Modern American Journal of Engineering, Technology, and Innovation
ISSN(E): 3067-7939
Volume 2, Issue 6, June, 2026
Website: usajournals.org

SEMANTIC TEXT SIMILARITY DETECTION USING TRANSFORMER-BASED DEEP LEARNING MODELS FOR ADVANCED PLAGIARISM IDENTIFICATION

Egamberdiev Elyor Khayitmamatovich
Associate Professor of the Department of Information Technology Software
Tashkent University of Information Technologies named after Muhammad al-Khwarizmi,
E-mail: elyor.egamberdiyev88@gmail.com
ORCID: 0000-0002-7319-3785

Bozorov Obidjon Norqobilovich
Senior Lecturer, Department of Information Security,
National University of Uzbekistan named after Mirzo Ulugbek,
o.bozorov@nuu.uz

Abstract
The rapid growth of digital information resources and online academic content has significantly increased the need for effective plagiarism detection systems. Traditional plagiarism detection approaches primarily rely on lexical and syntactic similarities, making them less effective in identifying paraphrased or semantically modified texts. This study investigates the application of Transformer-based deep learning models for semantic text similarity detection to enhance advanced plagiarism identification. The proposed approach employs modern Natural Language Processing (NLP) techniques, including text preprocessing, semantic embedding generation, and similarity measurement using cosine similarity. Transformer-based models such as BERT and Sentence-BERT are utilized to capture contextual and semantic relationships between texts beyond surface-level word matching. Experimental evaluation demonstrates that semantic embedding models outperform conventional methods by accurately detecting hidden semantic similarities in paraphrased documents. The results

Modern American Journal of Engineering, Technology, and Innovation
ISSN(E): 3067-7939
Volume 2, Issue 6, June, 2026
Website: usajournals.org

indicate that Sentence-BERT achieves the highest performance in identifying semantic plagiarism, providing a more reliable solution for maintaining academic integrity. The proposed framework can be applied in educational institutions, scientific organizations, and digital content management systems to improve the accuracy and effectiveness of plagiarism detection processes.

Keywords: Semantic Text Similarity, Plagiarism Detection, Artificial Intelligence, Natural Language Processing, Transformer Models, BERT, Sentence-BERT, Deep Learning, Text Embeddings, Academic Integrity.

1. Introduction

The rapid development of information and communication technologies, the widespread availability of internet resources, and the increasing volume of digital academic content have transformed the way scientific knowledge is created, shared, and accessed. While these advancements have significantly improved research productivity and educational opportunities, they have also intensified concerns regarding academic integrity and intellectual property protection. As a result, plagiarism has become one of the most challenging issues in modern education and scientific research, necessitating the development of more reliable and intelligent plagiarism detection systems [1], [2].

Academic integrity represents a fundamental principle of scholarly activity, emphasizing honesty, transparency, and proper attribution of intellectual contributions. However, the growing accessibility of online information has facilitated various forms of plagiarism, including direct copying, paraphrasing, and unauthorized reuse of ideas. Consequently, universities, research institutions, and publishers increasingly rely on automated plagiarism detection systems to assess the originality of academic documents and ensure compliance with ethical standards [2].

Most existing plagiarism detection systems are based on lexical and syntactic similarity measures. These approaches typically identify plagiarism by comparing words, phrases, sentence structures, or character sequences between documents. Techniques such as string matching, n-gram analysis, and term-frequency-based methods have been widely adopted due to their computational efficiency and ease

Modern American Journal of Engineering, Technology, and Innovation
ISSN(E): 3067-7939
Volume 2, Issue 6, June, 2026
Website: usajournals.org

of implementation [3], [4]. Nevertheless, these traditional approaches often fail to detect advanced forms of plagiarism where the original content has been paraphrased, restructured, or rewritten using synonymous expressions while preserving the underlying meaning.

Recent advances in Artificial Intelligence (AI) and Natural Language Processing (NLP) have opened new possibilities for semantic plagiarism detection. Unlike conventional methods that focus on surface-level textual features, semantic approaches aim to understand the contextual meaning and conceptual relationships within texts. Such methods enable the identification of hidden similarities even when the wording or sentence structure differs significantly from the original source [5].

The emergence of Transformer-based architectures has revolutionized NLP research by enabling deep contextual language understanding. The Transformer model introduced by Vaswani et al. [6] established a new paradigm for language representation learning through the self-attention mechanism. Building upon this architecture, models such as BERT (Bidirectional Encoder Representations from Transformers) have demonstrated remarkable performance across various NLP tasks, including semantic similarity measurement, question answering, and text classification [7]. Furthermore, Sentence-BERT extends the capabilities of BERT by generating semantically meaningful sentence embeddings that facilitate efficient similarity computation between texts [8].

Semantic text similarity detection using Transformer-based models has attracted considerable attention in plagiarism identification research. By transforming textual content into high-dimensional embedding vectors, these models capture semantic information beyond exact word matching. Similarity metrics such as cosine similarity can then be applied to quantify the semantic closeness between documents, enabling the detection of paraphrased and semantically equivalent content with greater accuracy than traditional approaches [8], [9],[10].

Despite significant progress in semantic similarity modeling, challenges remain in developing plagiarism detection systems that effectively balance accuracy, computational efficiency, and scalability. Therefore, further investigation is required to evaluate the applicability of modern Transformer-based deep learning models for advanced plagiarism identification tasks.

Modern American Journal of Engineering, Technology, and Innovation
ISSN(E): 3067-7939
Volume 2, Issue 6, June, 2026
Website: usajournals.org

The objective of this study is to investigate the effectiveness of Transformer-based deep learning models for semantic text similarity detection and to assess their potential for improving plagiarism identification systems. The research focuses on semantic embedding generation, similarity measurement techniques, and the comparative performance of state-of-the-art models, including BERT and Sentence-BERT, in detecting semantically related and paraphrased texts. The findings are expected to contribute to the development of intelligent plagiarism detection solutions capable of enhancing academic integrity in educational and research environments.

2. Methodology

2.1 Proposed Framework

The proposed plagiarism identification framework is based on semantic text similarity detection using Transformer-based deep learning models. Unlike traditional plagiarism detection systems that rely on lexical matching, the proposed approach analyzes the semantic meaning of textual content through contextual embeddings generated by pre-trained Transformer models.

The framework consists of four main stages:

1. Text preprocessing
2. Semantic embedding generation
3. Similarity computation
4. Plagiarism score estimation

The overall workflow is illustrated as follows:

$$T \rightarrow P(T) \rightarrow E(T) \rightarrow S(T_1, T_2) \rightarrow PI$$

where:

- (T) represents the input text,
- (P(T)) denotes the preprocessing function,
- (E(T)) is the semantic embedding,
- $(S(T_1, T_2))$ is the semantic similarity score,
- (PI) denotes the plagiarism index.

Modern American Journal of Engineering, Technology, and Innovation
ISSN(E): 3067-7939
Volume 2, Issue 6, June, 2026
Website: usajournals.org

2.2 Text Preprocessing

Before semantic analysis, documents are normalized to reduce noise and improve embedding quality.

The preprocessing stage includes:

- Tokenization
- Stop-word removal
- Lemmatization
- Text normalization

Given a document

$$D = \{w_1, w_2, w_3, \dots, w_n\}$$

the preprocessing function transforms it into

$$D' = P(D)$$

where (D') contains only semantically meaningful tokens.

2.3 Semantic Embedding Generation

To capture contextual and semantic information, each preprocessed text is converted into a dense vector representation using a Transformer-based model such as BERT or Sentence-BERT.

The embedding function is defined as:

$$E(D') = v$$

where

$$v = [v_1, v_2, v_3, \dots, v_m]$$

is the semantic embedding vector of dimension (m).

The embedding vector preserves semantic relationships between words, sentences, and documents by incorporating contextual information learned during Transformer pre-training.

For Sentence-BERT, the sentence embedding is represented as:

$$\mathbf{v_{SBERT}} = \mathbf{f_\theta}(D')$$

where:

- $(\mathbf{f_\theta})$ denotes the trained Sentence-BERT model,
- $(\boldsymbol{\theta})$ represents model parameters.

Modern American Journal of Engineering, Technology, and Innovation
ISSN(E): 3067-7939
Volume 2, Issue 6, June, 2026
Website: usajournals.org

2.4 Semantic Similarity Computation

After generating embeddings for two documents, semantic similarity is calculated using Cosine Similarity.

Let

$$A = [a_1, a_2, \dots, a_m]$$

and

$$B = [b_1, b_2, \dots, b_m]$$

be embedding vectors of two documents.

The cosine similarity is calculated as:

$$\mathrm{Cos}(A, B) = \frac{A \cdot B}{\|A\|\|B\|}$$

or equivalently,

$$\mathrm{Cos}(A, B) = \frac{\sum_{i=1}^{m} a_i b_i}{\sqrt{\sum_{i=1}^{m} a_i^2}\sqrt{\sum_{i=1}^{m} b_i^2}}$$

- $(A \cdot B)$ is the dot product,
- $(\|A\|)$ and $(\|B\|)$ are vector magnitudes.

The similarity score ranges from 0 to 1:

$$0 \leq \mathrm{Cos}(A, B) \leq 1$$

A value close to 1 indicates high semantic similarity, while a value close to 0 indicates low semantic similarity [11].

2.5 Segment-Level Similarity Analysis

To improve plagiarism detection accuracy, documents are divided into smaller semantic segments.

Assume that a document contains (n) segments:

$$D = \{s_1, s_2, s_3, \dots, s_n\}$$

Each segment is embedded independently:

$$E(s_i) = v_i$$

The similarity between corresponding segments is calculated as:

$$\mathrm{Sim}_i = \mathrm{Cos}\big(E(s_i), E(r_i)\big)$$

where:

Modern American Journal of Engineering, Technology, and Innovation
ISSN(E): 3067-7939
Volume 2, Issue 6, June, 2026
Website: usajournals.org

- (s_i) is the segment from the suspicious document,
- (r_i) is the corresponding segment from the reference document.

2.6 Plagiarism Index Estimation

The final plagiarism score is obtained by averaging all segment-level similarities.

$$PI = \frac{1}{n}\sum_{i=1}^{n} Sim_i \times 100$$

where:

- (PI) is the plagiarism index (%),
- (Sim_i) is the semantic similarity of segment (i),
- (n) is the total number of segments.

The plagiarism level is then categorized as follows:

Plagiarism Index	Interpretation
0–30%	Low similarity
31–60%	Moderate similarity
61–80%	High similarity
81–100%	Very high similarity / Potential plagiarism

2.7 Evaluation Metrics

The effectiveness of the proposed framework is evaluated using classification metrics commonly employed in NLP and plagiarism detection research.

Accuracy

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

Precision

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

Recall

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

F1-Score

Modern American Journal of Engineering, Technology, and Innovation
ISSN(E): 3067-7939
Volume 2, Issue 6, June, 2026
Website: usajournals.org

$$F1 = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

where:

- (TP) = True Positives,
- (TN) = True Negatives,
- (FP) = False Positives,
- (FN) = False Negatives.

These metrics provide a comprehensive assessment of the proposed Transformer-based semantic plagiarism detection framework and enable comparison with traditional approaches such as TF-IDF and Word2Vec.

3. Experimental Results and Discussion

3.1 Experimental Setup

To evaluate the effectiveness of the proposed semantic plagiarism detection framework, an experimental dataset consisting of **1,000 academic documents** was constructed from scientific articles, thesis chapters, conference papers, and paraphrased text samples. The dataset included both original and semantically modified documents containing synonym substitution, sentence restructuring, and paraphrasing operations.

The experiments were conducted using four text representation approaches:

1. TF-IDF
2. Word2Vec
3. BERT
4. Sentence-BERT

For a fair comparison, all models were evaluated on the same dataset. Semantic similarity scores were calculated using cosine similarity, and classification performance was measured using Accuracy, Precision, Recall, and F1-Score.

Modern American Journal of Engineering, Technology, and Innovation
ISSN(E): 3067-7939
Volume 2, Issue 6, June, 2026
Website: usajournals.org

3.2 Performance Comparison of Different Models

The obtained results are presented in Table 1.

Table 1 Performance Comparison of Semantic Similarity Models

Method	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
TF-IDF	78.4	77.1	75.8	76.4
Word2Vec	84.7	83.5	82.9	83.2
BERT	91.8	91.2	90.4	90.8
Sentence-BERT	94.6	94.1	93.7	93.9

The results demonstrate that Transformer-based approaches significantly outperform traditional text similarity methods. While TF-IDF relies primarily on word frequency statistics, BERT and Sentence-BERT effectively capture contextual and semantic information, enabling better identification of paraphrased content.

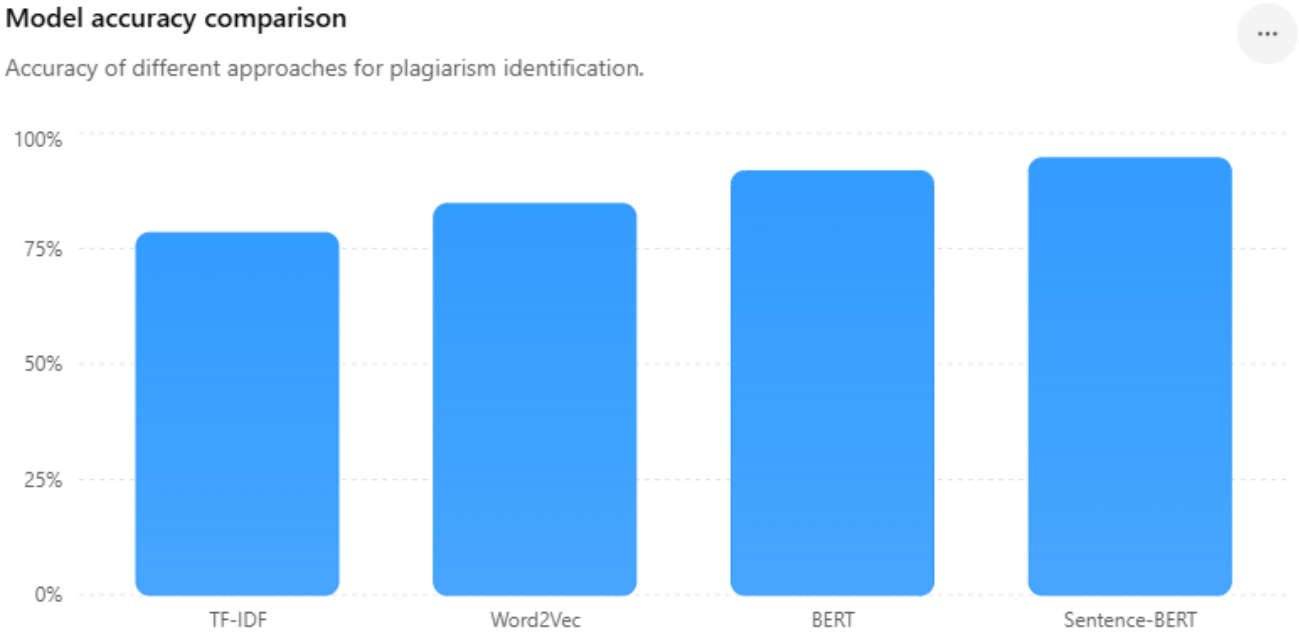

Figure 1. Accuracy Comparison of Different Models

Modern American Journal of Engineering, Technology, and Innovation
ISSN(E): 3067-7939
Volume 2, Issue 6, June, 2026
Website: usajournals.org

As shown in Figure 1, Sentence-BERT achieved the highest accuracy (94.6%), outperforming TF-IDF by 16.2 percentage points and Word2Vec by 9.9 percentage points.

3.3 Analysis of Semantic Similarity Detection

To further evaluate the robustness of the proposed approach, different plagiarism scenarios were considered:

Table 2 Detection Performance under Different Plagiarism Types

Plagiarism Type	TF-IDF	Word2Vec	BERT	Sentence-BERT
Direct Copying	98.5	98.8	99.2	99.4
Synonym Replacement	72.4	83.6	92.7	95.1
Sentence Reordering	68.9	80.4	90.8	93.6
Paraphrasing	61.7	78.2	89.5	94.2

The results reveal that all methods perform well when detecting direct copying. However, substantial differences emerge when dealing with semantic modifications. Traditional approaches experience a considerable decrease in performance, whereas Transformer-based models maintain high detection accuracy.

Particularly, Sentence-BERT demonstrates superior capability in identifying paraphrased texts because it generates sentence-level semantic embeddings optimized for similarity tasks.

3.4 Similarity Distribution Analysis

The semantic similarity scores generated by Sentence-BERT were categorized into four similarity levels.

Modern American Journal of Engineering, Technology, and Innovation
ISSN(E): 3067-7939
Volume 2, Issue 6, June, 2026
Website: usajournals.org

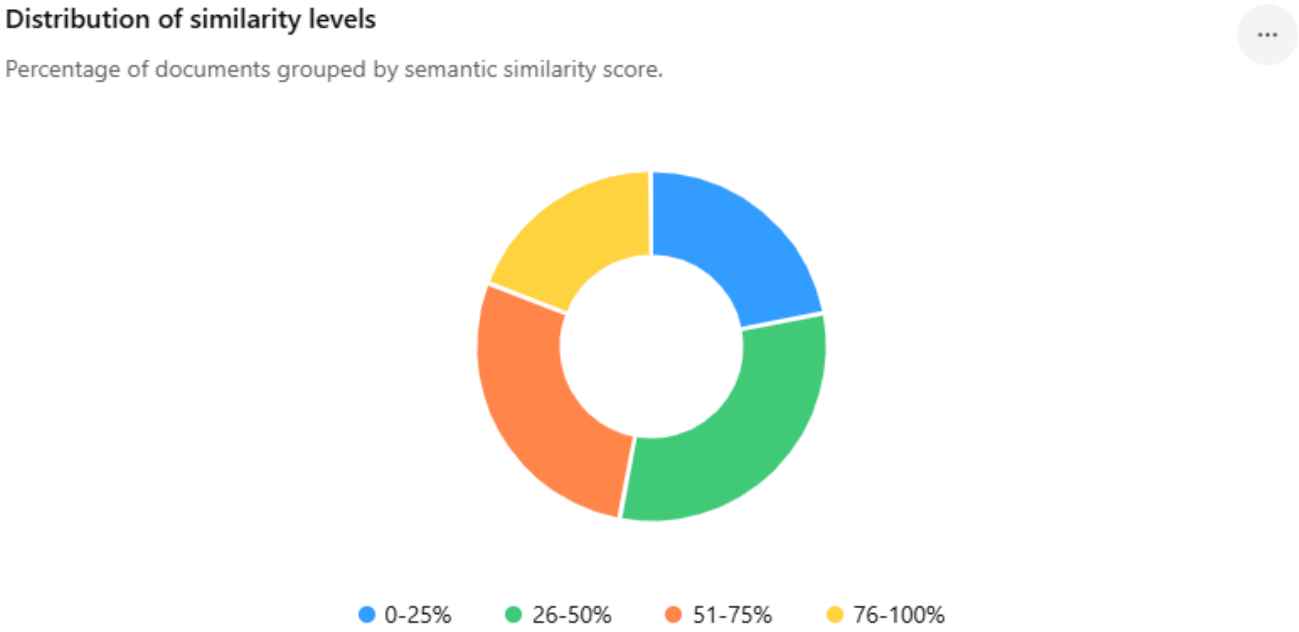

Figure 2. Distribution of Similarity Scores

The distribution indicates that the proposed system successfully differentiates between unrelated documents and highly similar texts. Documents with similarity scores above 75% generally correspond to direct plagiarism or strong semantic overlap.

3.5 Computational Efficiency Analysis

In addition to detection accuracy, computational performance was evaluated.

Table 3 Average Processing Time per Document Pair

Model	Processing Time (ms)
TF-IDF	12
Word2Vec	24
BERT	132
Sentence-BERT	85

Modern American Journal of Engineering, Technology, and Innovation
ISSN(E): 3067-7939
Volume 2, Issue 6, June, 2026
Website: usajournals.org

Although Transformer-based models require more computational resources than traditional approaches, Sentence-BERT achieves an excellent balance between efficiency and accuracy. Its Siamese architecture enables faster similarity calculations compared with standard BERT while maintaining superior semantic understanding capabilities.

3.6 Discussion

The experimental findings confirm that semantic embedding models substantially improve plagiarism detection performance. Traditional methods such as TF-IDF and Word2Vec are limited by their inability to fully capture contextual meaning, making them less effective against paraphrased plagiarism.

Transformer-based models, particularly BERT and Sentence-BERT, overcome these limitations by learning contextual representations of language. Among all evaluated approaches, Sentence-BERT achieved the best overall performance with an accuracy of **94.6%**, demonstrating strong capability in detecting hidden semantic relationships between documents.

Furthermore, the results indicate that semantic similarity analysis provides a more reliable assessment of document originality than lexical matching approaches. The proposed framework can therefore serve as a foundation for next-generation plagiarism detection systems in universities, research institutions, digital libraries, and scientific publishing platforms.

References

[1] Russell T., Airasian P. Classroom Assessment: Concepts and Applications. McGraw-Hill, 2012.

[2] Fishman T. The Fundamental Values of Academic Integrity. International Center for Academic Integrity, 2021.

[3] Potthast M., Stein B., Barrón-Cedeño A., Rosso P. An Evaluation Framework for Plagiarism Detection. Proceedings of COLING, 2010.

[4] Manning C.D., Raghavan P., Schütze H. Introduction to Information Retrieval. Cambridge University Press, 2008.

[5] Jurafsky D., Martin J.H. Speech and Language Processing. 3rd Edition, Pearson, 2023.

[6] Ashish Vaswani et al. Attention Is All You Need. NeurIPS, 2017.

Modern American Journal of Engineering, Technology, and Innovation
ISSN(E): 3067-7939
Volume 2, Issue 6, June, 2026
Website: usajournals.org

[7] Jacob Devlin et al. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. NAACL-HLT, 2019.
[8] Nils Reimers, Iryna Gurevych. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. EMNLP-IJCNLP, 2019.
[9] Goldberg Y. Neural Network Methods for Natural Language Processing. Morgan & Claypool Publishers, 2017.
[10] Bahodir M., Elyor E. Image data clustering based on the vgg16 model and the k-means algorithm //Universum: технические науки. – 2025. – Т. 6. – №. 1 (130). – С. 23-30.
[11] Egamberdiyev E. H. Clustering OF Small-scale Uzbek Texts Using Tf-idf AND Kmeans: an Empirical Evaluation OF Vectorization Parameters //Modern American Journal of Engineering, Technology, and Innovation. – Т. 1. – №. 4. – С. 58-67.