Volume 11 Issue 1 Year 2026 | Page 96-107 ISSN: 2527-9866
Received: ***16-12-2025*** | *Revised:* ***24-12-2025*** | *Accepted:* ***12-01-2026***

Plagiarism Detection in English Academic Documents using a Lexical Semantic Hybrid and Support Vector Machine

Callista Virginia[1], Derry Alamsyah[2]
[1,2] Multi Data Palembang University, Sumatera Selatan, Indonesia, 30113
e-mail: callistavirginia_2226250118@mhs.mdp.ac.id[1], derry@mdp.ac.id [2]
*Correspondence: callistavirginia_2226250118@mhs.mdp.ac.id

Abstract: *Detecting plagiarism in academic writing has become increasingly challenging due to advanced text modification strategies that reduce surface-level similarity while preserving the original meaning. This study proposes a hybrid plagiarism detection system that integrates lexical and semantic similarity features to distinguish between plagiarism and altered documents in academic texts. As a key contribution, this study provides a systematic evaluation of a lexical–semantic hybrid plagiarism detection approach using Support Vector Machine (SVM) on English-language academic documents, where all plagiarism cases across different obfuscation levels are consolidated into a single plagiarism class. Lexical similarity is modeled using Term Frequency–Inverse Document Frequency (TF–IDF), while semantic similarity is captured through Sentence-BERT embeddings. These features are combined into a two-dimensional hybrid similarity representation and classified using SVM. The proposed approach is evaluated on the PAN 2025 dataset using stratified 5-fold cross-validation. Experimental results show that the hybrid SVM-based model achieves an average accuracy of 92.5% with the optimal kernel, along with competitive precision, recall, F1-score, and AUC values. Kernel-based evaluation and cross-validation analyses further demonstrate the robustness and generalization capability of the proposed framework, indicating that the hybrid lexical–semantic representation is effective for distinguishing plagiarism and altered content in English academic writing.*

Keywords: *Plagiarism, lexical-semantic hybrid, SVM.*

1. Introduction

In academic environments, the information sources most frequently accessed by students and researchers are academic documents. Academic documents are often understood as scholarly outputs that represent the results of academic activities and contribute to the advancement of science and technology [1]. Generally, academic documents encompass various forms of scientific writing used in educational and research contexts, including journal articles, undergraduate theses, master's theses, dissertations, research reports, conference proceedings, and academic papers all of which serve to support scientific communication [2]. English-language academic documents hold a central role in the global scientific ecosystem because English has become the lingua franca of international scholarly publication and communication. These documents remain an essential component of modern scientific communication due to their contribution to expanding knowledge dissemination and improving research quality worldwide [3].

As an integral part of scientific communication, academic documents must uphold writing integrity, which emphasizes honesty and proper attribution of all utilized sources. This aligns with the principles of academic integrity that stress honesty and responsibility across educational and research activities [4]. Violations of these principles have led to various plagiarism cases in both academic and global contexts. The PlagiarismSearch report recorded an average plagiarism rate of 16.36% from 69.89 million documents in 2024, peaking at 18.79% in 2020. Copyleaks also reported that

plagiarism rates remain high across seven countries, even though they decreased from 35% to 17% between January 2023 and January 2024. These figures indicate that plagiarism continues to be a critical issue despite the advancement of tools that support scientific literacy.

Amid these challenges, plagiarism detection systems play an important role in maintaining academic integrity. Traditional approaches that rely heavily on lexical similarity, such as direct word matching, are highly sensitive to copy paste behavior [5]. However, these methods are less robust when facing paraphrased text that conveys the same meaning using different vocabulary [6]. This limitation arises because lexical methods depend solely on word overlap and lack contextual representation capabilities [7]. They cannot capture variations in sentence structure or phrase rearrangement, which makes them ineffective for identifying more subtle forms of plagiarism [8]. These limitations highlight the need for an approach capable of understanding the underlying meaning of text rather than simply detecting word occurrence.

To address these shortcomings, semantic similarity plays a crucial role, particularly in identifying paraphrased text where lexical or structural changes occur while the meaning remains intact [9]. This approach uses vector representations that incorporate contextual semantics, such as transformer-based models or semantic embeddings, to compare documents more comprehensively, enabling the system to detect paraphrasing even when the lexical surface differs [10]. One of the model developments that implements contextual semantic representation is Sentence-BERT (SBERT). However, semantic representations still require a classification model to differentiate between plagiarized and original text. Support Vector Machine (SVM) has proven effective in detecting various forms of plagiarism due to its ability to handle high-dimensional data with stable accuracy. SVM has demonstrated excellent performance in detecting text plagiarism, achieving high accuracy and F-measure scores of 92.91% and 92.95%, respectively [11].

Although SVM performs well, its effectiveness is strongly influenced by the quality of feature representation. Hybrid approaches that combine lexical and semantic features have shown superior performance compared to using a single type of feature. Research by Mehdi et al. demonstrated the effectiveness of hybrid approaches in improving detection accuracy, achieving 89.2% accuracy and an F1-score of 0.903 using a Linear SVM classifier. However, their dataset was limited to researcher-generated pairwise text in the general domain [12]. Moreover, cosine similarity, which provides a more accurate measurement of vector proximity than conventional weighting methods such as TF-IDF, was not incorporated into their feature design.

Furthermore, research evaluating the performance of lexical–semantic hybrid approaches using Support Vector Machine on English-language academic documents, particularly in the science and engineering domains, remains limited. Accordingly, this study aims to extend and enhance such hybrid approaches by systematically evaluating their reliability for plagiarism detection using the PAN 2025 dataset. The objective of this study is to evaluate the effectiveness of a lexical–semantic hybrid plagiarism detection approach using Support Vector Machine in distinguishing plagiarism and altered documents in English-language science and engineering academic texts.

2. Methods

A. Data Description

The dataset used in this study was collected through a digital document analysis approach by utilizing officially published secondary data. Specifically, the dataset was obtained from the PAN 2025 Generated Plagiarism Detection Spot-Check corpus, which was downloaded from the Zenodo Repository, the official data hosting platform for PAN datasets under the CLEF (Conference and Labs of the Evaluation Forum) initiative.

B. Research Methods

In conducting this research, the methodological design was first formulated to ensure that each stage proceeded systematically and aligned with the development objectives. The overall research workflow is illustrated in Figure 1.

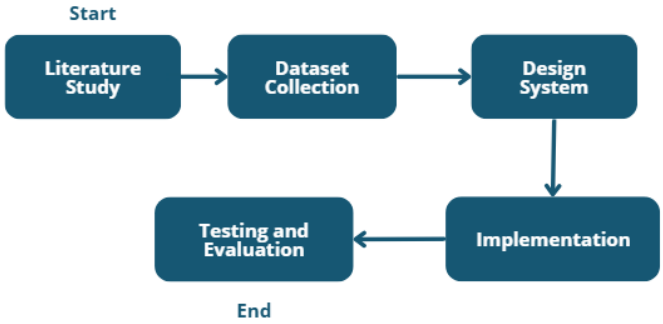

Figure 1. Research Method Workflow

1. Literature Study

The first stage of this research is the literature study, which aims to develop a comprehensive understanding of the theories, concepts, and prior studies related to text-based plagiarism detection. The main focus of this review includes studies on plagiarism detection systems and text similarity measurement approaches such as TF-IDF, *Cosine Similarity*, and SBERT. Through this phase, previous research findings are analyzed to identify strengths, limitations, and potential research gaps that can be addressed using a more efficient hybrid approach. This stage provides a solid theoretical foundation for the design and development of the plagiarism detection system proposed in this study.

2. Dataset Collection

The dataset consists of 151 files organized into several main components, including 50 source documents, 50 suspicious documents, 50 ground-truth annotation files, and an additional configuration file that defines document pair relationships. The src directory contains original reference documents, while the susp directory includes documents that have been modified or synthetically generated for evaluation purposes. Ground-truth annotations are stored in the *00-spot-check-truth* folder, which specifies text segments exhibiting similarity or plagiarism-related characteristics. Sample of the *ground-truth annotations* as shown below.

```
<document reference="suspicious-document020497.txt">
    <feature name="plagiarism" type="llm_prompted" llm="DeepSeek-R1" this_language="en"
    this_offset="17035" this_length="937" source_reference="source-document020497.txt"
    obfuscation="medium" source_offset="38921" source_length="937"/>
    <feature name="altered" type="llm_prompted" llm="DeepSeek-R1" this_language="en"
    this_offset="3620" this_length="127"/> … </document>
```

Although the PAN 2025 ground truth files contain 19 various types of annotations, this study only utilizes the essential information required for plagiarism detection. The adopted approach allows the proposed method to focus on the most relevant evidence while disregarding supplementary metadata that does not significantly influence detection results. The following section provides a brief explanation of the main attributes used to describe text similarity and the alignment between *suspicious* and *source* documents.

a. Document Reference : Identifies the suspicious document through its filename and serves as the primary reference for all annotations.
b. Feature Name : Indicates the annotation type, with *plagiarism* and *altered* being the most relevant as they represent text reuse or modification.

c. Source Reference : Specifies the source document used for comparison and is only provided for *plagiarism* cases, as *altered* texts do not rely on a specific source document.
d. Similarity: A score between 0 and 1 indicating the degree of similarity between the *suspicious* and *source* texts.
e. This_offset / This_length: Define the starting position and length of the detected segment within the suspicious document.
f. Source_offset / Source_length: Define the corresponding starting position and length of the matched segment in the source document.

In addition, the PAN 2025 dataset distinguishes between two main annotation categories, namely *plagiarism* and *altered*, which represent different forms of AI-generated text modification. The *plagiarism* category comprises text segments in suspicious documents that exhibit measurable similarity to their corresponding source documents, with transformation levels characterized by obfuscation types such as *simple*, *medium*, and *hard*. Conversely, the *altered* category includes text that has been substantially modified or entirely regenerated by Large Language Models without direct reference to a specific source document. In this study, all plagiarism instances across different obfuscation levels are consolidated into a single plagiarism class, as the objective is to evaluate the model's overall capability to recognize plagiarism-related content rather than to differentiate specific obfuscation types. Fully original documents are not explicitly modeled due to the pair-based structure of the dataset. Table 1 presents an example of the dataset used in this study.

Table 1. Sample Dataset

Suspicious Document	Source Document	Label	Sample Text
suspicious-document 020470.txt	source-document 020470.txt	altered	"Accretion onto stellar-mass black holes (BHs) is characterized by distinct state transitions, first identified through X-ray spectral studies of Cygnus X-1 by Tananbaum et al. (1972) and later understood as fundamental changes in the accretion flow structure…"
suspicious-document 020473.txt	source-document 020473.txt	altered	"Relation extraction (RE) is a crucial component of various NLP applications, including knowledge-base population and question answering… involves identifying relationships between two entity mentions in unstructured text."
suspicious-document 020468.txt	source-document 020468.txt	Plagiarism, simple	**Suspicious**: "Emission in spectral lines offers unique insights into interstellar turbulence… challenging to measure." **Source**: "Emission in spectral lines can provide unique information on interstellar turbulence… difficult to measure."
suspicious-document 020469.txt	source-document 020469.txt	Plagiarism, medium	**Suspicious**: "The art of instruction, a critical component in the ongoing evolution of contemporary human civilization, serves to equip students with the necessary knowledge and abilities." **Source**: "Teaching, which aims to help students learn new knowledge or skills effectively and efficiently, is important to advance modern human civilization."

suspicious-document 020481.txt	source-document 020481.txt	Plagiarism, hard	**Suspicious**: "A CPTP error on a physical qudit in a general resource state can lead to nonlinear noise propagation, preventing fully fault-tolerant operations through local measurements. **Source**: "For general resource states, not all physical errors can be represented as linear CPTP maps in the correlation space, which implies that noise propagation cannot always be modeled linearly"

3. System Design

The system design stage aims to construct and organize the workflow of the plagiarism detection system to be developed. The overall system design workflow is illustrated in Figure 2.

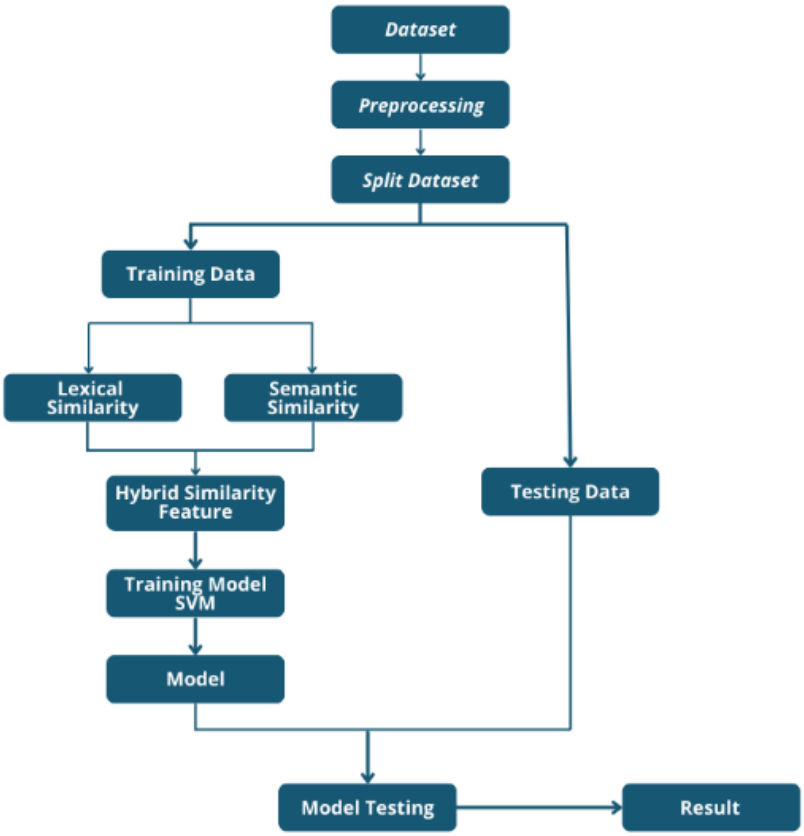

Figure 2. System Design Scheme

The process begins by loading the PAN 2025 dataset as the primary data source. All documents then undergo a series of preprocessing steps, including tokenization, lowercasing, stopword removal, and stemming or lemmatization, to normalize word forms and reduce noise. After preprocessing, the dataset is divided into 80% training data and 20% testing data to ensure that the model is trained and evaluated objectively using distinct portions of the data.

In the feature extraction stage, the training data are used to generate two types of text representations. First, TF–IDF vectorization is applied to construct lexical document representations, followed by the calculation of cosine similarity to obtain lexical similarity scores between document pairs. Second,

Sentence-BERT (SBERT) is employed to generate semantic embeddings, from which cosine similarity is computed to produce semantic similarity scores. During this process, the TF–IDF model is fitted on the training data to learn the vocabulary and IDF values, whereas SBERT is utilized directly as a *pre-trained* model. Both similarity scores are then combined into a two dimensional Hybrid Similarity Feature vector that represents the integrated lexical and semantic characteristics of each document pair.

In the model training stage, the Hybrid Similarity Feature vectors are used as input to train a Support Vector Machine (SVM), enabling the model to learn similarity patterns between document pairs. Once training is completed, the SVM is applied in the classification stage to categorize document pairs as either *plagiarism* or *altered* based on their Hybrid Similarity Feature values. During the testing phase, the test data are transformed using the feature extraction parameters learned from the training data to produce Hybrid Similarity Features, which are subsequently classified by the trained SVM model.

4. Implementation

The system implementation stage translates the designed architecture into an executable program. Python was chosen due to its rich ecosystem for text processing, including Pandas, NumPy, scikit-learn, NLTK, and the Transformers library for SBERT. The implementation follows the workflow shown in Figure 2, covering dataset loading, text preprocessing, feature extraction, and SVM-based classification. All components are integrated into a single pipeline to ensure consistent and reproducible similarity measurement and classification results. Parameter tuning is conducted by evaluating multiple SVM kernel functions, including linear, polynomial, radial basis function (RBF), and sigmoid kernels, under the same experimental configuration. The selection of the optimal kernel is based on cross-validation performance, with emphasis on balanced evaluation metrics such as F1-score and AUC to ensure robust classification performance.

5. Testing and Evaluation

The testing phase was conducted to evaluate the system's performance in identifying similarity levels between text documents. Several forms of evaluation were performed and presented in a series of tables to comprehensively assess the model's reliability. The evaluation results are summarized in the kernel-based evaluation table, the feature-based evaluation table, and the baseline comparison table. A train–test split with an 80%:20% ratio was initially applied to illustrate model behavior and support kernel, feature, and baseline-level comparisons. After obtaining the results from the kernel, feature, and baseline evaluations, a *k-fold cross-validation* analysis was conducted to ensure the consistency and generalizability of the model [13]. In this study, *Stratified 5-Fold Cross-Validation* was applied to maintain balanced class proportions across folds. The results indicate that evaluation based on a single split can produce performance estimates that are overly high for certain data partitions, leading to optimistic bias, where the model achieves perfect performance with an accuracy, F1-score [14], and AUC of 1.000. In contrast, the 5-fold cross-validation results provide more stable and realistic performance estimates by averaging results across multiple independent folds, yielding an average accuracy of 0.900 ± 0.094, an F1-score of 0.767 ± 0.200, and an AUC of 0.914 ± 0.070, which better reflect the model's generalization ability. Therefore, the use of *cross-validation* helps prevent training bias and provides a more objective assessment of the system's robustness in detecting plagiarism.

Subsequently, model performance was measured using several evaluation metrics, namely accuracy, precision, recall, and F1-score, calculated using Formula (1), (2), (3), (4).

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \times 100\% \qquad (1)$$

$$Precision = \frac{TP}{TP+FP} \times 100\% \qquad (2)$$

$$Recall = \frac{TP}{TP+FN} \times 100\% \quad (3)$$

$$F1 - Score = 2 \times \frac{Precision \times Recall}{Precision+Recall} \quad (4)$$

In these formulations, *True Positive (TP)* refers to the number of positive samples that are correctly classified by the system, *True Negative (TN)* indicates the number of negative samples correctly predicted, *False Positive (FP)* denotes the number of negative samples incorrectly classified as positive, and *False Negative (FN)* represents the number of positive samples incorrectly classified as negative.

The evaluation also included the calculation of the Area Under the Curve (AUC) to assess the model's overall ability to discriminate between positive and negative classes Click or tap here to enter text. The AUC value serves as a crucial indicator of model performance, especially under potential class imbalance conditions, and helps determine the most suitable model configuration Click or tap here to enter text. The AUC was computed using the trapezoidal rule, which estimates the area under the Receiver Operating Characteristic (ROC) curve based on numerical approximations of the True Positive Rate (TPR) and False Positive Rate (FPR). The formula for AUC is shown in Formula (5).

$$True\ Positive\ Rate\ (TPR) = \frac{TP}{TP + FN}$$

$$False\ Positive\ Rate\ (FPR) = \frac{FP}{FP + TN}$$

$$Area = \left(\frac{TPR_n + TPR_{n-1}}{2}\right) * (FPR_n + FPR_{n-1}) \quad (5)$$

The values TPR_n and TPR_{n-1} correspond to the true positive rates at threshold levels *n* and *n–1*, while FPR_n and FPR_{n-1} represent the false positive rates at the same respective thresholds.

Furthermore, a ROC curve visualization was generated to illustrate the relationship between TPR and FPR. This analysis provides a more comprehensive understanding of the model's capability to differentiate between plagiarism categories and assists in identifying the model with the best performance [15]. The ROC curve is shown in Figure 3.

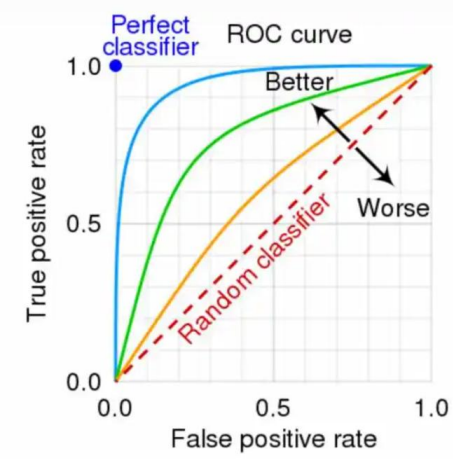

Figure 3. ROC Curve

3. Results and Discussion

A. System Implementation

The plagiarism detection system is implemented that aims to differentiate between *plagiarized* and *altered* documents. Each pair of documents, consisting of a suspicious text and its corresponding source, is represented through a hybrid similarity representation that captures both lexical and semantic information. These features are integrated into a unified input space and analyzed by a classification model to decide whether the suspicious document should be identified as *plagiarism* or considered *altered*.

B. Experimental Result

This research is limited to assessing the model's performance in distinguishing between plagiarism and altered documents. All plagiarism instances identified in the dataset, including those involving different obfuscation techniques, are consolidated into *plagiarism* class. The proposed model is not designed to explicitly learn or represent non-plagiarized class. The evaluation focuses on the model's overall ability to recognize plagiarism-related patterns rather than differentiating between specific obfuscation levels or modeling fully original documents.

In the experimental evaluation, two complementary evaluation settings are employed. First, a train–test split is used to illustrate the behavior of the proposed model and to support kernel-based, feature-based, and baseline comparisons under a fixed data partition. While this setting provides an intuitive comparison of different configurations, it is sensitive to data partitioning and may lead to optimistic performance estimates. Therefore, *k-fold cross-validation* is subsequently applied as the primary evaluation strategy to obtain a more reliable and generalizable assessment of model performance.

Table 2. Kernel Based Evaluation

Kernel	Accuracy	Precision	Recall	F1-Score	AUC
Linear	0.90	0.889	1.00	0.941	1.00
Polynomial	0.80	0.875	0.875	0.875	0.938
RBF	0.80	1.000	0.750	0.857	0.938
Sigmoid	0.50	0.714	0.625	0.625	0.938

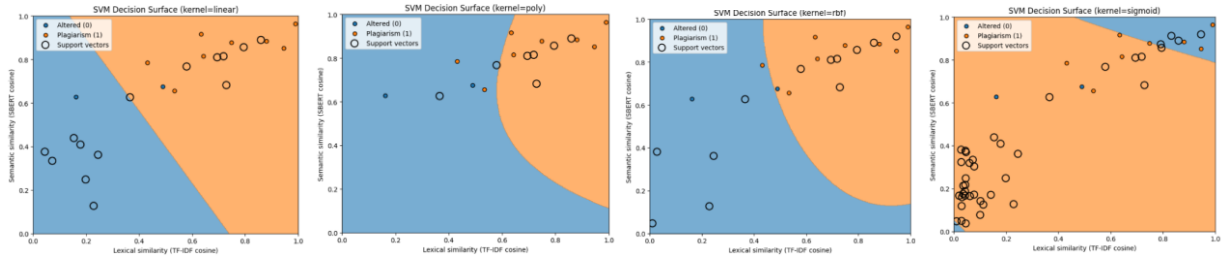

Figure 4. (a) Linear (b) Polynomial (c) RBF (d) Sigmoid

Based on Table 2 and the SVM decision boundary visualizations shown in Figure 4, the linear kernel achieves the highest overall performance, with an accuracy of 0.90, precision of 0.889, perfect recall of 1.000, an F1-score of 0.941, and an AUC of 1.00. As illustrated in Figure 4(a), the linear kernel forms a clear and stable decision boundary that effectively separates plagiarism and altered document pairs in the hybrid similarity feature space. The RBF and polynomial kernels also demonstrate competitive performance; however, their non-linear decision boundaries, as shown in Figure 4(b) and Figure 4(c), exhibit overlapping regions that lead to reduced recall compared to the linear kernel. In contrast, the sigmoid kernel performs poorly, which is consistent with the unstable and fragmented decision regions observed in Figure 4(d). Overall, these results indicate that a linear decision boundary is sufficient and more reliable for distinguishing plagiarism and altered documents using hybrid similarity features.

Table 3. Type of Features Evaluation

Feature	Accuracy	Precision	Recall	F1-Score	AUC
Lexical	0.90	1.000	0.875	0.933	0.938
Semantic	0.80	0.800	1.000	0.889	0.875
Hybrid	0.80	0.875	0.875	0.933	0.938

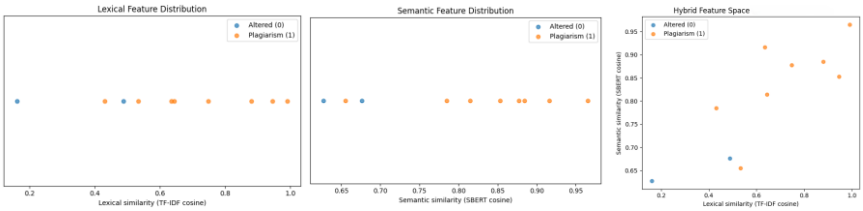

Figure 5. (a) Lexical (b) Semantic (c) Hybrid

Based on Table 3, the lexical approach using TF–IDF cosine similarity achieves strong performance, with an accuracy of 0.90, perfect precision (1.000), and an F1-score of 0.933, indicating its effectiveness in detecting plagiarism with high word overlap. The semantic approach based on SBERT embeddings attains lower accuracy (0.80) but perfect recall (1.000), demonstrating its ability to capture plagiarism cases with reduced lexical similarity at the expense of precision. The hybrid lexical–semantic feature combination provides balanced performance, achieving an F1-score of 0.933 and an AUC of 0.938. As visualized in Figure 5, the hybrid feature space exhibits clearer separation between plagiarism and altered documents compared to the individual lexical and semantic feature distributions, supporting the results in Table 5 and confirming the effectiveness of integrating surface-level and contextual similarity features for robust plagiarism detection.

Table 4. Baseline Comparison Evaluation

Baseline	Accuracy	Precision	Recall	F1-Score	AUC
Majority	0.20	0.00	0.000	0.000	0.500
Random	0.50	1.00	0.375	0.545	0.688
Threshold	0.50	1.00	0.375	0.545	0.938
SVM	0.80	0.87	0.875	0.875	0.938

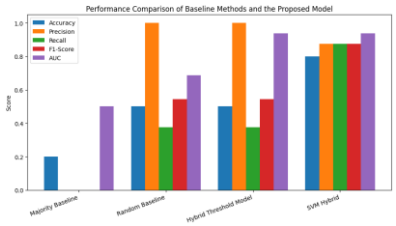

Figure 6. Visualization of Baseline Comparison

Based on Table 4, the proposed SVM model outperforms all baseline methods across all evaluation metrics. The Majority Baseline shows the weakest performance, with an accuracy of 0.20 and zero precision, recall, and F1-score, reflecting its inability to model document similarity. The Random Baseline and Hybrid Threshold Model achieve moderate accuracy (0.50) and high precision but suffer from low recall (0.375), indicating limited reliability in identifying plagiarism cases. In contrast, the SVM-based approach attains a substantially higher accuracy (0.80) with balanced precision (0.87), recall (0.875), and F1-score (0.875), demonstrating that its performance gains arise from learning discriminative patterns in the hybrid lexical–semantic feature space, as illustrated in Figure 6, rather than relying on fixed threshold heuristics. It is important to note that the results presented in Table 2, 3, and 4 are derived from a train–test split and are intended to provide illustrative comparisons across kernels, feature representations, and baseline methods. These results highlight relative performance differences but may vary depending on the chosen data partition. In contrast, the cross-validation results reported in Table 5 represent averaged performance across multiple folds and are therefore emphasized as the primary indicator of model robustness and generalization capability.

Table 5. Cross Validation

Kernel	Accuracy	Precision	Recall	F1-Score	AUC
Linear	0.875	0.567	1.000	0.700	0.94
RBF	0.925	0.700	1.000	0.800	0.94
Polynomial	0.900	0.667	1.000	0.767	0.91
Sigmoid	0.075	0.075	0.600	0.133	0.05

Based on Table 5, the polynomial kernel achieves the best cross-validation performance, with the highest mean accuracy (0.925) and F1-score (0.800), indicating stable classification across folds. The RBF and linear kernels follow with accuracies of 0.900 and 0.875 and F1-scores of 0.767 and 0.700, respectively, while all three kernels maintain perfect recall (1.000), demonstrating consistent identification of plagiarism cases. In contrast, the sigmoid kernel performs poorly, with accuracy and F1-score dropping to 0.075 and 0.133, confirming its unsuitability for the hybrid feature space. Overall, these results highlight the robustness and generalization capability of the proposed SVM approach, particularly when using polynomial and RBF kernels.

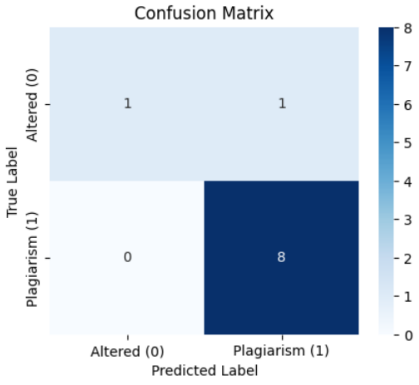

Figure 7. Confusion Matrix

Based on the confusion matrix on Figure 7, the model correctly classified 1 *altered* document as *altered* and successfully identified all 8 plagiarism cases as *plagiarism*. One altered document was misclassified as *plagiarism*, while no *plagiarism* cases were incorrectly labeled as *altered*. These results indicate that the model demonstrates strong capability in detecting *plagiarism*, achieving perfect recall for *plagiarism* cases. Overall, the confusion matrix suggests that the model prioritizes accurate plagiarism detection while maintaining reasonable discrimination between *plagiarism* and *altered* documents.

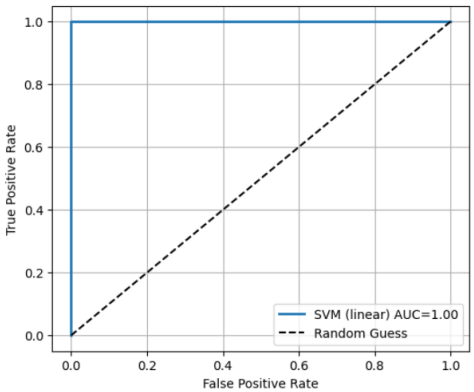

Figure 8. ROC Curve

Figure 8 shows the Receiver Operating Characteristic (ROC) curve, which provides a threshold-independent evaluation of the model's ability to distinguish between plagiarism and altered documents. The curve rises steeply toward the upper-left corner, with an Area Under the Curve (AUC) value of 1.00, indicating excellent discriminative performance. These results confirm that the classifier consistently assigns higher confidence scores to plagiarism cases than to altered documents and that its performance is robust across different decision thresholds.

4. Conclusions

The experimental results demonstrate that the proposed hybrid lexical–semantic approach combined with a Support Vector Machine classifier is effective in distinguishing between plagiarism and altered documents in English-language academic texts. Kernel-based evaluation reveals notable performance differences among SVM kernels, where the linear kernel achieves the strongest overall results in terms of accuracy and F1-score, while polynomial and RBF kernels exhibit competitive and stable performance, particularly under cross-validation. These findings suggest that both linear and nonlinear decision boundaries can effectively model the hybrid similarity space, depending on the evaluation setting. Feature-based evaluation further shows that lexical features remain highly effective for capturing plagiarism cases with strong surface-level similarity, while the integration of semantic information through a hybrid representation provides more balanced performance by incorporating contextual similarity. This combined feature approach enables the system to robustly distinguish plagiarism-related content from altered documents without relying on fixed heuristics. In comparison with baseline methods, the proposed SVM-based model consistently outperforms majority, random, and threshold-based approaches, and cross-validation results confirm its stability and generalization capability across multiple data partitions.

From a practical perspective, these results indicate that the proposed hybrid SVM-based framework can serve as an effective similarity-based screening component in academic plagiarism detection systems, particularly for identifying plagiarism-related content. The model is well suited to support academic integrity processes by providing consistent and interpretable classification outcomes that can assist educators and institutions in preliminary plagiarism assessment. Although the proposed model demonstrates reliable performance in distinguishing between plagiarism and altered documents, it does not explicitly differentiate between varying levels of obfuscation within plagiarism cases. Future work may extend this study by formulating the detection task as a multi-class classification problem, where different obfuscation levels (*simple*, *medium*, and *hard*) are modeled explicitly. In addition, further research may incorporate larger and more diverse academic corpora, explore more advanced semantic embedding models, or integrate deep learning–based classifiers to improve sensitivity to complex and fine-grained text transformations.

Acknowledgements

The author would like to thank Mr. Derry Alamsyah, S.Si., M.Kom., M.Pd. for his guidance and support throughout the research process. Appreciation is also extended to the PAN organizers and the CLEF initiative for providing the dataset used in this study, as well as to the affiliated institution for its support.

References

[1] A. Yanti, E. Salsabila, N. Khoirrun Nisa, A. Rapindo, S. Anggraini, and E. Tamara, "KARYA ILMIAH SCIENTIFIC WORK," *JIIC: JURNAL INTELEK INSAN CENDIKIA*, no. 10, pp. 6809–6817, Dec. 2024, [Online]. Available: https://jicnusantara.com/index.php/jiic

[2] H. Sanulita *et al.*, *PANDUAN PRAKTIS PENULISAN KARYA TULIS ILMIAH*, 1st ed. Daerah Istimewa Yogyakarta: PT. Green Pustaka Indonesia, 2024. [Online]. Available: www.greenpustaka.com

[3] A. F. Santosa and B. S. L. Barrantes, "Multilingual research dissemination: Current practices and implications for bibliometrics," Feb. 2025.

[4] M. Sozon, O. H. Mohammad Alkharabsheh, P. W. Fong, and S. B. Chuan, "Cheating and plagiarism in higher education institutions (HEIs): A literature review," *F1000Res*, vol. 13, p. 788, Jul. 2024, doi: 10.12688/f1000research.147140.1.

[5] M. Sajid, M. Sanaullah, M. Fuzail, T. S. Malik, and S. M. Shuhidan, "Comparative analysis of text-based plagiarism detection techniques," *PLoS One*, vol. 20, no. 4, Apr. 2025, doi: 10.1371/journal.pone.0319551.

[6] M. A. El-Rashidy, R. G. Mohamed, N. A. El-Fishawy, and M. A. Shouman, "An effective text plagiarism detection system based on feature selection and SVM techniques," *Multimed Tools Appl*, vol. 83, pp. 2609–2646, 2024, doi: 10.1007/s11042-023-15703-4.

[7] T. Vrbanec and A. Mestrovic, "Relevance of Similarity Measures Usage for Paraphrase Detection," in *International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management, IC3K - Proceedings*, Science and Technology Publications, Lda, 2021, pp. 129–138. doi: 10.5220/0010649800003064.

[8] J. P. Wahle, B. Gipp, and T. Ruas, "Paraphrase Types for Generation and Detection," in *Proceedings ofthe 2023 Conference on Empirical Methods in Natural Language Processing*, Association for Computational Linguistics, Oct. 2023, pp. 12148–12164. [Online]. Available: https://github.com/jpwahle/

[9] A. Al saqaabi, C. Stewart, E. Akrida, and A. Cristea, "A Paraphrase Identification Approach in Paragraph length texts," in *Proceedings of the 15th International Conference on Educational Data Mining, EDM 2022*, International Educational Data Mining Society, 2022, pp. 782–788. doi: 10.5281/zenodo.6852990.

[10] P. Abisheka, C. Deisy, and P. Sharmila, "T-SRE: Transformer-based semantic Relation extraction for contextual paraphrased plagiarism detection," *Journal of King Saud University - Computer and Information Sciences*, vol. 36, no. 10, Dec. 2024, doi: 10.1016/j.jksuci.2024.102257.

[11] M. Mehdi, S. Mushtaq, and G. Rabbani Butt, "A Hybrid TF–IDF and SBERT Approach for Enhanced Text Classification Performance," *Preprints.org*, pp. 1–7, Oct. 2025, doi: 10.20944/preprints202510.2427.v1.

[12] O. Tulak Bamba, Nur Vadila, Sri Fitrawati, V. W. Tedang, and Asrawati, "Naive Bayes dan Decision Tree: Studi Kasus Klasifikasi Kepuasan Pelanggan E-Commerce," *SIMKOM*, vol. 10, no. 2, pp. 254–262, Jul. 2025, doi: 10.51717/simkom.v10i2.897.

[13] Ratnawati, F., Siswanto, A., Effendy, A., & Tedyyana, A. (2023). Optimizing Pigeon-Inspired Algorithm to Enhance Intrusion Detection System Performance Internet of Things Environments. *JOIV: International Journal on Informatics Visualization*, *7*(4), 2215-2222.

[14] Tedyyana, A., Ratnawati, F., Syam, E., & Putra, F. P. (2022). Threat modeling in application security planning citizen service complaints. *Indonesian Journal of Electrical Engineering and Computer Science*, *28*(2), 1020.

[15] K. Lieberman, S. Yuan, S. K. Ravindran, and C. Tomasi, "Optimizing for ROC Curves on Class-Imbalanced Data by Training over a Family of Loss Functions," Jun. 2024, [Online]. Available: http://arxiv.org/abs/2402.05400