ҚАЗАҚСТАН РЕСПУБЛИКАСЫНЫҢ ҒЫЛЫМ ЖӘНЕ ЖОҒАРЫ БІЛІМ МИНИСТРЛІГІ
МИНИСТЕРСТВО НАУКИ И ВЫСШЕГО ОБРАЗОВАНИЯ РЕСПУБЛИКИ КАЗАХСТАН
MINISTRY OF SCIENCE AND HIGHER EDUCATION OF THE REPUBLIC OF KAZAKHSTAN

ХАЛЫҚАРАЛЫҚ АҚПАРАТТЫҚ ЖӘНЕ КОММУНИКАЦИЯЛЫҚ ТЕХНОЛОГИЯЛАР ЖУРНАЛЫ

МЕЖДУНАРОДНЫЙ ЖУРНАЛ ИНФОРМАЦИОННЫХ И КОММУНИКАЦИОННЫХ ТЕХНОЛОГИЙ

INTERNATIONAL JOURNAL OF INFORMATION AND COMMUNICATION TECHNOLOGIES

2025 (24) 4

қазан- желтоқсан

ISSN 2708–2032 (print)
ISSN 2708–2040 (online)

Халықаралық ақпараттық және коммуникациялық технологиялар журналы
ISSN 2708–2032 (print)
ISSN 2708–2040 (online)
Меншік иесі: АҚ «Халықаралық ақпараттық технологиялар университеті» (Алматы қ.).
Қазақстан Республикасы Ақпарат және қоғамдық даму министрлігіне мерзімді баспасөз басылымын есепке қою туралы куәлік № KZ82VPY00020475, 20.02.2020 ж. берілген
Тақырып бағыты: ақпараттық технологиялар, ақпараттық қауіпсіздік және коммуникациялық технологиялар, әлеуметтік-экономикалық жүйелерді дамытудағы цифрлық технология.
Мерзімділігі: жылына 4 рет.
Тираж: 100 дана.
Редакция мекенжайы: 050040 Алматы қ., Манас к., 34/1, каб. 709, тел: +7 (727) 244-51-09.
E-mail: ijict@iitu.edu.kz
Журнал сайты: https://journal.iitu.edu.kz

Международный журнал информационных и коммуникационных технологий
ISSN 2708–2032 (print)
ISSN 2708–2040 (online)
Собственник: АО «Международный университет информационных технологий» (г. Алматы).
Свидетельство о постановке на учет периодического печатного издания в Министерство информации и общественного развития Республики Казахстан **№ KZ82VPY00020475,** выданное от **20.02.2020 г.**
Тематическая направленность: информационные технологии, информационная безопасность и коммуникационные технологии, цифровые технологии в развитии социо-экономических систем**.**
Периодичность: 4 раза в год.
Тираж: 100 экземпляров.
Адрес редакции: 050040 г. Алматы, ул. Манаса 34/1, каб. 709, тел: +7 (727) 244-51-09.
E-mail: ijict@iitu.edu.kz Сайт журнала: https://journal.iitu.edu.kz

«International Journal of Information and Communication Technologies»
ISSN 2708–2032 (print)
ISSN 2708–2040 (online)
Owner: International Information Technology University JSC (Almaty).
The certificate of registration of a periodical printed publication in the Ministry of Information and Social Development of the Republic of Kazakhstan, Information Committee No. KZ82VPY00020475, issued on 20.02.2020.
Thematic focus: information technology, digital technologies in the development of socio-economic systems, information security and communication technologies
Periodicity: 4 times a year.
Circulation: 100 copies.
Editorial address: 050040. Manas st. 34/1, Almaty. +7 (727) 244-51-09. E-mail: ijict@iitu.edu.kz
Journal website: https://journal.iitu.edu.kz

INTERNATIONAL JOURNAL OF INFORMATION AND COMMUNICATION TECHNOLOGIES
ISSN 2708–2032 (print)
ISSN 2708–2040 (online)
Vol. 6. Is. 4. Number 24 (2025). Pp. 333–350
Journal homepage: https://journal.iitu.edu.kz
https://doi.org/10.54309/IJICT.2025.24.4.020

МРНТИ 28.23.37

DETECTING DUPLICATES IN KAZAKH TEXTS: A COMPARISON OF TF-IDF, WORD AND SENTENCE EMBEDDINGS

A.O. Tleubayeva*, S.V. Biloshchytska, O. Kuchanskyi, A.A. Mukhatayev, A.B. Nugumanova
Astana IT University, Astana, Kazakhstan.
E-mail: bsv@astanait.edu.kz

Arailym Tleubayeva — PhD student, senior lecturer at the School of Artificial Intelligence and Data Science, Astana IT University, Astana, Kazakhstan
E-mail: a.tleubayeva@astanait.edu.kz, https://orcid.org/0000-0001-9560-9756;
Svitlana Biloshchytska — Doctor of Technical Sciences, Professor at the School of Artificial Intelligence and Data Science, Astana IT University, Astana, Kazakhstan
https://orcid.org/0000-0002-0856-5474;
Oleksandr Kuchanskyi — Doctor of Technical Sciences, Professor at the School of Artificial Intelligence and Data Science, Astana IT University, Astana, Kazakhstan
https://orcid.org/0000-0003-1277-8031;
Aidos Mukhatayev, Candidate of Pedagogical Sciences, Professor at the School of General Education Disciplines, Astana IT University, Astana, Kazakhstan
https://orcid.org/0000-0002-8667-3200;
Aliya Nugumanova— PhD, Head of the Scientific and Innovation Center «Big Data and Blockchain Technologies», Astana IT University, Astana, Kazakhstan
https://orcid.org/0000-0001-5522-4421.

Abstract. This paper presents a comprehensive comparison of TF-IDF, word, and multilingual sentence embeddings for automatic duplicate detection in Kazakh texts. Experiments use the KazakhTextDuplicates dataset with labels for exact, paraphrase, contextual, and partial duplicates. All models were evaluated within a unified setup featuring standardized preprocessing, L2-normalized vectors, and validation-based threshold tuning. The Word2Vec model with TF-IDF weighting achieved the highest performance (F1 = 0.996; ROC-AUC = 0.9999; PR-AUC = 0.9999). The TF-IDF (1–3-grams) method remained competitive for exact and partial overlaps (PR-AUC = 0.932; ROC-AUC = 0.775), while FastText provided the best recall (R ≈ 0.99) at moderate precision. Among multilingual models, BGE-m3

and Snowflake Arctic achieved the best PR-AUC (≈0.614). In retrieval, the BM25 followed by dense re-ranking pipeline produced a small but consistent improvement over dense-only search (Recall@10: +0.04–0.12 pp; nDCG@10: +0.10–0.13 pp), confirming the effectiveness of combining lexical and semantic features for duplicate detection in morphologically rich, low-resource languages such as Kazakh.

Keywords: duplicate detection; Kazakh language; TF-IDF; word embeddings; sentence embeddings; semantic similarity; BM25; dense retrieval; hybrid reranking; low-resource NLP

For citation: A.O. Tleubayeva, S.V. Biloshchytsk***a***, O. Kuchanskyi, A.A. Mukhatayev, A.B. Nugumanov***a***. Detecting duplicates in kazakh texts: a comparison of tf-idf, word and sentence embeddings // International journal of information and communication technologies. 2025. Vol. 6. No. 24. Pp. 333–350. (In Eng.). https://doi.org/10.54309/IJICT.2025.24.4.020.

Conflict of interest: The authors declare that there is no conflict of interest.

Funding Information: 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.

ҚАЗАҚ ТІЛІНДЕГІ МӘТІНДЕРДЕГІ ДУБЛИКАТТАРДЫ АНЫҚТАУ: TF-IDF, СӨЗ ЖӘНЕ СӨЙЛЕМ ЭМБЕДДИНГТЕРІН САЛЫСТЫРУ

А.О. Тлеубаева*, С.В. Белощицкая, О.Ю. Кучанский, А.А. Мухатаев, А.Б. Нугуманова

Astana IT University, Astana, Kazakhstan.

E-mail: a.tleubayeva@astanait.edu.kz

Тлеубаева Арайлым — PhD докторант, «Жасанды интеллект және деректер ғылымы» мектебінің сеньор-лекторы, Astana IT University, Астана, Қазақстан
E-mail: a.tleubayeva@astanait.edu.kz, https://orcid.org/0000-0001-9560-9756;

Белошицкая Светлана — техника ғылымдарының докторы, «Жасанды интеллект және деректер ғылымы» мектебінің профессоры, Astana IT University, Астана, Қазақстан
https://orcid.org/0000-0002-0856-5474;

Кучанский Александр — техника ғылымдарының докторы, «Жасанды интеллект және деректер ғылымы» мектебінің профессоры, Astana IT University, Астана, Қазақстан
https://orcid.org/0000-0003-1277-8031;

Мухатаев Айдос — Педагогика ғылымдарының кандидаты, «Жалпы білім беру пәндері» мектебінің профессоры, Astana IT University, Астана, Қазақстан
https://orcid.org/0000-0002-8667-3200;

Нугуманова Алия — PhD, «Big Data and Blockchain Technologies», ғылыми-ин-новациялық орталығының жетекшісі, Astana IT University, Астана, Қазақстан
https://orcid.org/0000-0001-5522-4421.

Аннотация. Мақалада қазақ тіліндегі мәтін дубликаттарын автоматты түрде анықтау үшін TF-IDF, сөздік және көптілді сөйлем эмбеддингтері кешенді түрде салыстырылды. Эксперименттер KazakhTextDuplicates деректер жинағында жүргізілді, мұнда жұптар «нақты», «парафраз», «контекстік» және «ішінара» дубликат ретінде таңбаланған. Барлық модельдер бірыңғай эксперименттік ортада бағаланды: стандартталған алдын ала өңдеу, L2-нормаланған векторлар және валидация арқылы шек мәнін баптау. TF-IDF-пен салмақталған Word2Vec моделі ең жоғары нәтижелерге жетті (F1 = 0.996; ROC-AUC = 0.9999; PR-AUC = 0.9999). TF-IDF (1–3-грамма) әдісі нақты және ішінара сәйкестіктер үшін тиімді болды (PR-AUC = 0.932; ROC-AUC = 0.775), ал FastText жоғары толықтық (R ≈ 0.99) көрсетті. Көптілді модельдер арасында BGE-m3 және Snowflake Arctic PR-AUC бойынша үздік нәтижелерге (≈0.614) жетті. Іздеу міндетінде BM25 және кейінгі тығыз қайта ранжирлеу тәсілі тығыз іздеумен салыстырғанда аз болса да тұрақты өсім көрсетті (Recall@10: +0.04–0.12 п.б.; nDCG@10: +0.10–0.13 п.б.), бұл лексикалық және семантикалық белгілерді біріктірудің тиімділігін дәлелдейді.

Түйін сөздер: дубликаттарды анықтау; қазақ тілі; TF-IDF; сөздік эмбеддингтер; сөйлемдік эмбеддингтер; семантикалық ұқсастық; BM25; тығыз іздеу (dense retrieval); гибридті қайта ранжирлеу; ресурсы шектеулі тілдер үшін NLP

Дәйексөздер үшін: А.О. Тлеубаева, С.В. Белощицкая, О.Ю. Кучанский, А.А. Мухатаев, А.Б. Нугуманова. Қазақ тіліндегі мәтіндердегі дубликаттарды анықтау: tf-idf, сөз және сөйлем эмбеддингтерін салыстыру // Халықаралық ақпараттық және коммуникациялық технологиялар журналы. 2025. Том. 6. № 24. 333–350 бет. (Ағыл). https://doi.org/10.54309/IJICT.2025.24.4.020.

Мүдделер қақтығысы: Авторлар осы мақалада мүдделер қақтығысы жоқ деп мәлімдейді.

ОБНАРУЖЕНИЕ ДУБЛИКАТОВ В КАЗАХСКИХ ТЕКСТАХ: СРАВНЕНИЕ TF-IDF, ЭМБЕДДИНГОВ СЛОВ И ЭМБЕДДИНГОВ ПРЕДЛОЖЕНИЙ

А.О. Тлеубаева*, С.В. Белощицкая, О.Ю. Кучанский, А.А. Мухатаев, А.Б. Нугуманова
Astana IT University, Astana, Kazakhstan.
E-mail: a.tleubayeva@astanait.edu.kz

Тлеубаева Арайлым — PhD докторант, сеньор-лектор «Школы искусственного интеллекта и науки о данных», Astana IT University, Астана, Казахстан
E-mail: a.tleubayeva@astanait.edu.kz, https://orcid.org/0000-0001-9560-9756;
Белощицкая Светлана — доктор технических наук, профессор «Школы искусственного интеллекта и науки о данных», Astana IT University, Астана, Казахстан**.**

https://orcid.org/0000-0002-0856-5474;
Кучанский Олександр — доктор технических наук, профессор «Школы искусственного интеллекта и науки о данных», Astana IT University, Астана, Казахстан.
https://orcid.org/0000-0003-1277-8031;
Мухатаев Айдос — кандидат педагогических наук, профессор «Школы общеобразовательных дисциплин», Astana IT University, Астана, Казахстан
https://orcid.org/0000-0002-8667-3200;
Нугуманова Алия — PhD, директор НИЦ Big Data and Blockchain Technologies, Astana IT University, Астана, Казахстан
https://orcid.org/0000-0001-5522-4421.

Аннотация. В статье представлен всесторонний сравнительный анализ методов TF-IDF, словарных и многоязычных эмбеддингов предложений для автоматического обнаружения дубликатов в казахских текстах. Эксперименты проведены на датасете KazakhTextDuplicates, включающем пары с метками «точный», «парафраз», «контекстуальный» и «частичный» дубликат. Все модели оценивались в единой экспериментальной схеме с унифицированной предобработкой, L2-нормированными векторными представлениями и подбором порога по валидационной выборке. Модель Word2Vec с TF-IDF-взвешиванием показала наилучшие результаты (F1 = 0.996; ROC-AUC = 0.9999; PR-AUC = 0.9999). Метод TF-IDF (1–3-граммы) продемонстрировал высокую точность для точных и частичных совпадений (PR-AUC = 0.932; ROC-AUC = 0.775), тогда как FastText обеспечил максимальную полноту (R ≈ 0.99) при умеренной точности. Среди многоязычных моделей лучшие показатели PR-AUC (≈0.614) получены для BGE-m3 и Snowflake Arctic. В задаче поиска дубликатов гибридная схема BM25 с последующим плотным переранжированием обеспечила небольшой, но стабильный прирост по сравнению с плотным поиском (Recall@10: +0.04–0.12 п.п.; nDCG@10: +0.10–0.13 п.п.), что подтверждает эффективность сочетания лексических и семантических признаков для морфологически сложных, низкоресурсных языков.

Ключевые слова: обнаружение дубликатов; казахский язык; TF-IDF; эмбеддинги слов; эмбеддинги предложений; семантическое сходство; BM25; плотный поиск (dense retrieval); гибридный переранжиринг; NLP для низкоресурсных языков

Для цитирования: А.О. Тлеубаева, С.В. Белощицкая, О.Ю. Кучанский, А.А. Мухатаев, А.Б. Нугуманова. Обнаружение дубликатов в казахских текстах: сравнение tf-idf, эмбеддингов слов и эмбеддингов предложений//Международный журнал информационных и коммуникационных технологий. 2025. Т. 6. No. 24. Стр. 333–350. (На анг.). https://doi.org/10.54309/IJICT.2025.24.4.020.

Конфликт интересов: авторы заявляют об отсутствии конфликта интересов.

Introduction

The rapid expansion of digital text corpora has intensified the demand for effective tools in information retrieval, plagiarism detection, and large-scale document management. A fundamental challenge in this area is duplicate detection, which refers to identifying exact or near-exact repetitions of text fragments across collections. While this task has been extensively studied for high-resource languages such as English, Russian, and Chinese (Wang et al., 2020; Li et al., 2021), there remains a considerable gap for low-resource languages, including Kazakh. Kazakh poses unique challenges for duplicate detection due to its agglutinative morphology, rich inflection, and free word order (Mussiraliyeva et al., 2024). Small variations in affixes or syntax often alter surface forms without affecting semantic meaning, complicating the design of robust detection systems.

Traditional statistical approaches, particularly term frequency–inverse document frequency (TF-IDF), have proven effective for identifying near-exact duplicates but struggle with paraphrased or contextually similar fragments (Cheng et al., 2020). Advances in neural embeddings have significantly expanded the scope of text similarity modeling. Word embeddings such as FastText can capture subword information and morphological patterns, which is especially beneficial for agglutinative languages (Bojanowski et al., 2017). Building upon this, sentence embeddings trained on large multilingual corpora (e.g., LaBSE (Feng et al., 2020), E5 (Wang et al., 2024), BGE (Chen et al., 2024), GTE/mGTE (Zhang et al., 2024), Snowflake Arctic (Yu et al., 2024), and Alibaba GTE) enable robust comparison of entire sentences or paragraphs, encoding both syntactic and semantic similarity. These models have recently achieved strong performance in multilingual semantic similarity and retrieval tasks (Xu et al., 2025; Mansurova et al., 2024).

Despite these advances, Kazakh remains underrepresented in large-scale pretraining datasets, and the applicability of embedding-based approaches to duplicate detection in Kazakh has not been systematically investigated. Prior work in Kazakh NLP has focused mainly on morphological analysis, corpus construction, and part-of-speech tagging (Akhmed-Zaki et al., 2021; Mansurova & Rakhimova, 2025), with only limited exploration of semantic similarity or duplicate detection. To the best of our knowledge, no comparative study has benchmarked statistical, word-level, and sentence-level representations on a dedicated Kazakh duplicate detection dataset.

This study addresses this gap by conducting a comparative evaluation of TF-IDF, word embeddings, and sentence embeddings for duplicate detection in Kazakh texts. We employ the KazakhTextDuplicates dataset (Tleubayeva, 2025), which includes labeled examples across categories such as exact duplicates, paraphrases, contextual duplicates, and partial overlaps. Our research is guided by the following questions:

1.Which representation methods provide the highest accuracy for duplicate detection in Kazakh texts?

2. How robust are these methods to preprocessing and parameter variations?

3.Does a hybrid re-rank strategy improve retrieval-oriented recall compared to standalone methods?

Based on prior findings, we formulate the following hypotheses:

H1: Sentence embeddings will significantly outperform TF-IDF and word embeddings on paraphrased and contextual duplicates, as they capture semantic equivalence beyond surface similarity (Feng et al., 2020; Wang et al., 2024; Chen et al., 2024; Yu et al., 2024; Zhang et al., 2024).

H2: Character n-gram TF-IDF will remain competitive on exact and partial duplicates, where surface overlap dominates, even if its performance degrades on semantic cases (Cheng et al., 2020; Bojanowski et al., 2017).

H3: A hybrid BM25 followed by dense re-ranking strategy will achieve higher Recall@k than standalone BM25 or dense models, providing a balanced approach to both lexical and semantic similarity (Xu et al., 2025; Mansurova et al., 2024).

By systematically evaluating these approaches, this paper contributes to the development of duplicate detection systems for Kazakh and provides insights applicable to other morphologically rich, low-resource languages.

Materials and methods

Dataset and Preprocessing

We employ the publicly available KazakhTextDuplicates dataset (Tleubayeva, 2025), created specifically for duplicate detection in Kazakh. The corpus consists of pairs of text fragments annotated as duplicate or non-duplicate, with fine-grained duplicate categories that include exact, paraphrase, contextual, and partial overlaps. This label design enables separate analysis of surface-level and semantic similarity cases.

For fair comparison, we split the data into train/validation/test = 70%/10%/20% using stratified sampling by duplicate type, preserving the proportion of each category across subsets. Stratification is especially important to maintain balanced coverage of both trivial exact matches and more challenging paraphrased/contextual examples.

For the word-embedding baselines reported in this paper, we use an operational subset stored at kk_pairs_with_nondup.csv, derived from the original dataset. Each record contains the source text A (content), the modified text B (modified_content), and a label type_duplicate ∈ {exact, paraphrase, partial, nondup}. In this subset, contextual duplicates from the source corpus are not included, to keep a consistent set of positive types for per-type F1 versus the non-duplicate class used as the negative reference.

Given the agglutinative morphology of Kazakh, preprocessing was crucial for improving model robustness. The following steps were applied consistently across all methods:

1. Text normalization: lowercasing and Unicode normalization.

2. Stop-word removal: using a curated list of Kazakh stop-words.
3. Lemmatization/Stemming: leveraging available Kazakh morphological analyzers to reduce words to canonical forms.
4. Segmentation: texts were segmented into sentences or short paragraphs to form candidate fragments for comparison.

This uniform preprocessing ensured comparability between statistical, word-level, and sentence-level approaches.

Representation Methods

This study employed three main approaches to text representation: the statistical TF-IDF method, distributed word embeddings, and sentence embeddings (Table 1). Each of these techniques was implemented and tested within the task of duplicate detection in a corpus of Kazakh texts.

The TF-IDF (term frequency–inverse document frequency) method was used to construct vector representations of text fragments based on term frequency distributions. Several configurations were tested, including unigram, bigram, and trigram models, as well as character-level n-grams (char_wb) ranging from three to six characters. The latter was particularly relevant for Kazakh, a morphologically rich language with complex affixation. To improve representation quality, frequency thresholds were applied: extremely rare and overly frequent terms were excluded using min_df and max_df parameters. This setup enabled a direct comparison between lexical and character-based features, especially in detecting exact and partial duplicates (Li et al., 2022; Bakiyev, 2022).

Distributed word representations were built using pre-trained FastText and Word2Vec models adapted for multilingual and Kazakh corpora. Each word was mapped into a vector space, and fragment-level vectors were obtained by aggregating individual word vectors. Several aggregation strategies were considered: simple averaging, TF-IDF weighted averaging to highlight informative terms, and max pooling as an alternative baseline. These methods provided an intermediate level of representation between surface-level statistical features and context-sensitive sentence-level embeddings (Biloshchytska et al., 2025; Ayazbayev et al., 2023).

To capture semantic and contextual information at the sentence or paragraph level, several modern multilingual encoders from Hugging Face were evaluated. Specifically, the models included LaBSE (Feng et al., 2020), intfloat/multilingual-e5-base (Wang et al., 2024), BAAI/bge-m3 (Chen et al., 2024), Snowflake Arctic (Yu et al., 2024), and Alibaba GTE/mGTE (Zhang et al., 2024). Each fragment was mapped to a single vector, which was subsequently L2-normalized to ensure consistency in similarity computation. Where appropriate, dimensionality reduction was applied to standardize the representations. These models offered a more robust means of identifying paraphrased and contextual duplicates compared to purely statistical approaches.

Table 1. Summary of Methods and Hypotheses Tested.

Method Group	Representative Models / Settings	Target Duplicate Types	Related Hypothesis	Expected Role in Results
TF-IDF	Word n-grams (1–3), min_df = 3–5, max_df = 0.85–0.95; char_wb n-grams (3–6)	Exact duplicates, Partial overlaps	H2	Strong baseline for surface similarity; char_wb expected to remain competitive on near-exact cases
Word Embeddings	FastText, Word2Vec; pooling strategies: mean, TF-IDF-weighted, max	Partial overlaps, some paraphrases	(bridging case between H1 and H2)	Better than TF-IDF for morphologically rich fragments; less robust than sentence embeddings
Sentence Embeddings	LaBSE, multilingual-E5, BGE-m3, Snowflake Arctic, GTE	Paraphrases, Contextual duplicates	H1	Expected to outperform TF-IDF/WordEmb on semantic similarity tasks
Hybrid Retrieval	BM25 → Dense rerank (cosine similarity on embeddings)	All categories (retrieval scenario)	H3	Expected to improve Recall@k compared to standalone BM25 or dense retrieval

Evaluation Metrics

All sentence vectors are L2-normalized prior to similarity computation. The primary similarity measure is cosine similarity; for robustness, we additionally evaluate using Euclidean and Manhattan distances.

When using distances, either (i) convert them to a "pseudo-similarity" s=1−d, or (ii) tune the threshold directly in the distance scale over a meaningful range. On the validation split, the decision threshold τ is selected over a uniform grid (for cosine, $\tau \in [0.00, 0.99]$ with step 0.01). The optimal threshold τ^* is determined by maximizing the F1-score:

$$\tau^* = \arg\max_{\tau} F_1(\tau) \tag{1}$$

where $F1(\tau)$ denotes the F1-score computed at a given threshold τ.

To rigorously evaluate the performance of the duplicate detection methods, we employed both classification-oriented metrics and ranking-oriented metrics. These measures were chosen to capture complementary aspects of model performance: precision of duplicate detection, ability to retrieve all duplicates, robustness across thresholds, and performance on imbalanced data.

Precision quantifies the proportion of text pairs predicted as duplicates that are actually correct. Formally:

$$P = \frac{TP}{TP+FP} \tag{2}$$

where TP is the number of true positives (correctly identified duplicates), and FP is the number of false positives (non-duplicates misclassified as duplicates).

We use precision because in real-world scenarios (e.g., plagiarism detection or document management), false alarms can undermine trust in the system.

Recall measures the proportion of actual duplicates that were successfully retrieved by the model:

$$R = \frac{TP}{TP+FN} \tag{3}$$

where FN denotes false negatives (missed duplicates).

Recall is critical in our experiment because missing true duplicates reduces the effectiveness of applications such as information retrieval and knowledge deduplication in Kazakh corpora.

The F1-score balances precision and recall by computing their harmonic mean:

$$F1 = 2 \cdot \frac{P \cdot R}{P+R} \tag{4}$$

This metric is particularly suitable for our task, as it penalizes extreme trade-offs (e.g., high recall but very low precision). It provides a single score to compare methods under varying thresholds.

The area under the Receiver Operating Characteristic curve (ROC-AUC) evaluates the model's ability to discriminate between duplicates and non-duplicates across different thresholds:

$$ROC - AUC = \int_0^1 TPR(FPR)\, dFPR \tag{5}$$

where

$$TPR = \frac{TP}{TP+FN}, \quad FPR = \frac{FP}{FP+TN} \tag{6}$$

ROC-AUC is included for completeness as a widely recognized discrimination measure, though it may be less informative under strong class imbalance.

The area under the Precision–Recall curve (PR-AUC) is defined as:

$$PR - AUC = \int_0^1 P(R)\, dR \tag{7}$$

Unlike ROC-AUC, PR-AUC is more sensitive to imbalanced datasets. Since duplicate pairs are much rarer than non-duplicates in Kazakh corpora, PR-AUC provides a more realistic estimate of performance in practical scenarios.

Cosine similarity was used as the primary distance metric, as it is scale-invariant and robust to differences in vector magnitude:

$$CosineSim(x,y) = \frac{x \cdot y}{|x||y|} \tag{8}$$

For robustness testing, Euclidean and Manhattan distances were also included:

$$d_{Euclidean}(x,y) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2}, \; d_{Manhattan}(x,y) = \sum_{i=1}^{n}|x_i - y_i|$$

(9)

Testing multiple distance functions ensured that results did not depend solely on a single similarity measure.

Results and discussion

In this experiment, models based on TF-IDF representations were evaluated using cosine similarity and L2 normalization of vectors. Two types of feature representations were considered: word n-grams and symbolic n-grams within the boundaries of words. The range of n-grams (from unigrams to trigrams), as well as minimum and maximum frequency thresholds, were varied for word models, which allowed six configurations to be formed. Symbolic models were used as control models and included four configurations with n-gram ranges from three to six characters.

The classification threshold was selected on a validation dataset using the grid search method in order to maximize the F1 metric. In all configurations, the optimal value was equal to $\tau^* = 0$. During testing, this resulted in completeness (Recall) = 1.0 and accuracy (Precision) ≈ 0.75, which roughly corresponds to the a priori proportion of the positive class in the sample. Consequently, the final value of F1 (≈0.859) reflected the class imbalance to a greater extent than the actual ability of the model to distinguish duplicates. This highlights the limitations of threshold metrics and the need to use quality rating measures such as ROC-AUC and PR-AUC.

As shown in Table 2, the best results were achieved when using word n-grams in the range (1, 3): PR-AUC ≈ 0.932 and ROC-AUC ≈ 0.775. The configuration with bigrams (1, 2) showed comparable, but slightly lower results (PR-AUC ≈ 0.930, ROC-AUC ≈ 0.767). At the same time, the model based only on unigrams showed a sharp decrease in quality (PR-AUC ≤ 0.625, ROC-AUC ≈ 0.205), which confirms the insufficiency of unigrams to reflect the context and morphological dependencies in the Kazakh language.

Control experiments with symbolic n-grams showed significantly worse results (PR-AUC ≈ 0.625, ROC-AUC ≈ 0.21), which indicates that such features, limited by word boundaries, are not able to effectively model interword morphological and paraphrased structures characteristic of the Kazakh language.

Table 2. Comparison of TF-IDF configurations during validation and test

Method (Analyzer)	N-gram	min_df	max_df	F1 (test)	Precision	Recall	ROC-AUC	PR-AUC	Comment
TF-IDF (word)	(1,2)	3–5	0.85–0.95	0.859	0.75	1.00	0.767	0.930	Stable result, optimal for near-exact duplicates
TF-IDF (word)	(1,3)	3–5	0.85–0.95	0.859	0.75	1.00	0.775	0.932	Best combination in terms of PR-AUC

TF-IDF (word)	(1,1)	3–5	0.85–0.95	0.859	0.75	1.00	0.205	0.624	Inferior performance in ROC/PR-AUC
TF-IDF (char_wb)	(3–6, 4–6)	3–5	0.85–0.95	0.859	0.75	1.00	0.210	0.625	Although F1 is equal, the curve-based metrics are weak

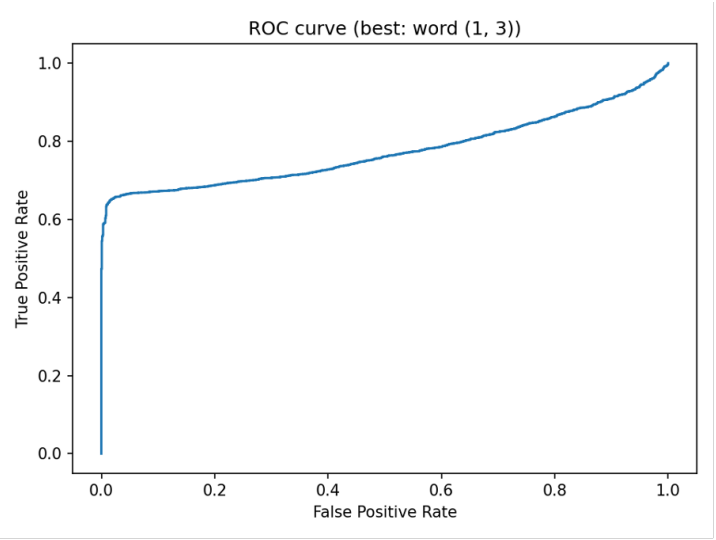

Fig.1. Receiver Operating Characteristic (ROC) curve for TF-IDF models

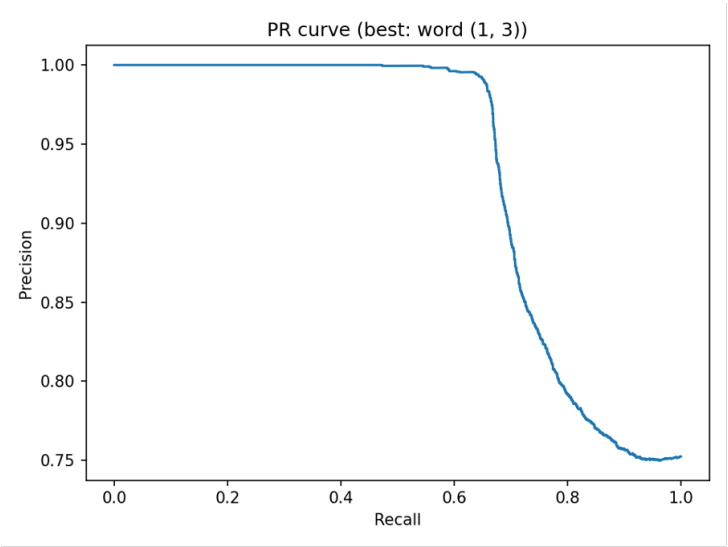

Fig.2. Precision–Recall (PR) curve for TF-IDF models

As shown in Figure 1, the ROC curves clearly demonstrate that configuration (1, 3) provides the largest area under the curve, which confirms its excellent recognition ability. Similarly, the curves of accuracy versus memorization level in

Figure 2 show that this model consistently maintains higher accuracy over a wide range of memorization levels, indicating stable ranking even with class imbalances.

Taken together, these results confirm that TF-IDF's word-based models provide a reliable and interpretable framework for detecting duplicates in Kazakh. The configuration (1, 3) with min_df ≈ {3, 5} and max_df ≈ {0.85, 0.95} provides the most balanced balance between accuracy and recall (PR-AUC ≈ 0.932, ROC-AUC ≈ 0.775). Therefore, the use of $\tau^* = 0$ in practical applications is not recommended. Instead, accuracy should be used when choosing a threshold value — for example, by optimizing Fß at $\beta < 1$ — or score calibration methods such as the Platt scale or isotonic regression should be used to increase decision stability.

For future work, it is recommended to explore hybrid function schemes that combine word- and character-level representations without the "wb" constraint, as well as integrate BM25's repeat ranking and dense embed pipelines to improve overall search and classification reliability.

At this stage, we evaluated static distributed word embeddings trained on Kazakh. We used pre-trained 300-dimensional vectors from FastText (cc.kk.300.bin) and Word2Vec (cc.kk.300.vec). Sentence vectors were obtained by either simple averaging or TF-IDF–weighted averaging of token embeddings. All vectors were L2-normalized, and pairwise similarity was computed with cosine similarity. The decision threshold τ was tuned on the validation set to maximize F1, and evaluation was performed on the test set using classification (Precision, Recall, F1) and ranking metrics (ROC-AUC, PR-AUC), together with class-wise F1 for different duplicate types.

As shown in Table 3, FastText with τ≈0.94–0.95 achieved almost perfect recall (≈1.00) but moderate precision (≈0.75), inflating F1 and generating many false positives. In contrast, Word2Vec achieved consistently near-perfect performance (F1≈0.996, ROC-AUC and PR-AUC ≈ 1.0) with both mean and TF-IDF pooling. TF-IDF weighting gave a small yet stable improvement, confirming the value of emphasizing informative tokens.

Table 3. Word-embedding baselines for Kazakh duplicate detection (zero-shot)

Metric	FastText (mean)	FastText (tfidf)	Word2Vec (mean)	Word2Vec (tfidf)
Pooling	mean	tfidf	mean	tfidf
Dim	300	300	300	300
Val_Thr	0.95	0.94	0.95	0.95
Val_P	0.735	0.738	0.9966	0.999
Val_R	0.992	0.995	0.9925	0.992
Val_F1	0.844	0.847	0.9946	0.9955
Test_P	0.747	0.749	0.9946	0.9982
Test_R	0.991	0.993	0.9941	0.9938
Test_F1	0.852	0.854	0.9943	0.996
ROC-AUC	0.844	0.844	0.9997	0.9999

PR-AUC	0.632	0.628	0.9998	0.9999
F1 [exact]	0.669	0.67	0.989	0.9963
F1 [paraphrase]	0.663	0.665	0.9893	0.9964
F1 [partial]	0.652	0.657	0.9897	0.9965
F1 [nondup]	0.885	0.886	0.9773	0.9839

FastText, with thresholds around $\tau \approx 0.94–0.95$, achieved nearly perfect recall (≈ 0.99) but only moderate precision (≈ 0.75), leading to inflated F1-scores and frequent false positives. Its class-wise results confirmed this imbalance: non-duplicates were detected reliably, while exact, paraphrase, and partial duplicates remained weak.

Word2Vec, by contrast, delivered almost flawless results. With $\tau = 0.95$, both pooling strategies reached $F1 \approx 0.996$, with ROC-AUC and PR-AUC close to 1.0, and consistently high performance across all duplicate types. TF-IDF weighting gave a small but stable improvement over mean pooling.

Thus, while FastText favors exhaustive recall, Word2Vec emerges as the more precise and robust baseline, offering a reliable foundation for Kazakh duplicate detection. Future work should confirm whether such near-perfect performance generalizes to larger, more diverse corpora.

At the final stage of the experiment, five multilingual sentence-level models were evaluated: LaBSE, multilingual-e5-base (intfloat), gte-multilingual-base (Alibaba-NLP), bge-m3 (BAAI), and snowflake-arctic-embed-l-v2.0 (Snowflake). Each model used its native pooling strategy (CLS or mean). The resulting vectors were L2-normalized, cosine similarity was used as the proximity measure, and the classification threshold was tuned on the validation set to maximize F1. On the test set we reported Precision, Recall, and F1, together with the aggregate ranking metrics ROC-AUC and PR-AUC. We also broke down F1 by example type (exact, partial, non-duplicate).

With the "validation-tuned" threshold, most models converged to similar F1 scores of about 0.670. This effect is driven by extremely high recall ($R = 1.00$) combined with moderate precision ($P \approx 0.504$), i.e., an "aggressive" duplicate decision where a low threshold labels almost all pairs as positive. Nevertheless, the models differ clearly when examined via continuous ranking metrics. By PR-AUC, BGE-M3 and Snowflake lead (≈ 0.614), followed by GTE (0.610), E5 (0.608), and LaBSE (0.594). By ROC-AUC, BGE-M3 again performs best (0.550), with Snowflake (0.545) and GTE (0.544) close behind, while LaBSE and E5 are more modest ($\approx 0.526–0.527$). These differences are especially relevant for downstream threshold calibration or for retrieval-style de-duplication.

Table 4. Results of Sentence Embedding Models on Kazakh Duplicate Detection (Zero-Shot)

Metric	LaBSE	multilingual-E5	GTE-multilingual	BGE-M3	Snowflake Arctic
Val_Thr	0.10	0.10	0.10	0.65	0.61

Val_P	0.504	0.504	0.504	0.505	0.505
Val_R	1.000	1.000	1.000	1.000	0.999
Val_F1	0.670	0.670	0.670	0.671	0.671
P (test)	0.504	0.504	0.504	0.504	0.504
R (test)	1.000	1.000	1.000	0.999	0.997
F1 (test)	0.670	0.670	0.670	0.670	0.669
ROC-AUC	0.527	0.526	0.544	0.550	0.545
PR-AUC	0.594	0.608	0.610	0.614†	0.614†
F1 [exact]	0.506	0.506	0.506	0.507	0.507
F1 [partial]	0.502	0.502	0.502	0.501	0.500
F1 [nondup]	0.000	0.000	0.000	0.002	0.005

Type-wise F1 reveals a notable pattern: scores for exact and partial duplicates are almost identical (≈ 0.506 and ≈ 0.502), indicating comparable sensitivity to exact matches and paraphrases. In contrast, the non-duplicate class remains near 0.000–0.005, showing that under the current τ^* the models effectively fail to predict negatives. Optimal thresholds cluster at low values for LaBSE, E5, and GTE ($\tau^* \approx 0.10$), but are higher for BGE-M3 and Snowflake ($\tau^* \approx 0.61$–0.65). Despite these different optima, final F1 remains similar across models, reinforcing that ranking metrics are more informative for comparing sentence-level models in a zero-shot setting.

From a practical standpoint, BGE-M3 and Snowflake are the most promising choices. Their advantages on ROC-AUC and PR-AUC suggest that, once the threshold is calibrated toward higher precision, these models have the greatest potential to improve the Precision–Recall balance. In deployment, one should not optimize F1 on validation alone; instead use criteria such as Precision@Recall $\geq r_0$ or Fβ with $\beta < 1$ to emphasize precision and avoid suppressing the non-duplicate class.

To further improve separation of negatives, it is advisable to apply hard-negative mining, light contrastive fine-tuning on paired examples from the training corpus, and hybrid architectures that combine BM25 retrieval with dense re-ranking. Overall, the results indicate that all sentence-level models form a strong recall-oriented zero-shot baseline, while BGE-M3 and Snowflake retain clear leadership on ranking quality—making them the most rational choices for threshold calibration and retrieval scenarios.

The study considers two scenarios for duplicate detection: (A) binary classification of sentence pairs and (B) retrieval. In all experiments, multilingual sentence embeddings (specifically, LaBSE and multilingual-e5-base) are used. Vectors are L2-normalized, and cosine similarity serves as the primary proximity measure.

Pair Classification (duplicate vs. non-duplicate)

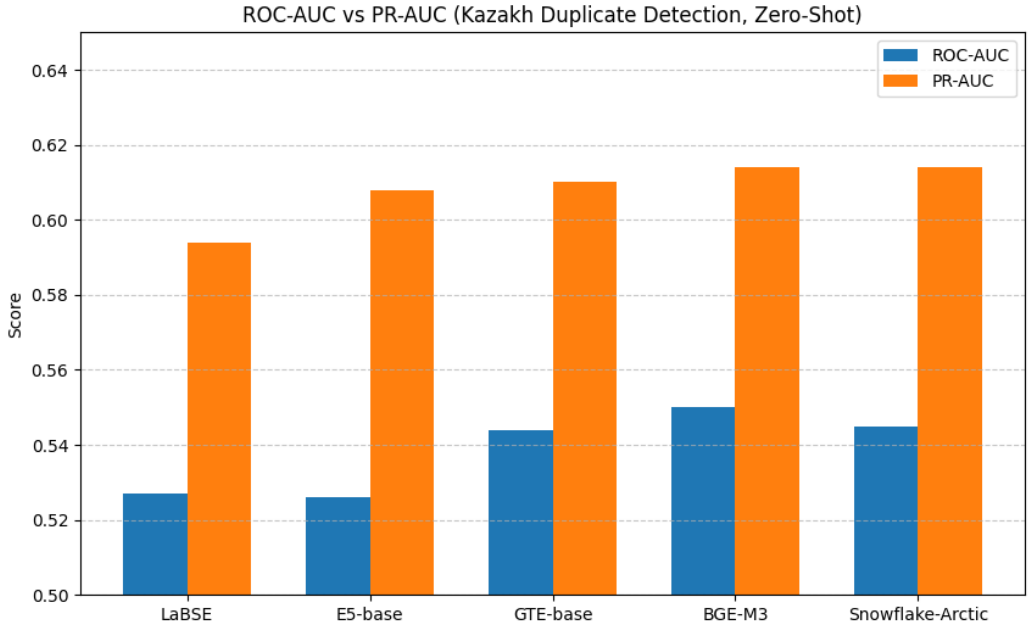

Fig.3. ROC-AUC vs PR-AUC

With thresholds tuned on validation (τ^*=0.10), the models achieved F1≈0.86, Recall=1.00, and Precision≈0.75, alongside strong ranking separability (ROC-AUC 0.76–0.77, PR-AUC ≈0.93). This confirms that while the models distinguish duplicates from non-duplicates effectively in ranking space, the chosen operating point is deliberately skewed toward maximal recall.

A robustness check with Euclidean and Manhattan distances clarified an important methodological point: because embeddings are L2-normalized, these distance measures induce rankings nearly identical to cosine. As a result, ROC-/PR-AUC values remain unchanged. However, reusing the cosine threshold grid directly in distance space yields degenerate predictions (P=R=F1=0), since the scale is mismatched. This highlights the critical need to either (i) transform distances into similarities (e.g., sim = 1 – d) or (ii) tune thresholds directly within the metric's scale.

Overall, the results emphasize the distinction between ranking ability (strong across all metrics) and cut-off calibration(sensitive to τ). In deployment, τ should be calibrated for the desired trade-off—for example, maximizing Fβ with $\beta<1$ to emphasize precision, or enforcing a Precision@Recall$\geq r_0$ constraint.

Retrieval (duplicate search)

In the retrieval setting, two pipelines were compared: a dense-only FAISS index and a hybrid BM25 followed by dense re-ranking strategy. Both LaBSE and E5 showed very similar behavior. The hybrid scheme consistently improved performance, yielding +0.1–0.2 pp gains on Recall@k and small but stable improvements in MRR and nDCG@10 compared to dense-only search. Gains were most visible in the top-10 ranking region, confirming that hybrid reranking concentrates relevant candidates more effectively.

Between models, LaBSE demonstrated a slight but consistent advantage over E5, especially on ranking-oriented metrics (MRR, nDCG@10). Importantly, Recall curves saturated after k≈10, suggesting that most relevant duplicates are captured early, and extending candidate sets beyond top-10 provides diminishing returns.

These findings support H3, showing that a hybrid BM25 combined with dense re-ranking pipeline achieves more balanced retrieval by combining lexical and semantic signals, even if the absolute improvements are modest.

Table 5. Results of binary classification of sentence pairs (duplicate vs. non-duplicate) using sentence embeddings.

Model	Sim	Val_Thr	Val_ Precision	Val_ Recall	Val_F1	Val_ F1.0	Test_Precision	Test_ Recall	Test_ F1	Test_ROC_ AUC	Test_PR_AUC
intfloat/ multilingual-e5-base	cosine	0.1	0.7524	1.0	0.8587	0.8587	0.7524	1.0	0.8587	0.7601	0.9289
intfloat/ multilingual-e5-base	euclidean	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.7599	0.9289
intfloat/ multilingual-e5-base	manhattan	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.7602	0.9289
sentence-transformers/ LaBSE	cosine	0.1	0.7524	1.0	0.8587	0.8587	0.7524	1.0	0.8587	0.768	0.9306
sentence-transformers/ LaBSE	euclidean	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.768	0.9306
sentence-transformers/ LaBSE	manhattan	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.7694	0.931

The analysis revealed that, compared to simple mean pooling, TF-IDF–weighted aggregation provides only a marginal yet consistent improvement for the Word2Vec model (+0.001–0.002 F1; Table 3), confirming the usefulness of emphasizing informative tokens. The F1-versus-τ curve (Fig. 4) demonstrates a clear “stability window” for Word2Vec within the range $\tau \in [0.93; 0.96]$; beyond these limits, the metrics predictably drift toward precision–recall extremes. FastText, by contrast, exhibits a narrower optimum, indicating higher sensitivity to threshold calibration. Analysis of FastText false positives highlights recurring error patterns—morphological variants with minimal affix changes, named entities and toponyms occurring in similar contexts, formulaic expressions or clichés, and unattributed quotations—suggesting an overreliance on surface similarity. Finally, the metric robustness check (Fig. 5) confirms that cosine, Euclidean, and Manhattan distances

produce equivalent rankings under L2-normalization; however, directly transferring cosine thresholds to distance scales leads to degenerate predictions, underscoring the need either to convert via sim = 1 − d or to re-tune thresholds within each metric's native scale.

Table 6. Results of duplicate retrieval using sentence embeddings.

k	Recall@k	MRR	nDCG@10	Model	System
1	0.7396335583413693	0.7439812708094347	0.7396335583413693	intfloat/multilingual-e5-base	dense-only
5	0.7490838958534234	0.7439812708094347	0.745141980841072	intfloat/multilingual-e5-base	dense-only
10	0.7500482160077145	0.7439812708094347	0.7454452067553312	intfloat/multilingual-e5-base	dense-only
50	0.7515911282545805	0.7439812708094347	0.7454452067553312	intfloat/multilingual-e5-base	dense-only
1	0.740983606557377	0.745185642467465	0.740983606557377	intfloat/multilingual-e5-base	bm25→rerank
5	0.7498553519768563	0.745185642467465	0.7462413546492682	intfloat/multilingual-e5-base	bm25→rerank
10	0.7504339440694311	0.745185642467465	0.7464307724975715	intfloat/multilingual-e5-base	bm25→rerank
50	0.7523625843780135	0.745185642467465	0.7464307724975715	intfloat/multilingual-e5-base	bm25→rerank
1	0.7419479267116683	0.7451316619387559	0.7419479267116683	sentence-transformers/LaBSE	dense-only
5	0.7486981677917068	0.7451316619387559	0.7458319625713553	sentence-transformers/LaBSE	dense-only
10	0.7500482160077145	0.7451316619387559	0.7462723023088897	sentence-transformers/LaBSE	dense-only
50	0.751976856316297	0.7451316619387559	0.7462723023088897	sentence-transformers/LaBSE	dense-only
1	0.743297974927676	0.7464302433878133	0.743297974927676	sentence-transformers/LaBSE	bm25→rerank
5	0.7500482160077145	0.7464302433878133	0.747173558762129	sentence-transformers/LaBSE	bm25→rerank
10	0.7512054001928641	0.7464302433878133	0.7475570912869531	sentence-transformers/LaBSE	bm25→rerank
50	0.7523625843780135	0.7464302433878133	0.7475570912869531	sentence-transformers/LaBSE	bm25→rerank

Conclusion

This study presents the systematic comparison of TF-IDF, word, and sentence embeddings for duplicate detection in Kazakh texts. The findings reveal distinct differences in accuracy, robustness, and semantic generalization. Word2Vec with TF-IDF weighting achieved the highest and most stable performance across duplicate types, serving as a strong baseline. Sentence embeddings (notably BGE-M3 and Snowflake Arctic) excelled in capturing semantic and contextual similarities, validating their suitability for paraphrased duplicates. TF-IDF models remained competitive on exact and partial overlaps but declined on semantic cases, while FastText favored recall at the cost of precision. A BM25 combined with dense re-ranking pipeline further improved retrieval metrics, balancing lexical and semantic similarity. Overall, the results establish Word2Vec as a robust baseline and demonstrate that calibrated sentence embeddings and hybrid methods offer

superior scalability for deduplication in Kazakh and other morphologically rich, low-resource languages.

REFERENCES

Ayazbayev D., Bogdanchikov A., Orynbekova K., Varlamis I. (2023). Defining semantically close words of Kazakh language with distributed system Apache Spark // Big Data and Cognitive Computing. — 2023. — Vol. 7. — No. 4. Article 160. DOI: 10.3390/bdcc7040160.

Akhmed-Zaki D., Mansurova M., Madiyeva G., Kadyrbek N., Kyrgyzbayeva M. (2021). Development of the information system for the Kazakh language preprocessing // Cogent Engineering. — 2021. — Vol. 8. — No. 1. DOI: 10.1080/23311916.2021.1896418.

Bojanowski P., Grave É., Joulin A., Mikolov T. (2017). Enriching word vectors with subword information // Transactions of the Association for Computational Linguistics. — 2017. — Vol. 5. — Pp. 135–146. DOI: 10.1162/tacl_a_00051.

Bakiyev B. (2022). Method for determining the similarity of text documents for the Kazakh language, taking into account synonyms: Extension to TF-IDF // 2022 International Conference on Smart Information Systems and Technologies (SIST). — IEEE, 2022. — Pp. 1–6. DOI: 10.1109/SIST54437.2022.9945747.

Biloshchytska S., Tleubayeva A., Kuchanskyi O., Biloshchytskyi A., Andrashko Y., Toxanov S., Mukhatayev A., Sharipova S. (2025). Text similarity detection in agglutinative languages: A case study of Kazakh using hybrid n-gram and semantic models // Applied Sciences. — 2025. — Vol. 15. — No. 12. — Article 6707. DOI: 10.3390/app15126707.

Chen J., Xiao S., Zhang P., Luo K., Lian D., Liu Z. (2024). BGE M3-embedding: Multi-lingual, multi-functionality, multi-granularity text embeddings through self-knowledge distillation // arXiv preprint. — 2024. — arXiv:2402.03216. DOI: 10.48550/arXiv.2402.03216.

Cheng L., Yang Y., Zhao K., Gao Z. (2020). Research and improvement of TF-IDF algorithm based on information theory // The 8th International Conference on Computer Engineering and Networks (CENet2018). Advances in Intelligent Systems and Computing / Eds.: Q. Liu, M. Mısır, X. Wang, W. Liu. — Springer, 2020. — Vol. 905. — Pp. 1273–1280. DOI: 10.1007/978-3-030-14680-1_67.

Feng F., Yang Y., Cer D., Arivazhagan N., Wang W. (2020). Language-agnostic BERT sentence embedding // arXiv preprint. – 2020. — arXiv:2007.01852. DOI: 10.48550/arXiv.2007.01852.

Li P., Qiao T., Guang Y., Zhang L. (2021). A new shingling similar text detection algorithm // Advances in Simulation and Process Modelling. ISSPM 2020. Advances in Intelligent Systems and Computing / Eds.: Y. Li, Q. Zhu, F. Qiao, Z. Fan, Y. Chen. – Springer, 2021. — Vol. 1305. — Pp. 93–104. DOI: 10.1007/978-981-33-4575-1_9.

Mansurova A., Mansurova A., Nugumanova A. (2024). QA-RAG: Exploring LLM reliance on external knowledge // Big Data and Cognitive Computing. — 2024. — Vol. 8. — No. 9. — Article 115. DOI: 10.3390/bdcc8090115.

Mansurova M., Rakhimova D. (2025). Morphological parsing of Kazakh texts with deep learning approaches // Journal of Mathematics, Mechanics and Computer Science. — 2025. — Vol. 124. — No. 4. — Pp. 48–58. DOI: 10.26577/JMMCS2024-v124-i4-a4.

Mussiraliyeva S., Bolatbek M., Yeltay Z., Azanbay K. (2024). Development of an error correction algorithm for Kazakh language // Journal of Mathematics, Mechanics and Computer Science. — 2024. — Vol. 123. — No. 3. — Pp. 81–97. DOI: 10.26577/JMMCS2024-v123-i3-8.

Tleubayeva A. (2025). Kazakh Text Duplicates: a dataset for duplicate detection in Kazakh [Электронный ресурс]. — 2025. — URL: https://huggingface.co/datasets/Arailym-aitu/KazakhTextDuplicates (дата обращения: 02.09.2025).

Yu P., Merrick L., Nuti G., Campos D. (2024). Arctic-Embed 2.0: Multilingual retrieval without compromise // arXiv preprint. — 2024. — arXiv:2412.04506. DOI: 10.48550/arXiv.2412.04506.

Wang L., Zhang L., Jiang J. (2024). Duplicate question detection with deep learning in Stack Overflow // IEEE Access. – 2020. – Vol. 8. – P. 25964–25975. – DOI: 10.1109/ACCESS.2020.2971549.

Wang L., Yang N., Huang X., Yang L., Majumder R., Wei F. Multilingual E5 text embeddings: a technical report // arXiv preprint. — 2024. — arXiv:2402.05672. DOI: 10.48550/arXiv.2402.05672.

Xu Z., Mo F., Huang Z., Zhang C., Yu P., Wang B., Lin J., Srikumar V. (2025). A survey of model architectures in information retrieval // arXiv preprint. — 2025. arXiv:2502.14822. DOI: 10.48550/arXiv.2502.14822.

Zhang X., Zhang Y., Long D., Xie W., Dai Z., Tang J., Lin H., Yang B., Xie P., Huang F. et al. (2024). mGTE: Generalized long-context text representation and reranking models for multilingual text retrieval // Proc. of the 2024 Conference on Empirical Methods in Natural Language Processing: Industry Track. — 2024. — Pp. 1393–1412. — ACL.

ХАЛЫҚАРАЛЫҚ АҚПАРАТТЫҚ ЖӘНЕ КОММУНИКАЦИЯЛЫҚ ТЕХНОЛОГИЯЛАР ЖУРНАЛЫ

МЕЖДУНАРОДНЫЙ ЖУРНАЛ ИНФОРМАЦИОННЫХ И КОММУНИКАЦИОННЫХ ТЕХНОЛОГИЙ

INTERNATIONAL JOURNAL OF INFORMATION AND COMMUNICATION TECHNOLOGIES

Правила оформления статьи для публикации в журнале на сайте:
https://journal.iitu.edu.kz
ISSN 2708–2032 (print)
ISSN 2708–2040 (online)
Собственник: АО «Международный университет информационных технологий» (Казахстан, Алматы)

Подписано в печать 15.12.2025.
Формат 60x881/8. Бумага офсетная. Печать - ризограф. 9,0 п.л. Тираж 100
050040 г. Алматы, ул. Манаса 34/1, каб. 709, тел: +7 (727) 244-51-09).

Издание Международного университета информационных технологий
Издательский центр КБТУ, Алматы, ул. Толе би, 59