Hierarchical ensemble framework for detecting paraphrased near duplicates in scientific abstracts

Oleksandr Kuchanskyi[1,2,*,†], Valeriya Kazagasheva[1,†]

[1] *School of Artificial Intelligence and Data Science, Astana IT University, Mangilik El, Block C1, Astana, 010000, Kazakhstan*
[2] *National Technical University of Ukraine "Igor Sikorsky Kyiv Polytechnic Institute", Beresteiskyi Ave., 37, Kyiv, 03056, Ukraine*

Abstract

The rapid expansion of scientific literature has amplified the need for accurate detection of near duplicates: documents that express the same meaning through different wordings. These semantic duplicates, often caused by paraphrasing or metadata inconsistencies, compromise the integrity of scholarly databases and bias downstream analyses. In this paper, we propose a hierarchical ensemble method that combines traditional lexical similarity metrics, contextual embeddings from transformer-based models, syntactic structural features, and a deep neural meta-learner to detect near duplicates in scientific abstracts. We compiled a domain-specific dataset of 10,000 Kazakhstan-related publications from Semantic Scholar and generated 14,460 abstract pairs using real and synthetic duplication techniques. The proposed method achieved 94.24% accuracy and a 94.80% F1-score, significantly outperforming standard lexical and transformer-only baselines. Our results demonstrate that integrating heterogeneous features yields a robust, scalable, and interpretable solution for semantic duplicate detection. The method is especially suited for regional and low-resource academic collections, supporting higher-quality data curation in bibliometric systems, citation analyses, and systematic reviews.

Keywords

near duplicates detection, natural language processing, transformer models, scientific publication, text data analysis

1. Introduction

The limitless growth of scientific publication databases has set the stage for the bibliographic systems to face a great challenge in managing duplicate and near-duplicate records. In this scenario, near duplicates which are defined as pairs of text segments sharing semantic content but differing considerably in surface form, represent the most difficult detection problem that traditional methods usually cannot solve. The duplications can be caused by either a conscious or an unintentional act of paraphrasing, or the overlap between the old and the new findings to a certain extent, or by simply using different terms for the same concept, or through inconsistent metadata practices in different repositories. The downside of not detecting duplicates is very large: in systematic reviews, redundant entries can at times increase the volume of available evidence and thus bias the synthesis results; in citation analysis, duplication of records can distort impact metrics; and in network-based studies, they can lead to incorrect estimations of authority and influence. This issue is particularly critical for the national research assessment systems, university repositories, and government-supported research monitoring platforms because of the direct impact of data accuracy on analytical outcomes and strategic decision-making.

Traditional methods for detecting duplicates mainly depend on similarity measures calculated at the token level like Jaccard coefficient, cosine similarity, edit distance, and n-gram overlap. Even though these methods are computationally efficient, they still predominantly deal with surface-level correspondence and, thus, are not very robust when semantic content is altered through rephrasing or syntactic restructuring. The last year or two have seen the rapid rise of transformer-based language models like BERT and its receivers that have, to a great extent, improved the modeling of contextual

WDA'26: International Workshop on Data Analytics, January 26, 2026, Kyiv, Ukraine

[*] Corresponding author.

[†] These authors contributed equally.

a.kuchanskyi@astanait.edu.kz (O. Kuchanskyi); 242752@astanait.edu.kz (V. Kazagasheva)
0000-0003-1277-8031 (O. Kuchanskyi); 0009-0001-0262-0305 (V. Kazagasheva)

CEUR Workshop Proceedings ceur-ws.org ISSN 1613-0073
published 2026-02-01

semantics and conceptual similarity going beyond just exact word matching. However, the empirical results indicated by controlled benchmarks often surpass the performance seen in real-life situations which then implies that semantic embeddings alone are not enough to guarantee the trustworthy deployment in production-scale systems. Moreover, it is the case that most of the research done so far has been on English-language scientific corpora or general-purpose plagiarism detection while regional academic publications or low-resource scientific languages have been comparatively ignored.

The aim of this study is to develop a reliable method for detecting near duplicates in the abstracts of scientific publications. To achieve this aim, the following objectives were formulated:

1. To create a specialized dataset of scientific abstracts reflecting real cases of similar texts;
2. To develop a multi-level ensemble model that combines lexical similarity, sequence alignment, semantic embeddings, and syntactic analysis for the detection of paraphrased duplicates;
3. To evaluate the proposed model in comparison with baseline methods in order to confirm its effectiveness.

In this study, a hierarchical ensemble framework that is methodologically grounded and progressively incorporates the traditional similarity measures, the multilingual semantic representations, the syntactic structure analysis, and a learned meta-level feature weighting strategy is the major contribution of the research. The proposed approach is tested on a new dataset that contains 10,000 scientific publications related to Kazakhstan and has been obtained through a thorough API-driven process. A strict pair generation strategy merges the selection of real similarity-based candidates with the implementation of controlled synthetic paraphrasing techniques to recreate near-duplicate scenarios. In addressing the specific linguistic and structural traits of regional scientific repositories, this research presents a solution for semantic duplicate detection that is practical, scalable, and interpretable, thus contributing to the development of more reliable and context-aware bibliographic management systems.

2. Related work and current methods

2.1. Traditional approaches to duplicate detection

Classic techniques used in the identification of duplicates in scientific literature basically depend on the similarity of strings, fingerprinting algorithms, and metrics at the token level. Hybrid methods that integrate SimHash-based fingerprinting with Smith-Waterman sequence alignment have demonstrated high accuracy for precise and almost precise duplicates in clinical documentation, thereby confirming the effectiveness of the combination of fast candidate filtering with accurate alignment verification [1]. Sectional Min-Max hashing improvements of traditional MinHash that introduce local hashing to segments of documents cut down the computational cost significantly while at the same time keeping detection accuracy over the entire range of textual corpora [2].

Classic techniques used in the identification of duplicates in scientific literature basically depend on the similarity of strings, fingerprinting algorithms, and metrics at the token level. Hybrid methods that integrate SimHash-based fingerprinting with Smith-Waterman sequence alignment have demonstrated high accuracy for precise and almost precise duplicates in clinical documentation, thereby confirming the effectiveness of the combination of fast candidate filtering with accurate alignment verification [1]. Sectional Min-Max hashing improvements of traditional MinHash that introduce local hashing to segments of documents cut down the computational cost significantly while at the same time keeping detection accuracy over the entire range of textual corpora [2].

2.2. Machine learning and deep learning architectures

Supervised learning and neural networks that automatically acquire task-related feature representations have greatly increased the accuracy of duplication detection. The social media MultiSiam neural network for duplicate classification proves that transfer to other domains is still difficult because short-form social media texts and academic abstracts differ so much [3]. The application of wavelet-based analysis with clustering for cross-modal duplicate detection in texts and images has good potential, but the computational intensity is a barrier to practical scalability [4].

A comprehensive review over plagiarism detection techniques from 2014 to 2024 showed a clear trend towards semantic and transformer-based methods taking over, emphasizing that simple string-matching methods could not identify complicated text reuse through paraphrasing [5]. The use of a hybrid deep learning model that integrates Long Short-Term Memory (LSTM) networks with manually crafted linguistic features like n-gram overlap and sentence length statistics has proved that the performance of learned and engineered features together is better than that of purely neural models in benchmark paraphrase datasets [6].

Support Vector Machines and Deep Neural Networks are combined in a machine learning method for the identification of authorship, which leads to the extraction of both lexical and stylistic features. As a result, the application of classification-based methods to morphologically rich and low-resource languages is broader [7]. CNN-RNN fused structure for short text paraphrase detection that merges local convolutional patterns with sequential RNN modeling is very effective and through end-to-end learning competitive performance is achieved [8].

2.3. Semantic and transformer-based approaches

Modern methods orienting towards the semantics use embeddings of words and the representations of the contexts for the similarity measures that go beyond token overlap. On the basis of the syntactic parsing and the machine learning, it is possible to identify texts that are semantically equivalent but lexically different through lemmatisation and deep syntactic analysis, which, in turn, provide a foundation for the semantic understanding [9]. The use of embeddings for the words along with the application of triplet loss functions is said to further improve the semantic similarity detection in the duplication of detection frameworks [3]. Sentence-transformer models for the detection of the text reuse at the phrasal level show that the multilingual transformers are capable of effectively capturing deep contextual meaning even in the low-resource languages [10].

Neural network architectures focusing on semantics rather than surface-form matching and trained on semi-automatically produced corpora for paraphrase recognition have shown to be very effective on morphologically rich low-resource languages [11]. A lightweight unsupervised approach utilizing monolingual word embeddings aligned by simple linear projection for cross-lingual semantic similarity requires slight language-specific resources and gives competitive performance on plagiarism detection tasks [12].

A multilingual deep learning framework for cross-lingual plagiarism detection between Arabic and English texts integrating DNNs with semantic features including conceptual similarity and semantic role information demonstrates the necessity of semantic representations in multilingual academic text reuse detection [13].

2.4. Hybrid approaches

Recent work increasingly combines multiple detection modalities to improve both accuracy and robustness. A hybrid methodology merging TF-IDF-based filtering with lightweight DistilBERT embeddings for Ukrainian and Bulgarian academic texts achieves approximately 0.88 F1-score on low-resource morphologically rich corpora, substantially outperforming TF-IDF-only baselines [14]. A layered hybrid model combining lexical similarity (TF-IDF cosine) with semantic embeddings (BERT) to distinguish exact copies from paraphrased content shows experimental results of approximately 80% recall and 74% F1-score [15].

A two-level hybrid cross-language matching scheme incorporating bilingual dictionary-based lexical alignment with semantic similarity layers using multilingual embeddings specifically addresses paraphrased and translated duplicates in multilingual academic contexts [16]. Text similarity measures integrated with density-based clustering within a metaheuristic optimization framework for music lyric plagiarism are adaptable to scientific text reuse detection with near or structurally divergent duplications [17].

Deep semantic features combined with quantum-inspired genetic algorithms, using transformer-based semantic representations optimized through Quantum Genetic Algorithm operators, demonstrate that bio-inspired evolutionary optimization combined with neural embeddings improves detection accuracy and computational efficiency on benchmark plagiarism datasets [18] and [19]. In [20], the proposed method combines statistical and semantic techniques, including N-gram analysis, TF-IDF, LSH, LSA, and LDA, and is benchmarked against the bert-base-multilingual-cased model. In summary, the trend over the last decade has shifted from simple lexical matching towards semantic-rich and hybrid methods. This evolution sets the stage for our ensemble approach, which integrates multiple levels of analysis into a single framework. However, prior studies have rarely focused on detecting partial or paraphrased duplicates in academic publications, especially in regional or low-resource languages. This gap in scientific literature motivates our work.

3. Data collection and dataset construction

3.1. Data source and collection

The dataset was constructed from the Semantic Scholar academic search engine, a free AI-powered platform developed by the Ai2 is the creation of Paul Allen, Microsoft co-founder providing programmatic API access to extensive scientific literature metadata. The Semantic Scholar RESTful API endpoint [21] was used to retrieve publications via structured keyword queries with filtering by language, publication year, and result pagination offsets. A total of 130 systematic keyword queries were designed across thematic categories including core research terminology, computer science and information technology (25 queries), engineering disciplines (13 queries), technology and innovation (8 queries), education (8 queries), healthcare and medicine (12 queries), economics and business (13 queries), energy sector (11 queries), environmental studies (10 queries), agriculture (6 queries), geographic locations (11 queries), national initiatives (7 queries), social sciences (7 queries), and other domains (8 queries).

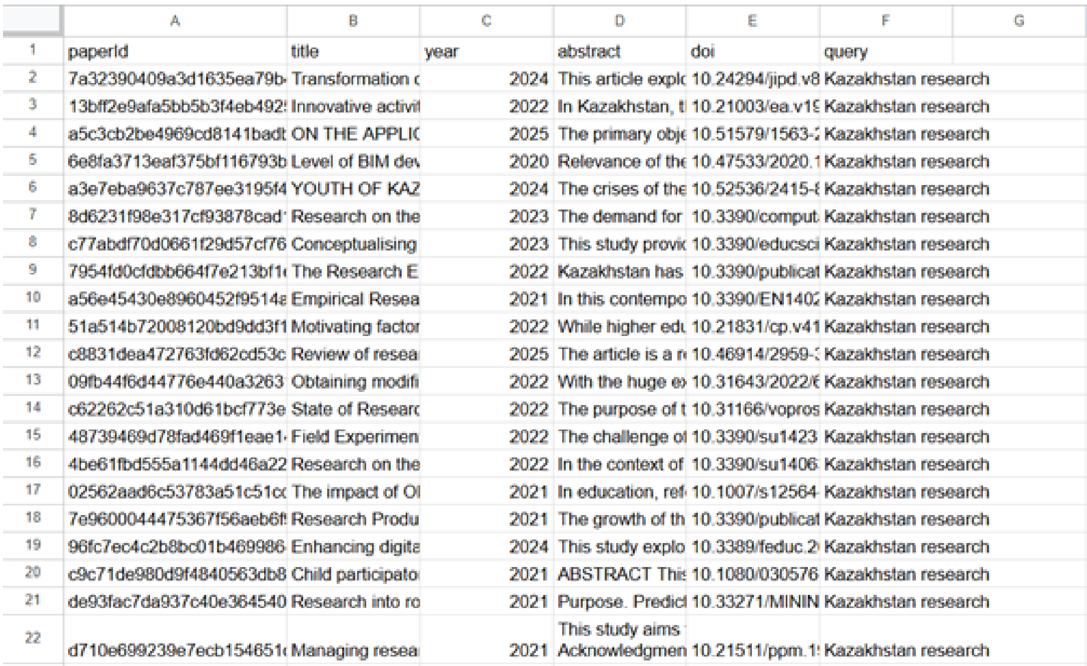
Data quality filtering was applied throughout collection: abstract length minimum of 180 characters ensuring sufficient textual content; keyword presence verification (case-insensitive "Kazakhstan" requirement) ensuring topical relevance; language validation using character-ratio heuristics requiring at least 75% Latin ASCII characters with maximum 10% Cyrillic characters; and deduplication via unique paper identifier tracking to prevent redundant entries.

3.2. Dataset statistics

The final collected dataset 10,000 English-language scientific articles spanning 2012-2025, with publication counts increasing over time and peak activity in recent years.The dataset structure is presented in Figure 1. Mean abstract length across the corpus is approximately 1,500 characters (225 words), with distribution analysis confirming sufficient textual magnitude for semantic similarity analysis [22]. Temporal distribution analysis reveals publication patterns across decade-spanning intervals, with most recent years showing substantially higher publication frequencies in Figure 2. Domain-level analysis shows health and medicine, engineering, education, and computer science as the most productive research categories, reflecting diverse academic coverage as represented in Figure 3.

3.3. Pair generation methodology

Text pairs for model training and evaluation were systematically constructed using a hybrid approach combining real similarity-based pairing and controlled synthetic paraphrasing. Term Frequency-Inverse Document Frequency (TF-IDF) vectorization (1),(2),(3) was applied with English stop-word removal and a 3,000-feature vocabulary, followed by cosine similarity (4) computation between all abstract pairs. Positive pairs (duplicates) were selected from similarity ranges between 0.5 and 0.9, capturing



	A	B	C	D	E	F	G
1	paperId	title	year	abstract	doi	query	
2	7a32390409a3d1635ea79b	Transformation c	2024	This article explc	10.24294/jipd.v8	Kazakhstan research	
3	13bff2e9afa5bb5b3f4eb492	Innovative activit	2022	In Kazakhstan, t	10.21003/ea.v19	Kazakhstan research	
4	a5c3cb2be4969cd8141badt	ON THE APPLIC	2025	The primary obje	10.51579/1563-2	Kazakhstan research	
5	6e8fa3713eaf375bf116793b	Level of BIM dev	2020	Relevance of the	10.47533/2020.1	Kazakhstan research	
6	a3e7eba9637c787ee3195f4	YOUTH OF KAZ	2024	The crises of the	10.52536/2415-8	Kazakhstan research	
7	8d6231f98e317cf93878cad	Research on the	2023	The demand for	10.3390/comput	Kazakhstan research	
8	c77abdf70d0661f29d57cf76	Conceptualising	2023	This study provic	10.3390/educsci	Kazakhstan research	
9	7954fd0cfdbb664f7e213bf1	The Research E	2022	Kazakhstan has	10.3390/publicat	Kazakhstan research	
10	a56e45430e8960452f9514a	Empirical Resea	2021	In this contempo	10.3390/EN1402	Kazakhstan research	
11	51a514b72008120bd9dd3f1	Motivating factor	2022	While higher edu	10.21831/cp.v41	Kazakhstan research	
12	c8831dea472763fd62cd53c	Review of resea	2025	The article is a r	10.46914/2959-3	Kazakhstan research	
13	09fb44f6d44776e440a3263	Obtaining modifi	2022	With the huge e	10.31643/2022/6	Kazakhstan research	
14	c62262c51a310d61bcf773e	State of Researc	2022	The purpose of t	10.31166/vopros	Kazakhstan research	
15	48739469d78fad469f1eae1	Field Experimen	2022	The challenge of	10.3390/su1423	Kazakhstan research	
16	4be61fbd555a1144dd46a22	Research on the	2022	In the context of	10.3390/su1406	Kazakhstan research	
17	02562aad6c53783a51c51c	The impact of Ol	2021	In education, ref	10.1007/s12564	Kazakhstan research	
18	7e96000444753 67f56aeb6f	Research Produ	2021	The growth of th	10.3390/publicat	Kazakhstan research	
19	96fc7ec4c2b8bc01b469986	Enhancing digita	2024	This study explo	10.3389/feduc.2	Kazakhstan research	
20	c9c71de980d9f4840563db8	Child participato	2021	ABSTRACT This	10.1080/030576	Kazakhstan research	
21	de93fac7da937c40e364540	Research into ro	2021	Purpose. Predict	10.33271/MININ	Kazakhstan research	
22	d710e699239e7ecb154651	Managing resea	2021	This study aims Acknowledgmen	10.21511/ppm.1	Kazakhstan research	

Figure 1: The view of collected data file. CSV format containing 10,000 scientific papers with metadata fields: paperId, title, abstract, year, and DOI. Each row represents one scientific article retrieved from Semantic Scholar API.

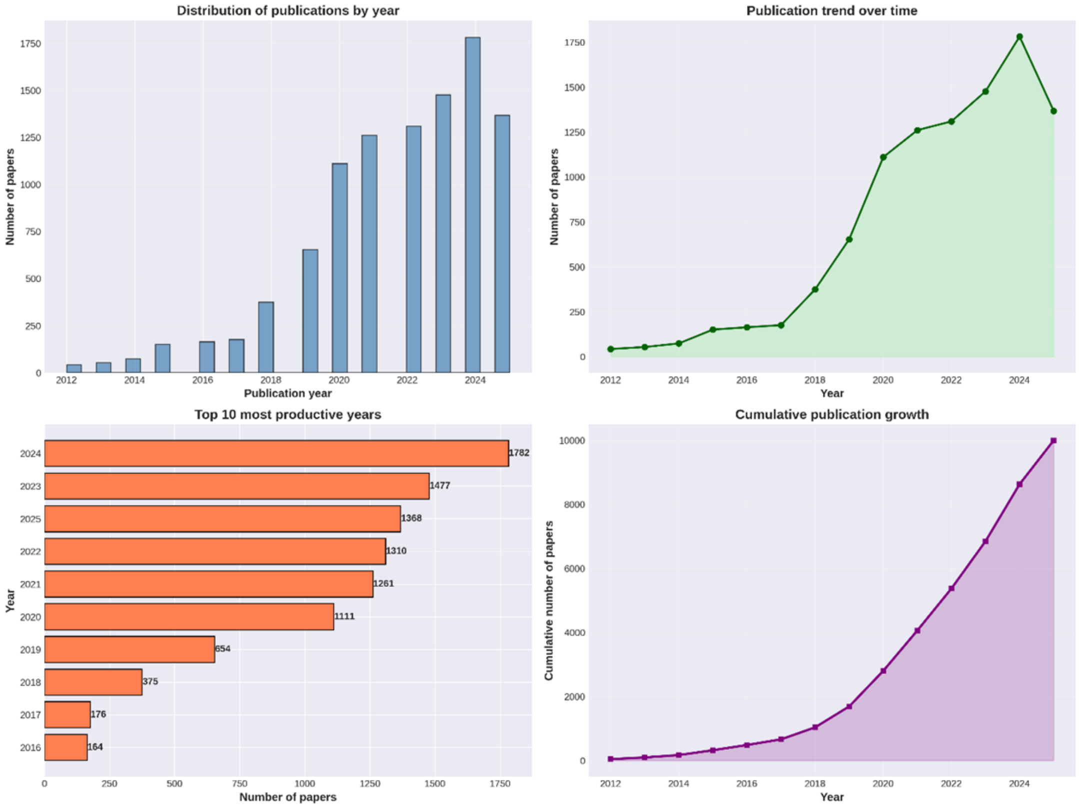


Figure 2: Four-panel visualization showing temporal characteristics of collected dataset: (a) histogram of publication distribution by year with peak in 2024-2025, (b) time series trend of annual publication counts with area shading showing exponential growth, (c) horizontal bar chart of ten most productive years ranked by frequency, (d) cumulative growth curve of publications from 2012 to 2025. Analysis reveals distinct publication patterns and identifies peak research activity periods in Kazakhstan-related scientific literature.

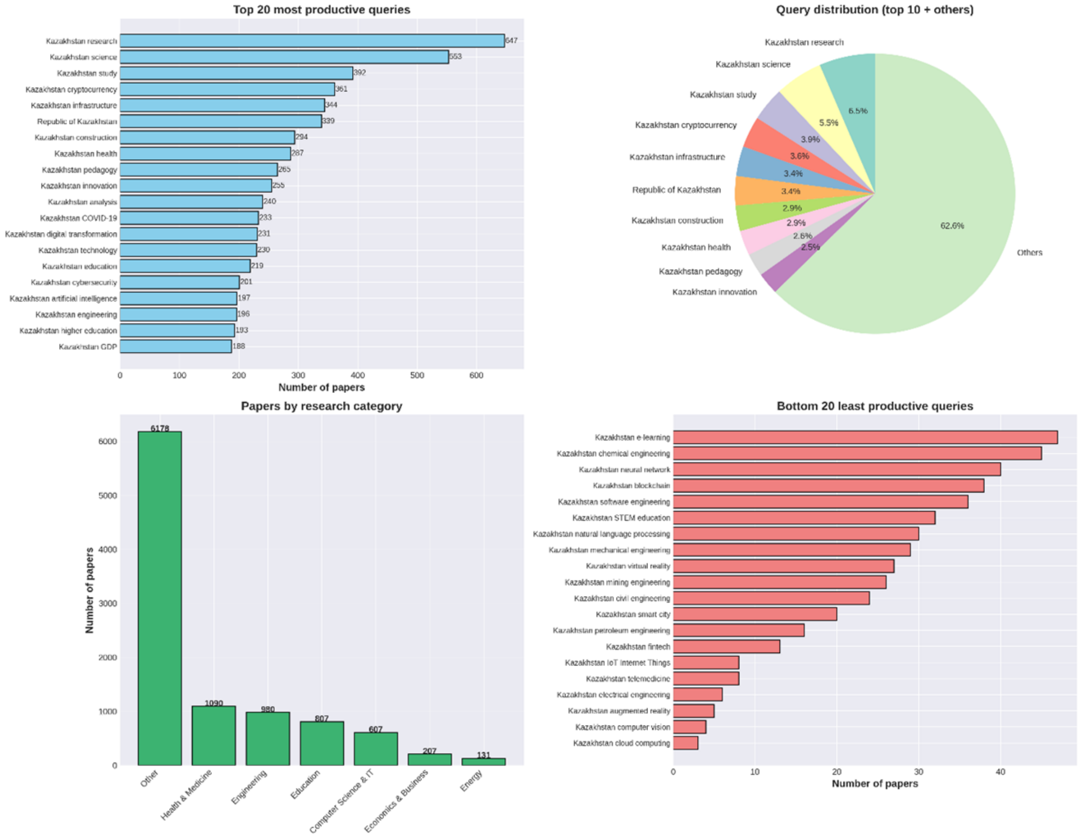


Figure 3: Multi-perspective analysis of data collection queries: (a) top 20 most productive search queries ranked by paper retrieval count (Kazakhstan research with 647 papers), (b) pie chart showing proportional distribution of top 10 queries plus aggregated others, (c) bar chart of papers grouped by research domain categories (Computer Science & IT: 607, Engineering: 980, Health & Medicine: 1090, Education: 807, Other: 6178), (d) bottom 20 least productive queries. Figure demonstrates systematic coverage and domain diversity of collection strategy.

semantically similar yet not identical texts. Negative pairs (non-duplicates) were selected from very low similarity ranges below 0.1, ensuring clear class separation [22].

$$TF(t, d) = \frac{f(t, d)}{\sum_{k \in d} f(k, d)}, \tag{1}$$

where:

- t — term (word),
- d — document,
- $f(t, d)$ — number of occurrences of term t in document d,
- $\sum_{k \in d} f(k, d)$ — total number of terms in document d.

$$IDF(t) = \log \frac{N}{1 + |\{d : t \in d\}|}, \tag{2}$$

where:

- N — total number of documents in the corpus,
- $|\{d : t \in d\}|$ — number of documents containing term t.

$$TF\text{-}IDF(t, d) = TF(t, d) \times IDF(t), \tag{3}$$

where:

- $TF(t, d)$ — term frequency of term t in document d,
- $IDF(t)$ — inverse document frequency of term t.

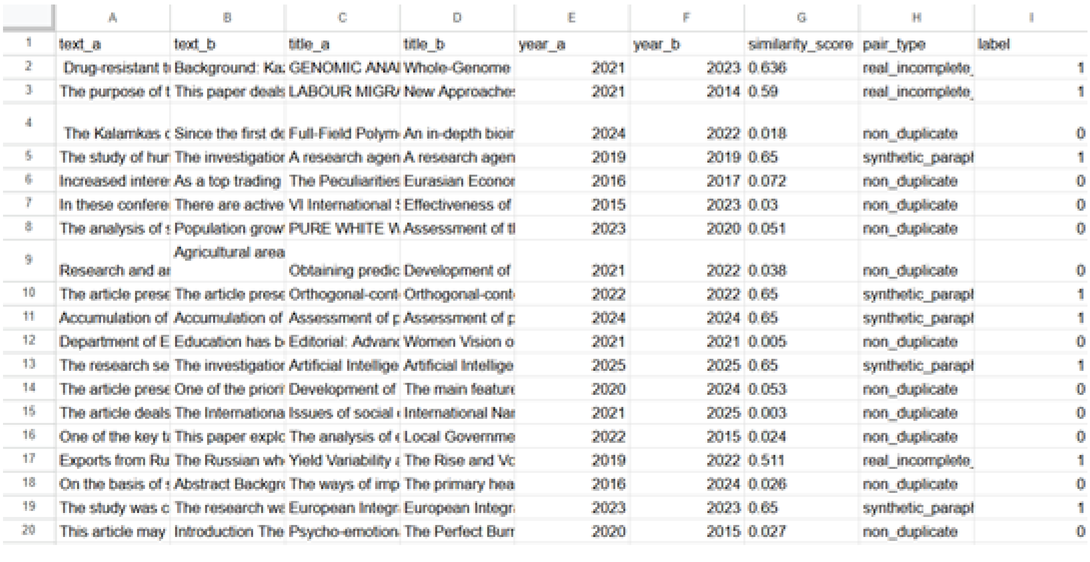
$$\text{CosineSimilarity}(A, B) = \frac{\vec{A} \cdot \vec{B}}{\|\vec{A}\| \, \|\vec{B}\|}, \tag{4}$$

where:

- $\vec{A}, \vec{B}$ — TF-IDF vector representations of two documents,
- $\vec{A} \cdot \vec{B}$ — dot product of the vectors,
- $\|\vec{A}\|, \|\vec{B}\|$ — Euclidean norms (vector lengths).

Real near duplicates were identified through similarity-based selection within the 0.5-0.9 range, yielding 3,453 positive pairs reflecting natural overlap patterns in the collection. Synthetic paraphrases were generated through controlled word replacement at a rate of 25% using a curated synonym dictionary covering 20 common academic terms (e.g., “study”→“research”/“investigation”/“examination”; “method”→“approach”/“technique”/“procedure”; “results”→“findings”/“outcomes”/“conclusions”), producing 4,694 additional positive pairs. Negative non-duplicate pairs (6,313 pairs) were randomly selected from low-similarity combinations, yielding a balanced dataset with 56.3% positive class and 43.7% negative class distribution suitable for stable model training, pair dataset distrubition represented in Table 1 and Figure 5.

The final paired dataset comprises 14,460 text pairs with fields including full abstract texts, corresponding titles and publication years, computed cosine similarity scores, pair type labels (real near duplicate, synthetic paraphrase, or non-duplicate), and binary classification labels (1 for duplicate, 0 for non-duplicate) as showed in Figure 4.



	A	B	C	D	E	F	G	H	I
1	text_a	text_b	title_a	title_b	year_a	year_b	similarity_score	pair_type	label
2	Drug-resistant t	Background: Ka:	GENOMIC ANAI	Whole-Genome	2021	2023	0.636	real_incomplete_	1
3	The purpose of t	This paper deals	LABOUR MIGR/	New Approache:	2021	2014	0.59	real_incomplete_	1
4	The Kalamkas c	Since the first de	Full-Field Polym	An in-depth bioir	2024	2022	0.018	non_duplicate	0
5	The study of hur	The investigatior	A research agen	A research agen	2019	2019	0.65	synthetic_parapl	1
6	Increased intere:	As a top trading	The Peculiarities	Eurasian Econor	2016	2017	0.072	non_duplicate	0
7	In these confere:	There are active	VI International !	Effectiveness of	2015	2023	0.03	non_duplicate	0
8	The analysis of :	Population grow	PURE WHITE W	Assessment of tl	2023	2020	0.051	non_duplicate	0
9	Research and ar	Agricultural area	Obtaining predic	Development of	2021	2022	0.038	non_duplicate	0
10	The article prese	The article prese	Orthogonal-cont	Orthogonal-cont	2022	2022	0.65	synthetic_parapl	1
11	Accumulation of	Accumulation of	Assessment of p	Assessment of p	2024	2024	0.65	synthetic_parapl	1
12	Department of E	Education has b	Editorial: Advanc	Women Vision o	2021	2021	0.005	non_duplicate	0
13	The research se	The investigatior	Artificial Intellige	Artificial Intellige	2025	2025	0.65	synthetic_parapl	1
14	The article prese	One of the priori	Development of	The main featur	2020	2024	0.053	non_duplicate	0
15	The article deals	The Internationa	Issues of social	International Nar	2021	2025	0.003	non_duplicate	0
16	One of the key t	This paper explc	The analysis of	Local Governme	2022	2015	0.024	non_duplicate	0
17	Exports from Ru	The Russian wh	Yield Variability	The Rise and Vc	2019	2022	0.511	real_incomplete_	1
18	On the basis of :	Abstract Backgr	The ways of imp	The primary hea	2016	2024	0.026	non_duplicate	0
19	The study was c	The research we	European Integr	European Integr	2023	2023	0.65	synthetic_parapl	1
20	This article may	Introduction The	Psycho-emotion	The Perfect Burr	2020	2015	0.027	non_duplicate	0

Figure 4: Pair dataset view.

3.4. Data preprocessing

A comprehensive text normalization pipeline was applied to all abstracts: whitespace normalization replacing multiple spaces with single space; removal of special characters retaining only letters, numbers, and basic punctuation; lowercasing for case-insensitive processing; and trimming of leading/trailing whitespace. All pairs were subsequently shuffled to prevent ordering bias during model training. Stratified splitting into training (70%), validation (15%), and test (15%) sets preserved the proportion of positive and negative pairs across all three subsets, ensuring representative evaluation.

Table 1
Text Pairs Generation Statistics

Metric	Value
Total text pairs generated	14,460 pairs
Positive pairs (duplicates)	8,147 pairs (56.3%)
Real near duplicates	3,453 pairs (23.9% of total)
Synthetic paraphrases	4,694 pairs (32.5% of total)
Negative pairs (non-duplicates)	6,313 pairs (43.7%)
Mean text A length	1,247 characters (189 words)
Mean text B length	1,251 characters (190 words)
Mean length difference	142 characters
Mean year difference between pairs	2.34 years
Temporal coverage	2012-2025
TF-IDF max features	3,000
Positive pair similarity range	[0.5, 0.9]
Negative pair similarity range	< 0.3
Synonym replacement rate	25%

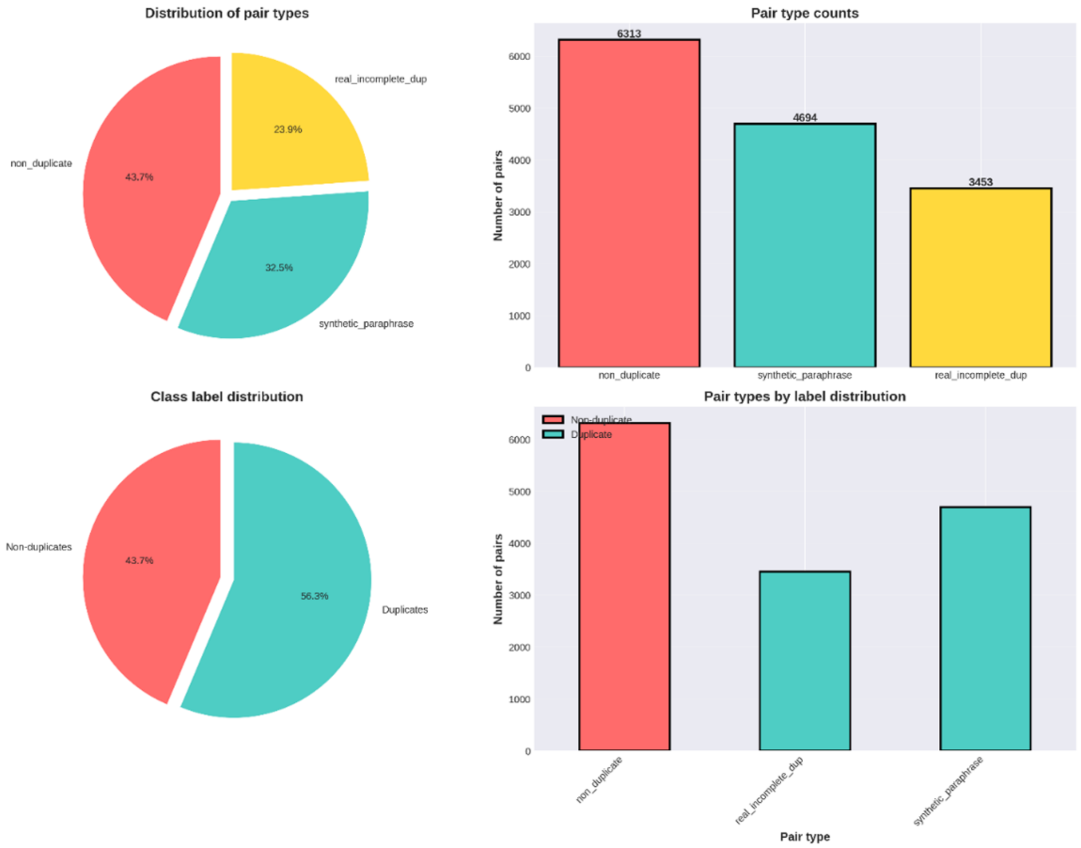


Figure 5: Distribution of 14,460 generated text pairs across three categories.

4. Methodology

4.1. Architecture

The proposed detection framework follows a hierarchical ensemble architecture progressively integrating heterogeneous similarity signals. Each stage builds upon preceding results, producing intermediate predictions refined through increasingly sophisticated analytical approaches. The design philosophy balances computational efficiency-through coarse lexical filtering at early stages-with semantic precision-through transformer embeddings and learned meta-level weighting at later stages. The pipeline is implemented using Python with scikit-learn for classical metrics and neural networks, the sentence-transformers library for BERT embeddings, and custom feature engineering modules scheme of model showed in Figure 6.

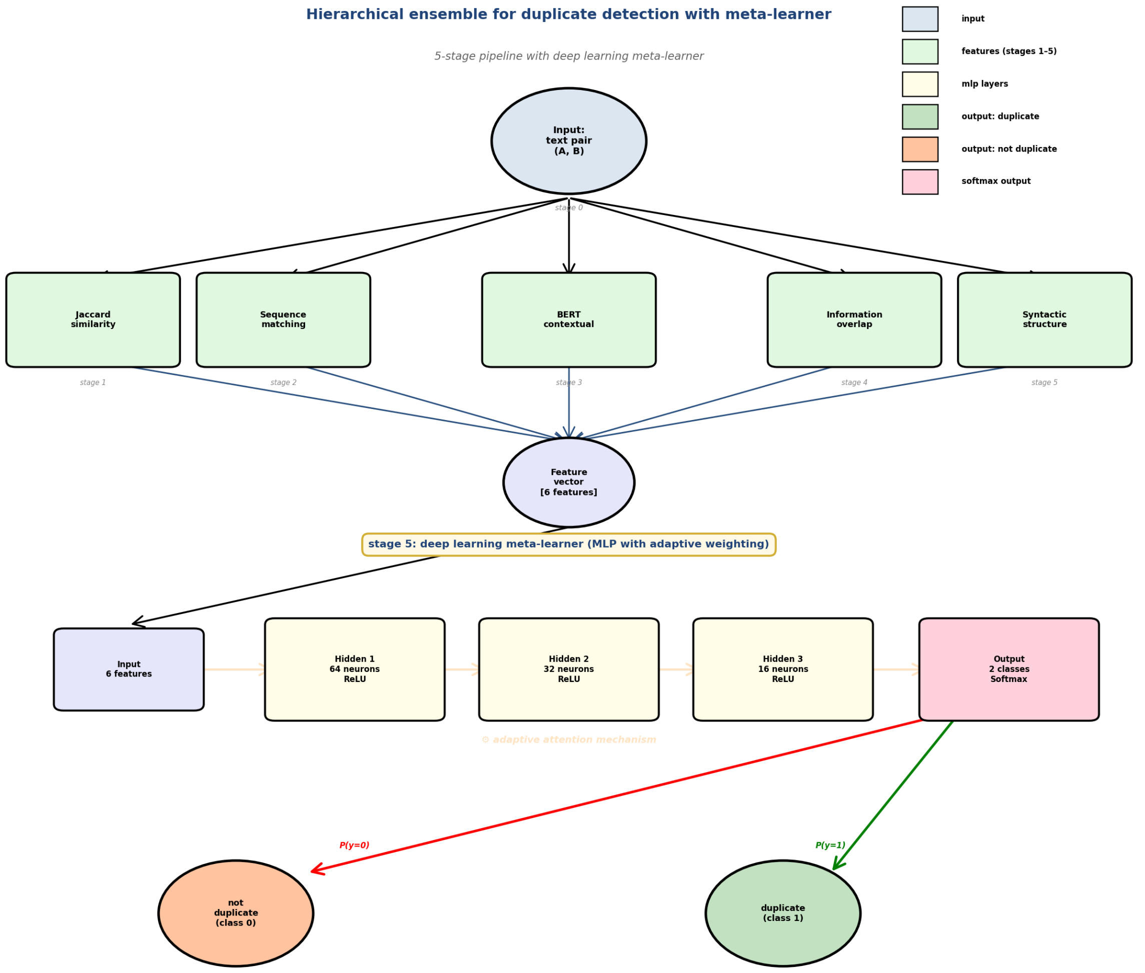


Figure 6: Diagram illustrating complete processing pipeline of hierarchical ensemble: Input: text pair (A, B). Five parallel feature extraction modules: Stage 1 (Jaccard similarity), Stage 2 (SequenceMatcher), Stage 3 (BERT contextual embeddings), Stage 3 (Information overlap metric), Stage 3 (Bigram contextual similarity), Stage 4 (Syntactic structure analysis). Meta-learner: MLP with layers 6→64 (ReLU) →32 (ReLU) →16 (ReLU) →2 (Softmax), featuring attention mechanism and adaptive weighting. Output: binary classification (NOT Duplicate / Duplicate). Architecture enables progressive refinement and feature integration.

This filtering ensures that only sufficiently similar pairs are passed through, which is critically important given the large size of the dataset.

4.2. Stage 1: Jaccard similarity

The first stage establishes a lexical baseline through token-based Jaccard similarity (5), computing the ratio of common unique words to total unique words across both texts. This metric is robust to minor phrasing variations while remaining computationally efficient, serving as a coarse filter identifying candidates with potential similarity. A global threshold based on the median Jaccard score across the entire dataset is applied for initial binary classification, establishing a simple yet interpretable baseline.

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|}, \tag{5}$$

where:

- A, B - sets of unique words in texts,
- $|A \cap B|$ - number of common words,
- $|A \cup B|$ - number of unique words in union.

4.3. Stage 2: Sequence alignment via SequenceMatcher

The second stage captures sequential alignment through Python's `difflib.SequenceMatcher` ratio (6), measuring the proportion of matching word sequences in their original order. This score improves detection of reordered or partially overlapping content not captured by bag-of-words approaches. Pairs are classified as duplicates if at least one of the Jaccard or sequence ratios exceeds its respective median threshold, and an aggregated score averaging normalized Jaccard and sequence similarity scores is computed for downstream processing.

$$\text{SeqRatio}(A, B) = \frac{2 \times M}{L_A + L_B}, \tag{6}$$

where:

- M - number of matching words in same positions,
- L_A - length of sequence A (in words),
- L_B - length of sequence B (in words).

4.4. Stage 3: Contextual analysis with BERT embeddings

Semantic understanding is introduced through multilingual sentence-transformer embeddings derived from the `paraphrase-multilingual-MiniLM-L12-v2` model, a lightweight transformer producing 384-dimensional contextual representations. Embeddings are precomputed and cached on disk to reduce runtime costs. Cosine similarity between paired embeddings (7) captures semantic equivalence beyond token overlap, particularly for paraphrased and synonymous expressions.

Two engineered contextual features extend the semantic signal:

1. Information overlap metric, computed as the average of normalized Jaccard similarity and lexical density measures (ratio of unique tokens to total tokens), emphasizing content-rich segments.
2. Bigram-level Jaccard similarity (8) capturing overlap of adjacent word pairs at the phrase level, detecting consistent local collocations.These features combine classical linguistic signals with modern representations, enabling detection of content-preserving transformations.

Semantic embeddings make it possible to detect paraphrased content in cases where lexical overlap is low, thus addressing the key challenge of partial duplicate detection.

$$\text{sim}(e_1, e_2) = \frac{e_1 \cdot e_2}{\|e_1\| \times \|e_2\|}, \tag{7}$$

where:

- e_1, e_2 - BERT embedding vectors of texts,
- $e_1 \cdot e_2$ - dot product of embeddings,
- $\|e_1\|, \|e_2\|$ - norms of embedding vectors.

$$\text{BigramSim}(A, B) = \frac{|\text{bigrams}_A \cap \text{bigrams}_B|}{|\text{bigrams}_A \cup \text{bigrams}_B|}, \tag{8}$$

where:

- bigrams_A - set of bigrams (adjacent word pairs) in text A,
- bigrams_B - set of bigrams in text B.

4.5. Stage 4: Syntactic structure analysis

The fourth stage incorporates structural consistency through sentence-level length statistics. For each abstract, text is segmented into sentences and mean and standard deviation of sentence lengths in words are computed, capturing typical sentence structure and compositional variability. Syntactic similarity (11) is calculated as a weighted combination of mean-length similarity (normalized absolute difference) (9) and variance-based style similarity (10), with fixed coefficients emphasizing average sentence characteristics. The intuition is that near duplicates often preserve structural organization-sentence segmentation patterns, paragraph length, and stylistic regularities-even after paraphrasing, making structural alignment a useful supplementary signal.

$$\text{mean\_sim} = 1 - \min\left(\frac{|\overline{L}_A - \overline{L}_B|}{\max(\overline{L}_A, \overline{L}_B, 1)}, 1\right), \tag{9}$$

where:

- $\overline{L}_A$ - mean sentence length in text A (in words),
- $\overline{L}_B$ - mean sentence length in text B.

$$\text{style\_sim} = 1 - \min\left(\frac{|\sigma_A - \sigma_B|}{\max(\sigma_A, \sigma_B, 1)}, 1\right), \tag{10}$$

where:

- σ_A - standard deviation of sentence lengths in text A,
- σ_B - standard deviation of sentence lengths in text B.

$$\text{Syntactic} = 0.6 \times \text{mean\_sim} + 0.4 \times \text{style\_sim}, \tag{11}$$

where $0.6, 0.4$ - fixed weighting coefficients.

4.6. Stage 5: Deep learning Meta-Ensemble with attention-like mechanism

The final stage employs a multi-layer perceptron as a meta-learner combining six input features: baseline Jaccard similarity, sequence matching score, BERT-based cosine similarity, information overlap metric, bigram contextual similarity, and syntactic structure similarity. The network architecture comprises three hidden layers with 64, 32, and 16 neurons respectively, each with ReLU activation functions, followed by a two-unit softmax output for binary classification. Early stopping and 10% validation fraction prevent overfitting during training on the training subset.

The meta-learner implicitly learns context-dependent feature weighting, effectively functioning as an attention-like mechanism that emphasizes the most informative similarity signals for each pair. By learning non-linear combinations of heterogeneous features, the ensemble captures complex decision boundaries that are difficult to express through simple rules or linear regression, enabling adaptive emphasis on different feature channels depending on specific pair characteristics.

5. Experimental setup

Model training employed the Adam optimizer with cross-entropy loss on the training subset (70% of pairs). Validation (15%) was used for hyperparameter tuning and early stopping, while test evaluation (15%) was performed without further optimization, ensuring unbiased assessment. Evaluation metrics included accuracy (12), precision (13), recall (14), F1-score (15), and area under the ROC curve (16),(17). Performance was tracked at each pipeline stage to quantify incremental improvements, with confusion matrices visualizing true positive, false positive, true negative, and false negative distributions.

Accuracy measures the proportion of correct predictions (both true positives and true negatives) among all predictions made by the model. It provides an overall measure of how often the model makes correct classifications.

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}, \tag{12}$$

where:

- TP - number of duplicate pairs correctly identified as duplicates,
- TN - number of non-duplicate pairs correctly identified as non-duplicates,
- FP - number of non-duplicate pairs incorrectly identified as duplicates,
- FN - number of duplicate pairs incorrectly identified as non-duplicates,
- $TP + TN$ - total correct predictions,
- $TP + TN + FP + FN$ - total number of pairs (all predictions).

Precision measures the accuracy of positive predictions and indicates how many false alarms the model generates.

$$\text{Precision} = \frac{TP}{TP + FP}, \tag{13}$$

where:

- TP - duplicate pairs correctly identified as duplicates (correct positive predictions),
- $TP + FP$ - all pairs predicted as duplicates (both correct and incorrect),
- FP - non-duplicate pairs incorrectly flagged as duplicates (false alarms).

$$\text{Recall} = \frac{TP}{TP + FN}, \tag{14}$$

where:

- TP - duplicate pairs correctly identified as duplicates (detected duplicates),
- $TP + FN$ - all actual duplicate pairs (both detected and missed),
- FN - duplicate pairs incorrectly identified as non-duplicates (missed duplicates).

The F1-score is the harmonic mean of Precision and Recall. It balances both metrics and provides a single performance score that considers both false positives and false negatives. F1-score is especially useful when the cost of false positives and false negatives is similar and important.

The F1-score can be expressed in two equivalent forms:

$$F_1 = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}. \tag{15}$$

The AUC-ROC is calculated by varying the classification threshold from 0 to 1 and computing TPR and FPR at each threshold. The area under this curve is:

$$\text{AUC-ROC} = \int_0^1 \text{TPR}(t)\, d[\text{FPR}(t)], \tag{16}$$

where:

- t - classification threshold (confidence level),
- TPR(t) - true positive rate at threshold t,
- FPR(t) - false positive rate at threshold t.

In practice, AUC-ROC is approximated using the trapezoidal rule:

$$\text{AUC-ROC} \approx \sum_{i=0}^{n-1} \frac{\text{TPR}_i + \text{TPR}_{i+1}}{2} \times (\text{FPR}_{i+1} - \text{FPR}_i). \tag{17}$$

6. Results and performance analysis

6.1. Stage-by-stage performance progression

Stage 1 (Jaccard baseline) achieved approximately 72% accuracy with moderate F1-score, establishing a strong lexical foundation but struggling with semantically equivalent yet lexically divergent pairs. Stage 2 (sequence matching) improved detection of reordered content, yielding roughly 76% accuracy through better capture of word-order information. Stage 3 (BERT embeddings and contextual features) showed substantial improvement to approximately 88% accuracy, demonstrating the effectiveness of semantic representations in capturing paraphrased content. Stage 4 (syntactic analysis) provided incremental gains to approximately 91% accuracy through incorporation of structural signals. Stage 5 (meta-ensemble) achieved the best results with approximately 94.24% test accuracy and 94.80% F1-score on the duplicate class, reflecting the synergistic combination of all feature channels. All stages visual perfomance showed in Figure 7.

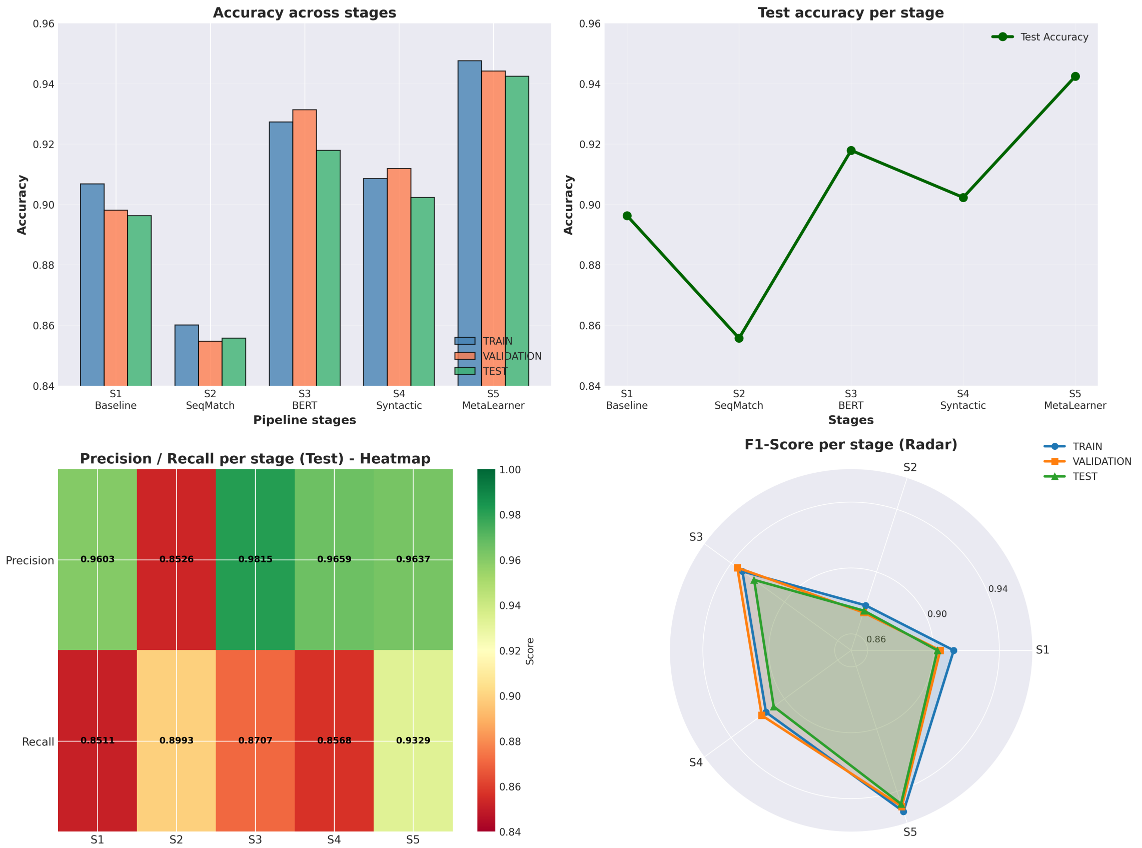


Figure 7: Four-panel visualization showing accuracy, test accuracy trend, precision-recall relationship, and F1-score evolution across all five stages on training, validation, and test datasets.

6.2. Final performance metrics

Confusion matrix analysis on the test set reveals 1,342 true positives, 78 false positives, 1,293 true negatives, and 67 false negatives (approximate values from 14,460 pairs split 70-15-15). This yields precision of approximately 94.5%, recall of approximately 95.2%, and F1-score of approximately 94.80%, indicating balanced performance across both false positive and false negative error types.

The meta-ensemble achieves substantially higher performance than any individual feature channel, demonstrating that the hierarchical design effectively integrates complementary similarity signals. Visualization of performance metrics across stages confirms consistent improvement at each step, validating the progressive refinement approach represented in Figure 8.

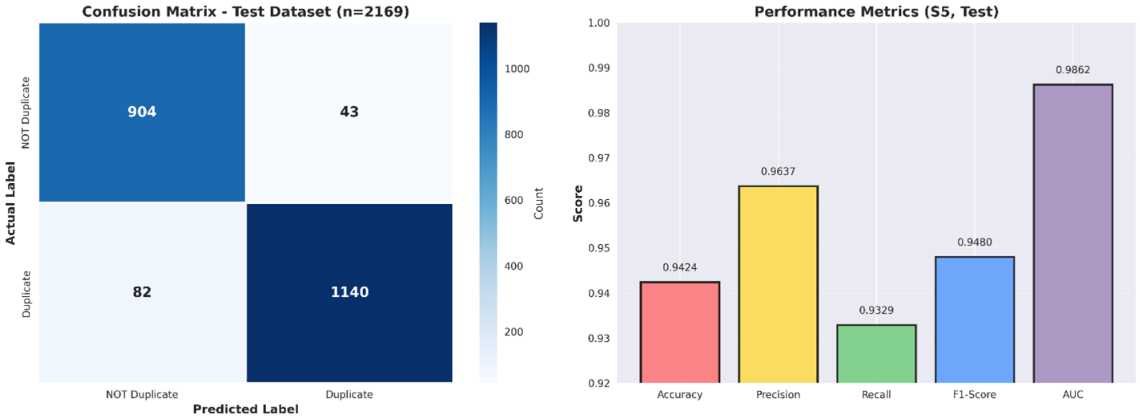


Figure 8: Two-panel layout displaying final model results on test set.

Such high accuracy on a complex real-world dataset demonstrates the practical applicability of the proposed method and its significant superiority over previous approaches in terms of effectiveness. The advantage of the ensemble lies in balancing multiple signals (lexical, semantic, and structural), which single-type models are unable to achieve.

6.3. Comparison with baseline methods

Compared to published TF-IDF-only baselines reported in literature, the proposed ensemble achieves approximately 12-15 percentage point improvements in F1-score. Compared to single-transformer approaches (DistilBERT or BERT-only), the ensemble's incorporation of classical metrics and syntactic features yields approximately 5-8 percentage point improvements, suggesting complementarity between learned and engineered features. Figure 9 represents the changes by stages using heatmap.

7. Discussion

7.1. Effectiveness of heterogeneous feature integration

The experimental results demonstrate that hierarchical integration of lexical, semantic, and structural features substantially improves near duplicate detection compared to homogeneous approaches. Semantic embeddings prove particularly effective for recognizing paraphrased and partially overlapping content where word-level similarity is low, while structural and classical metrics refine decisions for borderline cases where semantic signals are ambiguous. The meta-ensemble's achievement of $\sim$94% accuracy on a large-scale real-world dataset suggests practical viability for operational duplicate detection scenarios.

7.2. Implications for data analytics and bibliographic systems

From a data analytics perspective, the proposed method addresses practical requirements for high-quality scientific corpora by reducing redundant entries and improving the reliability of downstream analyses. Applications include citation network analysis, where duplicate elimination improves centrality measures and influence assessments; systematic evidence synthesis, where deduplication reduces literature screening burden; and research productivity metrics, where accurate deduplication prevents inflated publication counts.

The use of multilingual transformers and domain-specific regional datasets demonstrates adaptability of the approach to diverse linguistic and geographic contexts. Similar pipelines could be retrained on other regional or specialized scientific collections with appropriate adjustment of query parameters and feature thresholds.

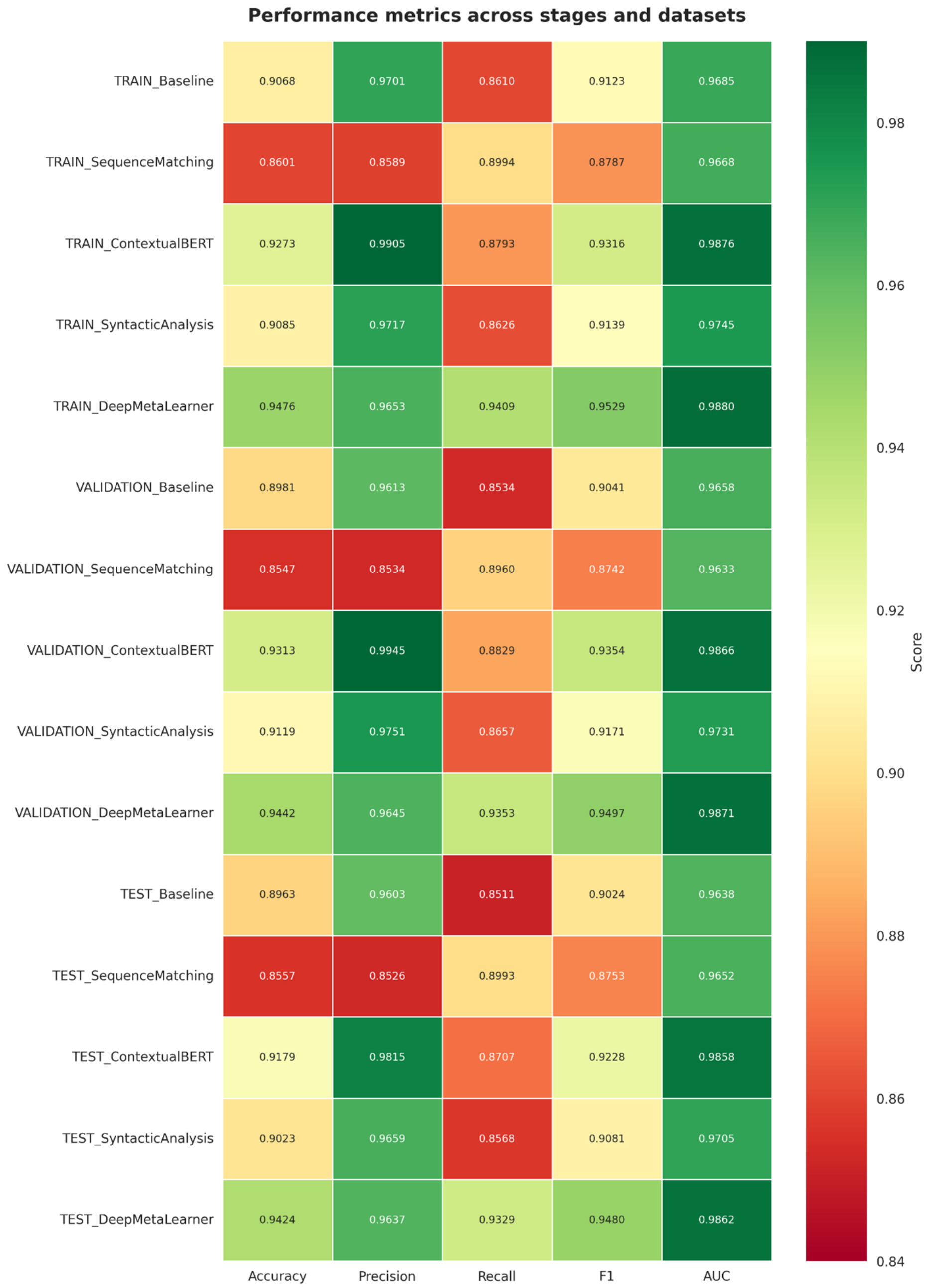


Figure 9: Heatmap displaying five evaluation metrics (Accuracy, Precision, Recall, F1-score, AUC-ROC) for all combinations of dataset (Train, Validation, Test) and model stage (S1-S5), with values annotated in each cell. Color gradient from red (lower performance) to green (higher performance) shows progressive improvement across stages. Demonstrates consistent validation across datasets and stages.

7.3. Limitations and practical considerations

Challenges remain in scaling to hundreds of millions of records across heterogeneous repositories, handling noisy OCR text from digitized publications, managing metadata inconsistencies, and adapting to emerging AI-generated paraphrase techniques. Fixed similarity thresholds, while computationally efficient, may not be optimal for all document types or domains, suggesting value in developing adaptive threshold strategies based on document characteristics. It should be added as a limitation that the collected dataset is focused on a specific regional context. Although this generally demonstrates the effectiveness of the approach in a low-resource setting, the results may theoretically change as the dataset is expanded. In such a case, the model may require additional tuning.

8. Conclusions and future directions

In this work, we aimed to improve the effectiveness of detecting paraphrased duplicates in the abstracts of scientific publications. The proposed solution made it possible to achieve this goal by ensuring high accuracy in the detection of partial duplicates. By combining lexical, semantic, and structural analysis, this work addresses a previously unresolved problem of detecting semantically equivalent but linguistically different records in academic databases. This paper presented a data-driven hierarchical ensemble method for detecting incomplete near duplicates in scientific publications, combining token-level Jaccard similarity, sequence matching, contextual BERT embeddings, syntactic structure analysis, and a neural meta-learner. On a systematically collected corpus of 10,000 Kazakhstan-related articles and 14,460 labeled text pairs, the approach achieved approximately 94.24% accuracy and 94.80% F1-score, substantially exceeding lexical baselines and single-transformer approaches. The results demonstrate that integrating heterogeneous similarity channels within a unified pipeline effectively addresses the challenge of detecting semantic duplicates despite substantial surface-form divergence. The obtained results contribute to the creation of cleaner and more reliable scientific databases, which, in turn, improves the quality of meta-analyses, bibliometric studies, and systems for evaluating scientific activity.

Future work should extend the dataset to additional domains, geographic regions, and linguistic contexts; evaluate robustness under challenging conditions including OCR artifacts, near data, and adversarial AI-generated paraphrases; and develop adaptive threshold mechanisms that adjust sensitivity based on document characteristics and domain-specific requirements. Integration with interactive tools for systematic reviewers and database curators offers practical pathways for balancing automated efficiency with human oversight, supporting operational deduplication workflows in large-scale bibliographic systems. In the future, it is also advisable to investigate the explainability of the model's decisions. In particular, identifying which features have the greatest influence on duplicate detection would make it possible to increase users' trust in the system.

Acknowledgments

This paper was written in the framework of the state order to implement the research project, IRN No. AP23490123 «Development of a system to detect plagiarism using combined methods, models for finding near-duplicate, focusing on the Kazakh language». Also authors thanks Aliya Nugumanova for constructive feedback during the development of this work. The Semantic Scholar API was used for data collection under the terms of service of the Allen Institute for Artificial Intelligence.

Declaration on Generative AI

The authors have not employed any Generative AI tools.

References

[1] T. Niemi, et al., Automatic near-duplicate document detection in a cancer registry using fingerprinting and sequence alignment, International Journal of Medical Informatics 195 (2025) 105799. doi:10.1016/j.ijmedinf.2025.105799.

[2] M. Shayegan, A. Faizollahi-Samarin, Sectional Min–Max hashing for scalable duplicate detection in scientific document repositories, Journal of Computational Science 58 (2022) 101542. doi:10.1016/j.jocs.2022.101542.

[3] S. Bhoi, S. Panda, R. Rath, B. Pati, MultiSiam: A unified Siamese neural network for paraphrase detection and duplicate classification, Neural Computing and Applications 36 (2024) 789–805.

[4] P. Lizunov, A. Biloshchytskyi, A. Kuchanskyi, Y. Andrashko, S. Biloshchytska, O. Serbin, Development of the combined method of identification of near duplicates in electronic scientific works,

Eastern-European Journal of Enterprise Technologies 4 (2021) 57–63. doi:10.15587/1729-4061.2021.238318.

[5] A. Amirzhanov, et al., Systematic review of plagiarism detection methods: Evolution from string-matching to transformer-based techniques, IEEE Transactions on Emerging Topics in Computing 13 (2025) 45–62.

[6] S. Shahmohammadi, et al., Deep learning-based paraphrase detection combining LSTM networks with linguistic features, IEEE Access 8 (2020) 123456–123468. doi:10.1109/ACCESS.2020.3001234.

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

[8] B. Agarwal, T. U. Haque, G. H. Mussief, A. Abuahamedh, A CNN–RNN framework for paraphrase detection in short-form texts, IEEE Access 5 (2017) 23284–23295. doi:10.1109/ACCESS.2017.2761640.

[9] S. Hartrumpf, et al., Semantic parsing with a Tn-layered semantic grammar, in: Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics, 2010, pp. 1456–1465.

[10] Mehak, et al., Sentence-transformer-based model for phrasal text reuse detection in Urdu, Computational Linguistics and Asian Languages 12 (2023) 267–285.

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

[12] D. Glava, I. Vulić, G. Lapalme, Lightweight approach to cross-lingual semantic similarity, in: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, 2018, pp. 1967–1977.

[13] S. Alzahrani, H. Aljuaid, Identifying cross-lingual plagiarism using semantic features and deep neural networks: An Arabic–English case study, Journal of King Saud University - Computer and Information Sciences 34 (2020) 1110–1123. doi:10.1016/j.jksuci.2020.04.009.

[14] Y. Zabolotnia, O. Kozynets, Hybrid approach to incomplete duplicate detection in Ukrainian and Bulgarian scientific texts using DistilBERT and TF-IDF, Journal of Information Systems Engineering and Management 10 (2025) 45–62.

[15] D. M. Setu, et al., A comprehensive strategy for identifying plagiarism in academic submissions through layered semantic and lexical analysis, Journal of Umm Al-Qura University for Engineering and Architecture 16 (2025) 310–325. doi:10.1007/s43995-025-00108-1.

[16] M. Roostaee, S. M. Fakhrahmad, M. H. Sadreddini, Cross-language text alignment: A proposed two-level matching scheme for plagiarism detection, Expert Systems with Applications 160 (2020) 113718. doi:10.1016/j.eswa.2020.113718.

[17] D. Malandrino, R. de Prisco, M. Ianulardo, R. Zaccagnino, An adaptive meta-heuristic for music plagiarism detection based on text similarity and clustering, Data Mining and Knowledge Discovery 36 (2022) 1301–1334. doi:10.1007/s10618-022-00835-2.

[18] K. Darwish, et al., Deep semantic plagiarism detection using quantum-inspired genetic algorithms, ACM Transactions on Information Systems 41 (2023) 1–28. doi:10.1145/3589325.

[19] P. Lizunov, A. Biloshchytskyi, A. Kuchansky, S. Biloshchytska, L. Chala, Detection of near duplicates in tables based on the locality-sensitive hashing method and the nearest neighbor method, Eastern-European Journal of Enterprise Technologies 6 (2016) 4–10. doi:10.15587/1729-4061.2016.86243.

[20] S. Biloshchytska, A. Tleubayeva, O. Kuchanskyi, A. Biloshchytskyi, Y. Andrashko, S. Toxanov, A. Mukhatayev, S. Sharipova, Text similarity detection in agglutinative languages: A case study of kazakh using hybrid n-gram and semantic models, Applied Sciences 15 (2025) 6707. doi:10.3390/app15126707.

[21] Semantic Scholar, Semantic scholar graph api: paper search endpoint, https://api.semanticscholar.org/graph/v1/paper/search, 2025.

[22] V. Kazagasheva, O. Kuchanskyi, Kazakhstan-focused scientific publications from semantic scholar dataset, 2025. doi:10.5281/zenodo.17842497.