DOI: **10.37943/25NVVS5297**

Valeriya Kazagasheva
Master student, School of Artificial Intelligence and Data Science
242752@astanait.edu.kz, orcid.org/0009-0001-0262-0305
Astana IT University, Kazakhstan
Oleksandr Kuchanskyi
Professor, School of Artificial Intelligence and Data Science
a.kuchanskyi@astanait.edu.kz, orcid.org/0000-0003-1277-8031
Astana IT University, Kazakhstan
Professor, Department of Biomedical Cybernetics
National Technical University of Ukraine
"Igor Sikorsky Kyiv Polytechnic Institute", Ukraine
Svitlana Biloshchytska
Professor, School of Artificial Intelligence and Data Science
bsv@astanait.edu.kz, orcid.org/0000-0002-0856-5474
Astana IT University, Kazakhstan
Dina Kantayeva
PhD student, School of Software Engineering
d.kantayeva@astanait.edu.kz, orcid.org/0000-0002-4097-5078
Astana IT University, Kazakhstan

SCALABLE NEAR-DUPLICATE DETECTION IN KAZAKH SCIENTIFIC TEXTS VIA SEMANTIC EMBEDDINGS AND OPTIMIZED CANDIDATE FILTERING

Abstract: This work considers the problem of efficient detection of near-duplicate documents in Kazakh scientific texts, which is particularly challenging due to the agglutinative nature of the language and the high computational cost of pairwise document comparison. Traditional approaches based on lexical similarity are ineffective under such conditions, while semantic models, although more accurate, are computationally expensive and scale poorly. To overcome these limitations, the study proposes a scalable framework that combines semantic similarity modeling with optimization techniques, including text canonicalization, efficient indexing, and multi-stage candidate filtering. The canonicalization process reduces morphological variability, increasing the stability of similarity estimation for Kazakh texts. The indexing mechanism, based on dense vector representations, enables efficient selection of candidate pairs using approximate nearest neighbor search. The hierarchical filtering strategy further reduces the number of comparisons, while a transformer-based model provides accurate semantic matching. The proposed approach is evaluated on a large-scale dataset of Kazakh scientific abstracts and near-duplicate pairs. The results demonstrate that the framework achieves high detection accuracy while significantly reducing computational costs compared to exhaustive pairwise comparison. The use of dynamic threshold adjustment allows effective handling of overlapping similarity distributions between duplicate and non-duplicate classes. The obtained results confirm that the combination of linguistic preprocessing and computational optimization is crucial for scalable near-duplicate detection in low-resource agglutinative languages such as Kazakh. The proposed framework can be applied in plagiarism detection, document deduplication, and large-scale text analysis systems.

Keywords: near-duplicate detection; semantic similarity; Kazakh language; agglutinative languages; text canonicalization; indexing; candidate filtering; optimization; transformer-based language models.

Received: 11.02.2026 Accepted: 25.02.2026 Published: 30.03.2026

Scientific Journal of Astana IT University
ISSN (P): 2707-9031 ISSN (E): 2707-904X
VOLUME 25, MARCH 2026

Introduction

Agglutinative languages are languages in which grammatical word forms are created by adding affixes to the base of a word and change depending on case, tense, mood, verb person, etc. Such languages include Finno-Ugric (Finnish, Hungarian, Estonian, etc.), Turkic (Turkish, Kazakh, Uzbek, Kyrgyz, etc.), Mongolic, Caucasian, Austronesian (Filipino, Malay, etc.), Japanese, Korean, and others [1]. The main idea of agglutination is that the root usually preserves its form, while each suffix adds a single meaning, and morphemes are easily detachable. That is, morphemes are unambiguous and clearly segmented, while long words are formed.

The rapid growth in the number of scientific publications and digital text repositories has significantly increased the need for effective methods of near-duplicate detection. Near-duplicate detection plays an important role in plagiarism detection, deduplication of bibliographic databases, and ensuring the integrity of scientific communication. Unlike exact duplicates, near duplicates include paraphrased or structurally modified texts that preserve the original meaning but change the surface forms of representation. Traditional lexical similarity methods, such as string matching or token overlap, are often unable to detect such cases, especially when complex paraphrasing techniques are used.

The problem of near-duplicate detection in texts written in agglutinative languages is complex. This is due to the combination of morphological and semantic factors. An example of an agglutinative language is Kazakh. It is characterized by intensive affixation, which leads to the formation of many word forms from a single root base. At the same time, texts that are semantically similar may differ significantly at the lexical representation level. Morphological variability complicates the process of near-duplicate detection. This reduces the effectiveness of traditional text analysis methods, which, for example, use n-gram representations. In addition, it is difficult to determine the presence of near duplicates due to the blurred boundary between what can be considered a duplicate and what can be considered an original text. Taking together, these factors provide the basis for the development and use of near-duplicate detection methods that can account for the semantics and context of texts. One such method that is well-suited for detecting near duplicates in texts written in agglutinative languages is BERT.

Another critical challenge is the computational complexity of comparing all pairs of documents in large corpora, which grows quadratically with the number of documents. This makes optimization strategies, including efficient indexing, candidate filtering, and text canonicalization, necessary for practical implementation.

Most existing studies focus primarily on analytical languages and do not account for the specifics of agglutinative languages. Modern approaches to near-duplicate detection are based on lexical or hybrid methods, the effectiveness of which is significantly reduced under conditions of high morphological variability. Therefore, the problem of near-duplicate detection in agglutinative languages, particularly in Kazakh, is relevant from both scientific and practical perspectives.

In addition to semantic accuracy, computational efficiency is a critical requirement for near-duplicate detection in large-scale text corpora. Naive pairwise document comparison has quadratic time complexity, which is unacceptable for practical applications. Therefore, optimization strategies such as efficient indexing, candidate filtering, and text canonicalization become necessary. These methods enable reducing the number of comparisons while preserving detection quality. This is especially important for agglutinative languages such as Kazakh, where morphological variability further increases computational complexity.

Literature Review

Agglutinative languages have specific characteristics that complicate the semantic comparison of texts, particularly the detection of near duplicates. This is because grammatical relations are formed by attaching numerous suffixes to word roots. As a result, the meaning of the text changes. This is the main difference between agglutinative languages and analytical languages, in which grammatical relations are determined based on separate function words. Thus, in agglutinative languages, a single lexical root can generate many word forms that depend on number, case, and other grammatical features. This means that texts can be semantically similar or even identical but may differ significantly at the lexical level. This important characteristic significantly complicates the application of traditional text comparison methods and is the main reason for the creation of new and the modification of existing methods and models that are

capable of more deeply considering the context of the text and its semantics. Such methods may be based on BERT-like approaches. In the work of Kessikbayeva G. et al. [2], the morphotactic of alternations for the agglutinative Kazakh language is formalized, and the construction of rules for two-level morphology is described. The work of Kessikbayeva G. et al. [2] provides an understanding of how normalization should be performed. In the work of Washington J. et al. [3], the principles and rules of morphological normalization for Kipchak languages, including Kazakh, are described. In general, morphological analysis for agglutinative languages is a separate preprocessing procedure, and this is related to the length of word forms. Such morphological analysis for the Kazakh language is described in works of Makhambetov O. et al. [4] and Yiner Z. et al. [5]. The work of Assylbekov Z. et al. [6] indicates that agglutination increases vocabulary sparsity. This complicates the implementation of near-duplicate detection methods.

Another feature of agglutinative languages is their morphological variability. This significantly complicates the task of near-duplicate detection in texts of such languages. Semantically similar statements may differ or even be completely opposite in meaning due to the productivity of affixation [7, 8]. Such languages are characterized by significant changes in cases, tenses, and aspects, which lead to the formation of many word forms with a single semantic core and different lexical representations. In this case, two texts may be close in semantics but show low lexical similarity. Therefore, methods based on surface-level text analysis for near-duplicate detection, such as n-gram methods, will almost certainly produce erroneous results. In contrast, context-oriented models such as BERT can capture a wide range of semantic relationships between words regardless of their morphological forms. This means that such methods are more suitable for the analysis of agglutinative languages. However, it should be noted that semantic similarity filtering may eliminate valid pairs in agglutinative languages [9], in particular Turkish and Kazakh, therefore, in this case, filtering should be performed with special control. That is, the filtering threshold should be calibrated on local data [10]. For the Kazakh language, near-duplicate detection is described in the work of Biloshchytska S. et al. [11], which compares the n-gram method, TF-IDF, BERT, and a hybrid method (a combination of statistical and semantic approaches), emphasizing agglutinativity.

In contrast, context-oriented models such as BERT are capable of capturing a wide range of semantic relationships between words regardless of their morphological forms. This means that such methods are more suitable for the analysis of agglutinative languages. However, it should be noted that semantic similarity filtering may eliminate valid pairs in agglutinative languages [9], in particular, Turkish and Kazakh, therefore, in this case, filtering should be performed with special control. That is, the filtering threshold should be calibrated on local data [10]. For the Kazakh language, near-duplicate detection is described in the work of Biloshchytska S. et al. [11], which compares the n-gram method, TF-IDF, BERT, and a hybrid method (a combination of statistical and semantic approaches), emphasizing agglutinativity. In addition to linguistic features, near-duplicate detection in agglutinative languages is associated with difficulties related to the calculation of statistical characteristics. Pairs of texts with similar values of similarity metrics may belong to different classes (“duplicate”, “non-duplicate”). This is because the analysis of similarity distributions between text pairs shows significant overlap between classes. That is, there is no clear boundary between classes, and near-duplicate detection has a nonlinear nature. Therefore, the use of simple heuristic rules and lexical metrics for near-duplicate detection in agglutinative languages is not rational.

Thus, there arises the problem of developing methods that can consider semantic and contextual dependencies between texts and handling the ambiguity of class boundaries. One of the solutions that allows improving the accuracy of near-duplicate identification for large data collections is the method described in the work of Reimers N. et al. [12]. In the work of Kuchanskyi O. et al. [13], a combination of lexical metrics, contextual BERT embeddings, and other syntactic features with adaptive thresholds is described. That is, Kuchanskyi O. et al. [13] indicate that for agglutinative languages, BERT-based solutions should be enhanced with additional features.

Other methods are also used for near-duplicate and plagiarism detection. In the work of Bhoi S. et al. [14], the MultiSiam network is described, which is applied to social media for duplicate classification. Despite this, it can be argued that short social media texts and scientific abstracts differ significantly in length and structure. Accordingly, constructing neural networks of this type may not produce the desired results. The use

Scientific Journal of Astana IT University
ISSN (P): 2707-9031 ISSN (E): 2707-904X
VOLUME 25, MARCH 2026

of wavelet analysis with clustering for cross-modal duplicate detection in texts and images has significant potential, but these methods are characterized by high computational complexity, which is a barrier to practical scalability [15]. A comprehensive review of plagiarism detection methods for the period 2014–2024 showed a clear trend toward semantic and transformer-based methods. At the same time, simple text comparison methods are not capable of detecting complex text reuse through substantial paraphrasing [16]. In general, Shahmohammadi H. et al. [17] show that the combination of machine learning and additional features determined by the type of language provides better performance. This is especially important for low-resource and morphologically rich languages, such as Kazakh [18]. Agarwal B. et al. [19] show that the combination of CNN-RNN can be effective due to end-to-end learning. Sentence-Transformer models are capable of effectively capturing deep contextual meaning even in low-resource languages [20, 21].

Thus, existing approaches to near-duplicate detection are generally oriented toward analytical languages (English, German, French, etc.) and do not consider the morphological characteristics of agglutinative languages, particularly Kazakh. Moreover, with regard to the Kazakh language, the problem of near-duplicate detection is poorly studied in the scientific literature. In particular, the effect of overlapping similarity distributions between classes has not been investigated.

Recent studies have shifted the focus toward semantic approaches based on neural networks and transformer models. Sentence-level vector representations generated by models such as BERT have demonstrated high effectiveness in detecting contextual similarity even in cases of substantial paraphrasing. Hybrid methods that combine lexical and semantic features have shown improved performance. Ensemble approaches further enhance robustness by integrating multiple similarity signals and learning optimal feature weights using machine learning models. Optimization methods for near-duplicate detection typically include candidate filtering, indexing, and approximate similarity search techniques. These approaches aim to reduce the number of pairwise comparisons, thereby improving scalability. However, most existing studies focus on high-resource languages. Research on low-resource and agglutinative languages remains limited, especially in the context of combining semantic modeling with computational optimization. Despite the growing number of studies in the field of semantic duplicate detection, several limitations remain. First, most existing methods focus primarily on improving detection accuracy, while computational efficiency and scalability often remain overlooked. Second, indexing strategies for efficient candidate selection are rarely integrated with semantic models. Third, text canonicalization methods adapted for agglutinative languages, such as Kazakh, are insufficiently studied. As a result, there is a lack of unified approaches that combine indexing, canonicalization, and semantic similarity evaluation within a single optimized framework.

In contrast to prior studies that focus either on semantic similarity modeling or on duplicate detection in high-resource languages, the present study combines three components within a single framework tailored to Kazakh scientific texts: canonicalization adapted to an agglutinative language; ANN-based candidate selection for scalability; transformer-based semantic matching with threshold calibration under overlapping similarity distributions. The contribution of the study is therefore not only empirical but also architectural, as it integrates linguistic preprocessing and computational optimization into a unified near-duplicate detection pipeline for a low-resource language.

The aim of this study is to develop a scalable and computationally efficient framework for detecting near duplicates in Kazakh scientific texts by integrating optimized indexing, text canonicalization, and semantic similarity modeling to minimize search time while maintaining high detection accuracy. To achieve this, the following tasks need to be performed:

1. Develop a canonicalization method that reduces morphological variability in Kazakh texts and improves similarity estimation.
2. Design an efficient indexing and candidate filtering mechanism that minimizes the number of pairwise comparisons.
3. Implement and evaluate a semantic similarity model for accurate near-duplicate detection while ensuring computational efficiency.

DOI: 10.37943/25NVVS5297

Data Collection and Dataset Construction

For the study, two datasets were collected: "Kazakh scientific publications dataset from Semantic Scholar (2000–2025)" [22]. The first dataset, "Kazakh academic abstracts corpus," contains a curated collection of 10,468 scientific article abstracts specifically focused on Kazakhstan and the Kazakh language. The data was programmatically collected from the Semantic Scholar API using targeted queries related to Kazakhstan.

Data structure:

- paperId: Unique identifier from Semantic Scholar.
- title: The title of the scientific paper.
- abstract_kk: The full text of the abstract in Kazakh.
- year: Publication year (ranging from 2000 to 2025).
- query: The search term used to retrieve the record.

The second dataset, "Kazakh abstract pairs for duplicate detection," consists of 11,851 pairs of Kazakh-language academic abstracts, developed specifically for research in duplicate detection. The dataset provides a balanced mix of positive (duplicate/paraphrased) and negative (distinct) pairs. To ensure high quality and complexity, the authors utilized a hybrid approach combining real-world data with controlled synthetic augmentation:

1. Near-duplicates: Generated using a custom paraphrasing engine that performs synonym replacement, sentence shuffling, and structural rephrasing based on Kazakh linguistics.
2. Similarity levels: Pairs are categorized into "high", "medium", and "low" similarity based on TF-IDF and Cosine Similarity scores.
3. Negative samples: Formed by pairing unrelated abstracts to provide "non-duplicate" labels for machine learning training.

Data structure:

- abstract_a / abstract_b: The pair of texts to be compared.
- similarity_score: The computed cosine similarity value (0.0 to 1.0).
- label: Binary indicator (1 for duplicates/paraphrases, 0 for different texts).
- pair_type: Qualitative description of similarity (e.g., high_similarity, different_abstract).
- rephrase_level: The intensity of the transformation (very_light, light, medium, strong, or strong_shorten).

Quality control was applied to the augmented near-duplicate pairs to reduce the risk of linguistically implausible or semantically distorted examples. First, rule-based constraints were imposed during generation to avoid excessive corruption of word order and to preserve core topic-related lexical units. Second, a manual inspection of a sample of generated pairs was carried out by native or proficient Kazakh speakers. The inspection focused on three aspects: grammatical acceptability, semantic consistency with the source abstract, and overall naturalness of the resulting text. Pairs that contained severe semantic drift, broken sentence structure, or obviously artificial transformations were excluded from the final dataset. This procedure was not intended as a full-scale linguistic annotation campaign, but as a practical verification step to increase the reliability of the synthetic training data.

This dataset represents a large-scale bibliometric collection of scientific publications compiled based on the Semantic Scholar Academic Graph. The dataset covers metadata of scientific publications related to scientific activity in Kazakhstan for the period from 2000 to 2025. The dataset integrates information from various scientific sources, including CrossRef, PubMed, and arXiv. During the dataset construction, publications were pre-filtered by affiliations, thematic sets, and also underwent data cleaning, duplicate removal, and normalization. A verification step was also performed to ensure that all publications fall within the defined time period.

Based on the preliminary analysis, it can be concluded that the dataset is generally multidisciplinary and includes scientific publications in engineering, natural sciences, computer science, social sciences, and medicine. The availability of citation information and reference lists makes it possible to construct scientific networks, including citation networks and bibliographic coupling networks. The presence of abstract data also

Scientific Journal of Astana IT University
ISSN (P): 2707-9031 ISSN (E): 2707-904X
VOLUME 25, MARCH 2026

allows for text analysis to detect near duplicates, specifically in scientific texts. The dataset has certain limitations, in particular the lack of full-text versions of articles, which complicates large-scale text analysis for near-duplicate detection. Nevertheless, the available volume is sufficient for implementing near-duplicate detection methods in Kazakh-language texts and includes the distribution of publications by year, as shown in Figure 1. This figure illustrates the year-over-year progression of Kazakh-language scientific output. A clear upward trend is visible, particularly after 2015, suggesting an intensification of local academic discourse. The distribution by the number of words in abstracts is shown in Figure 2.

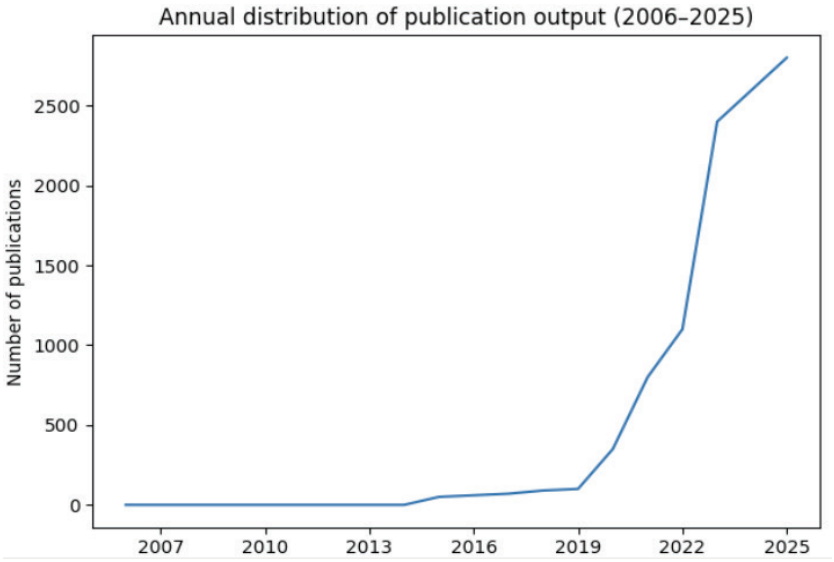


Figure 1. Annual distribution of publications (2006–2025)

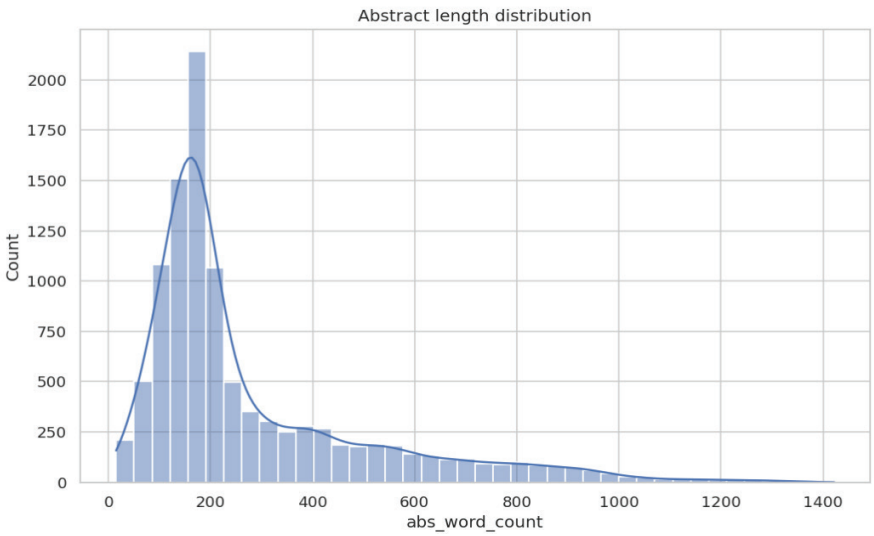


Figure 2. Distribution of word counts in scientific abstracts

The histogram (Fig. 3) reveals the density of abstract lengths, following a near-normal distribution with a slight positive skew. This indicates a standardized consensus on abstract length within the Kazakh academic community. The boxplot visualizes the variance in abstract length across different years. A slight increase in the median word count over time suggests a transition toward more detailed and informative summaries. Title lengths are concentrated between 6 and 11 words, demonstrating the linguistic constraints and stylistic preferences of scientific titling in the Kazakh language (Fig 4). The horizontal bar chart identifies

DOI: 10.37943/25NVVS5297

the most significant lexical units in the corpus after the removal of stop words. Terms such as «білім» and «талдау» dominate, reflecting the core thematic pillars of the dataset (Fig 5.).

Bigram analysis reveals common syntactic structures and collocations (e.g., 'research results', 'theoretical basis'), which are essential for understanding the formal register of Kazakh academic prose (Fig. 6).

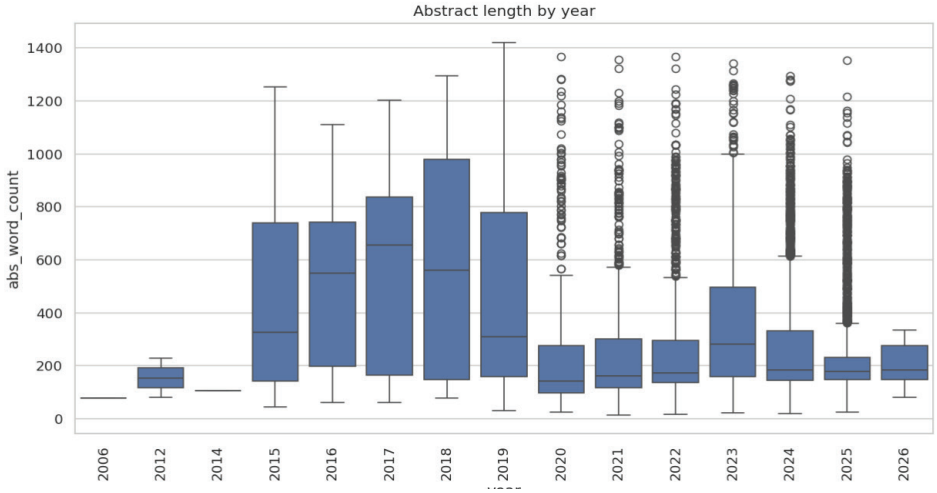


Figure 3. Temporal dynamics of abstract length: A boxplot analysis

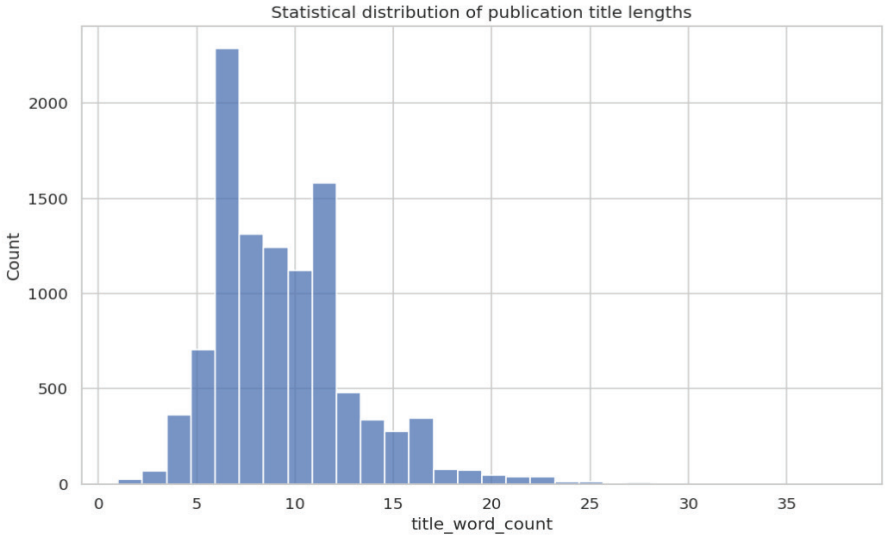


Figure 4. Statistical distribution of publication title lengths

Complete information about the collected dataset is provided at the link [22]. To improve transparency of the augmentation procedure, representative examples of generated positive pairs are provided in Appendix Tables A1–A4. These examples cover high-, medium-, low-, and very-low-similarity near-duplicate pairs and illustrate the main transformation types, including lowercasing, punctuation removal, stopword deletion, sentence reordering, lexical substitution, structural rephrasing, and perspective shift. The appendix is intended to help readers assess the linguistic plausibility and semantic consistency of the generated Kazakh paraphrases.

Scientific Journal of Astana IT University
ISSN (P): 2707-9031 ISSN (E): 2707-904X
VOLUME 25, MARCH 2026

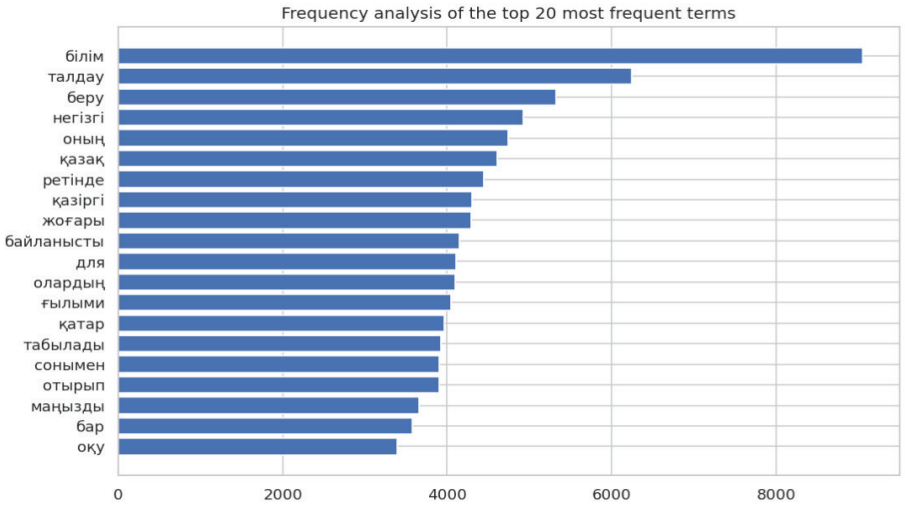


Figure 5. Frequency analysis of the top 20 most frequent terms

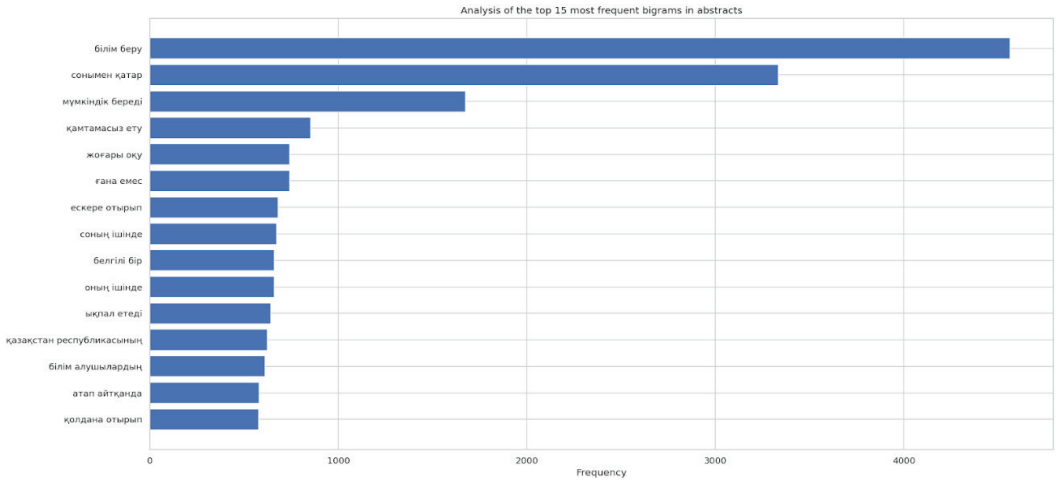


Figure 6. Analysis of the top 15 most frequent bigrams in abstracts

The synthetic augmentation procedure was used to enrich the diversity of positive near-duplicate examples and to simulate different levels of paraphrastic variation, but it did not replace real scientific abstracts. The resulting dataset combines authentic Kazakh academic texts with controlled transformations, which makes it possible to study near-duplicate detection under realistic lexical and semantic variability while preserving a link to genuine scholarly discourse.

Data Preparation and Methodology

To solve the problem, preliminary data analysis and preprocessing were performed. Figure 7 illustrates the quantitative balance of the corpus between the two primary classes: negative (distinct abstracts) and positive (near-duplicate pairs). The Kernel Density Estimation plot (Fig. 8) visualizes the distribution of lexical similarity scores for both classes. While the positive class exhibits a higher density at the upper end of the scale, the notable overlap between the two distributions in the mid-range underscores the necessity of moving beyond simple threshold-based lexical matching. Figure 9 provides a view of the dataset composition. The pie chart illustrates the quantitative prevalence of each generation type, where “medium_similarity” and “different_abstract” form the majority of the corpus. The heatmap (Fig. 10) presents the Pearson correlation coefficients among the engineered features. The high correlation between character and word counts is expected, but the low correlation between temporal gaps and similarity scores indicates that time is an independent factor in the occurrence of near-duplicates. Table 1 provides a detailed statistical breakdown of

DOI: 10.37943/25NVVS5297

the dataset used for training and validating the proposed hierarchical ensemble framework. The corpus consists of 11,851 unique abstract pairs, designed to represent a realistic distribution of both genuine scholarly content and potential near-duplicates.

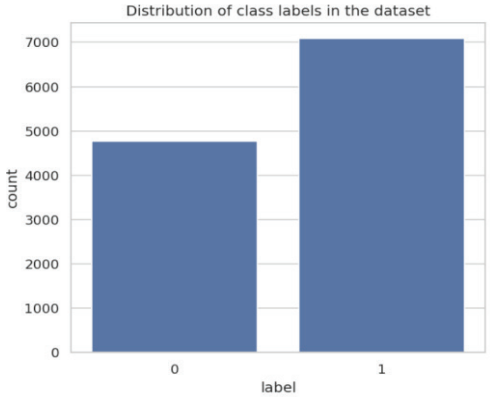


Figure 7. Distribution of class labels in the dataset

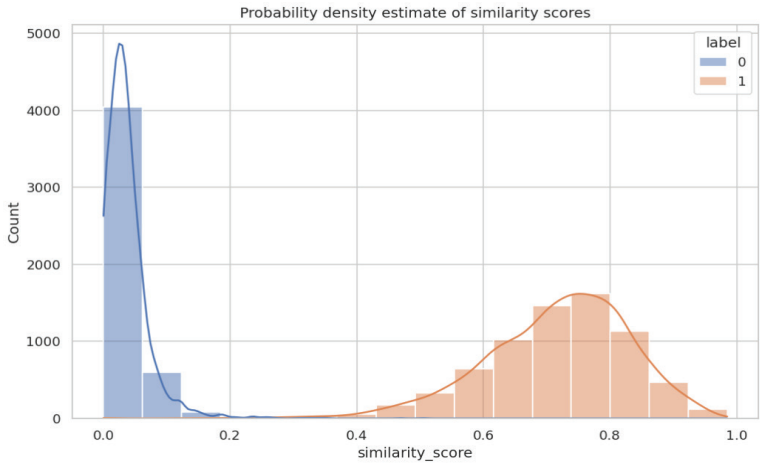


Figure 8. Probability density estimate of similarity scores

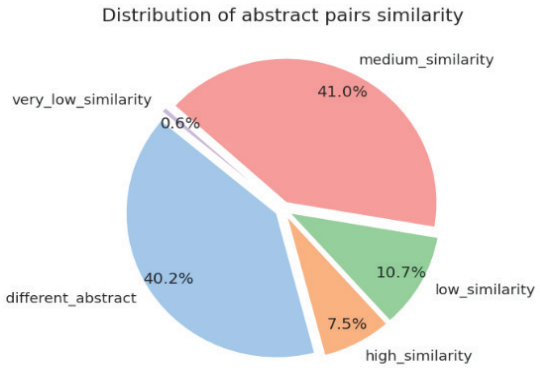
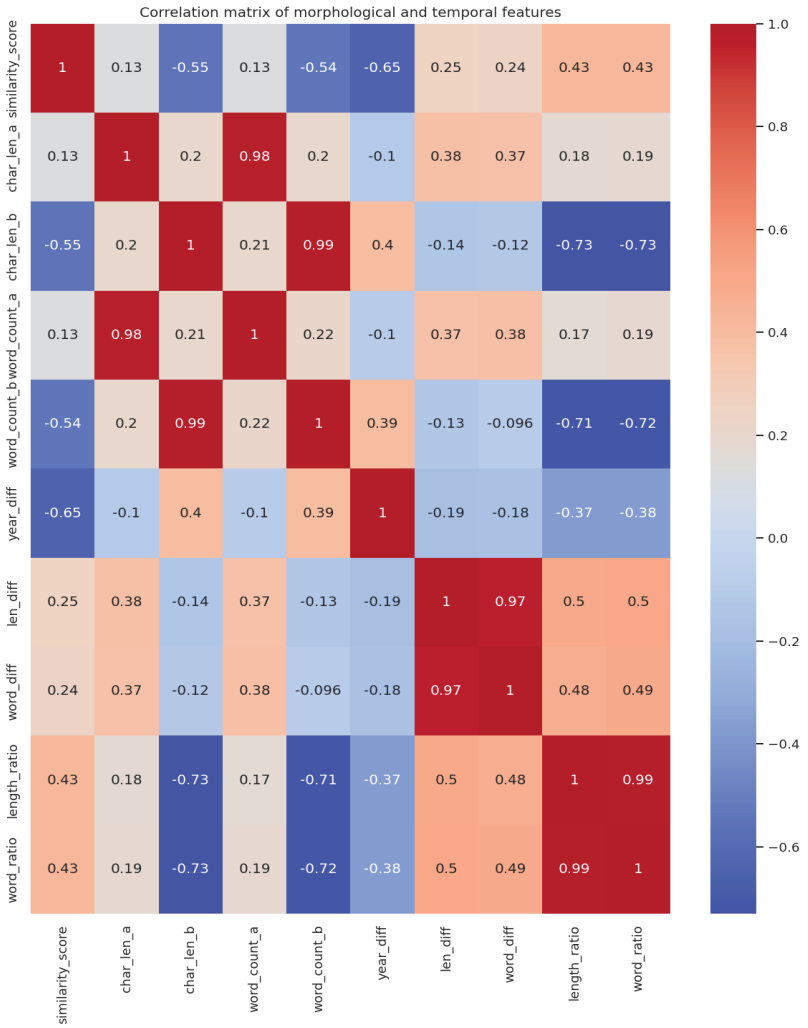


Figure 9. Distribution of abstract pair types

Scientific Journal of Astana IT University
ISSN (P): 2707-9031 ISSN (E): 2707-904X
VOLUME 25, MARCH 2026

Table 1. Distribution of pair types

Category	Subcategory / Pair type	Count (n)	Percentage (%)
Class label	Positive (Near-duplicates)	7,088	59.8%
	Negative (Distinct abstracts)	4,763	40.2%
Total	Full corpus	11,851	100.0%
Pair type	different_abstract	4,763	40.2%
	high_similarity	889	7.5%
	low_similarity	1,268	10.7%
	medium_similarity	4,860	41.0%
	very_low_similarity	71	0.6%
Total	Full corpus	11,851	100.0%



DOI: 10.37943/25NVVS5297

Figure 10. Correlation matrix of morphological and temporal features

In the next paragraph, we consider the formalization of the near-duplicate detection problem, considering optimization, canonicalization, and indexing.

Method for near-duplicate detection: optimization, canonicalization, indexing

Let $D = \{d_1, d_2, \ldots$ be a corpus of scientific documents (abstracts), where each document d_i is represented as a sequence of tokens:

$$d_i = \{w_1, w_2, \ldots \quad , \tag{1}$$

where w_k are tokens and L_i is the length of document d_i.

The dataset is transformed into a set of paired documents:

$$P = \{(d_i, d_j)\}, i \neq j, \tag{2}$$

Each pair is associated with a binary label: $y_{ij} \in \{0,1\}$, $y_{ij} = 1$ indicates a near-duplicate pair and $y_{ij} = 0$ indicates a non-duplicate pair.

The objective is to learn a function:

$$f : D \times D \to \{0,1\}, \tag{3}$$

that maps a pair of documents to a binary decision indicating whether they are near-duplicates. A transformer-based encoder is used to map each document into a dense vector space:

$$E : D \to \square \ , \tag{4}$$

for a given document d its embedding is defined as: $e = E(d)$, where $e \in \square$ is a fixed-length vector representation. For a pair of documents (d_i, d_j) the embeddings are: $e_i = E(d_i)$ and $e_j = E(d_j)$. The encoder corresponds to the Bi-Encoder architecture, where both documents are processed independently using the same model parameters.

The semantic similarity between two documents is computed using cosine similarity:

$$S(d_i, d_j) = \frac{e_i \cdot e_j}{\|e_i\| \cdot \|e_j\|}, \tag{5}$$

$\|e_k\|$ is the Euclidean norms.

The similarity score satisfies: $S(d_i, d_j) \in [-1,1]$ and normalized to $[0,1]$.

A threshold-based decision rule is applied to classify document pairs:

Scientific Journal of Astana IT University
ISSN (P): 2707-9031 ISSN (E): 2707-904X
VOLUME 25, MARCH 2026

$$y_{ij} \in \begin{cases} 1, \text{if } S(d_i, d_j) \geq T \\ 0, \text{if } S(d_i, d_j) < T \end{cases}, \quad (6)$$

Where T is decision threshold and the final classification function with indicator function, I is:

$$f(d_i, d_j) = I(S(d_i, d_j) \geq T), \quad (7)$$

or more formal

$$f(d_i, d_j) = I\left(\frac{E(d_i) \cdot E(d_j)}{\|E(d_i)\| \cdot \|E(d_j)\|} \geq T^*\right), \quad (8)$$

T^* is the optimal threshold.

The threshold T is selected based on validation data to maximize classification performance. Let $\Delta \subset P$ be the validation set. The optimal threshold T^* is defined as:

$$T^* = \arg\max_{T \in [T_{mn}, T_{mx}]} \frac{1}{|\Delta|} \sum_{(i,j) \in \Delta} I(y_{ij}(T) = y_{ij}), \quad (9)$$

In practice, the value $T \in (0.5, 0.9)$.

To address the scalability problem, the task of near-duplicate detection can be reformulated as an optimization problem. Let $D = \{d_1, d_2, \ldots$ be a corpus of scientific documents. The traditional approach requires computing similarity for all possible pairs of documents. The complexity of such an approach $O(N^2)$. Such complexity is unacceptable for large-scale corpora. However, preliminary filtering of candidate pairs can be performed. In this case, the number of pairs after filtering will be significantly smaller than the total number of pairs. However, it is important that sufficiently high near-duplicate detection performance is maintained on the filtered pairs, that is $\text{Recall} > R_{min}$, $\text{Precision} > P_{min}$.

To reduce the impact of morphological variability of the Kazakh language, a canonicalization function is introduced. Each document d_i is transformed into a normalized representation that includes the following steps: lowercasing, removal of punctuation and non-informative symbols, removal of stop words, and morphological normalization (stemming or lemmatization). Due to the agglutinative nature of the Kazakh language, a single root can generate many word forms. Canonicalization allows reducing lexical variability and improving the consistency of similarity computation.

To avoid exhaustive pairwise comparison, an indexing mechanism is introduced. Each document is encoded into a dense vector representation using a transformer-based model $\lambda_i = E(d_i)$. All vectors are stored in the index $\Lambda = (\lambda_1, \lambda_2, \ldots, \lambda_N)$. For each document, a neighborhood set is determined using approximate nearest neighbor search. This makes it possible to reduce the number of candidate pairs from $O(N^2)$ to approximately $O(N \log N)$. This ensures the scalability of the approach for large datasets.

A multi-stage candidate filtering strategy is applied to reduce the number of comparisons. It consists of three stages:

Stage 1: Lexical filtering. Pairs with low lexical similarity are discarded.

Stage 2: Index-based selection. The top k candidate documents are selected from the vector representation index using the nearest neighbor method.

DOI: 10.37943/25NVVS5297

Stage 3: Semantic similarity computation. The final similarity is computed using the cosine measure between vector representations.

Thus, the overall computational complexity of the proposed approach is $O(N \log N + kN)$, where k is the number of candidates selected for each document k . Compared to the exhaustive approach, in which the computational complexity is $O(N^2)$, the proposed approach provides a significant reduction in computational costs.

Architectural framework of BERT-based detection pipeline

The proposed methodology is based on a multi-stage computational pipeline for semantic near-duplicate detection in scientific abstracts. The architecture, illustrated in Figure 11, is based on a transformer-based Bi-Encoder framework and consists of five sequential stages: input data processing, embedding generation, vector similarity computation, dynamic threshold calibration, and final classification. The pipeline is designed to ensure both semantic expressiveness and computational efficiency through independent encoding of text pairs and reuse of precomputed embeddings.

The architecture is structured as follows:

1. Input stage. The process begins with the Input corpus, consisting of cleaned and paired Kazakh/multilingual scientific abstracts. The data is partitioned into training, validation, and test subsets to ensure unbiased performance evaluation and to prevent data leakage. Each instance represents a pair of texts (A, B), prepared for similarity analysis. The preprocessing pipeline includes: lowercasing, removal of non-informative characters, whitespace normalization, basic tokenization. This step ensures consistency of textual representations and reduces noise for downstream processing. The dataset is split into: training set, validation set, test set. A stratified sampling strategy is applied to preserve class distribution and prevent data leakage.

2. Embedding stage. A Bi-Encoder architecture serves as the primary feature extraction engine. Two identical encoders are applied: BERT encoder A (for text A), BERT encoder B (for text B). It utilizes the paraphrase-multilingual-MiniLM-L12-v2 transformer model to encode abstract pairs independently. To optimize computational resources and avoid redundant GPU operations, an Embedding serialization layer is integrated to cache and retrieve dense feature vectors.

3. For each text pair (d_i, d_j) the model generates embeddings $e_i = E(d_i)$ and $e_j = E(d_j)$. This independent encoding enables efficient processing, as each document embedding can be computed once and reused across multiple comparisons.

4. Embedding Serialization. As illustrated in Figure 11, an embedding serialization module is integrated into the pipeline. This component: stores precomputed embeddings, avoids repeated forward passes through the transformer, significantly reduces GPU usage. This design choice is critical for scalability, especially when processing large corpora.

5. Vector Pair Construction and Similarity Computation. After encoding, the embeddings are combined into a vector pair (e_A, e_B) which serves as the input for similarity computation. The semantic similarity between texts is calculated using cosine similarity (Cosine distance metric in Fig. 11): $S(A,B) = \frac{e_A \cdot e_B}{\|e_A\| \cdot \|e_B\|}$. Cosine similarity is chosen due to its effectiveness in high-dimensional embedding spaces and its robustness to lexical and structural variations.

Scientific Journal of Astana IT University
ISSN (P): 2707-9031 ISSN (E): 2707-904X
VOLUME 25, MARCH 2026

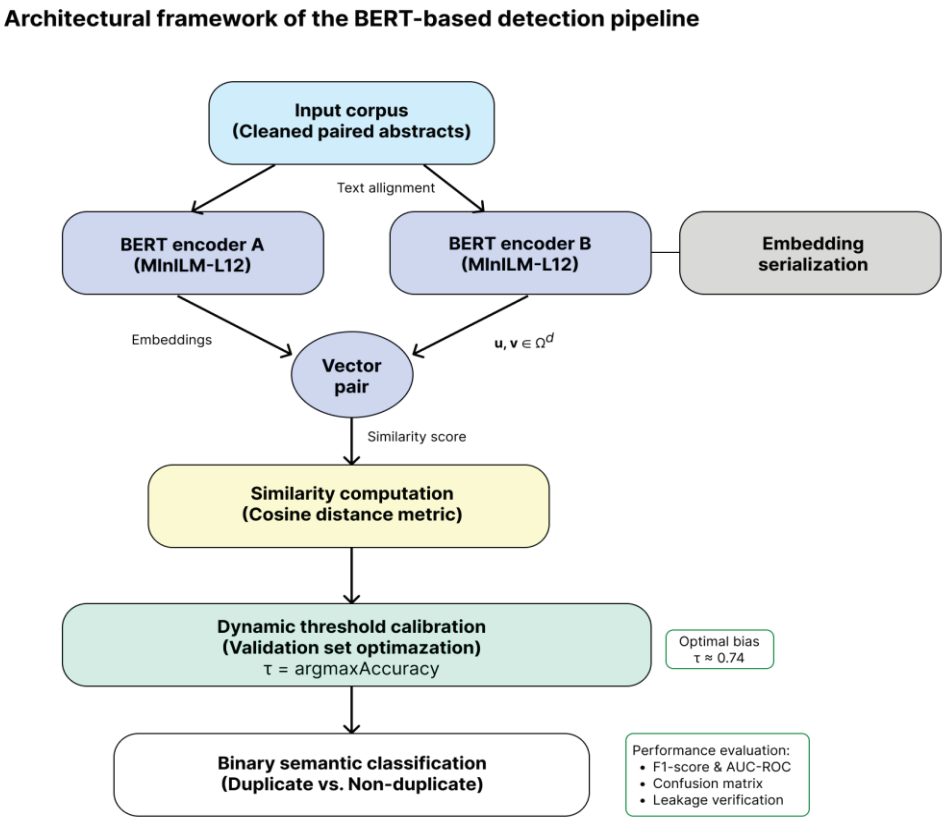


Figure 11. Architectural framework of BERT-based detection pipeline

6. Dynamic Threshold Calibration. A key component of the pipeline is the dynamic threshold calibration stage, which determines the optimal decision boundary for classification. Instead of using a fixed threshold, the system performs optimization on the validation set. The threshold T is selected by maximizing classification accuracy (9). As indicated on Fig. 11, the optimal threshold (e.g., $T \approx 0.74$) is determined empirically. This adaptive calibration is necessary due to the overlap in similarity distributions between duplicate and non-duplicate classes, which makes fixed thresholds unreliable.

7. Binary Classification. The final stage performs binary semantic classification (Fig. 11), assigning each text pair to one of two classes: duplicate and non-duplicate. The decision rule is defined as (6).

8. Performance Evaluation. The effectiveness of the system is evaluated using a comprehensive set of metrics, as indicated in the figure: F1-score, AUC-ROC, confusion matrix, leakage verification. These metrics provide a complete assessment of classification quality, including both precision-recall balance and robustness of predictions. The use of a separate validation set for threshold calibration ensures unbiased performance estimation.

Results

The paper proposes an approach that enables near-duplicate detection, taking into account the morphological variability of agglutinative languages, in particular the Kazakh language, using contextual semantic representations based on BERT and analyzing statistical properties, in particular the overlap of similarity distributions. This makes it possible to move from heuristic methods to more effective models, which is critically important for near-duplicate detection, particularly in scientific texts.

DOI: 10.37943/25NVVS5297

This bar chart compares the primary evaluation metrics (Accuracy, Precision, Recall, F1-score) for the training, validation, and test sets. It provides an empirical assessment of the model's generalization capability. Higher scores indicate better performance, while differences between training and test sets reveal potential overfitting or underfitting (Fig. 12).

The heatmap (Fig. 13) displays the classification performance on the test set, showing counts of true positives, true negatives, false positives, and false negatives. It allows visual evaluation of model errors, highlighting misclassified duplicates and non-duplicates. A balanced matrix along the diagonal indicates accurate predictions.

In addition to classification quality, the efficiency of the proposed optimization framework was evaluated. The number of candidate pairs after filtering was significantly reduced compared to full pairwise comparison. The proposed approach demonstrates a substantial reduction in computational costs while maintaining high detection accuracy. The analysis of execution time shows that the proposed method provides a significant speedup compared to the baseline approach based on exhaustive comparison. These results confirm that the integration of indexing, canonicalization, and candidate filtering effectively improves the scalability of the approach.

Discussion

An approach to near-duplicate detection in Kazakh-language scientific texts based on BERT-like models is proposed, which takes into account semantic relationships between texts regardless of their morphological form. A dataset of pairs of scientific texts in the Kazakh language with balanced similarity levels was formed and analyzed. An experimental evaluation of the effectiveness of the BERT-based approach was conducted, demonstrating its advantage over lexical methods under conditions of high morphological variability. The obtained results demonstrate that the proposed framework successfully achieves a balance between detection accuracy and computational efficiency. Although semantic similarity models based on transformer embeddings provide high-quality detection of near duplicates, their direct application to large-scale corpora is computationally expensive. The integration of indexing and candidate filtering significantly reduces the number of pairwise comparisons, making the approach scalable.

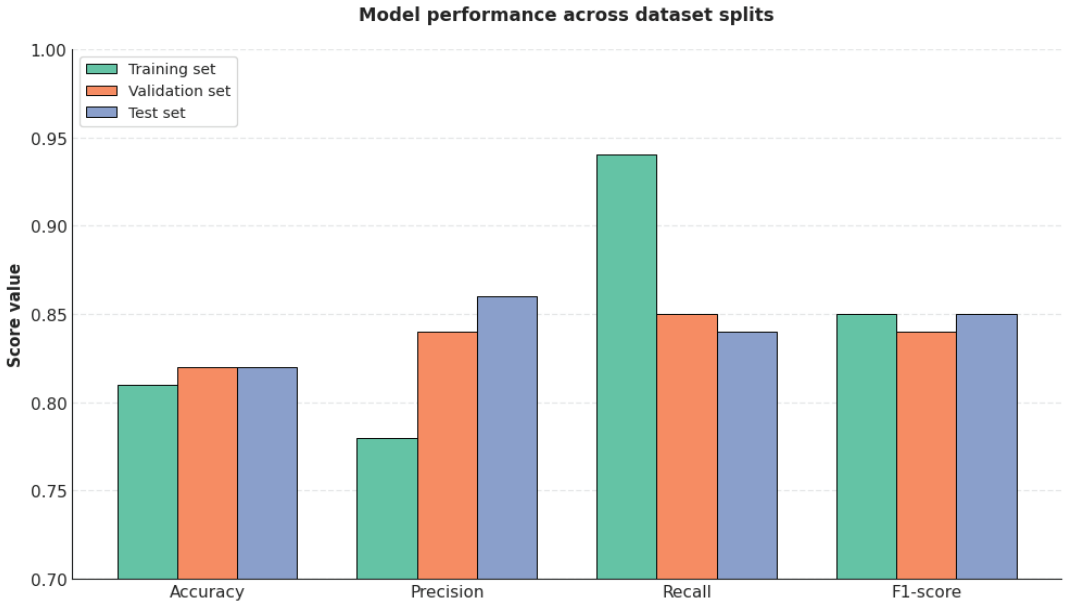


Figure 12. Model performance across dataset splits

Scientific Journal of Astana IT University
ISSN (P): 2707-9031 ISSN (E): 2707-904X
VOLUME 25, MARCH 2026

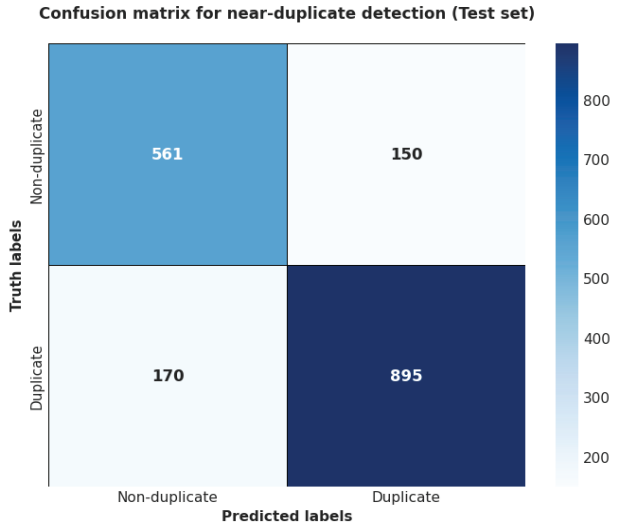


Figure 13. Confusion matrix for duplicate detection

An important observation is the trade-off between efficiency and recall. More aggressive filtering strategies reduce computational costs but may eliminate some valid near-duplicate pairs. Therefore, the selection of parameters, such as the number of nearest neighbors k and lexical filtering thresholds, should be carefully tuned depending on the application scenario. The analysis of similarity value distributions shows significant overlap between duplicate and non-duplicate classes. This confirms that the use of fixed classification thresholds is not optimal, and dynamic threshold calibration is required to achieve stable performance.

Canonicalization plays a key role in addressing the morphological variability of the Kazakh language. It increases the stability of lexical similarity metrics and improves the effectiveness of early filtering stages. At the same time, excessive normalization may lead to the loss of important linguistic information, which should be taken into account when designing preprocessing pipelines. The proposed multi-stage framework demonstrates that the combination of lexical filtering, embedding-based selection, and semantic similarity computation makes it possible to achieve both high accuracy and computational efficiency. This is especially important for low-resource agglutinative languages, where it is necessary to consider both linguistic complexity and data scarcity.

It should be noted that the study has certain limitations. The research focuses on scientific abstracts, which have a relatively formal and structured style. The proposed approach may require additional adaptation for other types of texts, such as social media or informal documents. In addition, the canonicalization process is based on general normalization methods and does not fully take into account advanced morphological analysis specific to the Kazakh language. More sophisticated linguistic processing may further improve the quality of the results.

Another limitation is related to the use of partially synthetic positive pairs. Although the augmentation procedure increases dataset diversity and enables controlled similarity levels, synthetic transformations may not capture the full range of naturally occurring paraphrasing strategies in real plagiarism or text reuse scenarios. For this reason, the reported results should be interpreted as evidence of the effectiveness of the proposed framework under a controlled experimental setting, while further validation on fully real-world annotated corpora remains an important direction for future work.

Conclusion

This paper presents a scalable framework for near-duplicate detection in Kazakh scientific texts, which combines semantic similarity modeling with optimization methods, including indexing, canonicalization, and candidate filtering. The proposed approach addresses two key problems: the high computational complexity of pairwise document comparison and the morphological variability of agglutinative languages. By combining

multi-stage filtering with transformer-based semantic representations, the framework significantly reduces the number of comparisons while maintaining high detection accuracy.

Experimental results demonstrate that the integration of indexing and reuse of embeddings improves scalability, while canonicalization increases the robustness of similarity estimation. The use of dynamic threshold calibration further improves classification quality under conditions of overlapping similarity distributions. The results obtained confirm that the combination of linguistic preprocessing and computational optimization is necessary for effective near-duplicate detection in low-resource agglutinative languages.

Future work will focus on integrating advanced morphological analysis for the Kazakh language, exploring more efficient indexing structures, and extending the approach to multilingual and cross-lingual duplicate detection scenarios.

Acknowledgement

This research work was carried out within the framework of the scientific project AP23490123 «Development of a system to detect plagiarism using combined methods and models for finding near-duplicate, focusing on the Kazakh language» for 2024-2026, financed by the Committee of Science, Ministry of Science and Higher Education of the Republic of Kazakhstan.

References

[1] Agglutinative language. (n.d.). In Glossary of Linguistic Terms. SIL International. Retrieved November 2, 2025. https://glossary.sil.org/term/agglutinative-language

[2] Kessikbayeva, G., et al. (2014). Rule based morphological analyzer of Kazakh language. In Proceedings of the Joint SIGMORPHON/SIGFSM Workshop (ACL Anthology). https://aclanthology.org/W14-2806.pdf

[3] Washington, J., et al. (2014). Finite-state morphological transducers for three Kypchak languages. In Proceedings of the 9th International Conference on Language Resources and Evaluation (LREC 2014). http://www.lrec-conf.org/proceedings/lrec2014/pdf/1207_Paper.pdf

[4] Makhambetov, O., et al. (2014). Toward a data-driven morphological analysis of Kazakh language. Türkiye Bilişim Vakfı Bilgisayar Bilimleri ve Mühendisliği Dergisi, 7(2). https://dergipark.org.tr/tr/download/article-file/395210

[5] Yiner, Z., et al. (2021). Two level Kazakh morphology. Acta Infologica. https://doi.org/10.26650/acin.842758

[6] Assylbekov, Z., et al. (2016). A free/open-source hybrid morphological disambiguation tool for Kazakh. In TurCLing 2016. https://doi.org/10.13140/RG.2.2.12467.43045

[7] Budur, E., et al. (2020). Data and representation for Turkish natural language inference. In Proceedings of EMNLP 2020. https://aclanthology.org/2020.emnlp-main.662.pdf

[8] Ercan, G., et al. (2018). AnlamVer: Semantic model evaluation dataset for Turkish – word similarity and relatedness. In Proceedings of COLING 2018. https://aclanthology.org/C18-1323.pdf

[9] Alkurdi, B., et al. (2022). Semantic similarity based filtering for Turkish paraphrase dataset creation. In Proceedings of ICNLSP 2022. https://aclanthology.org/2022.icnlsp-1.14.pdf

[10] Dehghan, S., et al. (2025). A Turkish dataset and BERTurk-contrastive model for semantic textual similarity. Journal of Information Systems and Telecommunication. https://doi.org/10.61186/jist.48127.13.49.24

[11] Biloshchytska, S., et al. (2025). Text similarity detection in agglutinative languages: A case study of Kazakh using hybrid n-gram and semantic models. Applied Sciences, 15(12), 6707. https://doi.org/10.3390/app15126707

[12] Reimers, N., et al. (2019). Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In Proceedings of EMNLP-IJCNLP 2019. https://public.ukp.informatik.tu-darmstadt.de/UKP_Webpage/publications/2019/2019_EMNLP_NR_SentenceBert.pdf

[13] Kuchanskyi, O., et al. (2026). Hierarchical ensemble framework for detecting paraphrased near duplicates in scientific abstracts. In CEUR Workshop Proceedings. https://ceur-ws.org/Vol-

Scientific Journal of Astana IT University
ISSN (P): 2707-9031 ISSN (E): 2707-904X
VOLUME 25, MARCH 2026

4155/paper05.pdf

[14] Bhoi, S., Markhedkar, S., Phadke, S., & Agrawal, P. (2024). MultiSiam: A Multiple Input Siamese Network For Social Media Text Classification And Duplicate Text Detection. https://doi.org/10.48550/arXiv.2401.06783

[15] Lizunov, P., Biloshchytskyi, A., Kuchanskyi, O., Andrashko, Y., Biloshchytska, S., & Serbin, O. (2021). Development of the combined method of identification of near duplicates in electronic scientific works. Eastern-European Journal of Enterprise Technologies, 4, 57–63. https://doi.org/10.15587/1729-4061.2021.238318

[16] Amirzhanov, A., Turan, C., & Makhmutova, A. (2025). Plagiarism types and detection methods: A systematic survey of algorithms in text analysis. Frontiers in Computer Science, 7, 1504725. https://doi.org/10.3389/fcomp.2025.1504725

[17] Shahmohammadi, H., Dezfoulian, M. H., & Mansoorizadeh, M. (2021). Paraphrase detection using LSTM networks and handcrafted features. Multimedia Tools and Applications, 80(4), 6479–6492. https://doi.org/10.1007/s11042-020-09996-y

[18] Zhang, Y. (2025). An ensemble deep learning model for author identification through multiple features. Scientific Reports, 15, 26477. https://doi.org/10.1038/s41598-025-11596-5

[19] Agarwal, B., Ramampiaro, H., Langseth, H., & Ruocco, M. (2018). A deep network model for paraphrase detection in short text messages. Information Processing and Management, 54(6), 922–937. https://doi.org/10.1016/j.ipm.2018.06.005

[20] Iqbal, H. R., Maqsood, R., Raza, A. A., & Hassan, S. U. (2023). Urdu paraphrase detection: A novel DNN-based implementation using a semi-automatically generated corpus. Natural Language Engineering, 30, 354–384. https://doi.org/10.1017/S1351324923000189

[21] Mehak, G., Muneer, I., & Nawab, R. M. A. (2023). Urdu Text Reuse Detection at Phrasal Level Using Sentence Transformer-Based Approach. Expert Systems with Applications, 234, 121063. https://doi.org/10.1016/j.eswa.2023.121063

[22] Kuchanskyi, O., & Kazagasheva, V. (2026). Kazakh scientific publications dataset from Semantic Scholar (2000–2025) [Data set]. Zenodo. https://doi.org/10.5281/zenodo.18672817

DOI: 10.37943/25NVVS5297

Appendix

Table A1. High-Similarity Pairs (score 0.840–0.985; n = 889, 7.5%). These pairs share the same source abstract. Augmentation applied light transformations: lowercasing, removal of punctuation marks and a small set of stopwords, occasional single-word synonym substitution, or minor word-order shifts. Semantic content and most surface tokens are preserved. Pair type: positive (near-duplicate). Similarity range: 0.840–0.985.

ID	Score	Rephrase level	Abstract A	Abstract B	Key differences
HS-01	0.985	light	Мақалада тың архив деректерінің негізінде Қазақстанның академиялық ғылымының қалыптасу тарихы және С. Асфендиаровтың Қазақ АКСР-дағы тұңғыш ғылыми-зерттеу орталықтары мен мекемелерін құрудағы рөлі зер...	мақалада тың архив деректерінің негізінде қазақстанның академиялық ғылымының қалыптасу тарихы асфендиаровтың қазақ акср дағы тұңғыш ғылыми зерттеу орталықтары мекемелерін құрудағы рөлі зерттелген күнг...	Capitalization removed; punctuation and a few connective words dropped; one synonym substitution (өзекті мәселе → өзекті қиындық)
HS-02	0.962	very light	Мақалада қазіргі қазақ прозасында көрініс тапқан ұлттық құндылық, ұлттық болмыс, ұлттық салт-дәстүр мәселесі кеңінен сөз болады. Зерттеу пәніне қазіргі қазақ прозасындағы туындылар сарапталып алынған....	ұлттық мақалада қазіргі қазақ прозасында көрініс тапқан құндылық болмыс салт дәстүр мәселесі кеңінен сөз болады анықтау пәніне қазіргі қазақ прозасындағы туындылар сарапталып алынған тәуелсіздік кезең...	Word-order resequencing at sentence start; removal of repeated ұлттық modifiers and superlative ең; punctuation stripped
HS-03	0.961	light	Қай кезеңде болмасын мерзімді басылымдар ақиқат өмірдің айнасы бола білді. Онда жарияланған публицистикалық шығармалардан сол кезеңнің бет-бейнесін, заман тарихын білеміз. Мақалада «Қазақстан коммунис...	қай кезеңде болмасын мерзімді басылымдар ақиқат өмірдің айнасы бола білді онда жарияланған публицистикалық шығармалардан кезеңнің бет бейнесін заман тарихын білеміз мақалада қазақстан коммунисі ақиқат...	Lowercase normalisation; hyphens and guillemets removed; determiner сол dropped; extra word ақиқат inserted in title
HS-04	0.954	light	Мақалада аударматанудағы ең күрделі мәселелердің бірі – аударма мәтінінің коммуникативтік әсер тудыру, реципиентке прагматикалық әсер ету қабілеті болып табылатын түпнұсқа мәтіннің прагматикалық әлеуе...	әсер прагматикалық мақалада аударматанудағы күрделі мәселелердің бірі аударма мәтінінің коммуникативтік тудыру реципиентке ету қабілеті болып табылатын түпнұсқа мәтіннің әлеуетін беру мәселесі қараста...	Word-order scrambling of noun phrase; superlative ең removed; synonym қарасталады for қарастырылады
HS-05	0.938	light	Мақалада қазіргі таңда мектептегі оқытудың маңызды міндеттерінің бірі болып табылатын оқушыларда функционалдық сауаттылықты қалыптастыру	мақалада қазіргі таңда мектептегі оқытудың маңызды міндеттерінің бірі болып табылатын оқушыларда функционалдық сауаттылықты	Second sentence merged; conjunction мен dropped between жолдары and тәсілдері; repeated Мақалада removed

Scientific Journal of Astana IT University
ISSN (P): 2707-9031 ISSN (E): 2707-904X
VOLUME 25, MARCH 2026

			мәселесі сөз болады. Мақалада функционалдық сауаттылықты оқушыл...	қалыптастыру мәселесі сөз болады функционалдық сауаттылықты оқушыларда қалып...	

Table A2. Medium-Similarity Pairs (score 0.620–0.839; n = 4,860, 41.0%). The largest category. Augmentation used medium-level rephrasing: sentence reordering, clause restructuring, substitution of content words, and foregrounding of different sections of the same abstract. The topic is clearly shared but surface overlap is substantially reduced. Pair type: positive (near-duplicate). Similarity range: 0.620–0.839.

ID	Score	Rephrase level	Abstract A	Abstract B	Key differences
MS-01	0.797	medium	Бұл мақалада техникалық эстетикадағы беткі қабат түсінігі және оның әрлеу мен қаптау материалдарының құрылымындағы рөлі талқыланады. Беткі қабаттың функционалдық, эстетикалық және символдық қасиеттері...	беткі қабаттың функционалдық эстетикалық символдық қасиеттері санаттарының негізінде анықталады мақалада уақытта жаңа сапалы деңгейге көшкен материалдардың техникалық жобалауымен байланысты әртүрлі өн...	Sentence reordering; introductory clause removed; result clause restructured as informational statement; punctuation stripped
MS-02	0.785	medium	Құқықтық саланы цифрландыру дәуірінде қылмыстық сот ісін жүргізуде қолданылатын дәлелдемелердің мазмұны мен іс жүргізу маңыздылығына әсер еткен елеулі өзгерістері орын алды. Қылмыстықістер бойынша дәл...	құқықтық әдістері жеке қазақстан қылмыстық нормативтік нормалардың салыстырмалы логикалық қолданылған зерттеу аналитикалық доктриналдық лексикалық түсіндірмесі мәселенің бірлігі процессуалдық нысаны к...	Heavy lexical substitution; methodology section foregrounded instead of background; topic keywords retained but restructured
MS-03	0.736	medium	Ұсынылып отырған зерттеу тәжірибесі қазіргі таңдағы – ғылым мен техниканың дамып, өркениеттің өркендеген шағында мектеп қабырғасындағы жасөспірімдерді рухани-адамгершілік тұрғысынан тәрбиелеудің маңыз...	жеткізе алмауы жүйелі ұсынылып отырған қарастыру тәжірибесі қазіргі таңдағы ғылым техниканың дамып өркениеттің өркендеген шағында қабырғасындағы мектеп жасөспірімдерді рухани адамгершілік тұрғысынан т...	Prefix clause inserted; word order shuffled; hyphen in compound adjective removed; conjunction мен dropped
MS-04	0.728	medium	Ұлтымыздың XIX ғасырдағы әдебиетінің шынайы көрінісі, ел тәуелсіздігі жолындағы өзекті арқауы – ұлы мүдде. Патшалық Ресейдің тұсында да, Советтік саясат кезінде	Көркем әдебиет пен қоғамдық-әлеуметтік құбылыстардың сабақтастығы, сонымен қатар ұлттық әдебиеттің ақтаңдақ беттері тұңғыш рет ғылыми түрде негізделіп, тарихи-	Topic (Mağjan Jumabayev) retained; framing shifted from national spirit to literary-historical analysis; new academic context added

DOI: 10.37943/25NVVS5297

			де ұлттық рухымызды сақтаған Мағжан Жұм…	мәдени контексте Мағжан Жұмабаев поэзиясы…	
MS-05	0.620	medium	Бұл ғылыми зерттеу жұмысы озонаторларды басқару жүйелерінің тиімділігін бағалау мен деректерді беру үшін сенсорлық желіні пайдалану тиімділігін зерттейді. Озон концентрациясын өлшейтін зондтар мен бақ…	Ғылыми жұмыс озонмен ауа немесе суды залалсыздандыру үрдісін бақылау кезінде сенсорлардың сымсыз желісінің конфигурациясындағы деректерді енгізу мен тарату тиімділігін қарастырады. Деректер берудің оң…	Application domain shifted (management → disinfection monitoring); sensor network described via different angle; conclusion clause added

Table A3. Low-Similarity Pairs (score 0.400–0.620; n = 1,268, 10.7%). Strong augmentation: major lexical substitution, removal of named entities (author names, country names), perspective shift, or addition of new methodological framing absent from the original. Domain keywords are the primary signal linking the two texts. Pair type: positive (near-duplicate). Similarity range: 0.400–0.620.

ID	Score	Rephrase level	Abstract A	Abstract B	Key differences
LS-01	0.587	strong	Бұл мақалада детектив шығармалардың өзіне тән жанрлық ерекшеліктері мен әлем әдебиетіндегі орны зерделеніп, қазіргі қазақ әдебиетінде детектив жанрында қалам тартып жүрген қаламгер Жадыра Шамұратованы…	кейіпкер сезімге беруші ерік басты түйсігі логикасы аналитикалық ойлауы ерекше дамыған тұлға бейнеленеді детективтік шығармалардың басы қасында әдетте жұмбақ оқиғаларды қылмыстарды тергеу орын алады ж…	Author name and work title removed; shifted to genre theory and character description; only domain keywords (детективтік, оқиғаларды) shared
LS-02	0.579	strong	Мемлекеттік қызмет атқару қоғам мен мемлекет тарапынан ерекше сенім білдіру болып табылады және мемлекеттік қызметшілердің моральдық әдептілік бейнесіне жоғары талаптар қойылады.	мемлекеттік қызмет атқару қоғам мемлекет тарапынан ерекше сенім білдіру болып табылады қызметшілердің моральдық әдептілік бейнесіне жоғары талаптар қойылады мемлекеттік қызметшілердің кәсіби міндеттер…	Opening retained with stopwords removed; additional clause on professional ethics added; conjunction dropped
LS-03	0.559	strong	Бұл мақалада комикс жанрының пайда болу тарихы, әлемдік әдебиет пен мәдениеттегі орны және оның қазіргі білім беру процесіндегі маңызы жан-жақты қарастырылады. Зерттеу барысында АҚШ, Еуропа, Жапония с…	мақалада комикс жанрының пайда болу тарихы әлемдік әдебиет мәдениеттегі орны қазіргі білім беру процесіндегі маңызы жан жақты қарастырылады сондай ақ комикстерді білім беруде қолданудың педагогикалық …	Country examples removed; pedagogical application foregrounded; argument reframed from historical survey to practical justification

Scientific Journal of Astana IT University
ISSN (P): 2707-9031 ISSN (E): 2707-904X
VOLUME 25, MARCH 2026

LS-04	0.549	strong	Мақалада жасанды интеллектінің ауыл шаруашылығындағы логистиканы болжау және басқарудағы рөлі қарастырылады. Заманауи ауыл шаруашылығында деректерге негізделген шешімдер қабылдаудың маңыздылығы негізд...	географиялық ақпараттық жүйе құралдары автоматтандыруға мүмкіндік береді стереофотограмметрияның көмегімен карталарды салу аэроғарыштық суреттерді өңдеу және деректерді талдау әдістері ауыл шаруашылығ...	AI-logistics framing replaced by GIS/remote-sensing framing; agriculture domain retained but methods diverge significantly
LS-05	0.402	strong	Мақалада елдің бәсекеге қабілеттілігіне тәуелділікке және Қазақстан Республикасының мемлекеттік бағдарламаларына сәйкес ұлттық адами капиталды қалыптастырудың мемлекеттік саясатты жетілдіру мәселелері...	тексеру шеңберінде факторлық қарау қазақстан өңірлерінің адами капиталының жетілу деңгейінің статистикалық көрсеткіштерінің жүйесі әзірленді адами капиталдың индексі есептелді аймақтар бойынша салысты...	Policy framing replaced by quantitative regional analysis; shared term адами капитал links both; different methodological angle

Table A4. Very-Low-Similarity Pairs (score 0.000–0.400; n = 71, 0.6%). The rarest positive category. Produced by the most aggressive augmentation: heavy word-order scrambling, large-scale deletion of informative tokens, or switching to a different methodological angle of the same underlying topic. Only a handful of domain-level keywords link A and B. Despite the low score, the ground-truth label is positive (near-duplicate) because both texts originate from the same source abstract. Similarity range: 0.000–0.400.

ID	Score	Rephrase level	Abstract A (excerpt)	Abstract B (excerpt)	Key differences
VL-01	0.378	strong	Қазақстандық білім сапасын бағалау жүйесі тест тапсырмаларын әзірлеуде көпжылғы тәжірибесі негізінде педагогикалық өлшеулер саласында үлкен жетістіктерге қол жеткізіп отыр. Халықаралық ұйымдар мен шет...	оқу сауаттылығының жеке тұлға қоғам өміріндегі рөлінің ерекше маңызы сипатталады оқу сауаттылығы білім алудың негізі өмірінің қажет адам кезеңінде әр болатын дағды болып есептеледі	Education domain retained; assessment system context fully absent in B; only сауаттылық and рөл link the texts; heavy condensation
VL-02	0.384	strong	Қазіргі уақытта макро және микроэлементтер тапшылығының себебі болып табылатын астық негізіндегі тамақ өнімдерінің тағамдық және биологиялық құндылығының жалпы төмендеуі байқалады. Сондықтан, қазіргі ...	қазіргі уақытта макро микроэлементтер тапшылығының себебі тамақ астық өнімдерінің табылатын болып тағамдық биологиялық құндылығының жалпы төмендеуі байқалады қарастыру барысында өнген астықтан сығынды...	Opening clause largely preserved but word order scrambled; B pivots to specific lab extraction method absent from A; biofortification goal dropped

DOI: 10.37943/25NVVS5297

VL-03	0.371	strong	Қазіргі таңда бүкіл әлемде жүріп жатқан жаһандық үдерістер экологиялық, саяси, экономикалық жүйелерде туындап отырған тұрақсыздық пен дағдарыстарды тудыруда. Тұрақты даму мақсаттарын жүзеге асырудағы ...	біріккен ұлттар ұйымының 2030 жылға дейінгі тұрақты өсу мақсаттары қоғамның барлық саласында жаһандық дамудың негізгі тетігіне айналды эмпирикалық тексеру SDG индикаторлары бойынша жүргізілді	Shared domain (global processes, sustainable development); B anchored to SDG framework and empirical testing; A more abstract/theoretical
VL-04	0.333	strong	Мақалада Жолдасбек Құрманқұловтың өмір жолы, шығармашылық кезеңі мен өнегелі ұстаз ретінде тәрбиелеп отырған шәкірттері жайында сөз қозғалады. Жолдасбек Мейрамбекұлы тарих ғылымының дамуына зор үлес қ...	Ұйымдастырушылық қабілеті мен экспедиция уақытындағы қырағылығы жастарды тәрбиелеуде үлкен жетістіктерге жетелегені айтылады. Сонымен қатар, Сырдариядағы экспедиция нәтижелері баяндалады.	Same subject (scholar biography); A covers life and pedagogy, B focuses on fieldwork and expedition results; distinct sentence structures
VL-05	0.277	strong	Ғылыми мақалада қазақ әйелдері киген кимешектің дәстүрлі қазақы ортада алатын орны, сипаты мен қолданыстық ерекшелігі туралы сөз болады. Кимешектің нышандық мағынасы мен ою-өрнек символикасы зерттелед...	Захарова Р.Д. ғалымдар еңбектерімен қатар, мақаланы жазу барысында зерттеушілер Н.И. Лобачева және Х. Арғынбаев еңбектері пайдаланылды. Карутц, кордиерит және отқа төзімді материалдар талқыланады.	A discusses kimeshek (traditional headdress) symbolism; B is a bibliographic/methodology note with unrelated material references — extreme topic drift