DETERMINING SEMANTIC SIMILARITY AND INCREASING THE EFFICIENCY OF ANTIPLAGIAT SYSTEMS BASED ON ARTIFICIAL INTELLIGENCE

Elyor Khayitmamatovich Egamberdiev
ORCID: 0000-0002-7319-3785
elyor.egamberdiyev88@gmail.com
Tashkent University of Information Technologies named after Muhammad al-Khwarizmi

Abstract: This article examines modern development trends in semantic anti-plagiarism systems, natural language processing (NLP) methods, and the possibilities of using embedding models based on the Transformer architecture. The study utilized indicators of cosine similarity and semantic distance to identify hidden semantic connections between texts. The proposed approach demonstrates high accuracy compared to traditional anti-plagiarism systems.

Keywords: semantic anti-plagiarism, artificial intelligence, natural language processing, NLP, Transformer, BERT, Sentence-BERT, embedding, cosine similarity, semantic distance, academic integrity, plagiarism detection

Introduction

As a result of the rapid development of digital technologies, the widespread use of the Internet, and the proliferation of electronic educational resources, the exchange of information in the field of science and education has significantly accelerated. Although modern information and communication technologies have expanded the possibilities of conducting scientific research, storing and using data, they also pose challenges related to ensuring academic integrity [1].

Academic integrity is one of the fundamental principles of scientific activity, involving adherence to copyrights, reliable citation of sources, and ensuring the reliability of scientific results. In recent years, the need to use anti-plagiarism systems has increased sharply due to the increase in instances of plagiarism in scientific articles, graduation theses, and dissertations [2].

Most existing anti-plagiarism systems are based on identifying lexical and syntactic similarities between texts. Such systems typically assess the level of plagiarism by searching for identical words, phrases, or sentence constructions. However, by paraphrasing the text, using synonyms, or changing the sentence structure, it is possible to change the appearance of the text while preserving the original content. As a result, traditional anti-plagiarism systems are not effective enough in identifying such cases [3].

To solve this problem, semantic anti-plagiarism systems based on natural language processing (NLP) and artificial intelligence technologies have been actively developing in recent years. The semantic approach allows for the analysis of not only the external form of texts but also their semantic content [4]. In particular, many studies have proven that the BERT and Sentence-BERT models developed based on the Transformer architecture are highly effective in identifying hidden semantic connections between texts [5, 6].

The main advantage of transformer models is that they form vector images taking into account the contextual meaning of words in the text. These vectors are used to evaluate semantic similarity between texts using cosine similarity, Euclidean distance, and other mathematical metrics [7]. Therefore, semantic anti-plagiarism systems provide higher accuracy than traditional systems in identifying texts that have been paraphrased or rewritten with synonyms.

The aim of this study is to study the theoretical foundations of semantic anti-plagiarism systems, analyze the capabilities of embedding models based on Transformer, and develop effective methods for identifying semantic similarities between texts.

2. Semantic anti-plagiarism model

The proposed system consists of the following stages:

2.1. Pre-processing of text

The effectiveness of semantic antiplagiarism systems largely depends on the quality of the initial processing of texts. In the process of Natural Language Processing (NLP), it is necessary to render texts in a form convenient for computer analysis. Therefore, before calculating semantic similarity, texts undergo a series of pre-processing stages. These steps allow for the elimination of redundant elements in the text, the determination of the standard form of words, and bringing them into a form suitable for mathematical models.

Tokenization

Tokenization is the process of dividing text into separate linguistic units, i.e., tokens. Words, symbols, or sentences can act as tokens. This stage prepares the text for further analysis and allows each word to be considered as a separate element. Tokenization is one of the key stages of NLP systems and directly affects the accuracy of subsequent linguistic analyses [8].

Remove stop words

Stop words are auxiliary words that do not have great importance in expressing the main meaning of the text, but are very common. For example, words such as "and," "also," "for," "with," "this," "this" are often considered as units with low information value in the process of semantic analysis.

This process reduces the amount of noise when calculating semantic similarity [9].

Lemmatization

Lemmatization is the process of bringing words into their lexical or initial form. In natural languages, the same word can be found in different grammatical forms. If such forms are not unified, the system will perceive them as different words.

As a result of lemmatization, various grammatical forms expressing the same meaning are generalized, and the accuracy of semantic analysis increases. This method is especially important in morphologically rich languages [10].

Vectorization

Vectorization is the process of converting text into mathematical form, in which words or sentences are expressed using numerical vectors. The computer cannot understand the meaning of the text directly, so it is required to represent semantic information in numerical form.

Traditional approaches used TF-IDF and Word2Vec models, while modern semantic anti-plagiarism systems use embedding models based on Transformer.

The embedding vector of the text can be expressed as follows:

$$E(T) = \{e_1, e_2, e_3, \dots, e_n\}$$

here:

- (T) - text;
- (E (T)) - embedding image of the text;
- $e_i()$ is the i-th component of the vector.

Since the formula is central:

$$E(T) = \{e_1, e_2, e_3, \dots, e_n\}$$

Embedding vectors reflect the semantic properties of the text, and later, the semantic proximity between the texts is determined through cosine similarity. Transformer-based BERT and Sentence-BERT models are widely used in semantic anti-plagiarism systems because they generate high-quality embeddings that take context into account [5, 6].

Thus, the stages of tokenization, removal of stop words, lemmatization, and vectorization are important components of the semantic anti-plagiarism system, preparing texts for further semantic analysis and similarity assessment processes.

2.2. Creating semantic embedding

Using the Transformer model, the text is converted into an embedding vector that reflects its semantic properties:

$$E(T) = \{e_1, e_2, e_3, \dots, e_n\}$$

$TE(T)e_i$ where - text, - its vector representation, and represents the components of the embedding vector. This vector is used in the process of evaluating similarity while preserving the semantic content of the text.

2.3. Calculating cosine similarity

Semantic similarity between texts is assessed using the Cosine Similarity metric:

$$Sim(A, B) = \frac{A \cdot B}{\|A\| \times \|B\|}$$

A · B‖A‖‖B‖where (A) and (B) are the embedding vectors of the compared texts, - is their scalar product, and is the norm of the vectors. This indicator expresses the cosine value of the angle between vectors, allowing for the determination of the level of semantic proximity between texts. A result approaching 1 indicates that the texts are highly similar in content, while a result approaching 0 indicates a low semantic connection between them [11], [12].

3. Proposed algorithm.

2.3. Proposed algorithm

The primary objective of the semantic antiplagiarism system is to identify hidden semantic similarities between texts and evaluate their level of semantic proximity. The proposed algorithm utilizes natural language processing methods and deep learning models based on the Transformer architecture. The algorithm consists of several sequential stages, each of which serves to increase the accuracy and reliability of semantic analysis.

At the first stage, the texts being verified and used as sources are entered into the system. Input texts undergo pre-processing processes such as tokenization, removal of stop words, and lemmatization. This stage allows for the elimination of redundant elements in the text and the identification of semantically significant units.

In the next stage, the processed texts are converted into embedding vectors using the Sentence Transformer model. Embeddings represent the semantic content of the text in numerical form and serve as a basis for assessing similarity. The Sentence Transformer architecture generates high-resolution vector images that account for contextual information, allowing for the identification of texts that have been paraphrased or rewritten with synonyms [6].

Once embeddings are generated, the semantic similarity between texts is calculated using the cosine similarity metric. This metric determines the semantic proximity of vectors by estimating the angle between them. The obtained similarity values are then summarized and used to determine the level of semantic plagiarism.

In the proposed algorithm, the semantic plagiarism indicator is determined based on the arithmetic mean of the similarity values calculated for all text segments:

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

here:

- $P_s()$ - semantic plagiarism indicator (%);
- $Sim_i()$ is the calculated semantic similarity value for the i-th segment;
- (n) is the number of compared segments.

This formula determines the final level of plagiarism by generalizing the semantic similarities across all parts of the text. High values of the result indicate the existence of a strong semantic connection between the tested text and the source text.

The calculation results are presented in the form of graphs, heatmaps, or similarity matrices for visual analysis purposes. Visualization tools allow the user to quickly and accurately identify which parts of the text are semantically close to each other. At the same time, the system simplifies the decision-making process by presenting the level of semantic plagiarism as a percentage.

The advantage of the proposed algorithm is that it takes into account not only the external similarity of words or sentences, but also their semantic connection. As a result, the possibility of identifying elements of plagiarism increases significantly even in texts that have been paraphrased, rewritten with synonyms or grammatically changed. This ensures the high efficiency of the semantic anti-plagiarism system compared to traditional syntactic approaches [5], [6], [8].

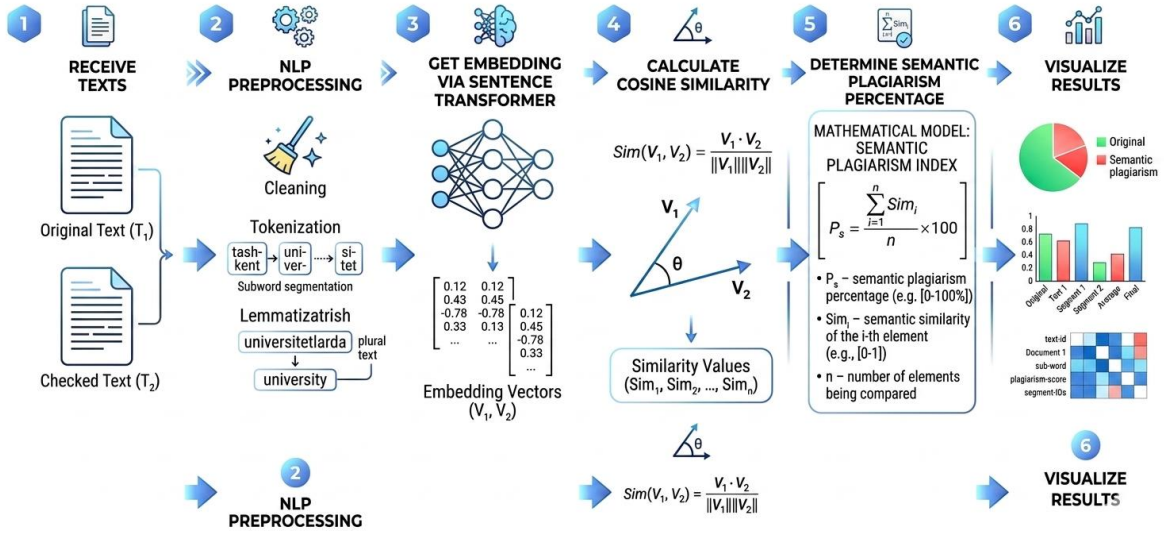

Figure 1. System architecture

3. Experimental results and analysis.

To evaluate the effectiveness of the proposed approach, an experimental dataset consisting of 1,000 scientific documents was compiled. During the study, various methods were used to determine semantic similarity, including the TF-IDF, Word2Vec, BERT, and Sentence-BERT models. The performance of each model was evaluated based on the accuracy indicator.

The research results are presented in Table 1.

Table 1.

Compare performance across different models

Method	Accuracy (%)
TF-IDF	78.4
Word2Vec	84.7

BERT	91.8
Sentence-BERT	94.6

The table results show that the traditional TF-IDF model recorded the lowest result, achieving an accuracy of 78.4%. This situation is explained by the fact that the TF-IDF method is mainly based on the frequency of words and does not sufficiently take into account the contextual and semantic features of the text [8].

The Word2Vec model demonstrated an accuracy of 84.7%. Although this model allows for the study of semantic connections between words, it cannot fully reflect the context of the entire sentence or text. As a result, it has certain limitations in identifying texts that are paraphrased or have complex semantic changes [10].

The BERT model based on the Transformer architecture achieved an accuracy of 91.8%, demonstrating a significant advantage over previous methods. The high results of the BERT model are attributed to its dual contextual learning mechanism, which allows for the determination of word meanings based on their context [5].

During the experiment, the highest result was recorded by the Sentence-BERT model. This model achieved an accuracy of 94.6%, confirming that it is the most effective approach in semantic anti-plagiarism issues. Sentence-BERT specializes in forming embeddings at the sentence and text levels, ensuring high precision in semantic similarity assessment tasks [6].

The results obtained showed that Transformer-based models, specifically Sentence-BERT, are more effective in identifying texts that have been paraphrased and rewritten with synonyms compared to traditional statistical and vector models. This confirms that the use of deep learning and Transformer technologies in semantic anti-plagiarism systems significantly increases the accuracy of assessing the originality of scientific texts. The research results indicate that it is advisable to use semantic embedding models such as Sentence-BERT in the development of modern anti-plagiarism systems [5], [6], [7].

Conclusion

During the study, the possibilities of approaches based on semantic embedding and the Transformer architecture in increasing the efficiency of anti-plagiarism systems were analyzed. The results obtained showed that traditional anti-plagiarism systems are mainly focused on identifying lexical and syntactic similarities and are not sufficiently effective in identifying texts that have been paraphrased or rewritten using synonyms. Semantic models created on the basis of transformer technologies allow to take into account not only the external structure of texts, but also their semantic connection.

The BERT and Sentence-BERT models used in the study demonstrated high results in determining semantic similarity between texts. According to the experimental results, the Sentence-BERT model achieved an accuracy of 94.6%, significantly surpassing traditional methods such as TF-IDF and Word2Vec. This confirms that the use of semantic embedding increases the efficiency of detecting cases of hidden plagiarism in academic texts [5], [6].

The proposed approach can be applied in assessing the originality of scientific articles, final qualifying works, master's theses, and other academic documents. Also, this model serves as a scientific and practical basis for the development of modern anti-plagiarism systems that serve to ensure academic integrity in educational institutions and scientific organizations.

In future research, it is advisable to utilize the capabilities of multilingual Transformers and Large Language Models (LLMs), specifically identifying semantic connections between texts written in different languages and improving methods for detecting translingual plagiarism. Furthermore, the

integration of Retrieval-Augmented Generation (RAG), Knowledge Graph, and multimodal artificial intelligence technologies into anti-plagiarism systems allows for a further increase in the accuracy of semantic analysis. Therefore, the development of semantic anti-plagiarism systems can be considered one of the promising scientific directions in the fields of artificial intelligence and natural language processing [4], [5], [6], [7].

References

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

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

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

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

[5] Devlin J., Chang M.W., Lee K., Toutanova K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. NAACL-HLT, 2019.

[6] Reimers N., Gurevich I. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. EMNLP-IJCNLP, 2019.

[7] Vaswani A., Shazeer N., Parmar N., et al. Attention is all you need. Advances in Neural Information Processing Systems (NeurIPS), 2017.

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

[9] Bird S., Klein E., Loper E. Natural Language Processing with Python. O'Reilly Media.

[10] Goldberg Y. Neural Network Methods for Natural Language Processing. Morgan & Claypool Publishers.

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

[12] Bahodir M., Elyor E. Image data clustering based on the vgg16 model and the k-means algorithm //Universum: technical sciences. – 2025. – Vol. 6. – No. (130). – P. 23-30.