Translating or Stealing? Probing the Limits of Cross-lingual Plagiarism Detection Systems in Literary Texts

DOI: https://doi.org/10.33806/ijaes1026

Abdulfattah Omar
Port Said University, Egypt
The Australian National University, Australia

Wafya Ibrahim Hamouda
Tanta University, Egypt

Waheed M. A. Altohami
Prince Sattam bin Abdulaziz University, Saudi Arabia
Mansoura University, Egypt

Received: 7.3.2025 Accepted: 23.11.2025 Published Online: 26.11.2025

Abstract: This study compares three plagiarism detection systems (Rabin-Karp, KNN, and Word2Vec) to measure their effectiveness in detecting cross-lingual plagiarism in Arabic literary texts translated from English. The dataset consisted of an Arabic translation of Daly Walker's '*I am the Grass'* (2012) conducted by the authors and evaluated by three translators with experience of more than ten years. It is divided into 60 percent directly translated, 30 percent paraphrased, and 10 percent original content. Findings showed that KNN achieved the highest precision in detecting cross-lingual plagiarism (26.7%), while Word2Vec performed best with paraphrased content (16.7%). Additionally, Rabin-Karp was most reliable in detecting original content (80% precision); however, all three systems demonstrated low overall accuracy (23–26%). These findings highlight the limitations of current systems when applied to Arabic texts, primarily due to the language's morpho-syntactic and lexical complexities. Given the limited scope of the study, as it analyzes a single text, it recommends expanding to multiple genres for broader generalizability. Furthermore, the study recommends the development of more sophisticated, hybrid plagiarism detection systems and the development of rich Arabic corpora to enhance their performance.

Keywords: Arabic, cross-lingual plagiarism, detection accuracy, detection precision, translation

1. Introduction

1.1 Background

Plagiarism detection systems (PDSs) are increasingly important in academia and beyond, as the internet makes it easier to pass off others' work as one's own through

copying and pasting. They offer reviewers all necessary information to assert complete or partial cases of plagiarism based on the copied resources. However, as Wall (2003) mentions, plagiarism now extends beyond individuals to institutions that proliferate platforms based on illegal text appropriation. Surprisingly, such unethical practices, including literary theft and infringing intellectual property rights, have become increasingly common in the age of the Internet with alarming numbers of replica texts, exacerbated by the spread of plagiarism via modern means of communication. PDSs employ different techniques for comparing documents against current databases to identify any similarities that may indicate plagiarism. Recently, more advanced algorithms and machine learning technologies have notably improved the accuracy and efficiency of such systems, and therefore they are now essential tools for researchers, educators, and content creators. In this regard, Sutherland-Smith (2008) affirms that these tools ensure work originality, proper citation, and academic integrity. As technology advances, PDSs will grow more sophisticated, thereby supporting intellectual property standards in the digital age.

Plagiarism generally refers to intellectual offenses that result from intentional or unintentional misuse of others' work. Intentional plagiarism includes deliberate acts such as translation without attribution, copying, and mosaic manipulation. Conversely, unintentional plagiarism results from cross-cultural differences in intellectual property practices, lack of awareness of citation norms, and poor paraphrasing skills (cf. Haitch 2016). Such distinction is essential for evaluating PDSs' performance since most of them are designed to detect textual similarities regardless of intent. However, it should be noted that PDSs may flag legitimate intertextual borrowing as plagiarism or miss subtle intentional obfuscations.

Within academia, this might happen directly (direct plagiarism) by copying text verbatim without attribution or accidentally (accidental plagiarism) due to negligence or misquoting. Further to this, in some cases, a writer might reuse parts of his previously published work without proper citation (i.e., self-plagiarism). A more cunning type of plagiarism is what Angelil-Carter (2014) refers to as 'mosaic plagiarism,' which involves paraphrasing and blending others' work as a new product. This means that mosaic plagiarism happens intentionally using the same language of the original sources. It is a kind of patchwork, as a mosaic plagiarist pieces together phrases, ideas, or small passages from diverse sources, paraphrases them, rearranges them, and presents them as original. However, the current study focuses on a rare but serious type of plagiarism – translation plagiarism – where the same literary elements (e.g., setting, characters, plot, etc.) of a particular literary work are borrowed and claimed original but in a different language (Teresa 2004). That is, while mosaic plagiarism and translation plagiarism involve disguising appropriation, mosaic plagiarism manipulates surface-level form within one language, whereas translation plagiarism masks ownership through cross-linguistic transfer.

Indeed, available PDSs have proven effective in detecting cross-lingual plagiarism using empirically tested and validated algorithms (e.g., Anguita,

Beghelli and Creixell 2011; El-Rashidy, et al. 2023). However, challenges persist in detecting plagiarism in Arabic texts by translation. Such a form of plagiarism is referred to as 'cross-language plagiarism' (cf. Aljohani and Mohd 2014; Hattab 2015; Alotaibi and Joy 2021) or 'language-transformation plagiarism' (cf. Zuo 2022). Still, it should be clarified that language-transformation plagiarism is a broader concept that involves rephrasing or restructuring a source text across languages, which may include both direct translation and paraphrased translation. Since cross-lingual plagiarism is concerned with the reuse of content across different languages (viz., the original and plagiarized texts are in different languages), translation plagiarism is a subtype of cross-lingual plagiarism. That is, a text is directly translated into another language without attribution.

Major challenges include Arabic's inflectional nature, elliptical structures, and lack of diacritization in most electronic or printed Arabic texts (Alshehri, Beloff and White 2024). Additionally, despite the large number of research on PDSs across languages, Arabic literary texts receive scant attention. In this regard, Sterman, Huang, Liu and Paulos (2020) highlighted the need for language-dependent methods to address controversies regarding issues of literary plagiarism. Further to this, computational linguistic tools could help in comparing controversial literary texts by identifying similarities, thereby guiding arbitrators to decide on plagiarism instances.

1.2 The present study

It is largely argued that translating literature from English into Arabic is notably challenging (cf. Classe 2000). Therefore, literary plagiarism via translation obfuscation is difficult to detect using common algorithms like text-similarity algorithms, shingling, fingerprinting, machine learning, or natural language processing (NLP). Such difficulty is attributed to Arabic's distinct morphological, syntactic, and stylistic features, which complicate identifying textual similarities, particularly by those unfamiliar with the language (Mesfar 2010). Without a deep understanding of Arabic linguistic conventions, distinguishing between faithful translations, plagiarized literary texts, and creative literary works built on patchwriting will remain a difficult task.

This study is inspired by the term 'translation plagiarism' that sparked heated debates in Egypt when the novelist Ahmed Mourad was accused of translation plagiarism in his best-selling novels '*Alfeel Al-Azraq'* (2012) and Ard *Al Elah*, originally published in Arabic in 2016 and then translated into English under the title *Land of the God* in 2020. Different critics claimed that parts of these novels were taken from English and foreign sources, while Mourad and other critics have defended his approach, arguing that the two novels were intertextual re-creations that require no citations. Indeed, Mourad's case highlights one of the big challenges in the literary world nowadays: the cultural and legal ambiguities surrounding intellectual property in many Arabic-speaking contexts. While Western academic norms typically demand rigorous source attribution, Arabic literary tradition

permits intertextual borrowing. Moreover, the inconsistent enforcement of copyright laws and the absence of a clear framework for defining literary plagiarism add to the complexity. Understanding this sociocultural drop is essential when designing and evaluating PDSs for Arabic literary texts. This notion is highlighted in Haitch (2016), who affirmed that cross-cultural differences in intellectual property practices might lead to unintentional plagiarism.

Therefore, the present study aims to explore which plagiarism detection systems (PDSs) work best for Arabic literary texts. In this respect, three PDSs, namely Rabin-Karp, KNN, and Word2vec, are assessed and compared against two metrics: precision and accuracy. The selection of these particular PDSs is based on their capability to identify similarities across languages (cf. Kumar et al. 2023). Each PDS focuses on different aspects of the literary text under exploration, and therefore they would capture any instance of plagiarism. That is, while Rabin-Karp assesses string matching, KNN captures similarity measurement, and Word2Vec focuses on semantic understanding. The diversity in their operational mechanisms makes them suitable for Arabic literary content, as they examine both surface-level and semantic plagiarism. Further to this, no previous studies addressed the efficacy of three PDSs when implemented for detecting plagiarism in literary texts. Other PDSs like Bert and Transformer-based models were tested before (e.g., Gharavi, Veisi and Rosso 2019), but they are less interpretable and require much training time. Also, Jaccard similarity treats text as bags of words and ignores syntax and semantics, which are critical in detecting cross-lingual plagiarism.

The study attempts to explore their capabilities identifying plagiarism found in in Arabic literary texts and to point out some of the challenges they encounter through performance analysis on a translated literary text. Hence, it seeks to answer two main questions:

1. Out of the three selected PDSs (Rabin-Karp, KNN, or Word2Vec), which one would most suitably be applied for plagiarism detection in Arabic literary works that have been translated into English?
2. What are the problems encountered by such systems during the processing of Arabic literary texts and how to mitigate them?

By answering these two questions, we assume that offering a systematic evaluation of the strengths and limitations of the target PDSs of Arabic literary texts would help decide on the best method that would technically avoid recurrent issues inherent in Arabic texts in general and Arabic literary texts in particular. By evaluating the performance and limitations of these PDSs, the study seeks to identify the best method for Arabic texts, highlighting each system's strengths and weaknesses. This will help researchers select the most effective PDS and inform the development of tailored solutions to improve detection in Arabic literary texts. Furthermore, taking into consideration the fact that Arabic is highly inflectional, as

a single root can generate many surface forms through affixes, suffixes, infixes, and clitic attachments, such variations might obscure similarity detection if algorithms depend on surface string matching as in Rabin-Karp. Also, translated or paraphrased content might go undetected due to a lack of diacritics in written Arabic texts. Moreover, the flexible word order of Arabic helps plagiarists to give semantically equivalent sentences with different syntactic structures.

The rest of the paper is structured as follows: Section 2 reviews literature on PDSs in general and cross-language/translation plagiarism in particular. Section 3 presents the theoretical framework, covering PDSs' philosophy, types of plagiarism, cross-language plagiarism, and its impact on intellectual property. Section 4 outlines the research design and procedure of analysis. Section 5 reports on the findings and relates them to previous studies. Part 6 summarizes the findings, addresses limitations, and suggests directions for future research.

2. Literature review

Plagiarism is using someone else's words, ideas, or even whole work without proper credit, harming both the author and the publisher responsible for disseminating the authored text (Sutherland-Smith 2008). Because certain literary works net their authors considerable profit over time, plagiarism is regarded as a crime that violates intellectual property rights and is as serious as theft. In some cases, it is worse than the theft of money, as the literary works are immortal forms of self-expression that engage readers' imagination, allowing them to experience events and emotional states they have not lived through. Hence, the value and benefit of these literary works must be credited to the author, the original creator (Sterman et al. 2020).

In general, literary works are subject to plagiarism, including poems, novels, and short stories, especially those with common themes and topics. For instance, classical Arabic poetry has sparked hectic debates on the acceptability of imitation as original literary work (Ulum 2023). Plagiarists seek to attract readers by professionally appropriating ideas to improve the quality of the writing and appeal to the readers. In this regard, related literature mentions three types of literary plagiarism: plagiarism of words, meaning, or both. Firstly, plagiarism of words involves appropriating others' words and combining them in one's own work. Secondly, meaning plagiarism is the most insidious, as the plagiarist alters a text while retaining its original meaning (Bouville 2008). Thirdly, as mentioned in Terry (2010), plagiarism of both words and meaning may include stealing entire parts, pages, or even an entire work. Such types of plagiarism are beyond detection by the current systems. Also, manual detection is further complicated, and therefore, more sophisticated systems should be developed.

More advanced NLP-based techniques (e.g., deep learning and machine learning) now help with yielding more accurate results through the identification of textual similarities and patterns in literary works. According to El-Rashidy et al. (2023), the integration of NLP techniques with PDSs offers more accurate results

based on the analysis of text structure and pattern. Likewise, Anguita et al. (2011) found that the integration of deep learning techniques improved PDS, particularly those that are based on convolutional and recurrent neural network architectures, being applied to sentiment analysis and image recognition. Additionally, based on different experimental studies, Quidwai, Li and Dube (2023) found that PDSs achieved very high rates of precision, recall, and overall accuracy. Such methodological enhancements always guarantee more reliable and efficient plagiarism detectors.

Early research on plagiarism detection in cross-lingual contexts focused on comparing Arabic documents based on similarities at various linguistic levels, from character n-gram to document structure. Such PDSs utilized a combination of document retrieval and detailed similarity analysis by means of machine learning or logical representations of document components, *i.e.*, words, sentences, and paragraphs. These methods have reached up to 75 percent f-score for document retrieval and around 70 percent for detailed similarity analysis (cf. Haikal 2012). Later on, intrinsic PDSs addressed scenarios where the original source is not accessible. Therefore, as Haikal (2012) mentions, it relied on detecting shifts in writing style within documents using machine learning, and such techniques have been benchmarked against bespoke Arabic corpora. Newer PDSs use either deep learning, semantic similarities based on word embeddings, or multilingual transformers. As a result, Alshehri et al. (2024) affirm that the performance of recent PDSs shows remarkable improvements with innovations in semantic similarity, word order encoding, and leveraging diacritical marks.

Still, cross-lingual plagiarism, particularly between English and Arabic, remains complex for several reasons. Firstly, Arabic is morphologically rich with inflections and complex word formations. Further to this, variations in diacritical marks and other typographical cues cause semantic shifts (cf. Alshehri et al. 2024). Secondly, translation varies considerably due to the use of synonyms, paraphrasing, syntactic variations, and idiomatic shifts. This renders direct text-matching approaches practically ineffective (Alotaibi and Joy 2021). Also, automatic translation-based PDSs are not fully reliable, as they may not preserve semantic nuances nor handle the complexities of Arabic, especially idioms and context-specific expressions. Thirdly, Alotaibi and Joy (2021) add that lack of high-quality annotated corpora and resources spanning Arabic and other languages represents another challenge for improving the performance of PDSs. Finally, Haikal (2012) mentions that most of the established PDSs target English, and necessary adaptations for Arabic are obviously lacking.

Based on available literature, we noted that the majority of modern PDSs have crucial limitations as far as complex literary works are concerned. A key challenge is their inability to distinguish among different forms of plagiarism, i.e., syntactic, lexical, or semantic. For instance, Mohabey et al. (2023) note that the traditional plagiarism detectors that use exact/approximate string-matching techniques cannot address semantic plagiarism. A string could be a sentence (shorter string) copied from a text (larger string). In addition, Son, Huong and

Thanh (2021) highlighted the challenge that emerges with the need to consider contextual and contextual cues to identify accurate similarities. Another challenging, more sophisticated form of plagiarism is referred to as 'translation obfuscation', where the translated text is intentionally obscured. Towards countering such limitations and improving the accuracy of these PDSs, Ahuja, Gupta and Kumar (2020) suggest the integration of specific techniques such as the use of hyperplane equations and support vector machines as well as the incorporation of semantic knowledge and linguistic features.

Much to our concern, Bouaine, Benabbou and Sadgali (2023) managed to address the issue of cross-lingual/translation plagiarism by incorporating techniques such as NLP-embedding techniques and deep learning into the current PDSs. Avetisyan et al. (2023) introduced another method that relies on open multilingual thesauri and pre-trained multilingual BERT-based models to offer a more detailed analysis using powerful NLP algorithms. However, these approaches are suitable for diverse, under-resourced languages, like French and Russian, since they do not depend on machine translation and word sense disambiguation (Mohtaj and Asghari 2022). Additionally, efforts are exerted to build up cross-language plagiarism detection corpora to support the performance of these systems and help with more sophisticated evaluation.

A plethora of studies have proposed different approaches for detecting Arabic cross-language plagiarism. Alotaibi and Joy (2021), for instance, developed a technique that integrates word embedding, term weighting techniques, and universal sentence encoder models to offer a more improved performance in Arabic-English plagiarism detection. Likewise, based on the Transformer architecture, Hourrane and Benlahmar (2022) introduced a graph-based approach that represents text passages in different languages using knowledge graphs, and their approach outperformed other existing approaches. Also, Jaber and Aliwy (2021) developed a reliable approach that employs an information retrieval system to calculate plagiarism percentages based on the intersection of grams (e.g., characters, words, syllables, etc.), thereby improving precision and recall. Also, Abdelhamid, Azouaou, and Batata (2022) compared eight PDSs for Arabic, French, and English academic texts in terms of their features, usability, technical aspects, and performance in detecting different levels of obfuscation, including cross-language plagiarism, verbatim, and paraphrase.

Despite extensive literature on the quality of the automatic PDSs in different languages, few studies evaluated their effectiveness in detecting cross-language plagiarism resulting from translation obfuscation in Arabic texts, particularly Arabic literary texts. Most prior studies focused on academic contexts due to the increasing demand for reliable detection systems for research papers and classroom assignments (e.g., Dougherty 2020; De Lima et al. 2021). Therefore, the present study addresses this gap by evaluating the performance of the three automatic PDSs (Rabin-Karp, KNN, and Word2vec) in light of two metrics – precision and accuracy

– with the aim of identifying the most suitable and effective system for detecting cross-language plagiarism in Arabic literary texts.

3. Theoretical framework

3.1 Plagiarism: An integrative perspective

In Arabic literature, the issue of literary theft has been common since ancient times (Naaman 2011). Many cases of plagiarism and literary theft have undermined writers' and researchers' credibility due to improper attribution, thereby prompting scholars and critics to verify the authenticity of Arabic books and poems throughout history (cf. Durakovic 2019). Nowadays, with the introduction of AI writing assistants, original texts are constantly subject to systematic modification, making them unrecognizable even to the original authors themselves. Furthermore, as Apter (2013) argues, many translators lack moral or judicial oversight and therefore allow unrestricted translation of dramas, cinematic works, novels, etc., into other languages.

Recently, different Arab authors have been accused of translation plagiarism, where they translate a text from another language into Arabic and claim it as their own, "with the intention of hiding its origin" (Gipp 2014:11). In this regard, Durakovic (2019) argues that many alleged literary thefts in the Arabic cultural heritage have been nothing but the intertextuality of images and ideas due to shared experiences. Similarly, Long (1991) notes that some forms of plagiarism are tolerated as intertextuality, where the poetic and prosaic intertwine and overlap. Supporters of this perspective claim that prior texts are intentionally or unintentionally inherent in all literary genres as authors draw on previous fiction and poetry, making writing a natural extension of previous works.

Conceptually, plagiarism and intertextuality are different in scope and hence should not be overlapped. Worton and Still (1990) note that intertextuality is a creative aspect of literary development, as it involves either explicit references to prior work (through quotations, allusions, parodies, pastiches, or archetypes) or implicit ones where the reference is unambiguous. Rigney (2019) states that an author referring to a previous work intends either to add a new layer of meaning, place the new work in the frame of the previous work, create humor, or reinterpret earlier works. Plagiarism, in contrast, is unethical and is historically and temporally situated, appropriating earlier work. It is often motivated by personal gain or academic dishonesty (Sutherland-Smith, 2008). Still, in Arabic literary tradition, intertextual borrowing has historically been viewed with more tolerance, especially in classical poetry (Ulum 2023). This cultural backdrop complicates the boundary between legitimate intertextuality and plagiarism. Practically, current plagiarism detection systems (PDSs) cannot distinguish between creative borrowing (intertextuality) and unethical copying (plagiarism), as they are designed to detect surface-level or semantic similarity only. This limitation means that PDS outputs must be supplemented by critical human judgment, especially in literary contexts where intertextuality is an accepted creative technique.

Plagiarism involves claiming credit for the words, ideas, and concepts of others. This form of data theft is often labelled 'academic dishonesty' or 'data fraud' (cf. De Lima et al. 2021). Haitch (2016) states that plagiarism can be either intentional, when a person knowingly copies others' works, or unintentional, being unaware of the conventions and rules. More recently, advanced forms of plagiarism, such as language-transformation and content-collage plagiarism, are becoming so common and persistent that they threaten academic integrity. Zuo (2022) mentions that while language-transformation plagiarism involves rephrasing a source text using different words while preserving the core ideas, content-collage plagiarism involves building up a new document based on various sources, thereby complicating the process of detection.

The advancement in technology and accessibility of diverse Internet-based datasets have exponentially increased plagiarism, creating a plagiarism pandemic that is difficult to discover and prevent. While it undermines the foundations of the literary world and hinders effective research, plagiarists unethically gain either money or academic acclaim. Curtis and Tremayne (2021) argue that despite the discernible increase in the number of plagiarism detection software, this menace is still on the rise in the literary world. In academia, institutions have tried diverse software, including Turnitin and iThenticate. However, new challenges and fraud techniques arise every day, such as obfuscated plagiarism, where data is changed, altered, and obscured in such a way that it is challenging to detect literary theft and even the original text from which data was taken. Research integrity is expected to face mounting challenges, especially with new advancements in AI.

3.2 Plagiarism detection, data mining, and artificial intelligence

Various PDSs have become popular in academic writing. Lancaster and Culwin (2005) classify PDSs based on their detection methodology, system metrics, article processing efficacy, and metrics complexity into intrinsic plagiarism detection (IPD) and extrinsic plagiarism detection (EPD). On the one hand, IPD involves detecting instances of ideas being copied and pasted as well as instances of verbal substitution, and therefore it is quite challenging to be automatically detected. EPD, on the other hand, compares the suspected text against a given reference dataset. Additionally, Alzahrani, Salim and Abraham (2012) mention that plagiarism detection methods include character-based methods, vector-based methods, syntax-based techniques, semantic-based strategies, fuzzy-based approaches, structural-based methods, and stylometry. Equally important, cross-lingual-based methods, semantic-based hybrid strategies and cluster-based approaches are gaining popularity. All these methods and techniques shown in Figure 1 can detect monolingual plagiarism, whereas the grammar-based methods can detect cross-lingual plagiarism.

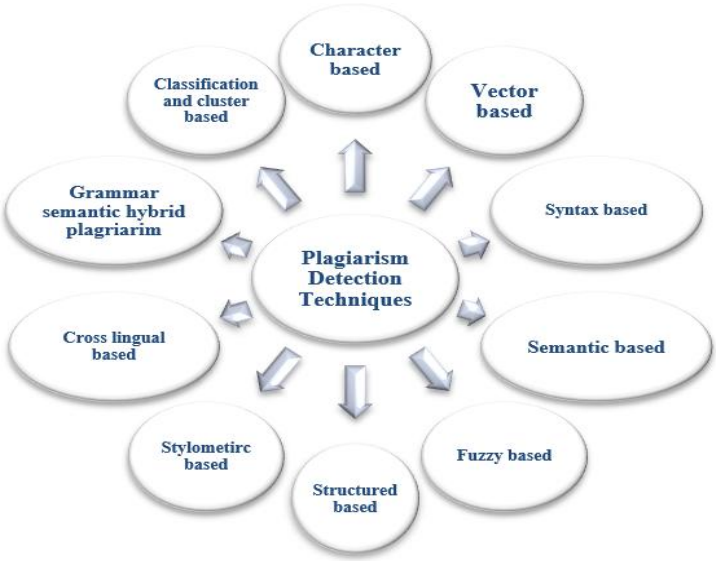

Figure 1. Common plagiarism detection methods, techniques, and strategies

Plagiarism detection is closely related to data mining, where the data sets include textual documents. Data mining helps to identify potential plagiarism by uncovering similarities in these documents. As shown in Figure 2, data mining methods are usually categorized into text mining, bi-grams, tri-grams, n-grams, and clustering methods (Nennuri et al. 2021).

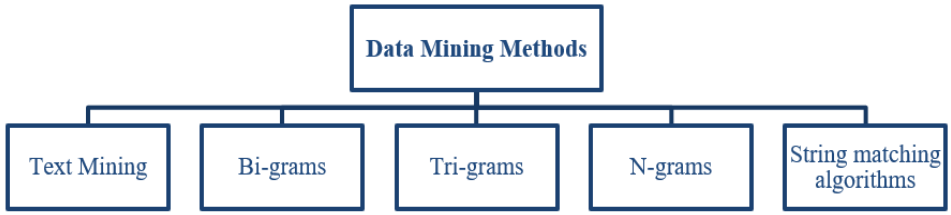

Figure 2. Common data mining methods for plagiarism detection

Text mining relies on a preprocessing process where punctuation marks are removed, and words are reduced to their base forms (stemming). N-grams achieve better results at the text-based level. However, Abid, Usman and Ashraf (2017) noted that a comparison of bi-gram, tri-gram, and n-gram reveals that bi-gram shows better recall and tri-gram shows better precision. Relatedly, Eka Diana and Hanana Ulfa (2019) found that string matching algorithms, like Jaccard, perform better than n-grams as far as accuracy and sensitivity are concerned, as they detect exact or similar matches. Also, Wijaya, Seputra and Parwita (2021) noted that Best Match 25 (BM25) and Rabin-Karp managed to detect both monolingual and cross-lingual plagiarism. While BM25 estimates the relevance of documents to a given search query, Rabin-Karp – as a hashing-based system – yields better performance and execution time through two major components: Term Frequency (i.e., the

number of times a term appears in a document) and Inverse Document Frequency (*i.e.*, how common a term is across the entire set of documents).

More recently, although artificial intelligence (AI) has emerged as a revolutionary approach to science and technology, it has also enabled a new form of literary fraud, where machines paraphrase texts (Wu, Chien, Chien and Yang 2021). Dehouche (2021) adds that AI-powered software can detect some cues from different materials and generate original text, and therefore threatens literary integrity. Conversely, AI-powered software has achieved an accuracy of 99 percent in detecting machine-paraphrased text (Foltýnek et al. 2020). Thus, AI can both generate and detect plagiarism. Table 1 below summarizes common plagiarism detection methods and their dominant features.

Table 1. A feature-based comparison of various plagiarism detection methods

Plagiarism detection methods	Dominant features
n-gram	Better detection of text-based plagiarism
bi-gram	Better recall
tri-gram	Better precision
Jaccard algorithm	Better accuracy and sensitivity
Rabin-Karp	Better execution time and better performance
BM25	Better specificity
AI powered systems	Accuracy of 99%

All these methods, techniques, and strategies are applicable to detecting plagiarism in different languages; however, the next section focuses on plagiarism detection in Arabic literary texts.

3.3 Plagiarism detection in Arabic literary texts

Arabic, a Semitic language, is written from right to left, with characters that take different forms (i.e., diacritics) depending on whether they are written in isolation or joined to make a word. Arabic's structural peculiarities made it difficult for software to detect plagiarized text in literature, especially in the academic context. Omar and Hilal (2022) found that identifying sentential and textual similarities in Arabic is empirically challenging due to processing complexities on the one hand and due to its distinct features on the phonological, graphical, morphological, syntactic, and semantic levels. Additionally, the presence of diacritics in Arabic can cause changes in meaning, further complicating plagiarism detection. Furthermore, classical Arabic's linguistic complexity complicates author attribution using traditional lexical or structural style markers (Al Duhayyim et al. 2022). These challenges highlight the need for specialized algorithms tailored to ease plagiarism detection given Arabic's unique features.

Still, fuzzy-based techniques have proved helpful in detecting plagiarism (Alzahrani et al. 2012). Similarly, semantic approaches using a multi-agent indexing system have generated promising results for detecting monolingual plagiarism in Arabic (Zouaoui and Rezeg 2019). However, most literary misconduct arises from translating foreign texts into Arabic. This cross-lingual plagiarism, combined with Arabic's character complexities, complicates plagiarism detection. In this regard, Hattab (2015) notes that latent semantic indexing has addressed this issue partially. Additionally, Arifin, Isa, Wulandhari and Abdurachman (2018) mentioned that methods such as winnowing (with an advanced preprocessing stage) and stemming produce good results in detecting both monolingual and cross-lingual, especially in Indonesian. Yet, Arabic urgently requires more advanced techniques to address the emerging challenges brought by the rapid advancement in technology.

In Arabic literature, paraphrasing remains a key challenge for PDSs' developers. Yet recurrent AI convolutional neural network technology, with its use of deep learning networks and global word, has achieved high precision and recall (Mahmoud and Zrigui 2020). Similarly, more advanced software has been designed to work without human oversight. Autonomous solutions, such as software integrating grey wolf optimization algorithms with modified abstract syntax trees, have shown high levels of precision and efficiency in detecting Arabic textual similarities (Zaher, Shehab, Elhoseny and Farahat 2020). Likewise, Jaber and Aliwy (2021) note that simple information retrieval systems can estimate the degree of literary misconduct, as semantic manipulations can be added for better performance.

In some instances, skip-gram and dice-based coefficient metrics, combined with corpus-based approaches, effectively identify textual similarities. Furthermore, deep natural learning processing helps to identify texts that go undetected by other methods (Ilyas et al. 2021). For instance, Chang et al. (2021) mention that Word2vec, with its word-to-word vector attachment technique, achieved marvellous performance in academic misconduct and plagiarism. Also, some studies (*e.g.,* Arabi and Akbari 2022) recommended the use of word-embedding networks for extrinsic plagiarism detection in Arabic texts, as it offers over 90 percent precision, thereby outperforming WordNet ontology. While accuracy is crucial for all PDSs, efficiency (viz., the ability to produce promising results with minimum utilization of available resources) is equally significant (Bellahsene, Bonifati and Rahm 2011).

With the growing interest in Arabic research, Arabic academic databases are expanding and therefore should be evaluated for literary fraud. Hence, the performance of PDSs is a key concern, although one that is mostly ignored. More importantly, there is a need to design a language-independent approach for detecting literary misconduct, thereby broadening plagiarism detection research and enabling cross-linguistic collaboration (Gharavi et al. 2019). Furthermore, as challenges grow, advanced research is essential to address literary misconduct. The

potential of AI and data mining approaches must also be utilized to support the integrity of literary research.

4. Methodology

4.1 Method

With the rapidly increasing incidences of plagiarism in literature, diverse methods have been developed to detect and address them. Research to devise new methods for the effective detection and termination of these illegal practices in Arabic texts is ongoing. However, no effective solutions have been offered so far, and countering plagiarism has emerged as a significant challenge in Arabic literature (Wali, Gargouri and Hamadou 2018). Therefore, following the tenets of the quantitative-qualitative comparative method (cf. Ragin 1998), the study explores the precision and accuracy of three selected automatic PDSs (Rabin-Karp, K-nearest neighbor (KNN), and Word2vec) for detecting cross-lingual/translation plagiarism in Arabic literary texts. See Figure 3.

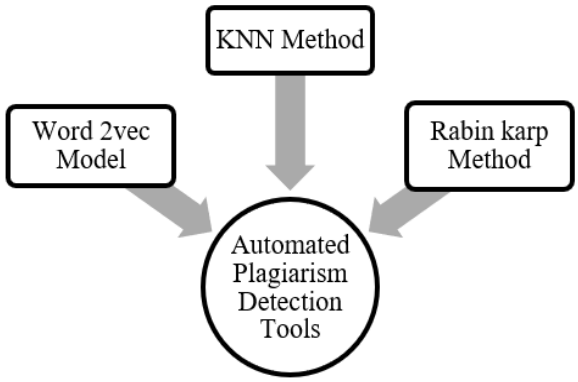

Figure 3. The plagiarism detection systems under investigation

The rationale for selecting the three PDSs is threefold. Firstly, Rabin-Karp is efficient in identifying verbatim copying and close paraphrasing in large data sets, as it is based on exact string matching (cf. Wijaya et al. 2021). Secondly, KNN is best known for its ability to detect semantic similarity or paraphrasing based on the comparison of patterns in training and target databases since it measures similarity among text chunks. Further to this, according to Adeniyi, Wei and Yongquan (2016), KNN can be used in various languages and genres, as it can be adapted to support a range of feature representations. Thirdly, Word2Vec can identify plagiarism instances that are based on rephrasing or using synonyms, as it implements NLP to capture the semantic associations and syntactic similarities among words. Nagoudi, Khorsi, Cherroun and Schwab (2018) also posit that Word2Vec employs deep learning techniques that treat words as vectors using

cosine similarity for the identification of similarity among word vectors with a high degree of precision.

In summary, the integration of the distinct features of three PDSs – i.e., Rabin-Karp's string matching, KNN's similarity measurement, and Word2Vec's semantic understanding – helps in offering a holistic assessment of recurrent PDSs of Arabic texts, particularly when exploring instances of cross-lingual or translation plagiarism.

4.2 Data description and data analysis procedure

The final dataset used for testing the three PDSs is an Arabic translation of Daly Walker's short story "*I am the Grass*" conducted by the authors, who hold advanced degrees in linguistics and translation studies. Then, it was cross-checked by the authors, with disagreements resolved collaboratively. Finally, it was evaluated by three independent translators with experience of more than ten years in translating literary works. The story addresses the themes of war, guilt, and redemption. Totalling 9180 words, it is about a nameless surgeon who served in the Vietnam War and is haunted by the atrocities he committed during the war. The English original text could be accessed through https://www.theatlantic.com/magazine/archive/2000/06/i-am-the-grass/378245/.

The rationale for selecting this short story in particular is threefold. Firstly, the short story is thematically rich and linguistically complex enough to allow for paraphrasing and translation challenges. Secondly, it is of manageable length to be feasibly pre-processed, translated, and analyzed within the study scope. Thirdly and finally, the short story represents a contemporary literary text in English which allows a controlled translation into Arabic, ensuring data consistency.

Given the scope of this study, we adopted a controlled translation protocol to balance fidelity and variation in the dataset. Specifically, 60 percent of the English story was directly translated into Arabic, 30 percent was paraphrased in Arabic to introduce syntactic and lexical variations, and 10 percent new content was added to the original text and then translated. Following Nennuri et al. (2021), the same data set was processed by the three PDSs (Rabin-Karp, KNN and Word2vec) for cross-lingual plagiarism detection. Figure 4 summarizes the procedure of comparing the three PDSs.

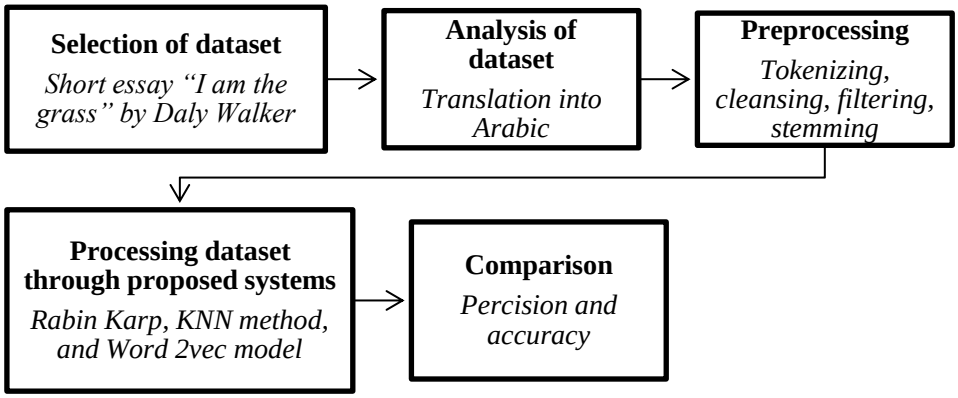

Figure 4. The procedure of data analysis

After preparing the Arabic test text, it was pre-processed to meet the technical requirements of the three selected PDSs to evaluate their performance using two metrics: precision and accuracy. These metrics measure how well the system identifies true positive (TP) cases, which are instances where plagiarism is correctly detected, and avoids false positives (FP), which are instances where plagiarism is incorrectly detected. The preprocessing phase comprised four stages: (1) tokenization, where the text is broken into individual units such as punctuation marks, words, etc. using the Python NLTK toolkit with adaptations for Arabic (Farasa segmenter for clitic separation); (2) stop-word removal, where punctuation marks, numbers, and stop words such as articles and conjunctions are removed based on the publicly available Arabic stop-word list developed by Khoja (2010) and later extended in Aljohani and Mohd (2014); (3) filtering, where semantically void words and irrelevant information are removed; and (4) stemming, where words are reduced to their roots/stems. In this regard, light stemming was applied using the NLTK implementation of the ISRI stemmer to reduce data sparsity while avoiding the overgeneralization that root-based stemming often causes in Arabic.

With regard to precision, we measured the proportion of correctly identified plagiarism instances out of all instances the system flagged as plagiarism using the following equation:

$$\text{Precision} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Positives}}$$

For accuracy, we measured the proportion of correctly identified plagiarism and non-plagiarism instances out of the total dataset using the following equation:

$$\text{Accuracy} = \frac{\text{True Positives} + \text{True Negatives}}{\text{Total Instances}}$$

The final dataset was then fed into each of the algorithms under evaluation, and the results were compared in light of their precision and accuracy.

5. Data analysis: Results and discussion

After preprocessing of the final dataset, the three target PDSs were tested for their precision and accuracy to detect cross-lingual/translated plagiarism, paraphrasing-based plagiarism, and the new content. Table 2 below offers rough statistics representing the plagiarism detection ratios for each system.

Table 2. Summary of the results

	Original content	Paraphrased content	Translation plagiarism
Dataset	10%	30%	60%
Detection ratio			
1. Rabin-Karp	8%	3%	13%
2. KNN	8%	2%	16%
3. Word 2vec	7%	5%	11%

Results show that, firstly, Rabin-Karp detected 13 percent out of 60 percent of cross-lingual/translation plagiarism, so that the precision percentage amounts to 21.7 percent. This percentage shows moderate precision, as Rabin-Karp may misclassify similar content as plagiarism due to its syntax-based nature. Furthermore, it detected only 3 percent out of 30 percent of the paraphrased content, with 10 percent precision. The reason is that Rabin-Karp is mainly designed for exact matches, and therefore it struggles with any semantic changes. However, it detected 8 percent out of 10 percent of the original content, which means that the precision percentage is 80 percent, indicating a high reliability for original content detection. In general, the overall accuracy of Rabin-Karb (24%) is considerably low, making it unsuitable for detecting paraphrasing and translation plagiarism.

Secondly, KNN detected 16 percent out of 60 percent of the processed text (cross-lingual plagiarism), with a precision percentage of 26.7 percent, which is the best performance in this category. Yet, like Rabin-Karp, KNN detected only 2 percent of 30 percent of the processed text (paraphrased plagiarism), with a precision percentage of 6.7 percent. Like Rabin-Karp, it detected 8 percent of 10 percent of the processed text (original content), with a precision percentage of 80 percent. Therefore, the overall accuracy of KNN is 26 percent. The higher accuracy in translated plagiarism detection makes KNN and Rabin-Karp viable options for this category.

Thirdly, Word2Vec detected 11 percent out of 60 percent of the processed text (cross-lingual plagiarism), with a precision percentage of 18.3 percent. This low precision percentage makes it the weakest of the three in this category. The reason might be that Word2Vec prioritizes semantic meaning over syntactic structures. As for paraphrased plagiarism, it detected 5 percent out of 30 percent of the processed text, with a precision percentage of 16.7 percent, thereby

outperforming both Rabin-Karp and KNN in this category due to its ability to identify semantic similarity. Regarding original content, Word2Vec detected 7 percent out of 10 percent (original content) of the processed text, with a precision rate of 70 percent that is slightly lower than Rabin-Karp and KNN for exact matches. Hence, the overall accuracy of Word2Vec is 23 percent.

The low precision percentages of the three PDSs in the category of paraphrased content could be attributed to the fact that Arabic allows the expression of the same original proposition of an utterance in diverse syntactic and lexical styles. The fact that Arabic words are highly inflected makes exact string matching through Rabin-Karp particularly ineffective, as morphological variations obscure similarities. In addition, Arabic allows multiple lexical choices to convey the same meaning, involving subtle stylistic or register shifts. That is why Word2Vec struggles despite being semantically oriented. Equally important, Arabic has a relatively free word order, and accordingly Arabic texts complicate algorithms that depend on positional similarity or n-gram overlap. This finally leads to misclassifications in both KNN and Rabin-Karp. Finally, the absence of diacritics in the Arabic text creates lexical ambiguities that confuse semantic similarity models and therefore produce either false positives or negatives. Therefore, paraphrasing strategies such as reordering, lexical substitution, deletion, or insertion often go undetected. The major differences among the three PDSs with regard to precision and accuracy are illustrated in Figure 5.

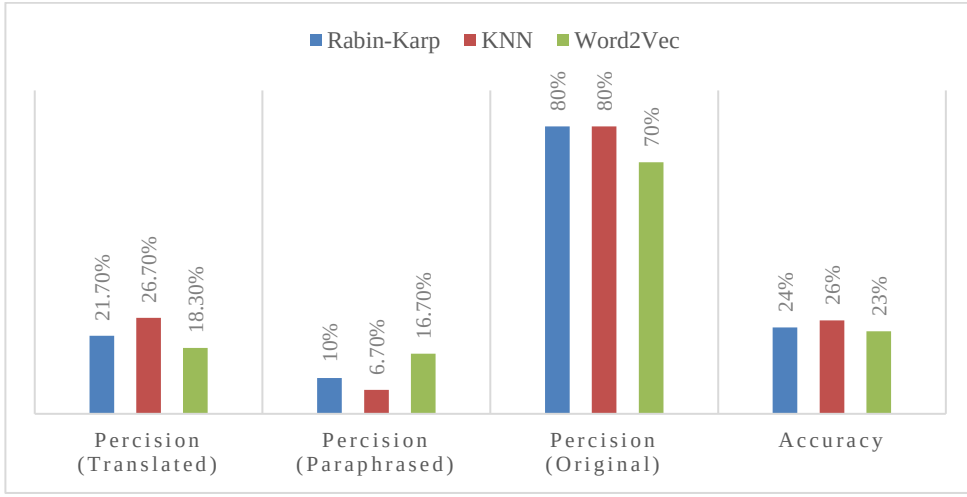

Figure 5. A comparative summary of the precision and accuracy of the three PDSs in detecting translation and paraphrased plagiarism

The minor differences among the three PDSs in detecting cross-lingual/translation plagiarism are attributed to the fact that the Arabic text

reproduced from the original English text through translation had diverse alternations, especially on the syntactic level. Also, these systems lack clear parameters to determine if a particular Arabic textual unit is likely a translation of another English textual unit. Further to this, current algorithms cannot scan more than two languages simultaneously.

In general, KNN exhibits the highest precision rate (26.7%) in detecting cross-lingual plagiarism, closely succeeded by Rabin-Karp. Yet, here it should be clarified that the superiority of KNN in detecting cross-lingual plagiarism is indicative rather than definitive. Due to its semantic analysis capabilities, Word2Vec proves to be the most effective system for detecting plagiarized paraphrase, achieving a precision rate of 16.7 percent. Both Rabin-Karp and KNN are highly recommended for detecting original content. Nonetheless, during the processing of Arabic literary texts, these PDSs face numerous challenges, including language complexity, absence of standardized resources, and dialectical variations. Hence, developing algorithms designed to cater for the distinct features of Arabic texts, building extensive Arabic language databases, and employing machine learning techniques for better precision and accuracy rates are all necessary to overcome these challenges, thereby supporting plagiarism detection in Arabic literature.

Although the three PDSs and the algorithms rely on the principles of data mining and artificial intelligence, results show that they failed to produce positive results in detecting different forms of plagiarism in Arabic texts, particularly cross-lingual plagiarism. While their performance in detecting plagiarism in English is satisfactory, their effectiveness in highlighting literary misconduct in Arabic texts is questionable. These results agree with Zouaoui and Rezeg (2019), who highlighted that Arabic's distinct writing system, semantic profile, and grammar create morphosyntactic and semantic complexities that hinder automated plagiarism detection. The distinct linguistic features of Arabic may also cause the evolution of new forms of plagiarism in the Arabic literature (Ibrahim, Saeed and Wakil 2017), and therefore urgent, dedicated efforts are much needed to minimize their prevalence. Likewise, in order to minimize reliance on the frequent manual checking during the process of plagiarism detection, more advanced methods are required. Further to this, the analysis relied on a single translated literary text, which restricts generalizability. Also, only precision and accuracy were used, without additional measures such as recall and F-1 score or statistical testing, which could have provided a robust evaluation of the three PDSs.

Taken wholly, the study's findings highlight the need for stringent measures that are necessary for protecting the integrity and originality of Arabic literature. Towards achieving this objective, new language-dependent algorithms ought to be developed, trained, and tested. In this regard, Nagoudi et al. (2018) recommended the development of a multilayered system that incorporates two or more algorithms. This procedure would make them complement one another for more accurate and precise findings. Their findings showed that such integration helped with upgrading accuracy in detecting plagiarism in Arabic texts. Accordingly, we recommend that

a hybrid approach that combines both Word2Vec for paraphrased content and KNN for cross-lingual plagiarism detection would improve overall efficacy. This recommendation is prospective rather than evidence-based, grounded in prior literature that shows hybrid models can outperform single algorithms in other contexts (e.g., Ahuja et al. 2020; Nagoudi et al. 2018).

Equally important, as noted by Gharavi et al. (2019), deep learning techniques (e.g., BERT) are effective in detecting illegal practices in Arabic literary texts. Further research can tackle the establishment of a global platform to develop language-independent methods for detecting different forms of plagiarism. This endeavour requires more advanced technology and skilled human resources to illustrate promising results. Also, Asghari et al. (2018) showed that other algorithms have proven to be accurate in addressing plagiarism issues in texts in languages similar to Arabic, and therefore they could be adapted to address plagiarism detection issues in Arabic literary texts, with special reference to cross-lingual plagiarism detection techniques.

6. Conclusion

Towards the identification of the most effective system for detecting cross-lingual plagiarism in Arabic literary texts, this study drew a comparison among Rabin-Karp, KNN, and Word2Vec. Although the three PDSs showed specific limitations that negatively influenced their general precision and accuracy, findings showed no significant differences among them. That is, none of the three systems under investigation detected all forms of plagiarism, especially translation plagiarism. Therefore, we argue that the current systems are practically unable to detect cross-lingual plagiarism in Arabic texts translated from English. What is more, they are still technically limited in identifying textual similarities between the original and reproduced texts, and hence they serve only as tools for assessors to identify plagiarism instances.

Dialectical variations also represent a key challenge in the automatic detection of plagiarism in Arabic texts. These dialectical variations cause linguistic analysis tools and text-matching software to struggle with detecting instances of cross-lingual plagiarism. Also, the lack of comprehensive Arabic databases and resources represents another challenge for the current PDSs. Equally important, plagiarism detection of content translated into Arabic is quite challenging, and this is ascribed to stark disparities in sentence structure, vocabulary, and morphology between Arabic and other languages. This presents a challenge to the efficiency of plagiarism detection software since word-for-word translations may not always produce dependable or easily detectable outcomes. Furthermore, the use of Arabic script and the absence of standardized spelling and punctuation rules in some dialects further complicate plagiarism detection.

The contribution of the present study resides in offering a systematic evaluation of the systems concerned with cross-lingual/translation plagiarism detection in Arabic literary texts. To our knowledge, this is the first to compare

three systems implementing different algorithms. Also, the dataset used as a benchmark is highly representative of literature, and the reproduced Arabic version involved different parameters for comparison: the original content, paraphrased content, and translated plagiarism. The results indicate that KNN shows higher precision trends compared to the other systems, but further research with larger datasets and statistical testing is needed to confirm these differences. The reason is that this study is limited to one translated literary prosaic text processed by only three PDSs. Therefore, future studies are recommended to follow the same approach to evaluate other recurrent methods with different algorithms employing larger datasets across multiple literary genres and apply robust inferential statistics (e.g., t-tests, ANOVA, bootstrapping methods) to ensure the reliability and generalizability of the findings. The overall dataset could be divided into subsets representing diverse literary genres such as poetry, dramas, and novels.

Acknowledgement
This study is supported via funding from Prince Sattam bin Abdulaziz University project number (PSAU/2024/R/1446)

Abdulfattah Omar
Associate Professor of Linguistics
Port Said University, Port Said, Egypt
The Australian National University, Australia
ORCID Number: 0000-0002-3618-1750
Email: a.a.omar2010@gmail.com

Wafya Hamouda
Associate Professor of Linguistics
Tanta University, Tanta, Egypt
ORCID Number: 0000-0002-9921-0125
Email: wafia.hamouda@edu.tanta.edu.eg

Waheed M. A. Altohami – Corresponding Author
Associate Professor of Linguistics
Prince Sattam bin Abdulaziz University, Al-Kharj, Saudi Arabia
Mansoura University, Mansoura, Egypt
ORCID Number: 0000-0001-8742-1366
Email: w.m.altohami@gmail.com

References

Abdelhamid, Mehdy, Faical Azouaou and Sofiane Batata. (2022). 'A survey of plagiarism detection systems: Case of use with English, French and Arabic languages.' *Arvix,* 1: 1-28.

Abid, Mahwish, Muhammad Usman and Muhammad W. Ashraf. (2017). 'Plagiarism detection process using data mining techniques.' *International Journal of Recent Contributions from* Engineering, *Science and IT,* 5(4): 68. https://doi.org/10.3991/ijes.v5i4.7869

Adeniyi, D.A., Z. Wei and Y. Yongquan. (2016). 'Automated web usage data mining and recommendation system using k-nearest neighbor (KNN) classification method'. *Applied Computing and Informatics,* 12(1): 90-108. https://doi.org/10.1016/j.aci.2014.10.001

Ahuja, Lovepreet, Vishal Gupta and Rohit Kumar. (2020). 'A new hybrid technique for detection of plagiarism from text documents.' *Arabian Journal for Science and Engineering,* 45(12): 9939-9952. https://doi.org/10.1007/s13369-020-04565-9

Al Duhayyim, Mesfer, Manal A. Alohali, Fahd N. Al-Wesabi, Anwer M. Hilal, Mohammad Medani and Manar A. Hamza. (2022). 'Securing Arabic contents algorithm for smart detecting of illegal tampering attacks.' *Computers, Materials and Continua,* 70(2): 2879-2894. https://doi.org/10.32604/cmc.2022.019594

Aljohani, Adel and Masnizah Mohd. (2014). 'Arabic-English cross-language plagiarism detection using winnowing algorithm'. *Information Technology Journal,* 13(14): 2349-2355. https://doi.org/10.3923/itj.2014.2349.2355

Alotaibi, Naif and Mike Joy. (2021). 'English-Arabic cross-language plagiarism detection'. *Proceedings of the Conference Recent Advances in Natural Language Processing - Deep Learning for Natural Language Processing Methods and Applications*. https://doi.org/10.26615/978-954-452-072-4_006

Alshehri, Mona, Natalia Beloff and Martin White. (2024). 'AraXLM: New XLM-Roberta based method for plagiarism detection in Arabic text'. *Lecture Notes in Networks and Systems*: 81-96. https://doi.org/10.1007/978-3-031-62277-9_6

Alzahrani, Salha M., Naomie Salim and Ajith Abraham. (2012). 'Understanding plagiarism linguistic patterns, textual features, and detection methods.' *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews),* 42(2): 133-149. https://doi.org/10.1109/tsmcc.2011.2134847

Angelil-Carter, Shelley. (2014). *Stolen Language? Plagiarism in Writing*. Harlow: Routledge.

Anguita, Angel, Alejandra Beghelli and Werner Creixell. (2011). 'Automatic cross-language plagiarism detection'. *2011 7th International Conference on Natural Language Processing and Knowledge Engineering.* https://doi.org/10.1109/nlpke.2011.6138189

Apter, Emily. (2013). *Against World Literature: On the Politics of Untranslatability*. Brooklyn: National Geographic Books.

Arabi, Hamed and Mehdi Akbari. (2022). 'Improving plagiarism detection in text document using hybrid weighted similarity'. *Expert Systems with Applications,* 207: 91-108. https://doi.org/10.1016/j.eswa.2022.118034

Arifin, Yulyani, Sani M. Isa, Lili A. Wulandhari and Edi Abdurachman. (2018). 'Plagiarism detection for Indonesian language using winnowing with parallel processing'. *Journal of Physics: Conference Series*, 978(1); 1-7. https://doi.org/10.1088/1742-6596/978/1/012082

Asghari, Habibollah, Salar Mohtaj, Omid Fatemi, Heshaam Faili, Paolo Rosso and Martin Potthast. (2018). *Algorithms and Corpora for Persian Plagiarism Detection*. Basingstoke: Springer International Publishing.

Avetisyan, Karen, Arthu Malajyan, Tsolak Ghukasyan, Arutyun Avetisyan and Chunxia Dou. (2023). '*A Simple and Effective Method of Cross-Lingual Plagiarism Detection'* (Pre-Print). https://doi.org/10.21203/rs.3.rs-3040948/v1

Bellahsene, Zohra, Angela Bonifati and Erhard Rahm. (2011). *Schema Matching and Mapping*. Berlin: Springer Science and Business Media.

Bouaine, Chaimaa, Faouzia Benabbou and Imane Sadgali. (2023). 'Word embedding for high performance cross-language plagiarism detection techniques.' *International Journal of Interactive Mobile Technologies* 17(10): 69-91. https://doi.org/10.3991/ijim.v17i10.38891

Bouville, Mathieu. (2008). 'Plagiarism: Words and ideas.' *Science and Engineering Ethics*, 14(3): 311-322. https://doi.org/10.1007/s11948-008-9057-6

Chang, Chia-Yang, Shie-Jue Lee, Chih-Hung Wu, Chih-Feng Liu and Ching-Kuan Liu. (2021). 'Using word semantic concepts for plagiarism detection in text documents.' *Information Retrieval Journal*, 24(4-5): 298-321. https://doi.org/10.1007/s10791-021-09394-4

Classe, Olive. (2000). *Encyclopedia of Literary Translation into English: A-L*. New York: Taylor and Francis.

Curtis, Guy J. and Kell Tremayne. (2021). 'Is plagiarism really on the rise? Results from four 5-yearly surveys.' *Studies in Higher Education*, 46(9): 1816-1826. https://doi.org/10.1080/03075079.2019.1707792

Dehouche, N. (2021). 'Plagiarism in the age of massive generative pre-trained transformers (GPT-3)'. *Ethics in Science and Environmental Politics*, 21, 17-23. https://doi.org/10.3354/esep00195

De Lima, Jorge Á., Áurea Sousa, Angélica Medeiros, Beatriz Misturada and Cátia Novo. (2021). 'Understanding undergraduate plagiarism in the

context of students' academic experience'. *Journal of Academic Ethics*, 20(2): 147-168. https://doi.org/10.1007/s10805-021-09396-3

Dougherty, M. V. (2020). *Disguised Academic Plagiarism: A Typology and Case Studies for Researchers and Editors*. Basingstoke: Springer Nature.

Durakovic, Esad. (2019). *The Poetics of Ancient and Classical Arabic Literature: Orientology*. London: Routledge.

Eka Diana, Nova and Ikrima Hanana Ulfa. (2019). 'Measuring performance of N-Gram and jaccard-similarity metrics in document plagiarism application'. *Journal of Physics: Conference Series, 1196*: 1-7. https://doi.org/10.1088/1742-6596/1196/1/012069

El-Rashidy, Mohamed A., Ramy G. Mohamed, Nawal A. El-Fishawy and Marwa A. Shouman. (2023). 'An effective text plagiarism detection system based on feature selection and SVM techniques.' *Multimedia Tools and Applications,* 83(1): 2609-2646. https://doi.org/10.1007/s11042-023-15703-4

Foltýnek, Tomáš, Terry Ruas, Philipp Scharpf, Norman Meuschke, Moritz Schubotz, William Grosky and Bela Gipp. (2020). 'Detecting machine-obfuscated plagiarism'. In Anneli Sundqvist, Gerd Berget, Jan Nolin and Kjell Ivar Skjerdingstad (eds.), *Sustainable Digital Communities*, 816-827. Basingstoke: Springer. https://doi.org/10.1007/978-3-030-43687-2_68

Gharavi, Erfaneh, Hadi Veisi and Paolo Rosso. (2019). 'Scalable and language-independent embedding-based approach for plagiarism detection considering obfuscation type: No training phase'. *Neural Computing and Applications,* 32(14): 10593-10607. https://doi.org/10.1007/s00521-019-04594-y

Gipp, Bela. (2014). *Citation-based Plagiarism Detection: Detecting Disguised and Cross-language Plagiarism Using Citation Pattern Analysis*. Basingstoke: Springer Vieweg.

Haikal, Walid. (2012). 'Detection of plagiarism in Arabic documents.' *International Journal of Information Technology and Computer Science*, 4(10): 80-89. https://doi.org/10.5815/IJITCS.2012.10.10

Haitch, Russell. (2016). 'Stealing or sharing? Cross-cultural issues of plagiarism in an open-source era'. *Teaching Theology and Religion,* 19(3): 264-275. https://doi.org/10.1111/teth.12337

Hattab, Ezz. (2015). 'Cross-language plagiarism detection method: Arabic vs. English'. *2015 International Conference on Developments of E-Systems Engineering (DeSE),* 141-144. https://doi.org/10.1109/dese.2015.25

Hourrane, Oumaima and El Habib Benlahmar. (2022). 'Graph transformer for cross-lingual plagiarism detection'. *International Journal of Artificial Intelligence (IJ-AI),* 11(3): 905-915. https://doi.org/10.11591/ijai.v11.i3.pp905-915

Ibrahim, Ribwar, Soran Saeed and Karzan Wakil. (2017). 'Plagiarism detection techniques for Arabic script languages: A literature review'. *Kurdistan Journal of Applied Research,* 2(3): 106- 111. https://doi.org/10.24017/science.2017.3.1

Ilyas, Muhammad, Nasreen Malik, Ahmad Bilal, Saad Razzaq, Fahad Maqbool and Qaisar Abbas. (2021). **'**Plagiarism detection using natural language processing techniques.' *Technical Journal,* 26(1): 90-101.

Jaber, Zahraa J. and Ahmed H. Aliwy. (2021). 'Design and implementation of Arabic plagiarism detection system'. In Valentina E. Balas, Vijender K. Solanki and Raghvendra Kumar (eds.), *Further Advances in Internet of Things in Biomedical and Cyber Physical Systems*, 347-358. Basingstoke: Springer International Publishing.

Kumar, D. P., Ananda Tiwari, B. S. Priya, M. G. Raghavendra and A. C. Raju. (2023). 'Plagiarism detection using KNN'. *Proceedings of the 1st International Conference on Frontier of Digital Technology Towards a Sustainable Society*. https://doi.org/10.1063/5.0130846

Lancaster, Thomas and Fintan Culwin. (2005). 'Classifications of plagiarism detection engines.' *Innovation in Teaching and Learning in Information and Computer Sciences,* 4(2): 1- 16. https://doi.org/10.11120/ital.2005.04020006

Long, Pamela O. (1991). 'Invention, authorship, "Intellectual property," and the origin of patents: Notes toward a conceptual history.' *Technology and Culture,* 32(4): 846-884. https://doi.org/10.2307/3106154

Mahmoud, Adnen and Mounir Zrigui. (2020). 'Semantic similarity analysis for corpus development and paraphrase detection in Arabic'. *The International Arab Journal of Information Technology,* 18(1): 1-7. https://doi.org/10.34028/iajit/18/1/1

Mesfar, Slim. (2010). 'Toward a cascade of morpho-syntactic tools for Arabic natural language processing'. In Alexander Gelbukh (ed.), *Computational Linguistics and Intelligent Text Processing*, 150-162. Basingstoke: Springer. https://doi.org/10.1007/978-3-642-12116-6_13

Mohtaj, Salar and Habibollah Asghari. (2022). 'A corpus for evaluation of cross language text re-use detection systems.' *Journal of Information Systems and Telecommunication,* 10(39): 169-179. https://doi.org/10.52547/jist.33583.10.39.169

Mohabey, Niraj, Yash Gavanang, Abubakkar Khan, Lavesh Singh Chib and Bhushan Patil. (2023). 'Plagiarism detection for project report using machine learning.' *International Journal of Engineering Technology and Management Sciences,* 7(3): 87- 93. https://doi.org/10.46647/ijetms.2023.v07i03.012

Naaman, Erez. (2011). 'Sariqain practice: The case of al-Ṣāḥib Ibn 'Abbād'. *Middle Eastern Literatures,* 14(3): 271-285. https://doi.org/10.1080/1475262x.2011.616712

Nagoudi, El Moataz, Ahmed Khorsi, Hadda Cherroun and Didier Schwab. '2L-APD: A two-level plagiarism detection system for Arabic documents.' *Cybernetics and Information Technologies*, 18(1): 124-138. https://doi.org/10.2478/cait-2018-0011

Nennuri, Rajashekar, M. Geetha Yadav, M. Samhitha, S. Sandeep Kumar and G. Roshini. (2021). 'Plagiarism detection through data mining techniques.' *Journal of Physics: Conference Series*, 1979(1): 1-6. https://doi.org/10.1088/1742-6596/1979/1/012070

Omar, Khaled and Ammar Hilal. (2022). 'Plagiarism detection in Arabic documents using word2vector and Arabic WordNet'. *2022 International Arab Conference on Information Technology (ACIT)*. https://doi.org/10.1109/acit57182.2022.9994090

Quidwai, Ali, Chunhui Li and Parijat Dube. (2023). 'Beyond black box AI generated plagiarism detection: From sentence to document level'. *Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications.* https://doi.org/10.18653/v1/2023.bea-1.58

Ragin, Charles C. (1998). 'The logic of qualitative comparative analysis'. *International Review of Social History*, 43(6): 105- 124. https://doi.org/10.1017/cbo9780511563874.006

Rigney, Ann. (2019). 'Texts and Intertextuality'. In Kiene B. Wurth and Ann Rigney (eds.), *The Life of Texts: An Introduction to Literary Studies,* 79-112. Amsterdam University Press.

Son, Nguyen V., Le T. Huong and Nguyen C. Thanh. (2021). 'A two-phase plagiarism detection system based on multi-layer long short-term memory networks.' *International Journal of Artificial Intelligence (IJ-AI)*, 10(3): 636-648. https://doi.org/10.11591/ijai.v10.i3.pp636-648

Sterman, Sarah, Evey Huang, Vivian Liu and Eric Paulos. (2020). 'Interacting with literary style through computational tools.' *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, 1: 1-12. https://doi.org/10.1145/3313831.3376730

Sutherland-Smith, Wendy. (2008). *Plagiarism, the Internet, and Student Learning: Improving Academic Integrity*. London: Routledge.

Teresa, Turell M. (2004). 'Textual kidnapping revisited: The case of plagiarism in literary translation.' *International Journal of Speech, Language and the Law - Forensic Linguistics*, 11(1): 1- 26. https://doi.org/10.1558/sll.2004.11.1.1

Terry, Richard. (2010). *The Plagiarism Allegation in English Literature from Butler to Sterne*. Basingstoke: Springer.

Ulum, Muhammad B. (2023). 'Plagiarism in classic Arabic poetics (Comparative study of al-jumahi and al-qairawany's thoughts).' *Jurnal CMES*, 16(1): 61-71. https://doi.org/10.20961/cmes.16.1.53447

Wali, Wafa, Bilel Gargouri and Abdelmajid Ben Hamadou. (2018). 'Using sentence similarity measure for plagiarism detection of Arabic documents.' In Ajith Abraham, Niketa Gandhi, Thomas Hanne, Tzung-Pei Hong, Tatiane Nogueira Rios and Weiping Ding (eds.), *Intelligent Systems Design and Applications,* 52-62. Basingstoke: Springer International Publishing.

Wall, David. (2003). *Crime and the Internet*. London: Routledge.

Wijaya, Indra., Andy Seputra and Wayan G. Parwita. (2021). 'Comparison of the BM25 and rabinkarp algorithm for plagiarism detection.' *Journal of Physics: Conference Series,* 1810(1): 1-10. https://doi.org/10.1088/1742-6596/1810/1/012032

Worton, Michael and Judith Still. (1990). *Intertextuality: Theories and Practices*. New York: Manchester University Press.

Wu, Jain-Shing, Ting-Hsuan Chien, Li-Ren Chien and Chin-Yi Yang. (2021). 'Using artificial intelligence to predict class loyalty and plagiarism in students in an online blended programming course during the COVID-19 pandemic.' *Electronics*, 10(18): 1-20. https://doi.org/10.3390/electronics10182203

Zaher, Mahmoud, Abdulaziz Shehab, Mohamed Elhoseny and Farahat F. Farahat. (2020). 'Unsupervised model for detecting plagiarism in internet-based handwritten Arabic documents.' *Journal of Organizational and End User Computing*, 32(2): 42-66. https://doi.org/10.4018/joeuc.2020040103

Zouaoui, Samia and Khaled Rezeg. (2019). 'Ontological approach based on multi-agent system for indexing and filtering Arabic documents.' *Journal of Digital Information Management*, 17(3): 145-163. https://doi.org/10.6025/jdim/2019/17/3/145-163

Zuo, Ziyu. (2022). 'On the determination of literary plagiarism in copyright law.' *PONTE International Scientific Researches Journal*, 78(6): 1-10. https://doi.org/10.21506/j.ponte.2022.6.4