Journal of COMPUTER AND FORENSIC SCIENCES

CFS 2025, Vol. 4, Issue 2, pp. 3–16
https://doi.org/10.5937/jcfs4-64419
Original research paper

A Hybrid Plagiarism Detection Framework Using Lexical and Semantic Similarity with Lightweight Sentence Transformers

Rahul Birwadkar[1*]

[1] Student, SRH University, Heidelberg, Germany

*Corresponding author: rbirwadkar6@gmail.com

Received: January 27, 2026 • Accepted: April 6, 2026 • Published: May 14, 2026

Abstract: Plagiarism detection has become increasingly challenging due to the widespread availability of paraphrasing tools and generative artificial intelligence systems. Traditional plagiarism detection techniques based on lexical similarity, such as TF-IDF and n-gram matching, often fail to identify semantically similar but lexically modified text. This paper presents a hybrid plagiarism detection framework that combines lexical similarity measures with semantic similarity derived from sentence transformer models. The proposed approach integrates TF-IDF–based cosine similarity with lightweight sentence embeddings generated using MiniLM and SBERT models. To enhance semantic detection performance, a MiniLM-based sentence transformer is fine-tuned on the PAN 2011 plagiarism detection corpus. Experimental evaluation demonstrates that the hybrid similarity approach significantly improves detection accuracy compared to purely lexical methods, particularly for paraphrased plagiarism cases. The framework is further validated using threshold-based analysis and real-world web content retrieved through automated scraping. The proposed system provides an efficient and scalable solution for plagiarism detection, balancing computational efficiency with semantic understanding, and is suitable for academic and real-world forensic applications.

Keywords: plagiarism detection, semantic similarity, sentence transformers, MiniLM, natural language processing.

1. INTRODUCTION

Plagiarism detection is a critical component of academic integrity, digital forensics, and content verification systems [1]. The rapid growth of digital content and the widespread availability of paraphrasing and generative text tools have significantly increased the complexity of identifying plagiarized material [1]. While traditional plagiarism detection techniques are effective in identifying exact text reuse, they often fail when content is modified through paraphrasing or structural rewriting [1], [2]. Conventional plagiarism detection systems primarily rely on lexical similarity measures such as n-gram matching, term frequency analysis, and string-based comparisons [2]. These methods are limited to surface-level text analysis and are highly sensitive to lexical changes, causing semantically equivalent but lexically altered text to remain undetected [1], [2]. Recent advances in natural language processing have introduced semantic representations capable of capturing

Citation: R. Birwadkar, "A Hybrid Plagiarism Detection Framework Using Lexical and Semantic Similarity with Lightweight Sentence Transformers," Journal of *Computer and Forensic Sciences*, vol. 4, no. 2, pp. 3–16, 2025, https://doi.org/10.5937/jcfs4-64419.

contextual meaning beyond exact word overlap [3]. Transformer-based language models enable the comparison of textual content at the semantic level, making it possible to identify paraphrased or meaning-preserving text reuse [3], [4]. However, purely semantic approaches may overlook exact textual matches and can be computationally expensive for large-scale applications [4], [5]. To address these limitations, this paper proposes a hybrid plagiarism detection framework that integrates lexical and semantic similarity measures. By combining TF-IDF–based similarity with sentence-level semantic embeddings generated by lightweight transformer models, the proposed approach aims to improve detection accuracy while maintaining computational efficiency. The main contributions of this work are as follows:

1) A hybrid similarity framework combining lexical and semantic plagiarism detection techniques.
2) A fine-tuned MiniLM-based sentence transformer optimized for plagiarism detection.
3) A comprehensive experimental evaluation on the PAN 2011 plagiarism detection dataset.
4) Validation of the proposed framework on real-world web content.

2. RELATED WORK

Plagiarism detection has been an active research area for several decades, particularly in academic integrity, digital forensics, and content verification systems. Existing approaches can broadly be classified into lexical-based methods, traditional machine learning approaches, and semantic or deep learning–based techniques. Each category exhibits distinct strengths and limitations, especially when applied to paraphrased or obfuscated plagiarism.

2.1. Lexical-Based Plagiarism Detection Approaches

Early plagiarism detection systems primarily relied on lexical similarity measures that compare surface-level textual features [1], [2]. Common techniques include string matching, n-gram overlap, term frequency analysis, and cosine similarity using vector space models. Among these, TF-IDF has been widely adopted due to its simplicity, interpretability, and effectiveness in detecting exact or near-exact text reuse [2]. Lexical approaches perform well when plagiarized content is copied verbatim or with minimal modifications. N-gram–based methods are effective at identifying localized overlap, while TF-IDF captures broader document-level similarity patterns, making these approaches computationally efficient and scalable [2]. However, lexical similarity methods are highly sensitive to word-level changes. Simple paraphrasing techniques such as synonym substitution, sentence restructuring, or word reordering can significantly reduce lexical overlap while preserving semantic meaning. As a result, purely lexical approaches often fail to detect paraphrased plagiarism, leading to high false-negative rates [1].

2.2. Traditional Machine Learning–Based Methods

To overcome the limitations of surface-level similarity measures, traditional machine learning approaches introduced classification-based plagiarism detection systems [1]. These methods typically rely on handcrafted features such as lexical overlap statistics, syntactic patterns, stylometric features, and similarity scores computed at different granularity levels. Although such approaches improved detection accuracy compared to purely lexical methods, they depend heavily on feature engineering and domain-specific heuristics. Designing effective features requires expert knowledge and often lacks generalization across datasets and languages, limiting robustness against advanced paraphrasing techniques [1].

2.3. Semantic Similarity and Embedding-Based Approaches

The introduction of distributed word representations marked a significant shift in plagiarism detection research [2]. Word embedding models enabled semantic comparison of text by capturing contextual relationships between words, allowing similarity computation beyond exact word overlap. Building upon word embeddings, deep learning architectures such as recurrent neural networks and encoder–decoder models enabled sentence- and document-level semantic representation [6], [7]. These approaches demonstrated improved performance in detecting meaning-preserving text transformations but often suffered from scalability limitations. Transformer-based models further advanced semantic similarity modeling by leveraging self-attention mechanisms to capture contextual dependencies across entire text sequences [3], [7]. Pre-trained transformer models provide strong general-purpose language representations that can be adapted to downstream tasks, including plagiarism detection.

2.4. Sentence-Level Embeddings for Plagiarism Detection

Sentence-level embedding models offer an efficient alternative to full sequence-to-sequence architectures by generating fixed-length vector representations for sentences or documents. These embeddings enable direct similarity computation using distance metrics such as cosine similarity, making them suitable for large-scale plagiarism detection systems [4]. Sentence transformer architectures explicitly optimize embeddings for similarity tasks using Siamese or triplet network structures. Lightweight variants, such as distilled transformer models, reduce computational cost while retaining strong semantic representation capabilities [5]. Despite these advantages, purely semantic approaches may overlook exact textual reuse or produce high similarity scores for conceptually related but non-plagiarized content.

2.5 Hybrid Plagiarism Detection Approaches

Hybrid plagiarism detection methods aim to integrate lexical and semantic similarity measures to leverage the strengths of both approaches [1], [4]. Lexical similarity contributes

precision by identifying exact text reuse, while semantic similarity improves recall by detecting paraphrased or meaning-preserving content. The proposed framework builds upon this line of research by introducing a lightweight hybrid approach that combines TF-IDF–based lexical similarity with sentence-level semantic embeddings generated by a fine-tuned MiniLM model. This design balances detection performance and computational efficiency, making it suitable for scalable plagiarism detection in academic and forensic contexts.

Earlier neural approaches based on recurrent architectures, including RNNs and encoder–decoder models, have been widely explored for sequence modeling tasks [6], [8], [9], [10]. More recently, transformer-based architectures such as BERT and its variants have significantly improved contextual text representation capabilities [3]. Generative transformer models, including BART and T5, have further advanced text generation and transformation tasks [11], [12]. In addition, knowledge distillation techniques have enabled the development of lightweight transformer models for efficient inference [13], supporting the use of compact architectures such as MiniLM. Earlier work on paraphrase generation and semantic similarity also includes statistical and rule-based approaches [14], [15].

3. PROPOSED HYBRID PLAGIARISM DETECTION FRAMEWORK

This section presents the proposed hybrid plagiarism detection framework, which integrates lexical and semantic similarity measures to improve the detection of plagiarized and paraphrased text. The framework is designed to address the limitations of purely lexical approaches while maintaining computational efficiency suitable for real-world deployment [16].

3.1. System Overview

The proposed framework follows a modular pipeline consisting of text preprocessing, lexical similarity computation, semantic similarity computation, hybrid similarity aggregation, and threshold-based classification. Given a suspicious document and a reference document, similarity scores are computed using both lexical and semantic approaches and combined into a unified hybrid similarity score used for plagiarism detection.

Figure 1 illustrates the workflow of the proposed hybrid plagiarism detection framework. Input document pairs (suspicious and reference documents) are first preprocessed using standard natural language processing techniques. The system then computes similarity through two parallel components.

In the lexical similarity branch, TF-IDF vectorization is applied, followed by cosine similarity to obtain the lexical similarity score. In the semantic similarity branch, sentence-level embeddings are generated using a MiniLM-based transformer model, and cosine similarity is applied to compute the semantic similarity score. The final hybrid similarity score is calculated as the average of the maximum lexical and semantic similarity scores for each document pair. This score is then used in a threshold-based classification mechanism to determine whether the document pair is plagiarized or non-plagiarized. This modular design enables flexibility by allowing individual components to be adapted or extended based on application requirements.

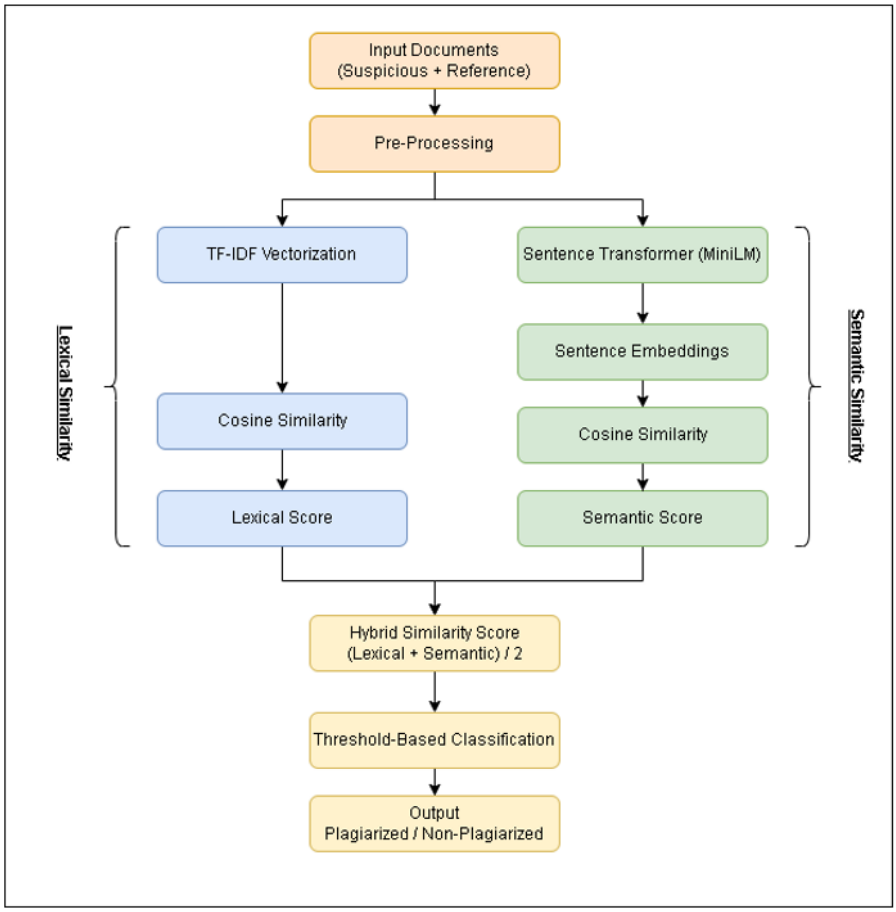

Figure 1: *Overview of the proposed hybrid plagiarism detection framework.*

3.2. Lexical Similarity Computation

Lexical similarity is computed using the Term Frequency–Inverse Document Frequency (TF-IDF) representation combined with cosine similarity. TF-IDF assigns higher weights to terms that are frequent within a document but rare across the corpus, making it effective for identifying exact or near-exact text reuse [2]. Each document is transformed into a TF-IDF vector, and cosine similarity is used to measure similarity between document pairs. To ensure fair comparison across documents of varying length, L2 normalization is applied to the TF-IDF vectors. This lexical similarity component serves as a strong baseline for detecting verbatim plagiarism but is limited in handling paraphrased or semantically altered text.

3.3. Semantic Similarity Using Sentence Transformers

To capture semantic similarity beyond surface-level word overlap, the framework employs sentence-level embeddings generated using a transformer-based language model.

Sentence transformer architectures are specifically designed to produce embeddings optimized for similarity comparison tasks [4]. In this work, a lightweight MiniLM-based sentence transformer is used due to its balance between semantic representation quality and computational efficiency [5]. Each document is encoded into a fixed-length dense vector representing its semantic content, and cosine similarity is computed between embedding pairs. To improve task-specific performance, the MiniLM model is fine-tuned using labelled document pairs derived from the PAN 2011 plagiarism detection dataset. Fine-tuning adapts the embedding space to better distinguish between plagiarized and non-plagiarized content while preserving the general linguistic knowledge of the pre-trained model.

3.4. Hybrid Similarity Score

The central contribution of this work is the integration of lexical and semantic similarity measures into a unified hybrid similarity score. Instead of relying on a single similarity metric, the proposed approach combines both perspectives to achieve more robust plagiarism detection.

The hybrid similarity score is defined as:

$$Hybrid\ Similarity = \frac{Lexical\ Similarity + \ Semantic\ Similarity}{2}$$

In the proposed implementation, the hybrid similarity score is computed as the average of the maximum lexical and semantic similarity scores obtained for each document pair. This formulation ensures that both exact lexical overlap and semantic equivalence contribute equally to the final similarity assessment. Lexical similarity provides precision for detecting copied text, while semantic similarity improves recall for paraphrased or meaning-preserving text reuse. The use of an arithmetic mean provides a simple and interpretable aggregation strategy while avoiding the need for additional hyperparameter tuning. While equal weighting is effective for the current study, future work may explore weighted combinations of lexical and semantic similarity, where optimal weighting factors can be learned or empirically determined to further improve detection performance.

3.5. Threshold-Based Classification

A threshold-based decision mechanism is applied to the hybrid similarity score to classify document pairs. If the computed hybrid similarity exceeds a predefined threshold, the document pair is classified as plagiarized; otherwise, it is classified as non-plagiarized. Different threshold values are evaluated experimentally to analyze their effect on detection performance. Lower thresholds favor recall by identifying a broader range of plagiarism cases, while higher thresholds improve precision by reducing false positives. This flexibility allows the framework to be adapted to different application scenarios, such as academic evaluation or forensic analysis.

3.6. Computational Complexity and Efficiency

Computational efficiency is an important consideration for plagiarism detection systems intended for large-scale or real-world deployment. The proposed hybrid framework is designed to balance detection accuracy with computational cost by combining efficient lexical similarity computation with a lightweight semantic model. TF-IDF vectorization and cosine similarity are computationally inexpensive and scale well with corpus size once document vectors are constructed. The MiniLM-based sentence transformer significantly reduces inference time and model size compared to larger transformer architectures while maintaining strong semantic representation capabilities. Fine-tuning is performed only once during training. During inference, embeddings for reference documents can be precomputed and stored, allowing similarity computation to be performed efficiently using vector operations. The hybrid similarity formulation avoids complex feature fusion or multi-stage classification pipelines, further reducing computational overhead. Overall, the proposed framework provides a practical and scalable solution for plagiarism detection that balances performance and efficiency.

4. EXPERIMENTAL SETUP

This section describes the dataset, preprocessing pipeline, model configuration, training strategy, evaluation protocol, and implementation details used to evaluate the proposed hybrid plagiarism detection framework. The goal is to ensure reproducibility and provide sufficient experimental detail for fair comparison with existing approaches.

4.1. Dataset Description

The experimental evaluation is conducted using the PAN 2011 plagiarism detection dataset, a widely used benchmark for external plagiarism detection research. The dataset consists of suspicious documents, corresponding source documents, and XML-based annotations identifying plagiarized text segments. The dataset includes various forms of plagiarism, ranging from exact copying to heavily paraphrased content. Each suspicious document may contain zero or more plagiarized segments originating from one or multiple source documents. This diversity makes the dataset suitable for evaluating both lexical and semantic plagiarism detection methods. The XML annotation files provide detailed metadata, including character offsets of plagiarized segments and plagiarism type information. These annotations are parsed to generate labelled document pairs for supervised training and evaluation.

4.2 Data Preprocessing

All documents undergo a standardized preprocessing pipeline prior to similarity computation. The preprocessing steps include tokenization, lowercasing, stop-word removal, and

normalization of punctuation and special characters. For lexical similarity computation, lemmatization is applied to reduce inflected word forms to their base representations. For semantic similarity computation, raw text is preserved as much as possible to retain contextual information required by transformer-based models. Documents that exceed the maximum token length supported by the model are truncated to ensure consistent input representation. This preprocessing strategy balances semantic fidelity with computational efficiency.

4.3. Lexical Similarity Configuration

Lexical similarity is computed using TF-IDF vectorization combined with cosine similarity. Unigrams and bigrams are used to capture both individual term usage and short contextual patterns. L2 normalization is applied to TF-IDF vectors to mitigate the influence of document length on similarity scores. The TF-IDF model is trained on the combined corpus of suspicious and source documents to ensure consistent term weighting across all document pairs. This configuration provides a reliable baseline for detecting exact and near-exact text reuse.

4.4. Semantic Similarity Model and Fine-Tuning Strategy

Semantic similarity is computed using a lightweight sentence transformer based on the MiniLM architecture. The model generates fixed-length dense embeddings that represent the semantic content of documents, enabling efficient similarity computation using cosine similarity. To adapt the model to the plagiarism detection task, fine-tuning is performed using labeled document pairs derived from the PAN dataset. Each training sample consists of a suspicious document, a corresponding source document, and a binary label indicating the presence of plagiarism. Fine-tuning optimizes the embedding space to increase similarity between plagiarized document pairs while reducing similarity for non-plagiarized pairs. A low learning rate is used to preserve general language understanding while enabling task-specific adaptation.

4.5. Training Configuration

The fine-tuning process is conducted using mini-batch training with a fixed batch size. The model is trained for a limited number of epochs to balance convergence and generalization. Dropout regularization is applied to reduce overfitting and improve robustness. The AdamW optimizer is employed due to its effectiveness and stability when fine-tuning transformer-based models. After training, the fine-tuned model is used exclusively in inference mode for embedding generation, allowing document embeddings to be precomputed and reused during similarity evaluation.

4.6. Evaluation Metrics

The framework is evaluated using standard classification metrics, including accuracy, precision, recall, and F1-score. These metrics provide complementary perspectives on detection performance, particularly in cases of class imbalance. Confusion matrices are additionally used to analyze false positive and false negative cases, providing insight into error characteristics across different similarity approaches.

4.7. Evaluation Protocol

The evaluation follows a document-pair comparison protocol. Lexical, semantic, and hybrid similarity scores are computed independently for each document pair. A threshold-based decision mechanism is then applied to classify each pair as plagiarized or non-plagiarized. Multiple similarity thresholds are evaluated to analyze robustness. Lower thresholds emphasize recall, while higher thresholds favor precision. This analysis supports adaptation of the framework to different application requirements.

4.8. Implementation Details

The experimental framework is implemented in Python using widely adopted natural language processing and deep learning libraries. Lexical similarity is computed using scikit-learn, including TF-IDF vectorization and cosine similarity. Semantic similarity is computed using Sentence-Transformers and Hugging Face Transformers, while model training and inference are carried out using PyTorch. Data handling and analysis are performed using pandas and NumPy, text preprocessing is performed using standard Python and NLTK utilities, and result visualization is performed using Matplotlib and Seaborn. Additional utilities such as tqdm are used for progress tracking.

5. RESULTS AND ANALYSIS

This section presents the experimental results of the proposed hybrid plagiarism detection framework. The performance of lexical, semantic, and hybrid similarity approaches is evaluated and compared using standard classification metrics. In addition, a threshold-based analysis is conducted to examine the robustness of the framework under different operating conditions.

5.1. Lexical Similarity Results

Lexical similarity based on TF-IDF and cosine similarity is evaluated as a baseline approach. The results show that lexical similarity performs reliably in cases of exact or near-exact text reuse. High precision is achieved when plagiarized content exhibits substantial lexical overlap with source documents. However, the performance of lexical simi-

larity decreases significantly for paraphrased plagiarism cases. Modifications such as synonym replacement, sentence restructuring, and word reordering reduce lexical overlap, leading to lower similarity scores and increased false negatives. These findings confirm the limitations of purely lexical plagiarism detection methods.

5.2. Semantic Similarity Results

Semantic similarity is evaluated using sentence embeddings generated by MiniLM-based sentence transformer models. Compared to lexical similarity, semantic approaches demonstrate improved performance in detecting paraphrased and meaning-preserving text reuse. The fine-tuned MiniLM model shows better recall and overall detection capability than the pre-trained variant. Fine-tuning enables the model to adapt to the characteristics of plagiarism detection, resulting in more reliable similarity scores for semantically equivalent document pairs. However, semantic similarity alone may produce higher similarity scores for conceptually related but non-plagiarized text, which can affect precision.

Table 1: *Performance comparison of pre-trained and fine-tuned MiniLM sentence embeddings under different similarity thresholds.*

Model	Approach	Accuracy	Precision	Recall	F1 Score	False Positive	False Negative
Pre-Trained (T=0.4)	Pre-trained Embeddings + Cosine Similarity	0.8825	0.9961	0.7678	0.8672	3	232
Pre-Trained (T=0.8)	Pre-trained Embeddings + Cosine Similarity	0.501	1	0.001	0.002	0	998
Sentence Transformer based (T=0.4)	Fine-tuned Embeddings + Cosine Similarity	0.494	0.493	0.4605	0.4762	473	539
Sentence Transformer based (T=0.8)	Fine-tuned Embeddings + Cosine Similarity	0.761	1	0	0	0	999
Direct Fine-Tune Model (T=0.4)	Fine-tuned MiniLM Model Cosine Similarity	0.761	0.6776	0.995	0.8062	473	5
Direct Fine-Tune Model (T=0.8)	Fine-tuned MiniLM Model Cosine Similarity	0.6355	0.9355	0.2903	0.4431	20	709

Table 1 compares the performance of pre-trained and fine-tuned MiniLM sentence embeddings under different similarity thresholds. The results indicate that fine-tuning significantly improves recall and overall F1-score, particularly at lower threshold values, enabling more effective detection of paraphrased plagiarism. In contrast, higher threshold values result in overly conservative behavior, leading to a sharp decline in recall for both pre-trained and fine-tuned models.

5.3. Hybrid Similarity Results

The proposed hybrid similarity framework combines lexical and semantic similarity scores into a unified metric. Experimental results indicate that the hybrid approach consistently outperforms individual similarity methods across evaluation metrics. By inte-

grating lexical precision with semantic robustness, the hybrid framework achieves a more balanced trade-off between precision and recall. This balance is particularly important in plagiarism detection scenarios, where both false positives and false negatives can have significant consequences. The hybrid approach effectively mitigates the weaknesses of individual similarity methods while preserving their strengths.

5.4. Threshold-Based Performance Analysis

To analyze the sensitivity of the framework, multiple similarity thresholds are evaluated. Lower threshold values increase recall by identifying a larger number of plagiarism cases, including heavily paraphrased content, but may introduce additional false positives. Conversely, higher thresholds improve precision by reducing false positives at the cost of lower recall. The results indicate that moderate threshold values provide the most effective balance between precision and recall. This flexibility allows the framework to be configured according to application requirements, such as strict academic evaluation or broader forensic screening.

5.5. Comparative Summary

A comparative analysis across lexical, semantic, and hybrid approaches highlights the advantages of the proposed framework. Lexical similarity performs well for detecting direct copying but struggles with paraphrased text. Semantic similarity improves detection of meaning-preserving modifications but may affect precision when used alone. The hybrid similarity framework leverages the complementary strengths of both approaches, resulting in improved overall performance. These results demonstrate that combining lexical and semantic similarity provides a more robust and reliable solution for plagiarism detection than using either method independently.

6. DISCUSSION

The experimental results demonstrate that combining lexical and semantic similarity measures provides a more robust approach to plagiarism detection than relying on either method independently. Lexical similarity techniques, such as TF-IDF, are effective for identifying direct text reuse but are limited when confronted with paraphrased or structurally modified content. Semantic similarity methods address this limitation by capturing contextual meaning, enabling the detection of meaning-preserving text reuse.

The proposed hybrid framework benefits from the complementary strengths of both approaches. Lexical similarity contributes precision by accurately identifying exact or near-exact overlap, while semantic similarity improves recall by detecting paraphrased plagiarism. The aggregation of these similarity signals results in a balanced detection strategy that reduces false negatives without introducing an excessive number of false positives.

The use of a lightweight MiniLM-based sentence transformer further enhances the practicality of the framework. Compared to larger transformer architectures, MiniLM offers reduced computational cost while maintaining strong semantic representation capability. Fine-tuning the model on a plagiarism-specific dataset enables better adaptation to the characteristics of plagiarism detection, improving task-specific performance without sacrificing efficiency. Threshold-based analysis highlights the adaptability of the framework. By adjusting similarity thresholds, the system can be configured for different application scenarios. Lower thresholds are suitable for exploratory or forensic analysis where recall is prioritized, while higher thresholds are appropriate for academic evaluation scenarios that require higher precision.

Despite these strengths, the framework has certain limitations. Similarity alone does not necessarily indicate plagiarism, as proper citation and source attribution play an important role in distinguishing legitimate referencing from unethical text reuse. It operates primarily at the document level and does not explicitly distinguish between cited and uncited text reuse. As a result, properly referenced quotations may still be flagged as similar. In addition, the performance of semantic similarity methods depends on the representativeness of the training data, which may affect generalization to domains with substantially different writing styles or vocabulary. Overall, the discussion confirms that the hybrid similarity strategy provides a practical and effective solution for plagiarism detection, balancing detection accuracy, interpretability, and computational efficiency.

7. CONCLUSION AND FUTURE WORK

This paper presented a hybrid plagiarism detection framework that integrates lexical and semantic similarity measures to improve the detection of plagiarized and paraphrased text. By combining TF-IDF–based lexical similarity with sentence-level semantic embeddings generated using a lightweight transformer model, the proposed approach addresses the limitations of traditional plagiarism detection techniques.

Experimental evaluation on the PAN 2011 plagiarism detection dataset demonstrates that the hybrid framework outperforms purely lexical and purely semantic approaches across multiple evaluation metrics. Fine-tuning a MiniLM-based sentence transformer further enhances detection performance while maintaining computational efficiency. Additional validation using real-world web content highlights the applicability of the framework in practical plagiarism detection scenarios.

Despite its effectiveness, the proposed framework focuses on textual similarity and does not explicitly distinguish between cited and uncited text reuse. Future work may incorporate citation-aware analysis to better differentiate legitimate, properly referenced reuse from plagiarism. Furthermore, extending the framework to support sentence- and paragraph-level analysis may improve fine-grained detection accuracy. Additional future research directions include cross-lingual plagiarism detection and the integration of contextual metadata to enhance robustness. These extensions could further improve the effectiveness of hybrid similarity-based plagiarism detection systems in diverse academic and forensic contexts.

FUNDING:
This research received no external funding.

INSTITUTIONAL REVIEW BOARD STATEMENT:
Not applicable.

INFORMED CONSENT STATEMENT:
Not applicable.

CONFLICTS OF INTEREST:
The author declares no conflict of interest.

REFERENCES

[1] K. T. Kalleberg, "Plagiarism detection using machine learning techniques," International Journal of Computer Applications, vol. 119, no. 13, pp. 1–6, 2015.

[2] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," in Proceedings of the International Conference on Learning Representations (ICLR), 2013.

[3] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in Proceedings of NAACL-HLT, 2019, pp. 4171–4186.

[4] N. Reimers and I. Gurevych, "Sentence-BERT: Sentence embeddings using Siamese BERT-networks," in Proceedings of the Conference on Empirical Methods in Natural Language Processing, 2019, pp. 3982–3992.

[5] W. Wang, F. Wei, L. Dong, H. Bao, N. Yang, and M. Zhou, "MiniLM: Deep self-attention distillation for task-agnostic compression of pre-trained transformers," in Advances in Neural Information Processing Systems (NeurIPS), 2020.

[6] I. Sutskever, O. Vinyals, and Q. V. Le, "Sequence to sequence learning with neural networks," in Advances in Neural Information Processing Systems (NeurIPS), 2014, pp. 3104–3112.

[7] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," in Proceedings of the International Conference on Learning Representations (ICLR), 2015.

[8] Y. Bengio, P. Simard, and P. Frasconi, "Learning long-term dependencies with gradient descent is difficult," IEEE Transactions on Neural Networks, vol. 5, no. 2, pp. 157–166, 1994, doi: 10.1109/72.279181.

[9] K. Cho et al., "Learning phrase representations using RNN encoder–decoder for statistical machine translation," in Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP), 2014, pp. 1724–1734.

[10] T. Luong, H. Pham, and C. D. Manning, "Effective approaches to attention-based neural machine translation," in Proceedings of EMNLP, 2015, pp. 1412–1421.

[11] M. Lewis et al., "BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension," in Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, 2020, pp. 7871–7880.

[12] C. Raffel et al., "Exploring the limits of transfer learning with a unified text-to-text transformer," Journal of Machine Learning Research, vol. 21, no. 140, pp. 1–67, 2020.

[13] J. Gou, B. Yu, S. J. Maybank, and D. Tao, "Knowledge distillation: A survey," International Journal of Computer Vision, vol. 129, pp. 1789–1819, 2021.

[14] C. Quirk, C. Brockett, and W. Dolan, "Monolingual machine translation for paraphrase generation," in Proceedings of EMNLP, 2004, pp. 142–149.

[15] D. Lin and P. Pantel, "Discovery of inference rules for question answering," Natural Language Engineering, vol. 7, no. 4, pp. 343–360, 2001.

[16] R. V. Birwadkar, "Plagiarism Detection and Paraphrasing based on Generative Artificial Intelligence," Master's thesis, Dept. of Information and Technology, SRH Hochschule Heidelberg, Heidelberg, Germany, 2025.