Journal Homepage: -www.journalijar.com

INTERNATIONAL JOURNAL OF ADVANCED RESEARCH (IJAR)

Article DOI:10.21474/IJAR01/24586
DOI URL: http://dx.doi.org/10.21474/IJAR01/24586

ISSN (O): 2320-5407
ISSN (P): 3107-4928

RESEARCH ARTICLE

INTELLIGENT SYSTEM FOR DETECTING PLAGIARISM BY OBFUSCATION IN ACADEMIC WORKS (ISDP)

Ousseini Madjiri Moctar[1,2], Gandou Abouba[1,2], Pr Mahamadou Issoufou Tiado[1,2] and Boukar Abatchia Nicolas[1,2]

1. Faculty of sciences and technology, University of Abdou Moumouni, Niamey, Niger.
2. Departement of Mathematics and Computer Science, Network and Telecommunications Research Team.

Manuscript Info

Manuscript History
Received: 12 September 2026
Final Accepted: 14 October 2026
Published: November 2026

Key words:-
Plagiarism detection,
DOUNG, contrastive fine-tuning,
Sentence-BERT, Obfuscation.

Abstract

Distance learning, particularly through the new generations of Open Digital Universities (DOUNG) (Abdoul aziz et al. 2022), faces the same challenges in terms of the integrity of university documents (articles, dissertations, etc.). Traditional methods, based on textual comparison, are unable to detect intelligent reformulations. This paper evaluates the contribution of a domain adaptation of a Sentence-BERT (Bidirectional Encoder Representations from Transformers) model, by contrastive fine-tuning on plagiarized source/passage pairs from the PAN (Plagiarism Analysis, Authorship Identification, and Near-Duplicate Detection) corpus, within a hybrid pipeline combining Winnowing fingerprints and semantic similarity. Through a controlled ablation study, we isolate the contribution of three levers, fine-tuning, similarity threshold and segment fusion, and show that enriched fine-tuning (multi-level and hard negative pairs) improves, with identical configuration and test set, the recall on paraphrases, and almost doubles the overall accuracy. We propose a hybrid approach combining the Winnowing algorithm for candidate source selection with a contrastivel y fine tuned SBERT model dedicated to semantic similarity assessment .this approach is based on a deep semantic analysis rather than a simple textual match, thus allowing us to respond effectively to sophisticated forms of plagiarism. Domain adaptation by contrastive fine-tuning of Sentence-BERT

"© 2026 by the Author(s). Published by IJAR under CC BY 4.0. Unrestricted use allowed with credit to the author."

Introduction:-

Plagiarism can be defined as the use of an original work without mentioning its author or as the act of attributing another person's work to oneself. This phenomenon, amplified by the high digitization of content and the easier access to online resources, threatens scientific integrity and compromises the fundamental values of learning. Therefore, the implementation of a plagiarism detection system appears to be a necessity to preserve the originality of students' work and to maintain the credibility of the diplomas awarded. Lexical fingerprint-based methods, such as Winnowing (Schleimer et al. 2003), effectively detect literal copying and light obfuscation, but fail on strong

Corresponding Author:-Ousseini Madjiri Moctar
Address:- 1. Faculty of sciences and technology, University of Abdou Moumouni, Niamey, Niger.2. Departement of Mathematics and Computer Science, Network and Telecommunications Research Team.
Email:-madjirimoctar93@gmail.com

paraphrasing, where vocabulary and word order are profoundly reworked. Semantic representations from sentence models such as Sentence-BERT (Reimers et al. 2019). offer a complementary lead: they bring together two formulations with similar meanings regardless of their lexical surface. However, a generic model is not specialized in the characteristics of plagiarized texts. We study the hypothesis that domain adaptation by contrastive fine-tuning, trained on pairs of passages linked by plagiarism, improves the detection of obfuscated cases. Our contribution is twofold: (i) a rigorous experimental evaluation, by evaluating, by means of comparative experiments, the contribution of fine-tuning as well as its contribution on performance according to the different parameters of the processing chain; (ii) an analysis of the factors influencing the ability of the system to detect truly plagiarized passages, which can be observed thanks to an evaluation metric (the recall), resulting in a prioritized progression roadmap. The measurements relate to the PAN-PC-11 corpus, which is a corpus of data developed and used as part of the PAN 2011 scientific campaign, organized during the CLEF (Conference and Labs of the Evaluation Forum) conference, and the PlagDet metric (Plagiarism Detection Score), which is a metric that measures the overall performance of an automatic plagiarism detection system.

The remainder of this paper is organized as follows. After reviewing the existing approaches, we present the adopted methodology and the experimental results obtained, before concluding with the perspectives arising from this work.

Related works: -

The authors of (Amirzhanovetal. 2025) and (Sajid et al. 2025) present two studies on the evolution of plagiarism detection methods, based respectively on the analysis of 41 and 189 scientific publications. They examine lexical, syntactic, semantic and machine learning and deep learning approaches. Their analyses highlight the limitations of traditional methods in the face of reformulated or translated texts and highlight the interest of hybrid approaches to improve plagiarism detection.

In (P.Abishekaetal. 2024), the authors propose T-SRE (*Transformer-based Semantic Relation Extraction*), a method based on Transformers to detect plagiarism by reformulation. It combines syntactic analysis, named entity recognition and contextual semantic representations. Evaluated on the Udacity dataset, it obtained an accuracy of 92%, a recall of 89%, an F1-score of 90.5% and a granularity of 1.28, showing the interest of combining syntactic and semantic information.

In (Saeed and Taqa, 2023), authors develop a system combining lexical (Jaccard and TF-IDF) and semantic (Doc2Vec and SBERT) methods. The source documents constituting the corpus of data are automatically retrieved from Google, Bing and Yahoo search engines. The experiments carried out on ten (10) suspect documents show better performance of lexical methods than of semantic methods in the context studied. This work nevertheless underlines SBERT's interest in the analysis of semantic similarity.

In (Wang et al. 2014), the authors propose a method combining the vector model (SVM) and the Jaccard coefficient to improve the detection of plagiarism on the PAN corpus evaluated in the framework of the CLEF. The results show that this combination improves performance compared to the individual use of these methods, particularly in terms of recall, precision and granularity. This approach highlights the interest of combining several similarity measures to improve the detection of plagiarized passages.

Yalcin et al. (2022) propose an external plagiarism detection system that combines syntactic and semantic features. Their approach primarily relies on Part-of-Speech (POS) tag n-grams, which are used to index sentences from source documents and retrieve candidate passages. Semantic similarity is subsequently assessed using Word2Vec and the Longest Common Subsequence (LCS) algorithm. Evaluated on the PAN-PC-11 corpus, the proposed method achieves PlagDet scores of 0.89 for Artificial None, 0.80 for Artificial Low, 0.54 for Artificial High, and 0.45 for Simulated plagiarism. Regarding recall, the system obtains 0.95, 0.83, 0.43, and 0.31, respectively, while the corresponding precision scores are 0.85, 0.89, 0.85, and 0.90. These results highlight the benefits of combining syntactic and semantic analysis, particularly for identifying passages that have undergone substantial linguistic transformations. However, the decrease in recall from 0.95 for Artificial None to 0.43 for Artificial High and 0.31 for Simulated plagiarism illustrates the persistent challenges associated with detecting heavily obfuscated and manually paraphrased passages.

In (Kaur et al. 2021), the authors propose an integrated plagiarism detection approach based on analyzing semantic relationships between source and suspicious passages. Their method constructs a dynamic relationship matrix by exploiting the Weighted Inverse Distance and GlossDice measures, as well as lexical relations derived from

WordNet. The authors combine several semantic features, including Weighted Inverse Distance and SynJaccard, to improve the discrimination of similar passages. Evaluated on the PAN-PC-11 corpus, their approach achieves a precision of 0.9459, a recall of 0.8861, and a PlagDet score of 0.8857. These results highlight the potential of hybrid approaches that exploit lexical and semantic relationships to identify extensively paraphrased passages.

In (Arabi etal. 2022), the authors propose two hybrid external plagiarism detection methods evaluated on the PAN-PC-11 corpus, achieving precision scores of 95.1% and 93.8%, respectively. Both approaches aim to improve the identification of texts that have undergone lexical and structural transformations. They incorporate an initial candidate selection phase based on the Bag-of-Words (BoW) model to reduce the search space by filtering documents and sentences. The first method combines FastText and TF-IDF to derive structural and semantic representations, whereas the second exploits WordNet and TF-IDF. Similarity scores are subsequently computed using the Dice coefficient and a weighted structural similarity measure. The results obtained by these two approaches highlight the benefits of combining lexical and semantic features with a candidate selection mechanism to improve plagiarism detection efficiency.

Materials and Methods: -

In this study, we propose a hybrid approach to plagiarism detection based on a combination of the Winnowing algorithm and a fine-tuned SBERT model. This approach utilizes a pipeline architecture consisting of interconnected, complementary modules that jointly leverage both the lexical and semantic characteristics of texts. It aims to improve plagiarism detection performance, addressing both direct textual similarities and more complex forms of reformulation and semantic paraphrasing.

Plagiarism Detection Methods:-

In this part we will discuss the different methods of plagiarism detection. Similarity between sentences occupies a very important place in Natural Language Processing (NLP) applications. We can mention two types of similarity between sentences: lexical similarity and semantic similarity.

Lexical similarity: Lexical similarity consists of comparing each word or each phase character by character. The expected result is within the range of [0,1] or can be expressed as a percentage.

Semantic similarity: Semantic similarity consists of comparing sentences and detecting whether they have the same meaning, i.e. evaluating the (semantic) proximity of their meaning. In (Raju et al. 2022), the authors present three different types of semantic similarities. They are essentially based on corpus, knowledge and strings.

Existing similarity measure: There are several equations that can be used to measure the similarity between sentences. We will mention a few:

- ✓ **Jaccard similarity** given by the formula below:

Jaccard's similarity measures the proportion of common words in relation to the set of distinct words in two sentences. For instance, for the two preprocessed sentences (Tokenization):

A = {Student, submits, his, dissertation, for, a, plagiarism, check}

B = {Student, submits, his, document, for, plagiarism, check}

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|}$$

With: $A \cap B = 6$ (The set of words common to the two sentences).$A \cap B$

$A \cup B = 9$ (The set of all the distinct words of the two sentences)

$$J(A,B) = \frac{6}{9} = 0{,}6667$$

66,67 % similarity

Dice coefficient: The Dice coefficient is identical to the Jaccard similarity, but it gives double weight to the intersection of the two sets.

$$\text{Dice(A, B)} = \frac{2 * |\text{ A} \cap \text{B }|}{(|A| + |B|)} = \frac{12}{15} = 0{,}8$$

80 % similarity

- ✓ **Cosine similarity:** Unlike the two previous methods, cosine similarity compares two numerical vectors by measuring the angle between them.

Its formula is: $$Sim = \frac{(A.B)}{(||A|| \times ||B||)}$$

Suppose that A and B are represented by the two vectors: A= (1,1,1,1,0,1,1,1,1) and B = (1,1,1,0,1,1,0,1,1)

$A.B = 6$; $||A|| = \sqrt{8}$ And $||B|| = \sqrt{7}$

So,

$$Sim = \frac{6}{\sqrt{56}} = 0{,}8018$$

80, 18 %Similarity

Transformer architecture

In (Vaswani et al. 2017), authors proposed the transform architecture. The transformer processes all the elements of the sequence simultaneously thanks to the attention mechanism.

For example, suppose the sentence: **The student submits his dissertation for plagiarism checking.**

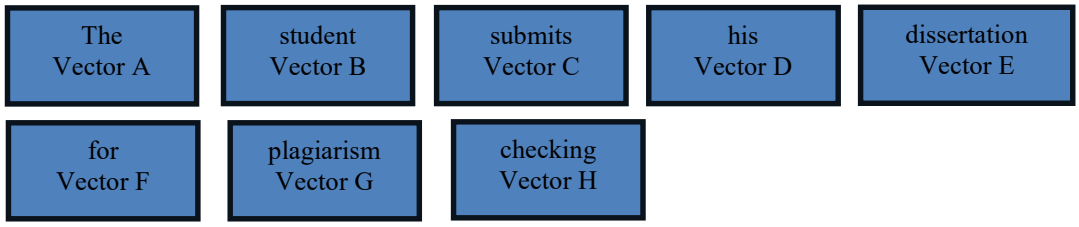
- ✓ **Step 1: Tokenization**

Thanks to Tokenization, this sentence will be cut into Tokens (words).

The - student – submits – his/her – dissertation – for – plagiarism – checking

- ✓ **Step 2: Initial Vectorization**

Each token receives a digital representation



- ✓ **Step 3: Parallel Calculations in the Transformer Layer**

From the initial representations, the Transformer performs all the necessary calculations for each token simultaneously. The self-attention mechanism comes into play at this stage. The self-attention mechanism allows each word to consider the other words in the sentence to better understand its context.

For example, to interpret the word "dissertation", the Transformer can consider:

- "**Student":** indicates that the person concerned is a student.
- "**Submits**": indicates that the work is being submitted or uploaded.

- "**his**": establishes a possessive relationship with the dissertation.
- "**plagiarism checking**": provides the context in which the dissertation is being used.

This will make it possible to represent this word taking into account the context in which it appears. Unlike recurrent networks (RNN, LSTM, GRU) which process sequences sequentially.

Taking the same sentence, here is the fundamental difference:

1. Treats the "**student**"
2. Milks "**submits**" taking into account the previous state
3. Treats "**his**"
4. Treatise "**dissertation**"
5. Continue until "**checking**"

Each step depends on the state conveyed by the previous step.

Architecture and Proposed Algorithm:-

Hybrid Discovery Pipeline Architecture:-

The proposed architecture is based on a hybrid approach consisting of a set of interconnected modules that work in the form of a pipeline for better plagiarism detection, including forms of reformulation and semantic paraphrases. This architecture is composed of:

Document acquisition module

This module receives the documents to be compared (source document and suspect document).Let be a corpus of N source documents. N = | D|. Each document is seen as an ordered sequence of sentences.

$$D = \{d_1, \ldots . d_N\}$$

Let us consider a suspicious document whose sentences are segmented into m sentences. d_S

$$\boldsymbol{d_S = (S_1, \ldots . S_m)} \quad \textbf{(1)}$$

Each source document for (1 ≤ i ≤ N) is segmented into sentences:$d_i n_i$

$$\boldsymbol{d_i = (t_{i,1}, \ldots \ldots t_{i,n_i})} \quad \textbf{(2)}$$

We gather all the source sentences in a single set T:

$$\boldsymbol{\mathcal{T} = \{t_{i,j} : 1 \leq i \leq N, 1 \leq J \leq n_i\}} \quad \textbf{(3)}$$

Word preprocessing module

After receiving the documents, a pre-processing follows through the Spacy tool by applying the following operations: cleaning, and segmentation by sentence.

Candidate Source Selection Module

Winnowing identifies candidate sources for each suspect document that share at least a minimum number of k-gram fingerprints and delineates areas that may contain plagiarism.

(Common seeds). For the source document d_i and the suspect document, we consider the set of seeds as being the set of pairs of positions where the same imprint appears in both documents:d_s

$$\boldsymbol{g_i = \{(p, q) : \exists h, h \in H(d_s) \text{ à } la\ position\ p, h \in H(d_i) \text{ à } la\ position\ q\}}. \quad \textbf{(4)}$$

(Candidate area). From the common seeds, we propose a spatial grouping rule to construct candidate areas. We then define the coverage of a sentence by a zone by means of the overlap of their intervals.

A zone for source i is a pair of intervals

$$\boldsymbol{z = (I_z^s, I_z^{src})}, \textbf{ with and } \boldsymbol{I_z^s \subseteq [0, |d_s|) I_z^{src} \subseteq [0, |d_i|)}, \text{ such that the region} \quad \textbf{(5)}$$

concentrates at least $\boldsymbol{(I_z^s, I_z^{src})}$**i** spatially contiguous seeds (deviation bounded by a margin). We note: $\boldsymbol{g_i}$

(6)

$$Z(d_s) = \bigcup_{i=1}^{N} \{z : z \text{ is an area of the source } i\}$$

The set of all the fields of the suspect document. A source i is a candidate if it has at least one zone (Schleimer et al. 2003).

(Coverage of a suspicious sentence). A suspicious sentence is covered by a z-zone $S_k$if its interval overlaps with the suspicious region of *z:*

$$couvre\ (z, S_k) \Leftrightarrow [a_k, b_k) \cap I_z^s \neq \emptyset. \quad (7)$$

Winnowing locates the alleged plagiarism as zones, both in the suspect document and in the source. The semantics are then calculated only for the suspect sentences covered by an area, which severely restricts the space for comparison.

Semantic Threshold Verification

(Pairing and detection in an area). So be it S_k a sentence covered by a zone z de source i = σ(z). We retain the most similar source sentence and its score:

$$j_{k,i}^* = \arg\max sim\ (S_k, t_{i,j}), \qquad .\rho_{k,i} = sim\ (S_k, t_{i,j_{k,i}^*}) \quad (8)$$

The couple (S_k, i) produces elementary detection if $\rho_{k,i} \geq$ T , where T is the similarity threshold. The set oı elementary detections is:

$$\Delta = \{(k, i, , [a_k, b_k)\rho_{k,i}) : \exists z \in Z(d_s), couvre\ (z, S_k), \sigma(z) = i, \rho_{k,i} \geq T\}. \quad (9)$$

Only sentences covered by a zone are submitted to the semantic judge. The decision is based solely on the cosine similarity and the threshold T.

Semantic representation module

Sentences in candidate areas are transformed into embeddings using the SBERT model.

(Semantic embedding): a function φ: X $\rightarrow R^d$ associates to any sentence x a normalized vector φ(x),$\|\phi(x)\|_2$ = 1.

Similarity Calculation Module

The vectors of the suspect sentences are compared to those of the candidate source sentences by the cosine similarity, to identify the closest matches. The similarity of two sentences is the cosine:

$$\mathbf{sim}(x, y) = \langle \phi(x), \phi(y) \rangle\ \epsilon[-1, 1]$$

The cosine similarity is a dot product because we have to make one of the normalized vectors. φ is the Sentence-Bert model. (Reimers et al. 2019).

Decision Module

At this level, we use an *S threshold*to determine whether a sentence is plagiarized or not.
We set the threshold *S* = 0.75.

Yes:

$$\mathbf{sim}(x, y) > 0.75$$

Then the sentence is considered plagiarized.x

Blending and Filtering Module

Neighboring detections pointing to the same source are repasted into coherent passages (fusion), reducing fragmentation. Anti-isolated filtering then discards short and isolated segments with a moderate score, which typically constitute false alarms.

Results generation module

The detected passages are formalized in localized annotations, specifying for each suspicious passage the corresponding source and their positions, allowing the evaluation of the detection.

This architecture allows us to better maintain our system and add new methods or modules. The base model is all-mpnet-base-v2, derived from the MPNet architecture.

Domain adaptation by contrastive fine-tuning

Fine-tuning does not modify the architecture: it moves the representations in the vector space in such a way as to bring the pairs closer together (source, plagiarized passage) and to distance the unrelated pairs. Our enhanced modelproduces pairs with two levels granularities (integer passage and aligned sentences) and adds to each anchor a hard negative, a source sentence thematically close but not related, obtained by searching in the source space. The training is based on 7,370 pairs of sentences and uses the Multiple Negatives Ranking Loss (Henderson and al. 2017), which exploits both the explicit and implicit negatives of the batch.

Table 1 presents the key hyperparameters selected for the fine-tuning phase of the pre-trained Sentence-BERT model (all-mpnet-base-v2), as part of its adaptation for the task of semantic plagiarism detection. Training was conducted in the Google Colab environment, using an NVIDIA Tesla T4 GPU. Given the GPU memory constraints of this hardware architecture, the batch size was set to 8 examples per training step to limit memory consumption and ensure the training process proceeded smoothly.

Table 1: Hyperparameter Configuration for SBERT Fine-Tuning

Hyperparameter	Parameter	Value	Description
Pre-trained model	BASE_MODEL	All-mpnet-base-v2	Pre-trained SBERT model used as the starting point for fine-tuning.
Number of epochs	EPOCHS	1	Number of complete passes through the training dataset.
Batch size	BATCH	8	Number of training examples processed simultaneously at each training step.
Learning rate	LR	2e-5	Controls the magnitude of weight updates during the optimization process
Warmup ratio	WARMUP_R	0,1	Proportion of training steps allocated to the gradual increase of the learning rate

Table 2 presents the distribution and configuration of the experimental data, divided into two sets: training and testing. This split, performed using a fixed random seed of 42, distinguishes the data used for fine-tuning SBERT from the data reserved for final evaluation. It should be noted that no separate validation set was created for this experiment.

Furthermore, Table 3 details the process of constructing the training data from the XML annotations of the PAN-PC-11 corpus. Specifically, it outlines the extraction of plagiarized passages, the creation of positive pairs at two levels of granularity (passage and sentence), their distribution across plagiarism types ("Artificial" and "Translation"), and the generation of "hard negatives" used for fine-tuning the pre-trained SBERT model.

Table 2: Experimental dataset distribution and configuration

Parameter	Training	Test
Suspicious documents	160	40
Source documents	204	45
XML annotation files	160	40
Proportion	80%	20%
Random seed	42	42
Validation set	Not used	Not used

Table 3 : Fine-tuning data construction

Parameter	Value
Extracted plagiarized passages	662
Passage-level positive pairs	662

Sentence-level positive pairs	6708
Total positive pairs	7370
Artificial pairs	7120
Translation pairs	250
Source sentence mining pool	6355
Hard negatives identified	7370
Final training triplets	7370

Ablation study: contribution of each lever

Table 4 isolates the cumulative effect of the three levers, added one after another, starting from the baseline configuration, on the "artificial" type.

Table 4 : Cumulative ablation on the artificial type (test using 40 documents)

Artificial				
	Precision	Recall	Granularity	plagdet
Base (all-mpnet, threshold 0,80)	0,9226	0,2894	1,148	0,3994
fine-tuned model (threshold 0.80)	0,9215	0,341	1,153	0,45
Fine-tuned model (threshold 0,75)	0,915	0,361	1,202	0,4546
Fine-tuned model + fusion (gap2500)	0,8921	0,3509	1,082	0,4760

Table 5 isolates the cumulative effect of the three levers, added one after another, starting from the baseline configuration, on the "Global" type.

Table 5 : Cumulative ablation on the global type (test using 40 documents)

Global				
	Precision	Recall	Granularity	plagdet
Base (all-mpnet, threshold 0,80)	0,7515	0,2417	1,118	0,3379
Fine-tuned model (threshold 0,80)	0,7986	0,2722	1,124	0,3737
Fine-tuned model (threshold 0,75)	0,7601	0,2865	1,165	0,3734
Fine-tuned model+ fusion (gap 2500)	0,7389	0,2801	1,067	0,3879

At a constant threshold (0.80), simply replacing the base model with the fine-tuned model increases recall from 0.289 to 0.341 for the "artificial" dataset and from 0.241 to 0.272 for the "global" dataset; this gain is attributable solely to the fine-tuning process. Lowering the threshold from 0.80 to 0.75 yields an additional gain in recall, and the two effects combine to improve detection coverage, despite a slight decrease in precision. Furthermore, increasing the merging parameter (GAP = 2500), while keeping the threshold constant at 0.75, results in improved granularity, which shifts from 1.202 to 1.082 in the "artificial" scenario and from 1.165 to 1.067 in the "global" scenario. This reduction is accompanied by an increase in the PlagDet score, from 0.4546 to 0.4760 and from 0.3734 to 0.3879, respectively, even though both recall and precision decrease slightly. These observations suggest that the merging process contributes primarily to better structuring of the detected passages by limiting their fragmentation, rather than directly increasing the overall coverage of detected plagiarism. The configuration combining the fine-tuned SBERT model, a threshold of 0.75, and a GAP value of 2500 offers the best overall compromise among the configurations evaluated.

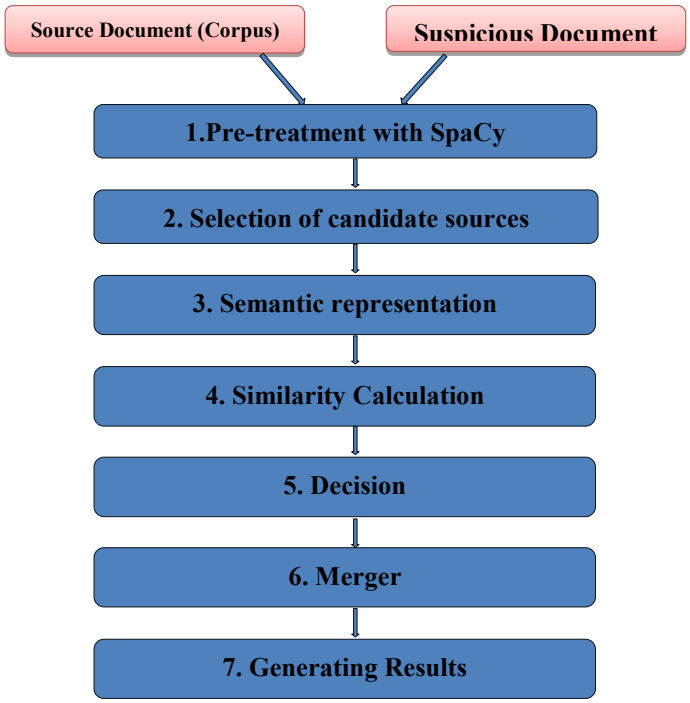


Figure 1 : General architecture of the proposed architecture

Results and Discussion: -

The experiments are based on the PAN Text Alignment corpus. The training partition and the test partitions are strictly disjointed at the level of the suspect document, guaranteeing the absence of leakage: no document used for the evaluation contributed to the fine-tuning. The metrics are those of PAN [1]: precision and recall at the character level, granularity (average number of detections per case), and plagdet = F1 / $\log_2(1 + \text{granularity})$. The results are broken down by type of obfuscation; the analysis focuses on the artificial type (monolingual obfuscation), which is the majority and statistically significant.

- **Precision**

VP: True Positive
FP: False Positive

$$Precision = \frac{VP}{VP + FP}$$

Accuracy reflects our system's ability to detect cases of plagiarism accurately and unambiguously.

- **Recall**

VP: True Positive
FN: False Negative

$$Recall = \frac{VP}{VP + FN}$$

After implementing our algorithm using the Python language, we tested our model on a corpus composed of 11093 source documents and 200 suspect documents. To optimize the computation time for similarity detection, we processed the 11093 source documents using the Spacy word processing tool and then we calculated and stored for each source document and for each sentence its embedding in a folder that we named document_source_embedding. We loaded the folder containing the 200 suspicious files, each file being processed with Spacy and segmented into sentences. Each sentence is transformed into a vector using fine-tuned SBERT for a comparison with the vectors of all the source files (11093 vectorized source files). The central result is established on a test set of 200 documents (987 cases, including 708 artificial and 158 translations), aligned and strictly disjointed from the training. The basic model and the fine-tuned model are compared in a rigorously identical configuration (threshold 0.75); only the model/embeddings pair changes. Any difference is therefore attributable to fine-tuning alone.

After calculating the following metrics — precision and recall — we obtained the following overall averages:

Table 6 : Presentation of the results of our method according to the 4 types of paraphrases present in PAN-PC-11

Sous-ensemble	PlagDet	Precision	Recall	Granularity	#cas	#detection
GLOBAL (all categories)	0,3708	0,6070	0,2939	1,097	987	562
GLOBAL (4 categories)	0,4735	0,9121	0,3499	1,097	829	374
ARTIFICIAL						
NONE	0,8336	0,9548	0,7397	1,000	89	66
LOW	0,5535	0,9040	0,4587	1,143	398	256
HIGH	0,1792	0,8317	0,1031	1,033	221	31
SIMULATED	**0,2703**	**0,9953**	**0,1564**	**1,000**	**121**	**21**

Table 6 presents the results obtained by our approach across the four obfuscation categories considered in the PAN-PC-11 corpus. Compared with the method proposed by Yalcin et al. (2022), our approach achieves higher precision in three out of the four categories, namely Artificial None, Artificial Low, and Simulated, while a slightly lower precision is observed for Artificial High. However, the recall achieved by our approach remains lower than that reported by these authors across all the considered categories. A similar trend is observed when comparing our results with those reported by Kaur et al. (2023), whose overall recall on PAN-PC-11 is higher than that of our approach, although the evaluation protocols are not strictly identical.

Furthermore, the analysis of the results reveals a progressive decrease in recall as the degree of obfuscation increases, from 0.7397 for Artificial None to 0.4587 for Artificial Low and 0.1031 for Artificial High. A similar decline is observed for the Simulated category, for which the recall reaches 0.1564. This degradation highlights the limitations of our system in detecting passages that have undergone substantial lexical and structural transformations. It may partly be attributed to the candidate source selection mechanism based on the Winnowing algorithm. Indeed, although this stage significantly reduces the search space by restricting comparisons to documents sharing common lexical fingerprints, it may also exclude relevant source documents when the passages have undergone extensive paraphrasing. Consequently, such documents are not forwarded to the subsequent semantic similarity comparison performed by the fine-tuned SBERT model. This limitation highlights the trade-off between reducing computational search costs and maintaining detection coverage, particularly in cases of heavily obfuscated plagiarism.

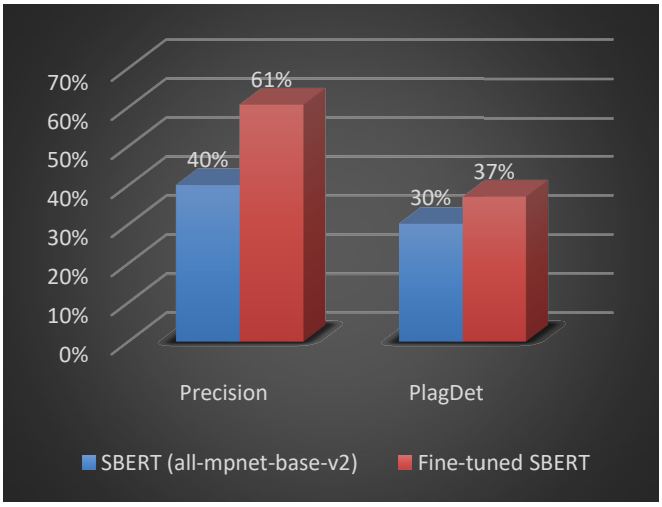


Figure 2 : Comparison between the Base model and the fine-tuned model at the global level

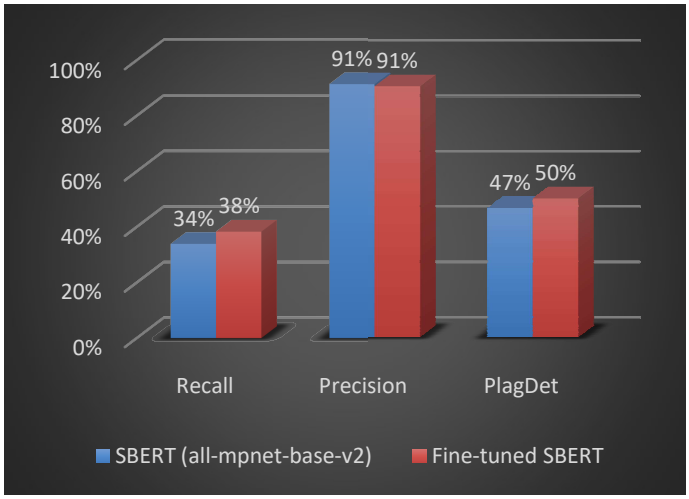


Figure 3 : Comparison between the Base model and the fine-tuned model at the artificial level

Figures 2 and 3 compare the baseline model (all-mpnet-base-v2) with the fine-tuned model. Fine-tuning increases recall for artificial plagiarism across the three categories (None, Low and High), from 0.339 to 0.383, while maintaining nearly unchanged precision. At the overall level, the fine-tuned model produces fewer detections (562 compared with 775), but these detections are more reliable, as evidenced by the increase in precision from 0.401 to 0.607. The overall PlagDet score also improves, rising from 0.302 to 0.371. These results demonstrate that model adaptation improves its ability to identify artificially paraphrased passages while reducing false-positive detections. Consequently, it achieves a better trade-off between recall and precision, despite a reduction in the total number of detected passages.

Conclusion: -

This study highlights the limitations of the baseline SBERT model in the field of plagiarism detection. To address these limitations, we propose a hybrid approach that combines the Winnowing algorithm, used for candidate source selection, with a contrastively fine-tuned SBERT model dedicated to semantic similarity assessment. The experimental results show that fine-tuning improves recall for artificial plagiarism across the three categories (None, Low and High), increasing it from 0.339 to 0.383, while maintaining nearly unchanged precision (0.914 compared with 0.907). At the overall level, the fine-tuned model produces fewer detections (562 compared with 775), but achieves substantially higher precision, increasing from 0.401 to 0.607. The overall PlagDet score also improves from 0.302 to 0.371, indicating a better balance between precision and recall.

Furthermore, the category-based evaluation reveals persistent limitations, particularly in detecting heavily obfuscated plagiarism. Recall decreases from 0.7397 for the *Artificial None* category to 0.4587 for *Artificial Low* and 0.1031 for *Artificial High*, reaching 0.1564 for *Simulated* plagiarism. This limitation may partly result from the lexical candidate selection mechanism employed by the Winnowing algorithm, which may exclude relevant source documents when substantial lexical and structural transformations reduce the overlap between source and suspicious passages. Consequently, such documents do not reach the subsequent semantic similarity comparison stage performed by SBERT.

Future work will therefore focus on improving candidate retrieval mechanisms to reduce the risk of excluding heavily paraphrased documents.

References: -

1. B. T. Abdoul Aziz, M. I. Tiado, N. A. Hassane, M. Harouna., 2022. Simulation Performance Evaluation of the Synchronous Course Transcription System in the DOUNG (Digital Open Universities - New Generation) Ad Hoc Network, Proceedings of the International Conference on Electrical, Computer and Energy Technologies (ICECET 20-22 July 2022) Prague-Czech Republic, INSPECT membership number:22028564, DOI**:** 10.1109/ICECET55527.2022.9873049,https://ieeexplore.ieee.org/document/9873049/
2. Shuai Wang, Haoliang Qi, Leilei Kong, Cuixia Du., 2013: Combining VSM and Jaccard Coefficient for External Plagiarism Detection,Proceedings of the 2013 International Conference on Machine Learning and Cybernetics, Tianjin, July 14-17.
3. Nils Reimers, Iryna Gurevych., 2019. Sentence-BERT: Embeddings of Sentences Using BERT Siamese Networks, Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th Joint International Conference on Natural Language Processing, pages 3982–3992, Hong Kong, China, November 3–7, 2019. 2019 Association for Computational Linguistics.10.18653/v1/D19-1410
4. Schleimer, S., Wilkerson, D. S., Aiken, A., 2003. Selection: local algorithms for document fingerprinting. Proceedings of the ACM SIGMOD.DOI : 10.1145/872757.872770
5. Altynbek Amirzhanov, Cemil Turan and Alfira Makhmutova, Types of Plagiarism and Detection Methods., 2025: A Systematic Study of Algorithms in Text Analysis, Frontiers in Computer Science, vol. 7, article 1504725. DOI: 10.3389/fcomp.2025.1504725
6. Muhammad Sajid, Muhammad Sanaullah, Muhammad Fuzail, Tauqeer Safdar Malik, and Shuhaida Mohamed Shuhidan., 2025. Comparative Analysis of Text-Based Plagiarism Detection Techniques, PLOS ONE, Vol. 20, No. 4, Item E0319551, 2025. DOI: 10.1371/journal.pone.0319551
7. Nora Raju T, Rahana P A, Raichel Moncy, Sreedarsana Ajay, Sindhya K Nambiar., 2022. Sentence Similarity – A State-of-the-Art Approach, International Conference on Computing, Communication, Security and Intelligent Systems (IC3SIS) | 978-1-6654-6883-1/22/$31.00 ©2022 IEEE | DOI: 10.1109/IC3SIS54991.2022.9885721.
8. Pon Abisheka, C. Deisy, and P. Sharmila, T-SRE., 2024: Transformer-Based Semantic Relationship Extraction for Contextual Paraphrased Plagiarism Detection, Journal of King Saud University – Computer and Information Sciences, Vol. 36, No. 10, Article 102257, December 2024. DOI: 10.1016/j.jksuci.2024.102257

9. Ayoub Ali M. Saeed and Alaa Yaseen Taqa., 2023. Using Retrieved Sources for Semantic and Lexical Plagiarism Detection, Iraqi Journal of Science, Vol. 64, No. 6, pp. 3136–3152, 2023. DOI: 10.24996/ijs.2023.64.6.41.
10. Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, Illia Polosukhin., 2017. Attention is All You Need, 31st Neural Information Processing Systems Conference (NIPS 2017), Long Beach, CA, USA.https://doi.org/10.48550/arXiv.1706.03762
11. Henderson, M., Al-Rfou, R., Strope, B., Sung, Y.-H., Lukács, L., Guo, R., Kumar, S., Miklos, B., & Kurzweil, R., 2017. Effective natural language response suggestion for intelligent response. arXiv: 1705.00652. DOI: 10.48550/arXiv.1705.00652
12. Kadir Yalcin *, Ilyas Cicekli , Gonenc Ercan., 2022.An external plagiarism detection system based on part-of-speech (POS) tagn-grams and word embedding,Available online 17 February 2022, 0957-4174/© 2022 Elsevier Ltd. All rights reservedhttps://doi.org/10.1016/j.eswa.2022.116677.
13. Manpreet Kaur, Vishal Gupta & Ravreet Kaur., 2021. Semantic-BasedIntegrated Plagiarism DetectionApproach for English Documents, IETE Journal of Research, DOI:10.1080/03772063.2021.2004383.
14. Hamed Arabi, Mehdi Akbari., 2022. Improving plagiarism detection in text document using hybrid weighted similarity, Available online 8 July 2022, 0957-4174/© 2022 Elsevier Ltd. All rights reserved, https://doi.org/10.1016/j.eswa.2022.11803