fingerprints** are extracted to provide a compact and efficient representation of the document's content. This process relies on the selection of distinctive textual units (e.g., *n*-grams), which are subsequently transformed into numerical fingerprints optimized for rapid comparison (Potthast et al., 2010).

These fingerprints are employed to index documents within the database and retrieve a limited set of candidate documents that are likely to exhibit similarity with the target text. The significance of this

stage lies in its role as a **primary filtering mechanism**, substantially reducing the number of comparisons required in later, more computationally intensive stages, and thus lowering the overall system cost (Clough, 2000).

Although fingerprinting techniques exhibit limited effectiveness in detecting deeply paraphrased plagiarism, they remain highly efficient in identifying explicit and near-explicit similarities. As such, they constitute an indispensable preparatory step in large-scale plagiarism detection systems (Barrón-Cedeño et al., 2013).

5.3 Semantic and Stylistic Analysis

Once candidate documents or segments have been identified, **AI-based models** are applied to perform deeper analyses targeting both meaning and style. In semantic analysis, distributed semantic representation models are employed to extract conceptual relationships between text segments and estimate degrees of semantic similarity, even in the absence of direct lexical overlap (Zhang & Clark, 2020). This capability enables the detection of indirect plagiarism, paraphrasing, and idea transfer accompanied by structural or lexical transformation (Barrón-Cedeño et al., 2013).

In parallel, **stylistic analysis** is utilized to identify inconsistencies or abrupt shifts in writing style, either within a single document or in comparison with previously established stylistic profiles of the author. Such approaches draw on advances in authorship analysis and stylistic modeling (Stamatatos, 2009). The combination of semantic and stylistic perspectives strengthens the reliability of the assessment by treating similarity as a **multidimensional phenomenon** encompassing form, meaning, and stylistic signature.

5.4 Decision-Making

The decision-making stage represents the culmination of the hybrid system, wherein outputs from previous stages are integrated into an ensemble model that generates a comprehensive report detailing similarity scores, types, and levels (Potthast et al., 2010). The report highlights suspicious segments and provides explicit justifications for their classification.

Importantly, this report is not intended to replace human judgment, but rather to **support expert evaluation** by offering interpretable quantitative and qualitative indicators. A clear distinction is maintained between automated detection and academic evaluation, ensuring that the final decision remains in the hands of domain specialists who can account for disciplinary norms, contextual factors, and the boundaries of legitimate citation practices (Clough, 2000).

In this manner, the plagiarism detection system evolves from a purely automated surveillance tool into an **assistive instrument** that promotes academic integrity and fosters a deeper understanding of scholarly writing practices.

6. Linguistic Specificities of Arabic in Text Fingerprinting–Based Plagiarism Detection

The Arabic language poses distinctive linguistic and technical challenges that render plagiarism detection more complex when compared to structurally simpler languages such as English. These challenges stem primarily from Arabic's rich morphology, syntactic flexibility, and orthographic variation, all of which directly affect the effectiveness of text fingerprinting algorithms that rely on surface-level string matching (Habash, 2010; Darwish, 2014).

6.1 Morphological Richness and Derivational Variability

Arabic is characterised by a root-based derivational morphology that enables the generation of a large number of word forms from a single root, with relatively stable semantic content but significant formal variation. For instance, the root *k–t–b* can yield forms such as *kataba* (wrote), *yaktubu* (writes), *maktūb* (written), *kitāb* (book), and *maktaba* (library). This morphological diversity renders literal string matching insufficient, as the same underlying idea may be reused through different inflectional or derivational forms without being detected by traditional fingerprinting systems.

Illustrative example:

- **Original text:** "The study relies on the analysis of written texts."
- **Rewritten text:** "The research is based on analysing textual writings."

Despite their semantic equivalence, literal text fingerprinting may fail to identify similarity without prior morphological processing, such as stemming or lemmatisation (Farghaly & Shaalan, 2009).

6.2 Orthographic Variation and Text Normalisation

Arabic orthography exhibits multiple forms for certain letters—most notably the various forms of *alif* (ا، أ، إ، آ) and the alternation between *yāʾ* (ي) and *alif maqṣūra* (ى)—as well as optional diacritics. These variations can result in different textual fingerprints despite identical semantic content.

Example:
"مسؤولية الباحث" vs. "مسئولية الباحث"

While linguistically trivial, such variations can significantly hinder automatic matching unless robust orthographic normalisation mechanisms are applied (Habash et al., 2012).

6.3 Syntactic and Rhetorical Flexibility

Arabic allows a high degree of syntactic flexibility, including word order variation, fronting and postponement, metaphorical expressions, and extensive synonymy. This flexibility increases the difficulty of detecting paraphrasing through surface-based methods.

Example:

- **Original text:** "Artificial intelligence contributes to the development of plagiarism detection tools."
- **Rewritten text:** "Plagiarism detection tools have witnessed significant development thanks to artificial intelligence techniques."

Although the propositional content remains unchanged, the restructured syntax and lexical variation substantially reduce surface similarity, posing challenges for traditional fingerprinting approaches (Stamatatos et al., 2015).

6.4 Proposed Solutions for Addressing Arabic-Specific Challenges

To mitigate the impact of Arabic linguistic characteristics on plagiarism detection, the literature recommends several complementary strategies:

- Adoption of advanced orthographic normalisation techniques;
- Integration of morphological analysis through stemming or lemmatisation;
- Combining text fingerprinting with semantic representations trained on Arabic corpora;

- Development of standardised Arabic benchmark datasets to support evaluation and comparative testing (Darwish & Magdy, 2014).

Such measures are essential for improving detection accuracy and ensuring fair and reliable plagiarism assessment in Arabic academic contexts.

7. Evaluation Protocols and Performance Metrics

Evaluation protocols constitute a fundamental component in assessing the effectiveness of plagiarism detection systems, as they enable objective, reproducible, and comparable performance measurement. These protocols rely on a combination of quantitative and qualitative metrics, which vary depending on the system's intended purpose and the type of plagiarism being targeted (Potthast et al., 2010; Gipp et al., 2015).

7.1 Core Evaluation Tasks

Evaluation tasks are typically divided into two principal categories (Potthast et al., 2014):

- **Source Retrieval:** The system's ability to identify potential source documents from a large reference corpus.
- **Text Alignment:** The system's ability to accurately localise and align plagiarised passages within the analysed document.

These two tasks reflect complementary stages of the detection process and are often evaluated independently to obtain a clearer understanding of system performance.

7.2 Quantitative Metrics

Commonly used statistical metrics include the following (Manning et al., 2008; Potthast et al., 2010):

- **Precision:** The proportion of correctly identified plagiarised passages relative to the total number of passages flagged by the system.
- **Recall:** The proportion of correctly detected plagiarised passages relative to the total number of passages that are actually plagiarised.
- **F1-score:** The harmonic mean of precision and recall, providing a balanced measure of detection performance.

Illustrative example: If a system identifies 80 passages as plagiarised, of which 60 are correct, and the true number of plagiarised passages is 100, then:

- Precision = 60 / 80 = 0.75
- Recall = 60 / 100 = 0.60

Such metrics allow for systematic comparison between systems under controlled experimental conditions.

7.3 Qualitative Metrics

In addition to quantitative indicators, qualitative evaluation metrics play a crucial role in assessing the practical usability of plagiarism detection systems (Sanchez-Perez et al., 2015). These metrics take into account:

- The accuracy of the detected passage boundaries;

- The system's ability to distinguish legitimate citation from unethical plagiarism;
- The clarity and interpretability of the generated reports for end users, such as editors and reviewers.

The importance of standardised evaluation lies in the fact that initiatives such as **PAN/CLEF** have become global benchmarks for plagiarism detection research. These initiatives provide shared datasets, well-defined evaluation protocols, and unified performance metrics, thereby enabling rigorous scientific comparison across different approaches and systems (Potthast et al., 2019).

8. Critical Discussion: Strengths and Limitations

A critical review of the scholarly literature highlights that plagiarism detection techniques cannot be evaluated in isolation from the contexts in which they are deployed—whether in terms of the nature of the texts, the scale of databases, linguistic characteristics, or prevailing academic policies. Each technical approach exhibits strengths that make it suitable for detecting particular forms of plagiarism, alongside inherent limitations that constrain its effectiveness in other scenarios. Understanding these strengths and limitations is therefore a prerequisite for the development of more accurate and equitable plagiarism detection systems.

8.1 Strengths of Text Fingerprinting

Text fingerprinting represents one of the earliest and most established techniques in plagiarism detection and has demonstrated robust performance in large-scale environments, particularly with the exponential growth of digital content. Its principal strengths include the following:

High Efficiency in Processing Large-Scale Textual Data

Text fingerprinting relies on compact representations of documents, enabling the rapid comparison of millions of texts through the use of inverted indexes. This makes it particularly well suited for large academic databases (Broder, 1997; Schleimer et al., 2003).

Ease of Implementation and Scalability

Algorithms such as *Shingling* and *Winnowing* are characterised by their relative simplicity of implementation and ease of integration into diverse information systems, while also offering high horizontal scalability.

Effectiveness in Detecting Verbatim Plagiarism

Empirical studies consistently demonstrate that text fingerprinting achieves very high accuracy rates in detecting direct copying and partial textual reuse, thereby constituting the first line of defence in most plagiarism detection systems (Barrón-Cedeño & Rosso, 2009).

8.2 Limitations of Text Fingerprinting

Despite its strengths, text fingerprinting suffers from structural limitations that restrict its ability to detect more sophisticated forms of plagiarism:

Limited Performance in Detecting Deep Paraphrasing

Because text fingerprinting primarily relies on surface-level similarity, it is largely ineffective in identifying paraphrasing that preserves meaning while substantially altering linguistic structure—a practice increasingly common in contemporary academic plagiarism (Potthast et al., 2019).

Sensitivity to Preprocessing Decisions

The performance of text fingerprinting is highly sensitive to preprocessing choices, including normalisation strategies, *k*-gram length selection, and stop-word removal policies. Inadequate parameter tuning may lead to elevated false positive or false negative rates.

8.3 Strengths of Semantic and Stylometric Analysis

With advances in artificial intelligence, semantic and stylometric analysis has emerged as a powerful alternative or complement to text fingerprinting, particularly in addressing obfuscated forms of plagiarism:

Detection of Meaning-Based Rather Than Form-Based Similarity

Modern semantic models rely on vector-based representations of textual content, enabling the detection of conceptual similarity even in the absence of lexical overlap. This capability has proven especially effective in identifying paraphrased plagiarism (Stamatatos et al., 2015).

Effectiveness Against Advanced Forms of Plagiarism

Stylometric analysis contributes to identifying anomalous shifts in writing style within a single document, making it particularly useful in cases of intrinsic plagiarism and plagiarism facilitated by automated text generation tools.

8.4 Technical and Ethical Limitations

Despite their advanced capabilities, semantic and stylometric approaches are subject to several notable constraints:

High Computational Cost

The extraction and processing of semantic representations demand substantial computational resources, particularly when applied to large-scale text collections. This requirement may limit their deployment in resource-constrained environments.

Limited Interpretability of Results

Some semantic models function as “black boxes,” making it difficult for non-specialist users to understand the rationale behind classifying a passage as plagiarised. This lack of transparency raises ethical and regulatory concerns related to accountability and explainability.

8.5 Toward an Integrative Approach

In line with recent trends in the literature, the findings of this study indicate that an integrative approach combining text fingerprinting with semantic and stylometric analysis offers the most effective balance between efficiency and accuracy (Potthast et al., 2019). While text fingerprinting can be employed to rapidly narrow the search space and retrieve candidate documents, semantic analysis can then be applied to ambiguous segments to refine the final decision. Such hybrid approaches are particularly well suited to addressing the complexities of contemporary academic plagiarism, especially in multilingual environments.

9. Ethical and Regulatory Considerations

The use of plagiarism detection systems is not limited to technical and algorithmic dimensions; rather, it raises a set of ethical and regulatory issues that directly affect the core of the research process and the principles of academic fairness. Although such systems play a vital role in safeguarding academic integrity, they may become instruments of exclusion or academic injustice if deployed without a clear normative framework that takes contextual factors into account and balances scientific rigor with researchers' rights. Accordingly, addressing ethical and regulatory dimensions constitutes a fundamental prerequisite for the effective and responsible implementation of plagiarism detection systems within academic institutions and scholarly journals (Bretag, 2016; Fishman, 2014).

9.1 Distinguishing Legitimate Citation from Plagiarism

One of the most common challenges in the use of plagiarism detection systems lies in the confusion between legitimate scholarly citation and unethical plagiarism. Academic citation is a lawful and essential practice for the accumulation of knowledge, provided that sources are properly acknowledged in accordance with established standards and that citations are employed within a clearly defined analytical or critical context (Pecorari, 2013).

However, automated detection systems—particularly those relying on similarity percentage scores—may misclassify properly cited material, widely used definitions, or standard methodological descriptions as instances of plagiarism if they lack sufficiently refined discrimination mechanisms. For example, high similarity scores frequently appear in "Methodology" or "Theoretical Framework" sections without necessarily indicating unethical conduct (Garner, 2011).

For this reason, the literature strongly recommends complementing quantitative similarity measures with qualitative analysis that takes into account the following criteria (Roig, 2015):

- The presence or absence of proper citations and references;
- The length of overlapping text segments and the context in which they are used;
- The functional nature of the text (e.g., definitional, methodological, or analytical).

9.2 The Role of Human Judgment

Contemporary studies consistently emphasise that plagiarism detection systems should be employed as **decision-support tools**, rather than as substitutes for human academic judgment (Sutherland-Smith, 2010). Regardless of their technical sophistication, automated systems remain incapable of fully grasping scientific context, assessing scholarly intent, or reliably distinguishing between unintentional error and deliberate misconduct.

The importance of human judgment becomes particularly evident in borderline cases, such as:

- Partial paraphrasing accompanied by indirect or incomplete attribution;
- Similarity arising from the use of standardised terminology within a specific discipline;
- Instances of self-plagiarism, which are governed by varying policies across journals and institutions.

Accordingly, the involvement of editors, supervisors, and academic integrity committees in interpreting similarity reports is not merely advisable but constitutes an ethical and regulatory necessity (COPE, 2019).

9.3 Privacy and Data Protection

The deployment of plagiarism detection systems raises sensitive concerns related to privacy protection and intellectual property rights, especially when dealing with unpublished works such as master's and

doctoral theses or manuscripts submitted for peer review. A major concern lies in the practice adopted by some systems of storing analysed texts within proprietary databases, potentially exposing them to unauthorised reuse or infringement of authors' rights (Gipp et al., 2015).

These concerns are further intensified when commercial platforms operating outside institutional governance frameworks are used. Consequently, this study recommends the following measures (European Commission, 2018):

- Ensuring transparency in data storage and processing policies;
- Clearly defining access rights to submitted texts and generated reports;
- Complying with national and international data protection regulations.

9.4 Institutional and Regulatory Frameworks

The fair and effective use of plagiarism detection systems necessitates the establishment of clear regulatory frameworks within academic institutions and scholarly journals. Such frameworks should be articulated through written and publicly accessible policies that explicitly define:

- Acceptable similarity thresholds, taking into account disciplinary differences and the nature of academic sections;
- Appeal and review mechanisms that grant researchers the right to contest results and provide scholarly justification;
- Guidelines governing the use of artificial intelligence tools, both for text generation and plagiarism detection, in order to ensure transparency and prevent misuse (Floridi et al., 2018).

The study further recommends that these policies be accompanied by awareness-raising and training programmes for researchers and students, with the aim of fostering a culture of academic integrity and transforming plagiarism detection systems from purely supervisory instruments into supportive mechanisms that enhance research quality.

10. Conclusion and Recommendations

Plagiarism detection has become one of the central issues in contemporary academic research, particularly in the context of rapid digital transformation, the widespread availability of electronic databases, and the emergence of generative artificial intelligence tools. This study has sought to provide an in-depth scholarly examination of the role of artificial intelligence in plagiarism detection, with a particular focus on text fingerprinting as one of the core technical foundations in this field.

First: Research Findings

Based on theoretical analysis and a systematic review of the relevant scholarly literature, the study arrives at several key findings, which may be summarised as follows:

The Importance of Text Fingerprinting as a Technical Foundation

The study demonstrates that text fingerprinting continues to represent a cornerstone of large-scale plagiarism detection systems, owing to its high computational efficiency and its ability to process vast volumes of text within short timeframes. Algorithms such as *Shingling* and *Winnowing* have proven particularly effective in detecting verbatim plagiarism and partial textual reuse.

Limitations of Text Fingerprinting in Addressing Advanced Plagiarism

The findings reveal that reliance on text fingerprinting alone becomes insufficient when dealing with advanced forms of plagiarism, including deep paraphrasing, translation plagiarism, and the appropriation of ideas or structural frameworks. These forms do not depend on lexical overlap but rather on semantic and conceptual similarity, which lies beyond the reach of surface-level matching techniques.

Effectiveness of AI-Based Hybrid Approaches

The study concludes that integrating text fingerprinting with AI-driven semantic and stylometric analysis results in a significant improvement in detection accuracy, while reducing both false positive and false negative rates. Hybrid systems are shown to be better equipped to cope with the growing complexity of contemporary academic texts.

The Specificity of the Arabic Language and Its Impact on Detection Accuracy

The analysis confirms that the Arabic language presents additional challenges when compared to Western languages, due to its rich morphology, syntactic diversity, and orthographic variability. Applying models developed for other languages to Arabic without linguistic adaptation leads to a noticeable decline in performance, thereby underscoring the need for language-specific solutions.

Escalating Plagiarism Challenges in the Era of Generative Artificial Intelligence

The study highlights that generative language models pose a novel challenge to plagiarism detection systems, as they enable the production of text that is formally original yet semantically close to multiple source materials. This development complicates detection processes and renders advanced semantic and stylometric analysis indispensable.

Second: Recommendations

In light of these findings, the study offers a set of scientific and practical recommendations aimed at improving plagiarism detection systems and strengthening academic integrity:

Adoption of Hybrid Detection Systems

Academic institutions and scholarly journals are encouraged to adopt hybrid plagiarism detection systems that combine text fingerprinting with semantic and stylometric analysis, rather than relying solely on similarity percentages derived from lexical matching.

Development of Arabic-Specific Solutions

The study strongly advocates investment in the development of Arabic or multilingual language models that explicitly account for the morphological and syntactic characteristics of Arabic, as well as the creation of standardised Arabic benchmark datasets to support research and system development.

Avoiding Sole Reliance on Similarity Scores

The study recommends that academic and editorial decisions should not be based exclusively on similarity percentages. Instead, qualitative analysis of the nature and context of overlapping passages is essential to distinguish legitimate citation practices from unethical plagiarism.

Ethical Governance of AI Tool Usage

The study emphasises the importance of establishing clear policies to regulate the use of artificial intelligence tools in scholarly research, both in text generation and in plagiarism detection, in order to ensure transparency and protect authors' intellectual rights.

Enhancing Academic Awareness

The study recommends embedding principles of academic integrity and responsible citation practices within university curricula, alongside training researchers and students in the informed and ethical use of plagiarism detection tools.

Third: Future Research Directions

This study opens several promising avenues for future research, most notably the development of cross-lingual plagiarism detection systems, the deployment of explainable artificial intelligence models capable of justifying their decisions, and the integration of deep conceptual analysis for detecting idea-level plagiarism. Such directions have the potential to significantly enhance the quality of scholarly research and to reinforce the principles of academic integrity within the contemporary digital environment.

References

Amirzhanov, A., Turan, C., & Makhmutova, A. (2025). *Plagiarism types and detection methods: A systematic survey of algorithms in text analysis*. Frontiers in Computer Science, 7, Article 1504725. https://doi.org/10.3389/fcomp.2025.1504725

Andriantsialo, E. F., Ratianantitra, V. M., & Mahatody, T. (2024). *Semantic exploration of textual analogies for advanced plagiarism detection*. In *Proceedings of the 16th International Conference on Computational Processing of Portuguese (PROPOR 2024) – Vol. 2* (pp. 130–133). Association for Computational Linguistics.

Barrón-Cedeño, A., & Rosso, P. (2009). On automatic plagiarism detection based on n-grams comparison. *Proceedings of the European Conference on Information Retrieval (ECIR 2009)*, 696–700. Springer. https://doi.org/10.1007/978-3-642-00958-7_76

Barrón-Cedeño, A., Gupta, P., Rosso, P., & Riedl, M. (2013). Plagiarism meets paraphrasing: Insights for the next generation in automatic plagiarism detection. *Computational Linguistics, 39*(4), 917–947. https://doi.org/10.1162/COLI_a_00153

Bretag, T. (2016). Challenges in addressing plagiarism in education. *PLoS Medicine, 13*(12), e1002183. https://doi.org/10.1371/journal.pmed.1002183

Broder, A. Z. (1997). On the resemblance and containment of documents. *Proceedings of the Compression and Complexity of Sequences 1997*, 21–29. IEEE Computer Society. https://doi.org/10.1109/SEQUEN.1997.666900

Broder, A. Z., Glassman, S. C., Manasse, M. S., & Zweig, G. (1997). Syntactic clustering of the Web. *Computer Networks and ISDN Systems*, 29(8–13), 1157–1166. https://doi.org/10.1016/S0169-7552(97)00031-7

Clough, P. (2000). Plagiarism in natural and programming languages: An overview of current tools and technologies. *Department of Computer Science, University of Sheffield*.

Committee on Publication Ethics (COPE). (2019). *COPE guidelines on plagiarism*. https://publicationethics.org/resources/guidelines

Darwish, K., & Magdy, W. (2014). Arabic information retrieval. *Foundations and Trends® in Information Retrieval, 7*(4), 239–342. https://doi.org/10.1561/1500000024

European Commission. (2018). *General Data Protection Regulation (GDPR): Regulation (EU) 2016/679*. Official Journal of the European Union. https://eur-lex.europa.eu/eli/reg/2016/679/oj

Farghaly, A., & Shaalan, K. (2009). Arabic natural language processing: Challenges and solutions. *ACM Transactions on Asian Language Information Processing (TALIP), 8*(4), Article 14. https://doi.org/10.1145/1644879.1644881

Fishman, T. (2014). *The fundamental values of academic integrity* (2nd ed.). International Center for Academic Integrity.

Floridi, L., Cowls, J., Beltrametti, M., et al. (2018). AI4People—An ethical framework for a good AI society: Opportunities, risks, principles, and recommendations. *Minds and Machines, 28*(4), 689–707. https://doi.org/10.1007/s11023-018-9482-5

Garner, H. R. (2011). Combating unethical publications with plagiarism detection services. *Urologic Oncology: Seminars and Original Investigations, 29*(1), 95–99. https://doi.org/10.1016/j.urolonc.2010.09.005

Gipp, B., Meuschke, N., & Beel, J. (2015). Comparative evaluation of text- and citation-based plagiarism detection approaches using GuttenPlag. *Journal of Informetrics, 9*(4), 920–941. https://doi.org/10.1016/j.joi.2015.08.004

Habash, N. (2010). *Introduction to Arabic natural language processing*. Morgan & Claypool Publishers. https://doi.org/10.2200/S00277ED1V01Y201008HLT007

Habash, N., Eskander, R., & Hawwari, A. (2012). A morphological analyzer for Arabic dialects. *Proceedings of the 12th Conference of the European Chapter of the ACL*, 182–189.

Manning, C. D., Raghavan, P., & Schütze, H. (2008). *Introduction to information retrieval*. Cambridge University Press.

Manzoor, M. F., Farooq, M. S., & Abid, A. (2025). *Stylometry-driven framework for Urdu intrinsic plagiarism detection: A comprehensive analysis using machine learning, deep learning, and large language models*. Neural Computing and Applications, 37, 6479–6513. https://doi.org/10.1007/s00521-024-10966-w

Manzoor, M. F., Farooq, M. S., Haseeb, M., Farooq, U., Khalid, S., & Abid, A. (2023). *Exploring the landscape of intrinsic plagiarism detection: Benchmarks, techniques, evolution, and challenges*. IEEE Access, 11, 140519–140545. https://doi.org/10.1109/ACCESS.2023.3338855

Opitz, J., Möller, L., Michail, A., Padó, S., & Clematide, S. (2025). *Interpretable text embeddings and text similarity explanation: A survey* (arXiv:2502.14862). https://doi.org/10.48550/arXiv.2502.14862

Pecorari, D. (2013). *Teaching to avoid plagiarism: How to promote good source use*. Open University Press.

Potthast, M., Barrón-Cedeño, A., Eiselt, A., Stein, B., & Rosso, P. (2010). Overview of the 2nd International Competition on Plagiarism Detection. *CLEF 2010 Working Notes*, 1–17.

Potthast, M., Gollub, T., Hagen, M., Kiesel, J., Michel, M., Oberländer, A., Tippmann, M., Barrón-Cedeño, A., Gupta, P., & Rosso, P. (2014). Overview of the 5th International Competition on Plagiarism Detection. *CLEF 2014 Working Notes Conference Labs and Workshops*.

Potthast, M., Hagen, M., Gollub, T., Tippmann, M., Kiesel, J., Rosso, P., Stamatatos, E., & Stein, B. (2019). Overview of the PAN 2019 Author Identification, Author Profiling, and Style Change Detection Tasks. *CLEF 2019 Labs and Workshops, Notebook Papers*.

Potthast, M., Stein, B., & Barrón-Cedeño, A. (2010). An evaluation framework for plagiarism detection. *Proceedings of the 23rd International Conference on Computational Linguistics (COLING 2010)*, 997–1005.

Roig, M. (2015). Avoiding plagiarism, self-plagiarism, and other questionable writing practices: A guide to ethical writing. *Office of Research Integrity*.
https://ori.hhs.gov/plagiarism-0

Saeed, S., Rajput, Q., & Haider, S. (2024). *SUMEX: A hybrid framework for semantic textUal siMilarity and EXplanation generation*. Information Processing & Management, 61(5), 103771.
https://doi.org/10.1016/j.ipm.2024.103771

Sajid, M., Sanaullah, M., Fuzail, M., Malik, T. S., & Shuhidan, S. M. (2025). *Comparative analysis of text-based plagiarism detection techniques*. PLOS ONE, 20(4), e0319551.
https://doi.org/10.1371/journal.pone.0319551

Sanchez-Perez, M. A., Gelbukh, A., Sidorov, G., & Gómez-Adorno, H. (2015). Plagiarism detection with genetic-based parameter tuning. *Expert Systems with Applications, 42*(3), 1858–1869.
https://doi.org/10.1016/j.eswa.2014.09.021

Schleimer, S., Wilkerson, D. S., & Aiken, A. (2003). Winnowing: Local algorithms for document fingerprinting. *Proceedings of the 2003 ACM SIGMOD International Conference on Management of Data*, 76–85. ACM.
https://doi.org/10.1145/872757.872770

Stamatatos, E. (2009). A survey of modern authorship attribution methods. *Journal of the American Society for Information Science and Technology, 60*(3), 538–556. https://doi.org/10.1002/asi.21001

Stamatatos, E., Potthast, M., Stein, B., & Rosso, P. (2015). Overview of the PAN/CLEF 2015 evaluation lab. *CLEF 2015 Working Notes Conference Labs and Workshops*.

Sutherland-Smith, W. (2010). Retribution, deterrence and reform: The dilemmas of plagiarism management in universities. *Journal of Higher Education Policy and Management, 32*(1), 5–16.
https://doi.org/10.1080/13600800903440519

Tang, J., Hu, Q., & Han, Z. (2025). *Efficient plagiarism detection via sentence embeddings and FAISS-based retrieval: Notebook for the Foshan University Artificial Intelligence Lab at CLEF 2025*. In *Working Notes of CLEF 2025* (CEUR Workshop Proceedings).

Wahle, J. P., Ruas, T., Meuschke, N., & Gipp, B. (2021). *Are neural language models good plagiarists? A benchmark for neural paraphrase detection*. In *Proceedings of the ACM/IEEE Joint Conference on Digital Libraries (JCDL 2021)* (pp. 226–229). IEEE. https://doi.org/10.1109/JCDL52503.2021.00065

Zhang, Y., & Clark, S. (2020). Sentence ordering and coherence modeling using contextualized word representations. *Proceedings of the AAAI Conference on Artificial Intelligence, 34*(05), 9370–9378.