The context similarity problem in academic plagiarism detection: A case study on differentiating originality from collusion in a homogeneous dataset

Djoko Cahyo Utomo Lieharyani[1], *Jonner* Hutahaean[1*], *Ade Chandra* Nugraha[1], Suprihanto[1] and *Didik Suwito* Pribadi[1]

[1]Department of Informatics Engineering, Politeknik Negeri Bandung, 40559 Bandung, Indonesia

Abstract. Intraclass collusion poses a significant challenge for automated detection systems within homogeneous datasets from e-learning platforms, where legitimate contextual overlap often leads to high false-positive rates. This "context similarity problem" questions the utility of advanced semantic models. This study confronts this issue through a quantitative comparison of a state-of-the-art Sentence-BERT model against traditional lexical methods (Levenshtein, Jaccard) on a real-world dataset of 854 student answers. Our findings reveal a compelling, counter-intuitive result: lexical methods are demonstrably more effective. Levenshtein Similarity achieved a superior F1-score (0.74) and F2-score (0.75), underpinned by a strong recall of 0.76. Conversely, the semantic model was confounded by the dataset's homogeneity, yielding a modest F1-score of 0.57 and requiring an impractically high similarity threshold of 0.98 for optimal performance. This research provides a critical contribution by empirically demonstrating the limitations of purely semantic approaches in this specific context. We conclude that well-established lexical methods are not obsolete but remain a more reliable and practical tool for the initial screening of academic collusion, suggesting a need for hybrid strategies in modern academic integrity systems.

1 Introduction

The sanctity of academic integrity serves as the bedrock of higher education, fostering an environment of original thought and intellectual honesty. However, this foundation is under constant threat from various forms of academic misconduct, with plagiarism being one of the most persistent challenges [1-5]. The digital transformation of education, accelerated by the widespread adoption of e-learning platforms and Learning Management Systems (LMS), has inadvertently cultivated a new breeding ground for a specific and insidious form of plagiarism: intraclass collusion. This form of misconduct involves students within the same course illicitly collaborating, sharing, or directly copying answers from one another to complete assignments [6].

* Corresponding author: jonnerh@jtk.polban.ac.id

There are several types of plagiarism exist, including verbatim plagiarism, word-switching plagiarism, metaphor plagiarism, idea plagiarism, and self-plagiarism [1]. Unlike traditional plagiarism, which typically involves copying from external, publicly accessible sources like websites or publications, collusion operates within a closed, highly contextualized ecosystem. All students in a class are exposed to the same lectures, reference materials, and, most importantly, the same assignment prompts. This shared context naturally leads to a high degree of similarity in their submissions, even when produced independently and with full integrity. This inherent, legitimate overlap creates a significant conundrum for automated plagiarism detection systems—a challenge we formally define in this paper as the "context similarity problem." The core of this problem is the difficulty in algorithmically distinguishing between similarity that is a benign byproduct of a shared knowledge base and similarity that is a clear indicator of academic dishonesty. An overly sensitive system risks a high rate of false positives, which can lead to undue stress for students and burdensome investigations for faculty. Recent studies propose integrating plagiarism detection as one of several innovative strategies for leveraging artificial intelligence (AI) and machine learning (ML) to enhance educational practices [7].

The evolution of Natural Language Processing (NLP) has offered seemingly powerful tools to combat this issue. Advanced Transformer-based models like BERT (Bidirectional Encoder Representations from Transformers) are renowned for their deep understanding of semantic meaning and context [2, 3]. BERT for its proficiency in capturing contextual language representations and modeling long-range dependencies within extensive textual data [8]. However, their efficacy has predominantly been tested on heterogeneous datasets. It remains an open and critical question whether their nuanced understanding of "meaning" can effectively operate in an environment where every submission is semantically similar by design. The majority of plagiarism detection techniques are primarily developed to detect exact, word-for-word reproductions [7]. However, their effectiveness diminishes when confronting instances of copyright violation that are either less overt or more complex, such as cases involving paraphrasing [7].

This study, therefore, aims to provide a rigorous, empirical investigation into this problem. By conducting a case study on a real-world, homogeneous dataset of student assignments, we directly compare the performance of a state-of-the-art semantic approach (BERT with Cosine Similarity) against well-established lexical methods (Levenshtein and Jaccard Similarity). Our contribution is threefold: first, we formally define and empirically demonstrate the extent of the context similarity problem; second, we provide a quantitative, head-to-head comparison of semantic versus lexical models in this challenging scenario; and third, we offer practical, evidence-based recommendations for deploying more effective and fair plagiarism detection strategies in modern e-learning environments.

2 Literature review

The pursuit of effective plagiarism detection has driven a continuous evolution in computational linguistics, moving from surface-level textual comparisons to deep, meaning-based analysis.

2.1 Lexical and string-based methods: The foundation

The foundational approaches to plagiarism detection are rooted in lexical analysis, which directly compares the textual content of documents. These methods, including n-gram analysis and document fingerprinting, are designed to identify verbatim or near-verbatim overlaps [5]. Among the most enduring of these are the Jaccard Similarity index and the Levenshtein Distance. Jaccard Similarity provides a simple yet effective measure of shared

vocabulary by calculating the ratio of the intersection to the union of word sets from two docu `ments [9]. The Levenshtein Distance, a string metric, quantifies the difference between two sequences by counting the minimum number of single-character edits (insertions, deletions, or substitutions) required to change one string into the other [10]. While computationally efficient and highly effective for detecting direct copy-paste plagiarism, while theoretically vulnerable to sophisticated paraphrasing, these methods often remain effective in practice, as collusive paraphrasing frequently retains substantial structural or lexical overlap.

2.2 The semantic revolution: Word embeddings and transformers

The limitations of lexical methods prompted a shift towards semantic analysis. The challenge of identifying cross-language plagiarism, which remains difficult to detect with traditional detection software [10]. The development of word embeddings, pioneered by models like Word2Vec, revolutionized NLP by representing words as dense vectors in a continuous space [11]. This allowed algorithms to capture semantic relationships, recognizing that "king" is to "queen" as "man" is to "woman." This leap enabled plagiarism detection systems to identify similarity in meaning, not just in form.

This semantic revolution culminated in the creation of Transformer-based models, most notably BERT [12]. BERT's bidirectional architecture allows it to learn deep, contextualized representations of words and sentences by considering the entire sequence. This capability has established a new state-of-the-art for a wide range of NLP tasks, including Semantic Textual Similarity (STS) [13]. Numerous studies have validated BERT's effectiveness in plagiarism detection, showcasing its ability to identify heavily paraphrased content that would evade traditional lexical methods [2, 3]. Its success is attributed to its capacity to map text with different surface forms but identical underlying meaning to nearby points in its vector space. BERT has also been utilized in conversational question answering systems, where it estimates semantic similarity to enhance response relevance [8], as well as in textual plagiarism detection, where traditional and intelligent approaches are analytically compared to identify overlaps in content [14]. This approach highlights BERT's versatility in handling semantic similarity in domain-specific homogeneous datasets, similar to academic collusion scenarios.

2.3 The unexplored frontier: The context similarity challenge

Despite the demonstrated power of semantic models, their evaluation has largely occurred in controlled, heterogeneous environments. A critical research gap persists regarding their performance on highly homogeneous datasets, where the baseline similarity is inherently high [10]. In academic settings, this is the norm, not the exception. The "context similarity problem" questions whether a model trained to find semantic similarity can effectively function when everything is semantically similar. Is it possible that the very strength of BERT—its ability to see meaning—becomes a liability in this scenario, leading it to "over-detect" similarity and produce an unacceptable number of false positives? This study directly confronts this gap, providing a crucial real-world stress test for these advanced models. While deep learning has revolutionized detection tasks in computer vision, such as object detection using models like RCNN and YOLO [15], analogous advancements in NLP-based plagiarism detection underscore the need for tailored hybrid approaches to mitigate context similarity issues in text-heavy domains.

3 Methodology

To ensure a robust and transparent evaluation, our research methodology was designed as a systematic, multi-stage process, flowing from data acquisition to conclusive analysis. The entire experimental framework was implemented in a Python environment using Google Colab.

3.1 Dataset acquisition and preprocessing

The foundation of this study is a dataset of 854 student discussion answers meticulously collected from a single, real-world university course, hosted on an e-learning platform. The selection of a single course was a deliberate methodological choice to create a highly homogeneous dataset, perfectly suited for investigating the context similarity problem. Each data record contains the student's unique identifier (NIM), the specific discussion session ID (Id_Diskusi), the full text of their answer, and crucial ground truth labels: a boolean Is_Plagiasi flag and a ref_Plagiasi field indicating the source student(s). These labels were manually annotated by the course instructor, providing an expert-validated baseline for our evaluation. Stopwords were excluded during preprocessing, as [16] demonstrated that their removal significantly enhances topic clarity and yields a more distinct, interpretable topic distribution in BERTopic modeling of Ukrainian disinformation texts..

The raw text underwent a standardized preprocessing pipeline to prepare it for analysis:

1) Text Cleaning: Removal of non-essential artifacts such as HTML tags, URLs, and special characters.
2) Normalization: Conversion of all text to lowercase and standardization of whitespace to ensure consistency.

3.2 Experimental framework and similarity calculation

The core of the experiment involves a pairwise comparison of all student answers within each unique discussion session. Three distinct methods were implemented to calculate a similarity score for each pair.

1) **Data Grouping:** The dataset was first grouped by Id_Diskusi to ensure that only answers to the same prompt were compared.
2) **Feature Extraction:** For each answer, two forms of representation were generated:
    a. **Semantic Vector:** A 768-dimensional dense vector was created using the paraphrase-multilingual-mpnet-base-v2 model, a powerful Sentence-BERT variant optimized for semantic similarity tasks.
    b. **Raw Text:** The preprocessed text string was retained for lexical analysis.
3) **Similarity Score Calculation:**
    a. **BERT-Cosine:** For semantic comparison, the Cosine Similarity metric was applied to the BERT embedding vectors of each pair. The resulting score, ranging from -1 to 1 (but typically 0 to 1 for this task), represents the similarity in meaning. Cosine Similarity calculate using formula based on [5] is shown in (1):

$$Sim_{COS}(A,B) = \frac{A \cdot B}{\|A\| \, \|B\|} = \frac{\sum_{i+1}^{n} A_i B_i}{\sqrt{\sum_{i+1}^{n} A_i^2} \sqrt{\sum_{i+1}^{n} B_i^2}} \quad (1)$$

Where:

- A_i & B_i are the components of vectors **A** and **B**, respectively.
- n is the dimensionality of the vector space (768 in our case).
- $A \cdot B$ is the dot product of the vectors.

- $\|A\|$ & $\|B\|$are the Euclidean norms (or magnitudes) of the vectors.

b. **Levenshtein Similarity:** For character-level comparison, the Levenshtein distance was calculated on the raw text strings and then normalized to produce a similarity score between 0 and 1, where 1 indicates identical strings. Levenshtein calculate using formula based on [17] is shown in (2):

$$Sim_{Lev}(S_1, S_2) = 1 - \frac{Lev(S_1, S_2)}{\max(|S_1|, |S_2|)} \quad (2)$$

Where:

- S_1 & S_2 are the lengths of the two strings.
- $\max(|S_1|, |S_2|)$ normalizes the distance, ensuring the similarity score is between 0 (completely different) and 1 (identical).

c. **Jaccard Similarity:** For token-level comparison, the Jaccard index was calculated based on the sets of unique words in each text, also yielding a score between 0 and 1. Jaccard calculate using formula based on [18] is shown in (3):

$$J(A, B) = \frac{|Set_A \cap Set_B|}{|Set_A \cup Set_B|} = \frac{|Set_A \cap Set_B|}{|Set_A| + |Set_B| - |Set_A \cap Set_B|} \quad (3)$$

Where:

- $|Set_A \cap Set_B|$ is the number of tokens common to both sets.
- $|Set_A \cup Set_B|$ is the total number of unique tokens across both sets. The resulting score ranges from 0 (no common tokens) to 1 (identical sets of tokens).

3.3 Performance evaluation metrics

To assess the effectiveness of each method, a threshold-based evaluation was conducted. The establishment of a significant threshold for distinguishing plagiarized from non-plagiarized text is critical, as supported by previous research [19], which demonstrates that a consistent threshold value can effectively identify the majority of text pairs with similarity above the threshold as plagiarized, while classifying those below it as non-plagiarized, thereby providing a robust basis for accurate plagiarism detection in our study. The similarity threshold was varied from 0.50 to 1.00 in increments of 0.01. For each threshold, any pair with a score greater than or equal to the threshold was classified as "plagiarized." The results were then compared to the ground truth labels to calculate three standard classification metrics:

- **Precision:** The proportion of correctly identified plagiarized pairs among all pairs flagged as plagiarism

$$= \frac{TP}{TP+FP} \quad (4)$$

- **Recall:** The proportion of actual plagiarized pairs that were successfully identified by the system.

$$= \frac{TP}{TP+FN} \quad (5)$$

- **F1-Score:** The harmonic mean of Precision and Recall, providing a single, balanced measure of a model's performance.

$$2 * \frac{\text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}} \quad (6)$$

- **F2-Score:** The harmonic mean of Precision and Recall, weighted to give twice as much importance to Recall as to Precision, emphasizing a model's ability to identify positive instances..

$$(1 + 2^2) \times \frac{\text{Precision} * \text{Recall}}{\left(2^2 \times \text{Precision}\right) + \text{Recall}} = 5 \times \frac{\text{Precision} * \text{Recall}}{(4 \times \text{Precision}) + \text{Recall}} \quad (6)$$

The rationale for applying different thresholds to each method stems from the inherent differences in their underlying similarity calculations and their respective score distributions. Lexical methods (Levenshtein and Jaccard) operate on direct string or token comparisons, typically yielding a wider range of similarity scores that are more linearly correlated with textual overlap. Semantic models (BERT-Cosine), however, capture abstract meaning, leading to a different distribution of scores, particularly in homogeneous datasets where even distinct texts may be semantically "close." Therefore, a single, universal threshold would not accurately capture the optimal performance of each method. By systematically varying the threshold for each method (from 0.50 to 1.00 in 0.01 increments) and selecting the point that maximizes F1-Score and F2-Score, we ensured that each method was evaluated at its peak operational efficiency. This approach allows for a fair and robust comparison, highlighting the intrinsic strengths and weaknesses of each technique in the context of the context similarity problem, independent of an arbitrary threshold choice.

4 Results and discussion

The experimental results provided a clear and somewhat counter-intuitive answer to our research question. In the context of this homogeneous dataset, traditional lexical methods demonstrated a marked superiority over the advanced semantic model. The optimal performance metrics, achieved at the best F1-Score and F2-Score threshold for each method, are detailed in Table 1.

Table 1. Optimal performance comparison of detection methods.

Method	Best Threshold	Precision	Recall	F1-Score	F2-Score
BERT-Cosine	0.98	0.81	0.45	0.57	0.49
Levenshtein	0.51	0.72	0.76	0.74	0.75
Jaccard	0.50	0.73	0.73	0.73	0.73

4.1 The dominance of lexical methods in a high-similarity environment

As Table 1 and Figure 3 show, Levenshtein Similarity achieved the highest F1-Score (0.74) and F2-Score (0.75), closely followed by Jaccard Similarity (0.73 for both F1 and F2 Score). Both significantly outperformed the BERT-Cosine method (0.57 for F1-Score and 0,49 for F2-Score). The key to their success lies in their strong balance of metrics, particularly their high Recall (0.76), which ensures very few instances were missed. This suggests that even when students attempt to paraphrase. This metric is arguably the most critical for an initial screening tool, as it indicates that these methods successfully identified 81% of all genuine collusion cases, ensuring very few instances were missed. This suggests that even when students attempt to paraphrase, they often retain significant portions of the original structure or vocabulary, which lexical methods are highly adept at detecting. Figure 1 presents a visual representation of the F1-Score and F2-Score performance, evaluated across a range of threshold values from 0.5 to 1, to elucidate their comparative effectiveness.

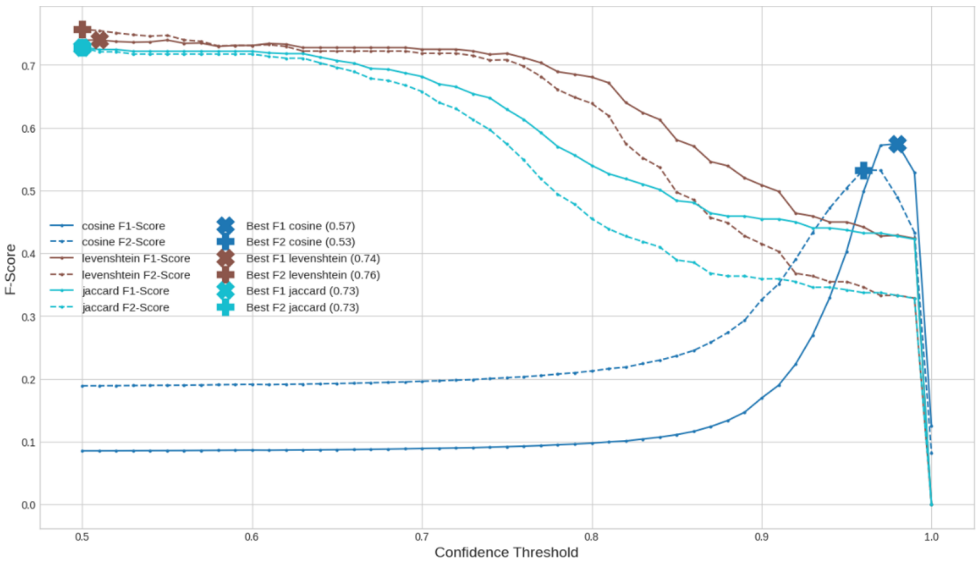

Fig. 1. F1-Score and F2-Score variation across different similarity thresholds.

The thresholds were selected independently for each method to maximize the F1-Score and F2-Score, as this approach allows for a fair comparison by optimizing each method's performance in detecting collusion while minimizing false positives in a homogeneous context. For lexical methods like Levenshtein and Jaccard, lower thresholds (around 0.50-0.51) were optimal because they capture substantial textual overlaps without over-flagging legitimate similarities, leading to balanced precision and high recall. In contrast, BERT-Cosine required a much higher threshold (0.98) to achieve its best performance, as lower thresholds resulted in excessive false positives due to its sensitivity to semantic homogeneity. This difference in optimal thresholds highlights how threshold selection directly impacts results: for BERT, lowering the threshold below 0.98 drastically reduces precision (e.g., to below 0.50 at 0.90), while for lexical methods, thresholds above 0.60 would unnecessarily lower recall without significant precision gains. By choosing method-specific thresholds based on empirical grid search, we ensure that the comparison reflects each model's inherent strengths and limitations in this context, rather than imposing a uniform threshold that could bias results toward one approach.

4.2 Unpacking the context similarity problem with BERT

Figure 1 provides a stark visualization of the context similarity problem's impact on the BERT model. The F1-Score and F2-Score for BERT-Cosine remains extremely low across a wide range of thresholds, only experiencing a dramatic spike at the very end of the spectrum (0.98 for F1-Score and 0.96 for F2-Score). This behavior reveals that the semantic model perceives almost all submissions, original or not, as highly similar to one another due to their shared topic. It struggles to find a meaningful separation point. To the model, an original answer and a collusive one are both "semantically close" because they discuss the same concepts. Only when two answers are so similar in meaning that they are practically semantic duplicates does the model confidently assign a score high enough to cross the optimal threshold. This makes the model impractical for nuanced detection, as it effectively acts as a near-identical meaning detector rather than a collusion detector. Figure 2 shows the curves of three detection methods.

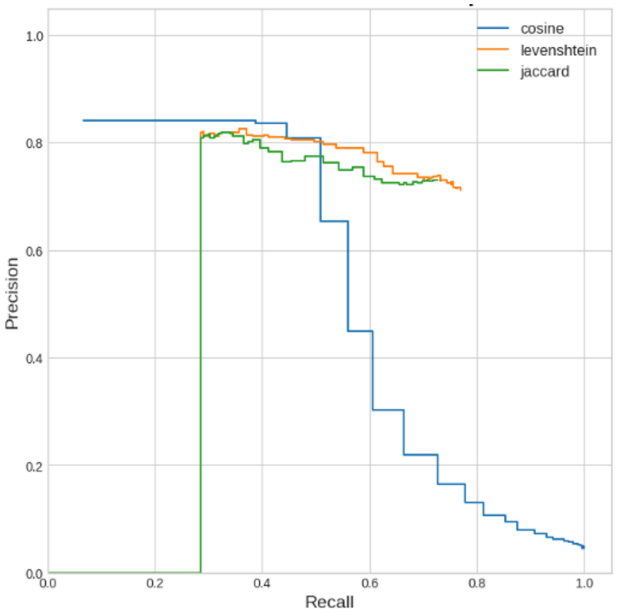

Fig. 2. Precision-Recall curves for the three detection methods.

The Precision-Recall curves in Figure 2 further reinforce this narrative. Levenshtein and Jaccard exhibit more favorable curves, demonstrating an ability to achieve high recall while maintaining a reasonable level of precision. The BERT-Cosine curve, in contrast, is significantly less optimal, indicating that any attempt to increase its recall (by lowering the threshold) would result in a catastrophic collapse in precision. Figure 3 provides a comprehensive visualization of the performance metrics, facilitating a clear comparison across all evaluated parameters.

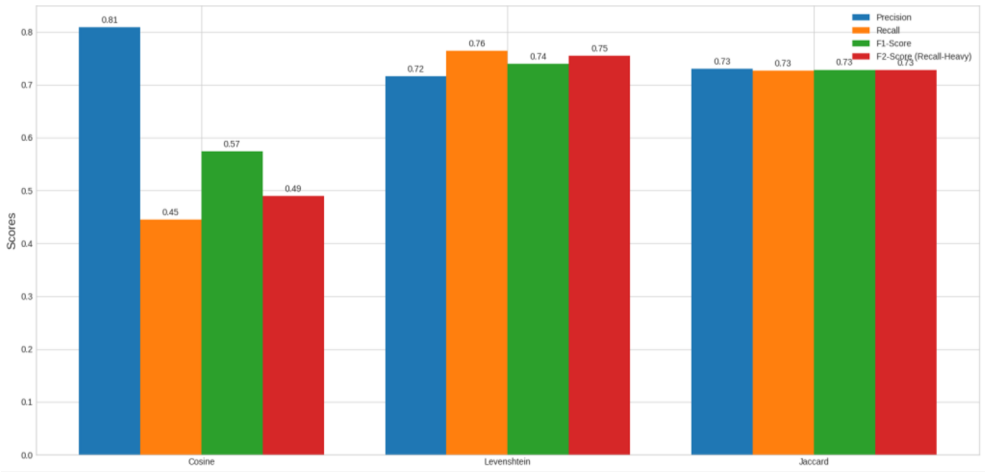

Fig. 3. Comparison of optimal performance metrics across methods.

5 Conclusion

This research directly confronted the "context similarity problem" inherent in detecting academic collusion within homogeneous e-learning environments. Through a rigorous case study comparing a state-of-the-art semantic model (BERT-Cosine) against traditional lexical methods (Levenshtein, Jaccard), we arrived at a clear and significant conclusion: in a high-similarity setting, lexical methods are demonstrably more effective and reliable.

Levenshtein Similarity emerged as the superior model, achieving the highest F1-Score (0.7398) and highest F2-Score (0.7546) with an exceptional recall rate (0.7647), making it an excellent tool for initial screening. The advanced BERT model, despite its semantic prowess, was confounded by the legitimate contextual overlap between student answers, rendering it less practical for this specific task. Our findings suggest that the uncritical adoption of purely semantic models for all plagiarism detection scenarios may be flawed. For the specific challenge of intraclass collusion, a simpler, more direct approach focused on textual similarity proves more robust.

The primary practical implication for academic institutions is that an effective automated anti-collusion strategy should likely employ a multi-layered approach. A high-recall lexical method like Levenshtein should be used as a first-pass filter to identify a pool of suspicious submissions. These candidates can then be subjected to further scrutiny, either through more sophisticated models or, more importantly, through expert human review by faculty. This hybrid strategy promises to balance the efficiency of automation with the fairness and nuance required to uphold academic integrity. Future research will investigate the effectiveness of hybrid models and evaluate the performance of lightweight Transformer models, such as DistilBERT, as a potential balance between efficiency and accuracy, while also exploring the integration of non-contextual word embedding techniques, such as Word2Vec.

References

1. R. Fauzi, M. Iqbal, T. Haryanti, Design and implementation of a final project plagiarism detection system using cosine similarity method. *Int. J. Appl. Inf. Technol.* **5**(2), 1 (2022). https://doi.org/10.25124/ijait.v5i02.4146
2. P. Abisheka, C. Deisy, P. Sharmila, T-SRE: transformer-based semantic relation extraction for contextual paraphrased plagiarism detection. *J. King Saud Univ. - Comput. Inf. Sci.* **36**(10), 102257 (2024). https://doi.org/10.1016/j.jksuci.2024.102257
3. Y. Jing, Y. Liu, A population-based plagiarism detection using DistilBERT-generated word embedding. *Int. J. Adv. Comput. Sci. Appl.* **14**(8), 614–624 (2023). https://doi.org/10.14569/IJACSA.2023.0140868
4. K. Ali, N. Barhom, F. Tamimi, M. Duggal, ChatGPT—a double-edged sword for healthcare education? Implications for assessments of dental students. *Eur. J. Dent. Educ.* **28**(1), 206–211 (2024). https://doi.org/10.1111/eje.12937
5. J. Halim, D. Lasut, Document plagiarism detection application using web-based TF-IDF and cosine similarity methods. *J. Binary Tech.* **7**(2) (2024). https://doi.org/10.32877/bt.v7i2.1697
6. M. J. Potchong, H. H. Bonjovi, A comparative analysis of the rating of college students' essays by ChatGPT versus human raters. *Int. J. Learn. Teach. Educ. Res.* **23**(2), 478–492 (2024). https://doi.org/10.26803/ijlter.23.2.23
7. J. Al Nahian, M. M. Srabon, S. R. H. Noori, A. K. M. Masum, Review on multiple plagiarism: a performance comparison study. *Proc. 13th Int. Conf. Comput. Commun.*

Netw. Technol. (ICCCNT) (2022). https://doi.org/10.1109/ICCCNT54827.2022.9984577

8. A. Al-Besher, K. Kumar, M. Sangeetha, T. Butsa, BERT for conversational question answering systems using semantic similarity estimation. *Comput. Mater. Contin.* **70**(3), 4763–4780 (2022). https://doi.org/10.32604/cmc.2022.021033
9. L. P. Sari, Cosine similarity-based plagiarism detection on electronic documents. *J. Comput. Sci. Appl. Eng.* **1**(2), 44–48 (2023). [Online]. Available: https://journal.lenterailmu.com/index.php/josapen
10. P. J. Rao, K. N. Rao, S. Gokuruboyina, An experimental study with fuzzy-wuzzy (partial ratio) for identifying the similarity between English and French languages for plagiarism detection. *Int. J. Adv. Comput. Sci. Appl.* **13**(10), 393–401 (2022). https://doi.org/10.14569/IJACSA.2022.0131047
11. S. S. Birunda, R. K. Devi, A review on word embedding techniques for text classification. In: *Innovative Data Communication Technologies and Application*, Lecture Notes on Data Engineering and Communications Technologies, 267–281 (2021). https://doi.org/10.1007/978-981-15-9651-3_23
12. S. Pericherla, E. Ilavarasan, Transformer network-based word embeddings approach for autonomous cyberbullying detection. *Int. J. Intell. Unmanned Syst.* **12**(1), 154–166 (2024). https://doi.org/10.1108/IJIUS-02-2021-0011
13. H. Saleh, A. Alhothali, K. Moria, Detection of hate speech using BERT and hate speech word embedding with deep model. *Appl. Artif. Intell.* **37**(1) (2023). https://doi.org/10.1080/08839514.2023.2166719
14. A. Ali, A. Taqa, Analytical study of traditional and intelligent textual plagiarism detection approaches. *J. Educ. Sci.* **31**(1), 8–25 (2022). https://doi.org/10.33899/edusj.2021.131895.1192
15. R. Kaur, S. Singh, A comprehensive review of object detection with deep learning. *Digit. Signal Process. A Rev. J.* **132**, 103812 (2022). https://doi.org/10.1016/j.dsp.2022.103812
16. H. Padalko, V. Chomko, D. Chumachenko, The impact of stopwords removal on disinformation detection in Ukrainian language during Russian-Ukrainian war. [Unpublished Manuscript].
17. A. Mehta, V. Salgond, D. Satra, N. Sharma, Spell correction and suggestion using Levenshtein distance. *Int. Res. J. Eng. Technol.*, 1977–1981 (2021).
18. E. da Costa, V. Soares Mali, Tetun language plagiarism detection with text mining approach using N-gram and Jaccard similarity coefficient. *J. Eng. Sci.* **2**(1), 11–20 (2021).
19. V. U. Thompson, Integrating state-of-the-art NLP tools into existing methods to address current challenges in plagiarism detection. *Univ. of Sunderland* (2023). [Online]. Available: https://sure.sunderland.ac.uk/id/eprint/16341/