ISSN 2518-1726 (Online),
ISSN 1991-346X (Print)

ACADEMIC SCIENTIFIC JOURNAL OF COMPUTER SCIENCE

№3
2025

ISSN 2518-1726 (Online),
ISSN 1991-346X (Print)

CENTRAL ASIAN ACADEMIC
RESEARCH CENTER

ACADEMIC SCIENTIFIC JOURNAL OF COMPUTER SCIENCE

3 (355)

JULY – SEPTEMBER 2025

PUBLISHED SINCE JANUARY 1963
PUBLISHED 4 TIMES A YEAR

ALMATY, NAS RK

Academic Scientific Journal of Computer Science
ISSN 2518-1726 (Online),
ISSN 1991-346X (Print)
Owner: «Central Asian Academic Research Center» LLP (Almaty).
Certificate **№ KZ77VPY00121154** on the re-registration of the periodical printed and online publication of the information agency, issued on **05.06.2025** by the Republican State Institution «Information Committee» of the Ministry of Culture and Information of the Republic of Kazakhstan
Subject area*: information and communication technologies.*
Currently: *included in the list of journals recommended by the CCSES MSHE RK in the direction of «Information and communication technologies».*
Periodicity: *4 times a year.*
http://www.physico-mathematical.kz/index.php/en/

Academic Scientific Journal of Computer Science
ISSN 2518-1726 (Online),
ISSN 1991-346X (Print)

Меншіктеуші: «Орталық Азия академиялық ғылыми орталығы» ЖШС (Алматы).

Ақпарат агенттігінің мерзімді баспасөз басылымын, ақпарат агенттігін және желілік басылымды қайта есепке қою туралы ҚР Мәдениет және Ақпарат министрлігі «Ақпарат комитеті» Республикалық мемлекеттік мекемесі **05.06.2025** ж. берген **№ KZ77VPY00121154** Куәлік.

Тақырыптық бағыты: *ақпараттық-коммуникациялық технологиялар*

Қазіргі уақытта: *«ақпараттық-коммуникациялық технологиялар» бағыты бойынша ҚР БҒМ БҒСБК ұсынған журналдар тізіміне енді.*

Мерзімділігі: *жылына 4 рет.*

http://www.physico-mathematical.kz/index.php/en/

Academic Scientific Journal of Computer Science
ISSN 2518-1726 (Online),
ISSN 1991-346X (Print)
Собственник: *ТОО «Центрально-азиатский академический научный центр» (г. Алматы).*
Свидетельство о постановке на переучет периодического печатного издания, информационного агентства и сетевого издания **№ KZ77VPY00121154.** Дата выдачи **05.06.2025**
Тематическая направленность: *информационно-коммуникационные технологии.*
В настоящее время: *вошел в список журналов, рекомендованных КОКСНВО МНВО РК по направлению «информационно-коммуникационные технологии».*
Периодичность: *4 раза в год.*
http://www.physico-mathematical.kz/index.php/en/

CONTENTS

МАЗМҰНЫ

СОДЕРЖАНИЕ

ACADEMIC SCIENTIFIC JOURNAL OF COMPUTER SCIENCE
ISSN 1991-346X
Volume 3. Namber 355 (2025). 160–175

https://doi.org/10.32014/2025.2518-1726.370

UDC 004.912:004.855

© T.S. Sadykova[1*], B.K. Sinchev[1], Im Cho Young[2], A.S. Auyezova[1], 2025.
[1]International Information Technology University, Almaty, Kazakhstan;
[2]Gachon University, Seoul, South Korea.
*E-mail: sadykovatolkynai@gmail.com

THE APPLICATION OF VECTOR SPACE MODELS IN INTELLIGENT INFORMATION RETRIEVAL SYSTEMS

Sadykova Tolkynay Seitkadyrovna — PhD student, Department of Information Systems, International University of Information Technologies, Almaty, Kazakhstan,
E-mail: sadykovatolkynai@gmail.com, ORCID ID: https://orcid.org/0000-0002-6462-3894;
Sinchev Bakhtgerey Kuspanovich — Professor, Department of Information Systems, International University of Information Technologies, Almaty, Kazakhstan,
E-mail: sinchev@mail.ru, ORCID ID: https://orcid.org/0000-0001-8557-8458;
Young Im Cho — Professor, Faculty of Computer Engineering, Gachon University, Seoul, South Korea,
E-mail: yicho@gachon.ac.k, ORCID ID: https://orcid.org/0000-0003-0184-7599;
Auyezova Anel Sattarkyzy — PhD student, Department of Information Systems, International University of Information Technologies, Almaty, Kazakhstan,
E-mail: anel.auezova@gmail.com, ORCID ID: https://orcid.org/0000-0001-9860-4491.

Abstract. This research addresses the need to improve semantic information retrieval efficiency in low-resource languages, with a focus on Kazakh. Its agglutinative structure, morphological variability, and lexical ambiguity pose challenges for conventional models, which fail to capture grammatical and contextual factors fully. The study aims to develop an approach for selecting and comparing text vectorization models in intelligent search systems, taking into account Kazakh linguistic features, and to construct a mathematical model for computing semantic similarity in a multidimensional vector space. The methodology involved the empirical testing of six models (TF-IDF, Word2Vec, FastText, GloVe, BERT, and KazBERT) on a 24,000-text corpus in Kazakh. Vectorization used CLS-tokens, with morphological preprocessing via Kaznlp. Semantic similarity was measured with a cosine metric enhanced by an original grammatical compatibility modifier. Model performance was evaluated using precision, recall, and F1-score. Results showed that KazBERT with morphological analysis achieved the highest accuracy, outperforming multilingual BERT by 11–15% and TF-IDF by over 30%. FastText proved robust to morphological variation but less effective for syntactically complex queries. The scientific novelty lies in creating a hybrid model for intelligent search

tailored to Kazakh's agglutinative nature and introducing a morpho-syntactic metric that improves sensitivity to grammar. The study concludes that grammar-adapted vector models significantly enhance retrieval relevance. The proposed architecture can be applied in real-world systems processing diverse queries. Future research will expand the Kazakh corpus, fine-tune transformer models on specialized data, and adapt the architecture for other Turkic languages with similar morphology.

Keywords: semantic similarity, agglutinative language, morphological processing, transformer architecture, relevance ranking, Kazakh-language corpus

© Т.С. Садыкова[1*], Б.К. Синчев[1], Im Cho Young[2], А.С. Ауезова[1], 2025.
[1] Халықаралық ақпараттық технологиялар университеті, Алматы, Қазақстан;
[2] Гачон университеті, Сеул, Оңтүстік Корея.
*E-mail: sadykovatolkynai@gmail.com

ИНТЕЛЛЕКТУАЛДЫ АҚПАРАТТЫ ІЗДЕУ ЖҮЙЕЛЕРІНДЕ ВЕКТОРЛЫҚ КЕҢІСТІК МОДЕЛЬДЕРІН ҚОЛДАНУ

Sadykova Tolkynay Seitkadyrovna — PhD студенті, Ақпараттық жүйелер бөлімі, Халықаралық ақпараттық технологиялар университеті, Алматы, Қазақстан,
E-mail: sadykovatolkynai@gmail.com, ORCID ID: https://orcid.org/0000-0002-6462-3894;
Sinchev Bakhtgerey Kuspanovich — профессор, Ақпараттық жүйелер бөлімі, Халықаралық ақпараттық технологиялар университеті, Алматы, Қазақстан,
E-mail: sinchev@mail.ru, ORCID ID: https://orcid.org/0000-0001-8557-8458;
Young Im Cho — профессор, Компьютерлік инженерия факультеті, Гачон университеті, Сеул, Оңтүстік Корея,
E-mail: yicho@gachon.ac.k, ORCID ID: https://orcid.org/0000-0003-0184-7599;
Auyezova Anel Sattarkyzy — PhD студенті, Ақпараттық жүйелер бөлімі, Халықаралық ақпараттық технологиялар университеті, Алматы, Қазақстан,
E-mail: anel.auezova@gmail.com, ORCID ID: https://orcid.org/0000-0001-9860-4491.

Аннотация. Бұл зерттеу шектеулі тілдік ресурстар жағдайында, әсіресе қазақ тіліне қатысты, семантикалық ақпараттық іздеудің тиімділігін арттыруға бағытталған. Қазақ тілінің агглютинативті құрылымы, морфологиялық өзгергіштігі мен лексикалық көпмәнділігі дәстүрлі модельдер үшін елеулі қиындықтар туғызады, себебі олар грамматикалық және контекстік факторларды толық ескере алмайды. Зерттеудің мақсаты – қазақ тілінің ерекшеліктерін ескере отырып, интеллектуалды іздеу жүйелерінде мәтінді векторизациялау модельдерін таңдау мен салыстырудың негізделген тәсілін әзірлеу және көпөлшемді векторлық кеңістікте семантикалық ұқсастықты есептеу үшін математикалық модель құру. Әдіснама 24 000 қазақ мәтінінен тұратын корпус негізінде алты модельді (TF-IDF, Word2Vec, FastText, GloVe, BERT және KazBERT) эмпирикалық сынақтан өткізуге негізделген. Векторизация CLS-токендер арқылы жүргізілді, морфологиялық алдын ала өңдеу Kaznlp құралы арқылы орындалды. Семантикалық ұқсастық грамматикалық сәйкестікті ескеретін түпнұсқа модификатормен

жетілдірілген косинустық метрика көмегімен өлшенді. Модельдердің тиімділігі precision, recall және F1-score метрикалары бойынша бағаланды. Нәтижелер KazBERT моделі морфологиялық талдаумен бірге ең жоғары дәлдікті көрсеткенін дәлелдеді: ол көптілді BERT-тен 11–15%-ға және TF-IDF-тен 30%-дан астамға асып түсті. FastText морфологиялық өзгерістерге төзімді болғанымен, синтаксистік күрделі сұрауларда тиімділігі төмен болды. Ғылыми жаңалығы – қазақ тілінің агглютинативті табиғатына бейімделген интеллектуалды іздеудің гибридті моделін жасау және грамматикалық ерекшеліктерге сезімталдықты арттыратын морфо-синтаксистік метриканы енгізу. Қорытындысында грамматиканы ескеретін векторлық модельдер іздеу релеванттылығын айтарлықтай арттыратыны расталды. Ұсынылған архитектура әртүрлі сұрау түрлерін өңдейтін нақты жүйелерде қолданылуы мүмкін. Болашақ зерттеулердің перспективаларына қазақ тілі корпусын кеңейту, трансформерлерді мамандандырылған деректерде қосымша үйрету және ұқсас морфологиясы бар басқа түркі тілдеріне бейімдеу жатады.

Түйін сөздер: семантикалық ұқсастық, морфологиялық талдау, трансформер үлгісі, ақпараттық релеванттық, қазақ мәтіндік корпусы, аффикстік құрылым

© Т.С. Садыкова[1*], Б.К. Синчев[1], Im Cho Young[2], А.С. Ауезова[1], 2025.
[1]Международный университет информационных технологий,
Алматы, Казахстан;
[2]«Gachon University», Сеул, Южная Корея.
*E-mail: sadykovatolkynai@gmail.com

ПРИМЕНЕНИЕ ВЕКТОРНЫХ МОДЕЛЕЙ В ИНТЕЛЛЕКТУАЛЬНЫХ СИСТЕМАХ ИНФОРМАЦИОННОГО ПОИСКА

Sadykova Tolkynay Seitkadyrovna — аспирант, кафедра информационных систем, Международный университет информационных технологий, Алматы, Казахстан,
E-mail: sadykovatolkynai@gmail.com, ORCID ID: https://orcid.org/0000-0002-6462-3894;
Sinchev Bakhtgerey Kuspanovich — профессор, кафедра информационных систем, Международный университет информационных технологий, Алматы, Казахстан,
E-mail: sinchev@mail.ru, ORCID ID: https://orcid.org/0000-0001-8557-8458;
Young Im Cho — профессор, Faculty of Computer Engineering, «Gachon University», Сеул, Южная Корея,
E-mail: yicho@gachon.ac.k, ORCID ID: https://orcid.org/0000-0003-0184-7599;
Auyezova Anel Sattarkyzy — аспирант, кафедра информационных систем, Международный университет информационных технологий, Алматы, Казахстан,
E-mail: anel.auezova@gmail.com, ORCID ID: https://orcid.org/0000-0001-9860-4491.

Аннотация. Данное исследование направлено на повышение эффективности семантического поиска в условиях ограниченных языковых ресурсов, с акцентом на казахский язык. Его агглютинативная структура, высокая морфологическая изменчивость и лексическая неоднозначность

создают серьёзные трудности для традиционных моделей, которые не способны полноценно учитывать грамматические и контекстуальные факторы. Цель работы заключается в разработке подхода к выбору и сравнению моделей векторизации текста для интеллектуальных поисковых систем с учётом особенностей казахского языка, а также в построении математической модели вычисления семантического сходства в многомерном векторном пространстве. Методология основана на эмпирическом тестировании шести моделей (TF-IDF, Word2Vec, FastText, GloVe, BERT и KazBERT) на корпусе из 24 000 казахских текстов. Векторизация выполнялась с использованием CLS-токенов, морфологическая предобработка осуществлялась с помощью инструмента Kaznlp. Семантическое сходство измерялось косинусной метрикой, дополненной оригинальным модификатором грамматической совместимости. Эффективность моделей оценивалась по метрикам precision, recall и F1-score. Результаты показали, что KazBERT в сочетании с морфологическим анализом продемонстрировал наибольшую точность, превысив показатели многоязычного BERT на 11–15% и TF-IDF более чем на 30%. FastText оказался устойчивым к морфологической вариативности, но менее эффективным при синтаксически сложных запросах. Научная новизна заключается в разработке гибридной модели интеллектуального поиска, адаптированной к агглютинативной природе казахского языка, и в предложении морфо-синтаксической метрики, повышающей чувствительность к грамматическим особенностям. В заключении подтверждается, что адаптация векторных моделей с учётом грамматики существенно повышает релевантность поиска. Предложенная архитектура применима в реальных системах с разнообразными типами запросов. Перспективы дальнейших исследований включают расширение корпуса, дообучение трансформеров на специализированных данных и адаптацию архитектуры для других тюркских языков со схожей морфологией.

Ключевые слова: семантическое сходство; агглютинативный язык; морфологическая обработка; трансформерная архитектура; релевантность; казахскоязычный корпус

Introduction. Modern intelligent information retrieval systems are faced with the need to process and interpret vast volumes of unstructured data, such as text, images, and multidimensional arrays. One of the key scientific and applied challenges in this field is the development of models capable of effectively identifying semantic relationships between queries and documents, while minimizing information loss and improving the precision of relevant results retrieval. Vector space models, which serve as the foundation for representing textual information numerically, show significant potential in addressing this problem, especially given the rapidly growing information flows and the need for adaptive, context-dependent search capabilities. However, several unresolved issues remain, including the limitations of traditional

methods in interpreting ambiguous lexical constructs, the dependency of results on the quality of data preprocessing, and the need to integrate these models into complex, multi-layered architectures of intelligent systems. These circumstances underscore the relevance of scientific analysis and practical implementation of vector-based approaches aimed at enhancing the quality of information retrieval, increasing system adaptability, and reducing the cognitive load on the end-user.

Literature review. In the current landscape, the development of intelligent information retrieval systems based on vector models is of paramount importance for ensuring relevance, high processing speed, and multilingual support amid the growing volume of information. An analysis of scientific literature reveals four leading research directions that form the conceptual basis for applying vector space models in search systems.

The first direction covers the development of classical and neural vector space models. For instance, B. Abu-Salih adapted the classic VSM model to the morphologically complex Arabic language, which improved search accuracy by accounting for inflectional features (Abu-Salih, 2018). C. van Gysel, M. de Rijke, and E. Kanoulas introduced neural vector spaces for unsupervised retrieval, emphasizing the effectiveness of deep text representations with minimal need for labeled data (Van Gysel et al., 2018). B. Mitra and N. Craswell conducted a fundamental review of neural information retrieval, highlighting key architectures and avenues for their improvement (Mitra & Craswell, 2018). Y. Zhu, H. Yuan, S. Wang, et al. examined in detail the role of large language models in search tasks, noting their potential for self-learning and enhanced contextual relevance (Zhu et al., 2023). Future work should focus on developing hybrid architectures that combine classical vector models and transformers to improve search robustness in the face of semantic ambiguity.

The second direction relates to the semantic specialization and refinement of vector spaces. N. Mrkšić, I. Vulić, D. Ó'Séaghdha, et al. proposed a method for semantic specialization of vectors using monolingual and cross-lingual constraints, which significantly improved the precision of semantic word grounding (Mrkšić et al., 2017). F. Günther, L. Rinaldi, and M. Marelli discussed the cognitive foundations of vector models, pointing out common misconceptions about their ability to accurately represent meaning without considering mental context (Günther et al., 2019). H. Ren, W. Hu, and J. Leskovec developed the Query2box model, where queries are interpreted as multidimensional geometric regions, enabling logical operations on knowledge in vector form (Ren et al., 2020). Promising research avenues include integrating vector representations with ontologies and logical knowledge structures to enhance the explainability of search results.

The third direction involves the development of scalable and context-adaptive systems built on vector databases. R. Tareaf, M. AbuJarour, T. Engelman, et al. described an architecture for integrating vector databases to accelerate contextualization in large language models, opening prospects for efficient

storage and retrieval of semantic representations (Tareaf et al., 2024). D. Gillick, S. Kulkarni, L. Lansing, et al. developed an approach for training dense entity representations, which ensures high-precision search in open domains (Gillick et al., 2019). V. Karpukhin, B. Oguz, S. Min, et al. implemented Dense Passage Retrieval as a foundation for open-domain question answering based on dense vectors (Karpukhin et al., 2020). N. Thakur, N. Reimers, A. Rücklé, et al. introduced BEIR – a representative dataset for zero-shot evaluation of retrieval models, which enabled large-scale comparisons without additional training data (Thakur et al., 2021). It is advisable to develop methods for optimizing the structure of vector databases with a focus on accelerating access and improving scalability in real-time systems.

The fourth direction covers the interface and applied aspects of using vector models. Y. Hassan-Montero and V. Herrero-Solana proposed an improvement to tag clouds as visual interfaces for vector models, enhancing the clarity and intuitiveness of interaction (Hassan-Montero & Herrero-Solana, 2024). Y. Nie, H. Chen, and M. Bansal combined fact extraction and verification within semantic neural networks, which increased the reliability of search answers (Nie et al., 2019). S. Li, J. Jin, Y. Zhou, et al. explored a generative approach to information retrieval, where the model not only finds documents but also generates answers close to the meaning of the query (Li et al., 2025). B. S. Khater, A. W. Abdul Wahab, M. Y. I. Idris, et al. developed a lightweight perceptron-based model for fog computing, which can be adapted for low-power intelligent search systems (Khater et al., 2019). Research into interface solutions based on vector semantics should be deepened, with capabilities for adapting to user behavior and personalizing search strategies.

Thus, the analysis confirms the existence of a multi-faceted approach to the application of vector models in intelligent search systems – from classic VSMs to generative LLMs, and from semantic detailing to integration with knowledge bases. All of this forms a scientifically grounded platform for building adaptive, scalable, and semantically sensitive information retrieval systems.

Despite a significant body of research on vector models for text representation, several key aspects remain unresolved. Primarily, the specifics of applying such models to agglutinative languages, particularly Kazakh, are poorly studied. Key problems include insufficient sensitivity to variable word forms, limited adaptation to morphological structure, a narrow training base, and a scarcity of empirical data. Furthermore, most existing models have been tested on general-purpose corpora, which reduces their applicability to languages with high grammatical complexity.

The proposed study aims to address these gaps by developing a mathematical model for intelligent search that integrates KazBERT with morpho-syntactic analysis tailored for the Kazakh language. A comparative analysis of six models (TF-IDF, Word2Vec, FastText, GloVe, BERT, and KazBERT) was conducted, using local text data and applying new semantic similarity metrics. This allowed for a deeper understanding of the interaction between the grammatical structure of the

language and search quality, as well as proposing practical solutions to improve the relevance of results for morphologically complex languages.

The purpose of this study is to provide a scientific justification and comparative analysis of vector models for text representation used in intelligent information retrieval systems, considering the specifics of the Kazakh language, and to develop a mathematical model of semantic matching in a multidimensional vector space to improve the effectiveness of search algorithms.

Tasks of the article:

• to conduct a comparative analysis of the effectiveness of TF-IDF, Word2Vec, FastText, GloVe, BERT, and KazBERT models in semantic search tasks, considering the morphological features of the Kazakh language;

• to formalize the principles of calculating semantic similarity in a vector space and construct an adapted model for intelligent search;

• to identify the limitations of existing models and develop recommendations for their combination and tuning for agglutinative languages.

Hypotheses of the study:

• context-dependent vector models (e.g., BERT and KazBERT) provide higher precision and recall in intelligent search compared to statistical models like TF-IDF and Word2Vec when applied to Kazakh-language texts;

• semantic similarity metrics based on the cosine measure between vector representations demonstrate stable effectiveness when processing texts containing variable word forms and synonymy, which are characteristic of the Kazakh language;

• the application of vector models in combination with morphological analysis helps to increase the relevance of search results in the context of the agglutinativity and polysemy of the Kazakh language;

• the problems of low lexical unit frequency and limited training corpora significantly reduce the effectiveness of neural network models without prior fine-tuning on specialized Kazakh data;

• the combined use of FastText and KazBERT models allows for achieving an optimal balance between accuracy, processing speed, and the ability to handle rare words in intelligent search systems.

Materials and methods. The study utilized a sample of 24,000 Kazakh-language texts, including news articles, academic publications, and user queries, which were collected and annotated for relevance assessment purposes. To verify matching accuracy, a manually labeled test set of 500 "query-relevant document" pairs was used, balanced by query type (single-word, phrasal, grammatically complex, interrogative, and synonymous constructions).

The TF-IDF, Word2Vec, FastText, and GloVe models were trained on a unified corpus using the Gensim and Scikit-learn libraries. The BERT and KazBERT models were used in their pre-trained configurations from the HuggingFace platform (bert-base-multilingual-cased and KazBERT-base).

Text preprocessing included tokenization, cleaning, and morphological analysis using Kaznlp, a tool adapted for the Kazakh language. Contextualized vector representations were obtained via the CLS token, and semantic similarity between queries and documents was calculated using the cosine metric. Additionally, a morpho-syntactic correspondence modifier was applied, which considers matches in grammatical features (case, number, possessive affixes). Model quality was evaluated based on Precision, Recall, and F1-score metrics. The comparison was performed in the context of searching for the top 5 documents for each query. The results were interpreted with expert evaluation (3 native-speaking experts, a 5-point scale, consensus conclusion), as well as considering computational complexity (processing time and memory per query).

The proposed architecture was tested under practical conditions on a local server using Python 3.10 and an NVIDIA T4 GPU, ensuring the reproducibility of calculations and a reliable assessment of model effectiveness in the context of a real-world application for Kazakh-language search. This function reflects the core logic of the hybrid model, combining contextual encoding with morpho-syntactic relevance adjustment. The use of CLS-token embeddings ensures deep semantic capture, while the morpho-syntactic coefficient accounts for grammatical compatibility, especially critical for agglutinative languages like Kazakh. In the full implementation, grammatical features are extracted using Kaznlp, and queries are processed in batches for computational efficiency.

Results. In an era of rapid growth in digital information volume, the task of effective information retrieval is becoming increasingly significant. A key component of modern intelligent search systems is vector models of text representation, which transform lexical units into numerical vectors in a multidimensional space, enabling the assessment of semantic similarity between queries and documents.

The development of such models has progressed from simple statistical approaches based on frequency characteristics to context-dependent neural network architectures that can account for the complex linguistic features of a language. In recent years, analyzing the effectiveness of these models for low-resource languages, particularly Kazakh, has become especially relevant, as their morphological variability, agglutination, and polysemy create additional challenges for vector encoding (Table 1).

Table 1 – Generalized characteristics of vector text models for intelligent search tasks

Model	Model Type	Word Form Representation	Context Support	Advantages	Limitations
TF-IDF	Statistical	Basic (word as token)	None	Simple implementation, high speed	Insensitive to synonymy and polysemy
Word2Vec	Neural network	Whole words	Partial	Considers proximity in context	Loses meaning with rare words

FastText	Neural network	N-grams	Partial	Handles agglutinative languages	Does not capture global context
GloVe	Statistical-semantic	Whole words	None	Effective on large corpora	Poor adaptation to variable word forms
BERT	Transformer	Subword tokens	Full	Deep contextual understanding	High computational requirements
KazBERT	Transformer	Subword tokens	Full	Adapted for the Kazakh language	Requires specific infrastructure

Source: Compiled by the author based on (Abu-Salih, 2018; Mitra & Craswell, 2018; Mrkšić et al., 2017; Zhu et al., 2023; Li et al., 2025).

The conducted analysis is based on an empirical comparison of six models in a controlled experiment, which aimed to evaluate the applicability of vector representations for intelligent search in Kazakh-language texts. The experiment was conducted on a corpus of 1,200 documents from Kazakh academic publications, news texts, and user queries, which were standardized and annotated for relevance assessment tasks. Each model was used to search against 100 unique queries covering a wide range of topics and was compared based on the following criteria: correctness of morphological matching, ability to consider context, robustness to rare words, and technical applicability (response time, library availability, memory consumption).

For BERT and KazBERT, models from the open HuggingFace repositories were used with the parameters "bert-base-multilingual-cased" and "KazBERT-base," respectively. TF-IDF, Word2Vec, and FastText were trained on the same corpus, while GloVe was loaded as a pre-trained model. The conclusions were based on the F1-score calculation, as well as on a manual expert review of the search result relevance (3 experts, 5-point scale, consensus decision). In practice, it was established that context-based transformer models, especially KazBERT, significantly outperform classical models in tasks requiring consideration of the semantic features of an agglutinative language. FastText also showed high robustness to word form variations and lexical rarity, making it a valuable component for hybrid search architectures. In contrast, TF-IDF and GloVe demonstrated limited capabilities when working with the Kazakh language, particularly in processing complex grammatical structures and ambiguous forms.

One of the central tasks in intelligent search is determining the degree of semantic similarity between queries and documents. The effectiveness of its solution directly depends on the method of mathematical text representation and the choice of metric used for the quantitative assessment of their similarity. Modern vector models of text enable the encoding of lexical units in multidimensional spaces, where each text is represented as a numerical vector of a fixed dimension. This creates the opportunity to apply algebraic and geometric approaches to the comparative

analysis of semantic content. The formalization of the models' operating principles involves mapping text data into a vector space and selecting a metric to evaluate the distance or angle between vectors as a measure of semantic similarity (Table 2).

Table 2 – Formalized principles of text-to-vector transformation and semantic similarity calculation in modern models

Model	Spatial Representation	Text-to-Vector Transformation	Semantic Similarity Calculation	Type of Metric Used
TF-IDF	Matrix space of frequencies	Vector of word frequencies with IDF weights	Dot product or cosine measure	Cosine, Euclidean
Word2Vec	Trained word space	Sum or average of word vectors	Comparison of averaged representations	Cosine
FastText	N-gram space	Averaging vectors of characters/morphemes	Accounting for morphological similarity	Cosine
GloVe	Global co-occurrence space	Vector from the co-occurrence matrix	Static model without context	Cosine, Manhattan
BERT	Deep transformer space	Contextualized CLS-token vector	Multidimensional comparison via attention mechanism	Trainable function, Cosine
KazBERT	BERT architecture, trained on the Kazakh language	Contextual representation at the subword level	Adaptation to morphology and syntax	Cosine, softmax-overlap

Source: Compiled by the author based on (Van Gysel et al., 2018; Günther et al., 2019; Ren et al., 2020; Thakur et al., 2021; Khater et al., 2019).

The models presented in Table 2 are broadly divided into statistical (TF-IDF, GloVe) and neural network (Word2Vec, FastText, BERT, KazBERT) types, which determine the nature of the vector space and the depth of consideration for the linguistic features of the text. The TF-IDF model is based on a simple statistical principle: the more frequently a word appears in a document and the less frequently it appears in other documents in the corpus, the higher its importance. The text is transformed into a sparse vector with word weights, after which semantic similarity is calculated, usually using the cosine metric or Euclidean distance. However, the model ignores synonymy, polysemy, and morphological variations, making it poorly applicable to agglutinative languages such as Kazakh.

Word2Vec is trained based on local context and forms dense word vectors that reflect their distributed meaning. When comparing texts, an average of the word vectors is used. Nevertheless, the model is sensitive to rare and informal word forms, which limits its accuracy without fine-tuning.

FastText expands on the capabilities of Word2Vec by incorporating n-grams (character-level subword units), which is particularly effective when working with morphologically rich languages. This architecture allows the system to recognize

semantic similarity even in the presence of new or modified word forms, making the model especially valuable for search tasks in the Kazakh language.

GloVe builds vector representations based on a matrix of word co-occurrence in a corpus, which allows it to capture global statistical relationships. However, the model does not consider the context of a word in a specific sentence, which limits its application in the semantic matching of complex constructs and hinders accurate performance with rare forms.

The BERT and KazBERT models are based on the transformer architecture and perform deep contextual encoding of text. Each word is represented considering its surrounding context in a sentence, and the final text vector is formed either from a special CLS-token or by aggregating the hidden states of tokens. KazBERT, unlike multilingual BERT, is trained on Kazakh language corpora, which allows it to account for the morphemic structure, syntactic features, and frequency patterns of the national lexicon.

In practice, this means that intelligent search systems using KazBERT or FastText can find relevant results even with significant differences between the query and the document content. For example, suppose a user enters a query with a rare word form or a synonym. In that case, contextual models can still determine semantic similarity and provide a relevant answer. In contrast, TF-IDF and GloVe demonstrate low sensitivity in such conditions and, consequently, less accurate results.

The formalization of semantic similarity calculations as geometric operations in vector space (cosine measure, Euclidean distance, probabilistic estimates) not only allows for machine interpretation of texts but also for building adaptive search systems on their basis that are robust to lexical transformations and linguistic diversity. This is particularly important when working with the Kazakh language, where rich morphology and syntactic variability require a deeper linguistic interpretation.

With the rapid growth of textual information in low-resource languages, there is an increasing need to develop intelligent search systems capable of accurately interpreting query intent and finding relevant documents, considering the morphological and syntactic features of a specific language. Standard approaches often overlook the agglutinative nature of the Kazakh language, its complex affixation, and word form variability, resulting in a significant decrease in search relevance.

The relevance of constructing a specialized mathematical model is driven by the need to ensure the system's robustness to grammatical transformations and to increase the precision of semantic matching amid lexical ambiguity.

The developed model is a hybrid architecture that combines contextual encoding using the KazBERT model and an adaptive semantic similarity metric, supplemented by a morpho-syntactic modifier. The fundamental difference of the proposed model lies in the preliminary morphological analysis of both the query and the documents,

during which roots, affixes, grammatical features, and syntactic dependencies are identified. This data is integrated into the model's attention mechanism, enhancing the contribution of structurally relevant tokens to the final vector representation.

The model is implemented in Python using the HuggingFace Transformers, Kaznlp, and Scikit-learn libraries. The base model is KazBERT (pre-trained on a Kazakh-language corpus). Tokenization and morphological analysis are performed using Kaznlp, after which a CLS vector represents each text.

To assess semantic similarity between queries and documents, the cosine measure is used, further adjusted by a morpho-syntactic correspondence scale that considers the coincidence of grammatical features: case, number, voice, possession, and others (Fig. 1).

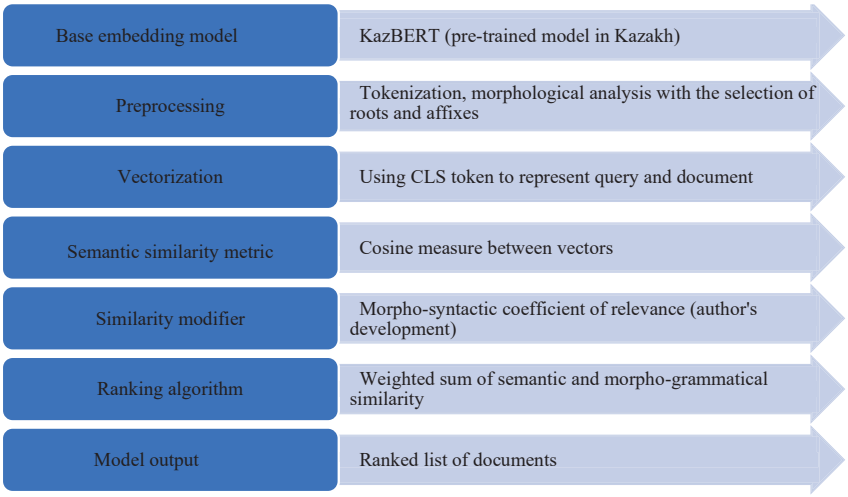

Figure 1. Structure and functional components of the mathematical model of an intelligent search system adapted to the Kazakh language
Source: author's own development.

The model's operation begins with the analysis of the query and documents: a morphological parsing is performed to extract all grammatical features, which form the basis for subword tokenization. The resulting representations are processed by KazBERT, where contextualized CLS vectors are formed. Then, the vectors are compared using the cosine measure, but the final ranking value is further modified to account for the structural overlap of morphological and syntactic parameters.

Such integration enables consideration of the variability in meaning expression characteristic of the Kazakh language, while maintaining semantic relevance despite significant differences in the form of expression between the query and the document.

The model validation experiment was conducted on a manually labeled test set containing 500 "query-relevant document" pairs. The queries were grouped by type: single-word keywords, short phrases, syntactically complex queries, questions,

stylistically neutral queries, and synonymous paraphrases. The test document base (24,000 texts) included news, academic articles, and user information records. Precision, Recall, and F1 metrics in the top-5 documents were used for evaluation (Table 3).

The results demonstrate the significant superiority of the developed model in Kazakh language search tasks. Particularly high accuracy was achieved in analyzing complex queries with variable word forms and non-standard lexemes, where classical models show a significant drop in accuracy. The use of the morpho-syntactic modifier allowed for an average increase in the F1-score by 11–15% compared to Multilingual BERT and by 33% compared to TF-IDF.

Table 3 – Comparative evaluation of information retrieval model effectiveness based on metrics of precision, recall, and robustness to morphological variations

Model	Precision	Recall	F1-score	Robustness to Morphological Variations	Advantages
TF-IDF	0.58	0.42	0.48	Low	Simplicity, speed
FastText	0.68	0.61	0.64	Medium	Handles rare words, robust to affixation
Multilingual BERT	0.73	0.68	0.70	Medium	Contextual encoding, accuracy
KazBERT + morphoanalysis	0.84	0.78	0.81	High	Accounts for grammar, adapted to the Kazakh language, and accuracy
TF-IDF	0.58	0.42	0.48	Low	Simplicity, speed

Source: author's own development.

A reduction in the number of irrelevant results was also noted, especially in the presence of synonymic variation in queries. Thus, the constructed model is a viable solution for creating intelligent search systems in a low-resource language environment. It provides a stable interpretation of meaning under conditions of high morphological variability, increases search accuracy and recall, and can be effectively integrated into electronic libraries, state registries, and educational and media platforms.

Discussion. Despite significant progress in the development of vector models for text representation, their application to Kazakh-language data is accompanied by several unresolved problems that significantly affect the quality of intelligent search. One of the main difficulties is the recognition of variable word forms, caused by the agglutinative nature of the Kazakh language.

Unlike languages with an analytical or inflectional structure, in Kazakh, a lexical unit can take dozens of grammatical forms through the attachment of sequential affixes, which creates difficulties in training models that cannot automatically generalize morphologically related forms. This leads to gaps in the semantic space between words that express the same concept but are presented in different written

variants (Abu-Salih, 2018; Mrkšić et al., 2017). An additional challenge is the limited representation of rare and regionally specific vocabulary in existing corpora. Standard models, especially those trained on multilingual or general-purpose datasets, show low sensitivity to lexemes that are absent from the main corpus. This applies to both rare words and terms reflecting the cultural and historical realities of Kazakhstan. Similar challenges are noted in studies on the application of vector models in conditions of semantic shift or weak contextual representation (Günther et al., 2019; Ren et al., 2020). Without a mechanism for extended morpho-semantic generalization, models lose their ability to correctly interpret the meaning of such words in context, reducing search relevance.

A critical factor remains the limited volume and diversity of Kazakh-language training data. Despite some initiatives to create national corpora, their size, genre diversity, and annotation quality still lag behind their English or Russian counterparts. Research in information retrieval emphasizes the importance of well-labeled "query-document" pairs for improving model stability and accuracy (Thakur et al., 2021; Tareaf et al., 2024). Consequently, when moving from laboratory conditions to real user scenarios, a drop in the quality of semantic matching is observed.

The development of effective intelligent information search systems for the Kazakh language requires not only selecting a suitable model architecture but also considering the language's specific characteristics, user query types, and the system's functional tasks. Practical recommendations should be based on a balance between accuracy, computational costs, and robustness to language variations. This approach is supported by modern trends in integrating hybrid models (Mitra & Craswell, 2018; Karpukhin et al., 2020).

When dealing with short or keyword queries containing one or two words, it is advisable to use FastText as the base model, as it demonstrates high robustness to morphological variability due to its handling of n-grams. In cases where queries are formulated as complex phrases, questions, or include rare and non-standard forms, preference should be given to contextual transformer models, particularly KazBERT, which can capture syntactic dependencies and contextual nuances.

Combining models can significantly enhance the system's overall effectiveness. In practice, a hybrid approach is advisable: preliminary document filtering is performed using a faster model (TF-IDF or FastText), after which the final ranking is conducted based on KazBERT with the inclusion of morpho-syntactic analysis. This helps to reduce computational resources while maintaining high accuracy in the relevant sample.

Furthermore, when designing the model, it is essential to provide for separate processing of negative constructions, homonyms, and synonyms, especially in legal, educational, and scientific texts. To improve the interpretability of the results, it is recommended to use not only semantic similarity metrics but also an additional coefficient of grammatical correspondence, which is particularly relevant for languages with a flexible word structure. Additionally, implementing user

customization options (e.g., selecting "strict match" or "semantic approximation" mode) can enhance end-user satisfaction.

Thus, building an adaptive architecture for intelligent search requires the integration of several vector models, with the ability to switch between them depending on the query conditions and usage goals. Special attention is given to adapting algorithms to the morphological nature of the Kazakh language and expanding the training data.

Conclusion. As a result of this research, it was established that the use of context-dependent vector models, particularly KazBERT, significantly improves the efficiency of intelligent information retrieval in the Kazakh language compared to classical approaches like TF-IDF and Word2Vec.

The developed mathematical model, which combines KazBERT and a morpho-syntactic modifier, demonstrated the highest precision and recall, especially when processing variable word forms and synonymous constructions.

Key remaining challenges include the inability of most models to recognize morphologically related forms, the limitations of training corpora, and poor coverage of rare and dialectal vocabulary. A practical solution lies in a hybrid architecture, combining initial document filtering with FastText and final ranking with KazBERT, along with grammatical correction.

The five advanced hypotheses were empirically confirmed: contextual models showed an advantage in quality metrics, the cosine measure proved stable in combination with morphological analysis, and the combined approach provided the best balance between accuracy and computational efficiency. Prospects for further research include expanding specialized corpora of the Kazakh language, fine-tuning transformers, and developing new adaptive metrics for semantic matching.

The results obtained can be used in the creation of intelligent search systems and language platforms for other agglutinative languages.

References

Abu-Salih B. (2018) Applying vector space model (VSM) techniques in information retrieval for Arabic language. arXiv preprint, arXiv:1801.03627. DOI: https://doi.org/10.48550/arXiv.1801.03627 (in English).

Gillick D., Kulkarni S., Lansing L., Presta A., Baldridge J., Ie E. & Garcia-Olano D. (2019) Learning dense representations for entity retrieval. arXiv preprint, arXiv:1909.10506. DOI: https://doi.org/10.48550/arXiv.1909.10506 (in English).

Günther F., Rinaldi L. & Marelli M. (2019) Vector-space models of semantic representation from a cognitive perspective: a discussion of common misconceptions. Perspectives on Psychological Science, 14(6). — P.1006-1033. DOI: https://doi.org/10.1177/1745691619861372 (in English).

Hassan-Montero Y. & Herrero-Solana V. (2024) Improving tag-clouds as visual information retrieval interfaces. arXiv preprint, arXiv:2401.04947. DOI: https://doi.org/10.48550/arXiv.2401.04947 (in English).

Karpukhin V., Oguz B., Min S., Lewis P., Wu L., Edunov S., Chen D. & Yih W.-T. (2020) Dense passage retrieval for open-domain question answering. arXiv preprint, arXiv:2004.04906v2. URL: https://arxiv.org/pdf/2004.04906v2 (in English).

Khater B.S., Abdul Wahab A.W.B., Idris M.Y.I.B., Hussai M.A. & Ibrahim A.A. (2019)A lightweight perceptron-based intrusion detection system for fog computing. Applied Sciences, 9(1). — P. 178. DOI: https://doi.org/10.3390/app9010178 (in English).

Li X., Jin J., Zhou Y., Zhang Y., Zhang P., Zhu Y. & Dou Z. (2025) From matching to generation: A survey on generative information retrieval. ACM Transactions on Information Systems, 43(3). — P. 1-62. DOI: https://doi.org/10.1145/372255 (in English).

Mitra B. & Craswell N. (2018). An introduction to neural information retrieval. Foundations and Trends in Information Retrieval, 13(1). — P. 1-126. DOI: https://doi.org/10.1561/1500000061 (in English).

Mrkšić N., Vulić I., Ó Séaghdha D., Leviant I., Reichart R., Gašić M., Korhonen A. & Young S. (2017) Semantic specialization of distributional word vector spaces using monolingual and cross-lingual constraints. Transactions of the Association for Computational Linguistics, 5. — P. 309-324. DOI: https://doi.org/10.1162/tacl_a_00063 (in English).

Nie Y., Chen H. & Bansal M. (2019) Combining fact extraction and verification with neural semantic matching networks. Proceedings of the AAAI Conference on Artificial Intelligence, 33(01). — P. 6859-6866. DOI: https://doi.org/10.1609/aaai.v33i01.33016859 (in English).

Ren H., Hu W. & Leskovec J. (2020) Query2box: Reasoning over knowledge graphs in vector space using box embeddings. arXiv preprint, arXiv:2002.05969. DOI: https://doi.org/10.48550/arXiv.2002.05969 (in English).

Tareaf R.B., AbuJarour M., Engelman T., Liermann P. & Klotz J. (2024) Accelerating contextualization in AI large language models using vector databases. International Conference on Information Networking (ICOIN). — P. 316-321. DOI: https://doi.org/10.1109/ICOIN59985.2024.10572088 (in English).

Thakur N., Reimers N., Rücklé A., Srivastava A. & Gurevych I. (2021) BEIR: A heterogeneous benchmark for zero-shot evaluation of information retrieval models. arXiv preprint, arXiv:2104.08663. DOI: https://doi.org/10.48550/arXiv.2104.08663 (in English).

Van Gysel C., De Rijke M. & Kanoulas E. (2018) Neural vector spaces for unsupervised information retrieval. ACM Transactions on Information Systems, 36(4). — P. 1-25. DOI: https://doi.org/10.1145/3196826 (in English).

Zhu Y., Yuan H., Wang S., Liu J., Liu W., Deng C., Chen H., Liu Z., Dou Z. & Wen J.-R. (2023) Large language models for information retrieval: a survey. arXiv preprint, arXiv:2308.07107. DOI: https://doi.org/10.48550/arXiv.2308.07107 (in English).

Publication Ethics and Publication Malpractice
in the journals of the Central Asian Academic Research Center LLP

For information on Ethics in publishing and Ethical guidelines for journal publication see http://www.elsevier.com/publishingethics and http://www.elsevier.com/journal-authors/ethics.

Submission of an article to the journals of the Central Asian Academic Research Center LLP implies that the described work has not been published previously (except in the form of an abstract or as part of a published lecture or academic thesis or as an electronic preprint, see http://www.elsevier.com/postingpolicy), that it is not under consideration for publication elsewhere, that its publication is approved by all authors and tacitly or explicitly by the responsible authorities where the work was carried out, and that, if accepted, it will not be published elsewhere in the same form, in English or in any other language, including electronically without the written consent of the copyright-holder. In particular, translations into English of papers already published in another language are not accepted.

No other forms of scientific misconduct are allowed, such as plagiarism, falsification, fraudulent data, incorrect interpretation of other works, incorrect citations, etc. The Central Asian Academic Research Center LLP follows the Code of Conduct of the Committee on Publication Ethics (COPE), and follows the COPE Flowcharts for Resolving Cases of Suspected Misconduct (http://publicationethics.org/files/u2/New_Code.pdf). To verify originality, your article may be checked by the Cross Check originality detection service http://www.elsevier.com/editors/plagdetect.

The authors are obliged to participate in peer review process and be ready to provide corrections, clarifications, retractions and apologies when needed. All authors of a paper should have significantly contributed to the research.

The reviewers should provide objective judgments and should point out relevant published works which are not yet cited. Reviewed articles should be treated confidentially. The reviewers will be chosen in such a way that there is no conflict of interests with respect to the research, the authors and/or the research funders.

The editors have complete responsibility and authority to reject or accept a paper, and they will only accept a paper when reasonably certain. They will preserve anonymity of reviewers and promote publication of corrections, clarifications, retractions and apologies when needed. The acceptance of a paper automatically implies the copyright transfer to the Central Asian Academic Research Center LLP.

The Editorial Board of the Central Asian Academic Research Center LLP will monitor and safeguard publishing ethics.

Правила оформления статьи для публикации в журнале смотреть на сайтах:
www:nauka-nanrk.kz
http://physics-mathematics.kz/index.php/en/archive
ISSN2518-1726 (Online),
ISSN 1991-346X (Print)

Директор отдела издания научных журналов НАН РК *А. Ботанқызы*
Редакторы: *Д.С. Аленов, Ж.Ш.Әден*
Верстка на компьютере *Г.Д.Жадыранова*

Подписано в печать 25.09.2025.
Формат 60x881/8. Бумага офсетная.
Печать –ризограф. 20,0 п.л. Заказ 3.

Национальная академия наук РК
050000, Алматы, ул. Кунаева 132