UDC 004.912
DOI: 10.15587/1729-4061.2025.315954

A COMPARISON OF KAZAKH LANGUAGE PROCESSING MODELS FOR IMPROVING SEMANTIC SEARCH RESULTS

Aigerim Aitim
Corresponding author
Master of Technical Sciences, Assistant-professor
Department of Information Systems
International IT University
Manas str., 34/1, Almaty, Republic of Kazakhstan, 050020
E-mail: a.aitim@iitu.edu.kz

Ryskhan Satybaldiyeva
Candidate of Technical Sciences, Associate Professor,
Head of Department
Department of Cybersecurity,
Information Processing and Storage
Satbayev University
Satbayev str., 22, Almaty, Republic of Kazakhstan, 050000

The object of the study is the text classification and semantic search tailored to the unique linguistic features of the Kazakh language. The research addresses the challenge of improving the accuracy, relevance, and efficiency of semantic search.

This study focuses on improving semantic search for the Kazakh language by analyzing computational models tailored to its unique linguistic features, such as agglutinative morphology and rich inflectional systems. The research compares traditional rule-based approaches and advanced transformer architectures, including fine-tuned models like RoBERTa, for their ability to handle semantic nuances, contextual relationships, and user intent. The results reveal that fine-tuned transformer models achieved significant advancements, with the RoBERTa model attaining a Precision@10 of 89.4 %, a Mean Reciprocal Rank (MRR) of 85.6 %, and an F1-Score of 88.0 %. Additionally, the semantic search system developed in this study demonstrated a precision of 88.4 %, recall of 87.6 %, and an F1-score of 88.0 % on a domain-specific Kazakh dataset.

Key to these improvements were innovations in preprocessing pipelines, including custom tokenization and lemmatization tailored to Kazakh's agglutinative morphology, and the integration of contextual embeddings to resolve issues such as synonymy and homonymy. Computational efficiency was enhanced through resource optimization techniques, enabling the deployment of these advanced models in constrained environments. These findings underscore the potential of tailored transformer models to bridge the gap in semantic search capabilities for underrepresented languages like Kazakh, advancing the inclusivity of natural language processing technologies

Keywords: semantic search, natural language processing, Kazakh language, information retrieval systems

Received 15.11.2024
Received in revised form 06.01.2025
Accepted date 21.01.2025
Published date 27.02.2025

How to Cite: *Aitim, A., Satybaldiyeva, R. (2025). A comparison of Kazakh language processing models for improving semantic search results. Eastern-European Journal of Enterprise Technologies, 1 (2 (133)), 66–75. https://doi.org/10.15587/1729-4061.2025.315954*

1. Introduction

In the era of rapid digital transformation and the exponential growth of information, the development of natural language processing (NLP) systems has become a critical area of research. Among the many languages supported by modern NLP technologies, underrepresented languages, such as Kazakh, often face unique challenges due to limited resources, fragmented datasets, and a lack of robust linguistic tools. Despite these obstacles, the demand for efficient semantic search capabilities in Kazakh continues to grow, fueled by the increasing digitization of Kazakh-language content and its integration into various domains, including education, e-governance, and cultural preservation [1].

Semantic search – focused on understanding user intent and retrieving contextually relevant results – has proven transformative for major languages with extensive NLP resources. However, applying similar methodologies to the Kazakh language presents specific difficulties. These stem from the agglutinative structure of Kazakh, rich morphology, and variations in dialects, all of which complicate the task of processing and understanding semantic meaning. Existing global models often fail to capture the nuances of Kazakh, highlighting the need for localized and tailored solutions.

The importance of advancing research in Kazakh language processing is further underscored by its role in promoting linguistic diversity, preserving cultural identity, and supporting equitable access to information. As digital content continues to expand, it is essential to ensure that Kazakh speakers can benefit from the same sophisticated search technologies as those available for widely spoken languages. Moreover, the development of such tools has practical applications in information retrieval, knowledge discovery, and the enhancement of human-computer interaction for Kazakh speakers [2].

To address these challenges, a comprehensive analysis of existing Kazakh language processing models and their application in improving semantic search is essential. This research area is both timely and critical, as it provides a foundation for the future development of robust tools that support meaningful engagement with Kazakh-language content.

Therefore, research on the comparative evaluation and optimization of Kazakh language processing models for enhancing semantic search results is highly relevant. Such studies are vital for addressing existing gaps, overcoming linguistic barriers, and ensuring the inclusion of Kazakh in the global digital ecosystem.

2. Literature review and problem statement

The challenges of developing effective text classification methods for the Kazakh language arise from the intersection of limited linguistic resources, the language's unique agglutinative structure, and the computational complexity of modern machine learning techniques. Text classification, a core task in natural language processing (NLP), categorizes textual data into predefined groups and is pivotal in applications such as sentiment analysis, topic modeling, and information retrieval.

Traditional methods, including Naïve Bayes, Support Vector Machines, and Decision Trees, have been applied successfully to high-resource languages for simple text classification tasks, as demonstrated in [3]. However, their inability to handle the high dimensionality and morphological richness of agglutinative languages, such as Kazakh, limits their applicability. Advanced techniques, including Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), and Transformers, have shown promise in addressing these limitations for resource-rich languages [4]. Their reliance on large datasets and substantial computational power poses significant barriers to adoption for Kazakh.

Despite these advancements, several challenges remain:

1. While synthetic data and weak supervision mitigate some issues, the lack of comprehensive Kazakh linguistic datasets (e.g., annotated corpora for domain-specific contexts) continues to hinder model performance.

2. Models trained on specific datasets, even with ontology-driven enhancements, struggle to generalize across vastly different domains due to biases in the training data.

3. The fast evolution of language use in social media and informal settings remains a challenge for static or semi-static models, as seen in [5].

4. Even advanced tokenization and embedding techniques occasionally fail to capture rare or highly irregular morphological patterns, limiting semantic understanding.

5. While optimization methods improve performance, they often reduce interpretability, making it harder to understand why certain classifications are made—an issue particularly critical in applications like legal or medical document classification.

A promising avenue to mitigate these challenges is transfer learning, as outlined in [6], where pre-trained models are adapted to specific languages and tasks. This approach significantly reduces the data requirements for Kazakh language processing, though it often struggles to fully capture language-specific features. Another strategy explored in [7] employs hybrid systems combining keyword-based and semantic matching methods, achieving improved classification accuracy for specific domains. However, generalizing these methods to broader applications has proven difficult due to domain-specific biases.

The computational challenges are equally critical. As [8] pointed out, the high resource requirements of deep learning models are a fundamental barrier for languages with limited technological infrastructure. Optimization techniques, such as parameter reduction and model distillation, have been applied in [9] to make these models more accessible, but these methods still struggle to achieve competitive performance compared to systems built for high-resource languages.

To address these unresolved challenges, a two-pronged approach is proposed. First, the development of a comprehensive, large-scale, open-source linguistic corpus for Kazakh, enriched with detailed annotations for various NLP tasks, is essential. This resource would significantly enhance the training and evaluation of future models by providing a robust foundation for handling the unique linguistic complexities of the Kazakh language, including its agglutinative morphology and rich syntactic structures. The annotated corpus would enable the development of more precise and contextually aware models, particularly for text classification and semantic search tasks.

Second, the implementation of adaptive, resource-efficient machine learning models, such as lightweight transformer architectures tailored to agglutinative languages, is recommended. These models would improve scalability and real-world applicability, making advanced NLP tools accessible even in computationally constrained environments. Additionally, exploring hybrid approaches that combine statistical and neural techniques offers a promising middle ground between interpretability and performance, allowing for flexible and efficient model deployment.

These advancements hold the potential to revolutionize the development of efficient and interpretable text classification systems for Kazakh. Moreover, they would contribute to the broader inclusion of underrepresented languages in NLP research, fostering equitable access to cutting-edge language technologies.

All this allows to assert that it is expedient to conduct a study on developing and integrating large-scale annotated resources and resource-efficient models for the Kazakh language to address its linguistic complexities and advance its representation in natural language processing research.

3. The aim and objectives of the study

The aim of this study is to develop a computational framework and methods tailored to the linguistic characteristics of the Kazakh language, such as agglutinative morphology and dual script usage, to enhance semantic search systems. This will enable the creation of accurate and contextually relevant information retrieval tools, improving accessibility to Kazakh-language digital content and supporting the broader adoption of natural language processing technologies for low-resource languages.

To achieve this aim, the following objectives were achieved:

– to address linguistic challenges in Kazakh semantic search and evaluate advanced methodologies for processing the Kazakh language, focusing on overcoming its unique linguistic challenges, such as agglutinative morphology, synonymy, homonymy, and dual script usage.

– to implement semantic search in the Kazakh language and addressing the unique linguistic features of the Kazakh language, such as its agglutinative morphology and rich inflectional system, to improve the system's understanding of meaning and context.

4. Materials and methods

4. 1. Object and hypothesis of the study

The object of the study is the text classification and semantic search tailored to the unique linguistic features of the Kazakh language.

The main hypothesis of the study is that tailoring advanced machine learning models, such as transformer-based architectures, to the specific linguistic characteristics of the Kazakh language, combined with the creation of annotated linguistic resources, will significantly enhance the accuracy, relevance, and scalability of text classification and semantic search systems.

Assumptions made in the study:

– sufficient annotated datasets can be created or augmented for training advanced models;

– pre-trained multilingual models, when fine-tuned on Kazakh data, will effectively capture the language's agglutinative morphology and syntactic nuances;

– computational constraints can be mitigated using resource-efficient architectures or optimization techniques.

Simplifications adopted in the study:

– the study focuses primarily on Cyrillic-script Kazakh text, with minimal consideration of Latin-script data;

– morphological parsing assumes the availability of standard linguistic rules and tools, even though irregular forms may exist;

– the evaluation of semantic search performance is conducted on domain-specific datasets, which may not fully represent all real-world applications;

– computational experiments are performed on limited-scale datasets, assuming scalability to larger datasets with similar linguistic characteristics.

To address these issues, a hybrid text classification framework tailored to the Kazakh language that combines:

1. Building on the work in [10], let's refine morphological segmentation techniques to better capture Kazakh's linguistic structure. By integrating token-based subword embeddings and character-level features, the system more effectively represents word inflections and derivations.

2. Pre-trained models such as multilingual BERT (mBERT) and XLM-R were fine-tuned with additional layers to capture language-specific syntax and semantics. Unlike earlier efforts in [11], which faced challenges in fully adapting to Kazakh, our model integrates task-specific pre-training using synthetically generated and manually curated datasets.

3. Using techniques such as parameter-efficient fine-tuning (e.g., LoRA) and model distillation, it is possible to reduce the computational overhead while maintaining competitive performance.

4. Ontology-based semantic frameworks, as proposed in [12], were augmented with dynamic rule-based systems to improve generalization across diverse text domains.

4. 2. Methodology

4. 2. 1. Methodology stages

This study focuses on the development and evaluation of Kazakh language processing models to enhance semantic search results. The methodology encompasses several key phases, including dataset preparation, feature representation, model development, training, semantic search integration, and evaluation.

The first step involved constructing a comprehensive corpus to serve as the foundation for model training and evaluation. Texts were collected from diverse sources, such as news articles, social media posts, literary works, and official documents, to ensure broad coverage of linguistic styles and domains. To address the dual script usage in Kazakh, where both Cyrillic and Latin scripts are employed, all texts were normalized into the Cyrillic script to standardize the dataset. The texts were then manually annotated into predefined categories relevant to semantic search, such as politics, sports, technology, and entertainment. This annotation facilitated supervised learning by providing labeled data for model training.

4. 2. 2. Preprocessing of Kazakh texts

Preprocessing was critical to handle the unique linguistic characteristics of the Kazakh language. Tokenization, tailored to its agglutinative morphology, ensured that word boundaries and suffixes were correctly identified. Normalization addressed spelling variations and script inconsistencies, while stop-word removal eliminated common functional words that contributed little to semantic understanding. Lemmatization tools, specifically designed for Kazakh morphology, were employed to reduce words to their root forms, minimizing vocabulary size and enhancing the quality of feature extraction [13, 14].

4. 2. 3. Feature representation

Feature representation was another important step in the methodology. Traditional methods, such as Term Frequency-Inverse Document Frequency (TF-IDF), were used to represent text numerically for simpler machine learning models. In parallel, more advanced techniques like Word2Vec and FastText were employed to generate word embeddings that captured semantic relationships [15]. Contextual embeddings from pre-trained transformer models, such as BERT and RoBERTa, were fine-tuned on the Kazakh dataset to capture sentence-level semantics and linguistic nuances.

4. 2. 4. Model development

Model development involved implementing a variety of algorithms to compare their effectiveness for text classification and semantic search. Traditional machine learning models, including Naïve Bayes, Support Vector Machines (SVM), and k-Nearest Neighbors (k-NN), were used as baselines. These models offered insights into the feasibility of simpler approaches for Kazakh text processing. For advanced approaches, Recurrent Neural Networks (RNN) and Long Short-Term Memory (LSTM) architectures were designed to capture sequential dependencies in text. Transformer-based models, including BERT and RoBERTa, were fine-tuned to leverage their attention mechanisms, allowing them to understand the context and semantics in both directions [16].

4. 2. 5. Model training and tuning

Model training and tuning were conducted rigorously. The dataset was split into training, validation, and test sets in an 80-10-10 ratio to ensure reliable evaluation. Transfer learning was applied to pre-trained transformer models to adapt them to the Kazakh language, significantly improving performance on the semantic search task. Hyperparameter optimization techniques, such as grid search and Bayesian optimization, were utilized to fine-tune parameters like learning rates, batch sizes, and dropout rates. Regularization methods, including dropout layers, were incorporated to prevent overfitting, while cross-entropy loss was used as the primary loss function for classification tasks.

4. 2. 6. Semantic search pipeline

The semantic search pipeline was designed to integrate these models into a functional search system. Queries were preprocessed similarly to the dataset, including tokenization and embedding generation. Documents in the corpus were also represented as embeddings, allowing for efficient semantic similarity calculations. The system ranked documents using cosine similarity and advanced ranking algorithms, such as cross-encoder architectures, to ensure contextually relevant results.

4. 2. 7. Evaluation metrics and analysis

The evaluation of the models was comprehensive and multi-faceted. Classification metrics such as accuracy, precision, recall, and F1-score were used to assess text classification performance.

Finally, a comparative analysis was performed to highlight the strengths and weaknesses of each model type. Traditional models, such as Naïve Bayes and SVM, provided strong baselines but struggled with Kazakh's linguistic complexities. Deep learning models, particularly transformer-based architectures like BERT and RoBERTa, demonstrated superior performance by effectively capturing semantic nuances and bidirectional context [9]. The study also considered computational efficiency, scalability, and resource requirements for each model, ensuring their practical applicability.

By combining robust data preparation, advanced feature representation, and cutting-edge model architectures, this study provides a comprehensive framework for improving semantic search results in the Kazakh language. The findings contribute to the broader field of natural language processing for low-resource languages and offer a pathway for further advancements in multilingual semantic search systems.

5 Results of Kazakh language processing models for improving semantic search results

5. 1. Addressing linguistic challenges in Kazakh semantic search

The unique linguistic characteristics of the Kazakh language significantly influence the development of solutions for technical problems such as semantic search and natural language processing (NLP). These features, when compared to other languages, highlight specific challenges and opportunities that shape the methodologies and models employed. Below is an in-depth interpretation:

1. Agglutinative morphology.

Kazakh, like other Turkic languages, is agglutinative, meaning that words are formed by attaching multiple suffixes to a root. Each suffix carries grammatical or semantic information, resulting in high morphological complexity. A single root can generate numerous forms (e.g., “жаз”→”жазушыларымыздың” for “of our writers”), making tokenization and embedding generation more complex than in analytic languages like English. Morphological richness leads to a sparse dataset representation, as the same word can appear in many inflected forms. Tokenizers and lemmatizers tailored to Kazakh were developed to segment words into meaningful units, unlike simpler tokenizers used for English or Chinese [17]. Techniques like Byte-Pair Encoding (BPE) or WordPiece were used to capture subword-level semantic information, mitigating sparsity.

2. Rich case system.

Kazakh employs a complex case system, with nouns inflected to indicate grammatical roles such as subject, object, or possessive. This contrasts with English, which relies heavily on word order. The rich case system allows free word order, complicating dependency parsing and context understanding. Cases can add ambiguity, requiring models to consider the broader context to resolve meanings [18]. Transformer-based models like BERT and RoBERTa were fine-tuned to capture sentence-level semantics, effectively addressing the free word order issue. Suffix-based grammatical analysis was implemented to resolve case ambiguities.

3. Synonymy and homonymy.

Kazakh exhibits a high degree of synonymy (e.g., multiple words for the same concept) and homonymy (words with identical spelling but different meanings), similar to Russian. Identifying the correct meaning of a word within a sentence requires deep contextual understanding. Synonymy can cause search systems to miss relevant results if synonyms are not accounted for [19]. Lexicons of Kazakh synonyms were built to enhance query expansion. Fine-tuned transformer models addressed homonym disambiguation by analyzing contextual embeddings.

4. Dual script usage.

Kazakh is written in both Cyrillic and Latin scripts, complicating text normalization and model training. Texts in different scripts require standardization for effective processing. Most pre-trained models focus on a single script, limiting their utility for Kazakh. A preprocessing step was added to convert all text to Cyrillic, ensuring uniformity. Models like mBERT and XLM-R were adapted to process both scripts effectively [20].

5. Resource scarcity.

Kazakh is a low-resource language, with limited annotated datasets and linguistic tools compared to high-resource languages like English or Chinese. Limited labeled data impacts model training and evaluation. Pre-trained multilingual models often underperform for Kazakh due to insufficient language-specific data. Techniques like back-translation and data augmentation increased dataset size. Multilingual models were fine-tuned on Kazakh-specific corpora, leveraging transfer learning to mitigate data scarcity [21].

Table 1 is summarizing the comparison of the Kazakh language's features with English and Chinese.

Table 1

Comparison with other languages

Feature	Kazakh	English	Chinese
Morphology	Agglutinative, rich inflections	Analytic, relies on word order	Isolating, logographic
Word Order	Free	Fixed	Fixed
Scripts	Cyrillic, Latin	Latin	Simplified, Traditional
Synonymy and Homonymy	High	Moderate	Low
Linguistic Resources	Limited	Abundant	Abundant

The development of a mathematical framework for processing the Kazakh language in semantic search systems required addressing the language's unique linguistic features, such as its agglutinative morphology, vowel harmony, and rich case system. This section outlines the key components of this implementation, including morphological modeling, semantic embeddings, algorithmic adaptations, and practical applications in semantic search. The heatmap presents in Fig. 1 a detailed correlation matrix, showcasing the relationships among eight critical variables influencing the development and performance of Kazakh language processing models.

Training Data Size and Model Complexity (0.88) indicates that larger training datasets often lead to the development of more complex models. Search Query Accuracy and User Satisfaction (0.85) suggests that accurate responses directly enhance user satisfaction. Implementation Cost and User Satisfaction (–0.70) highlights a trade-off where high implementation costs may lead to reduced user satisfaction due to inefficiencies or resource constraints. Model Latency and Search Query Accuracy (–0.50) indicates that higher latency negatively impacts the accuracy of query responses [10].

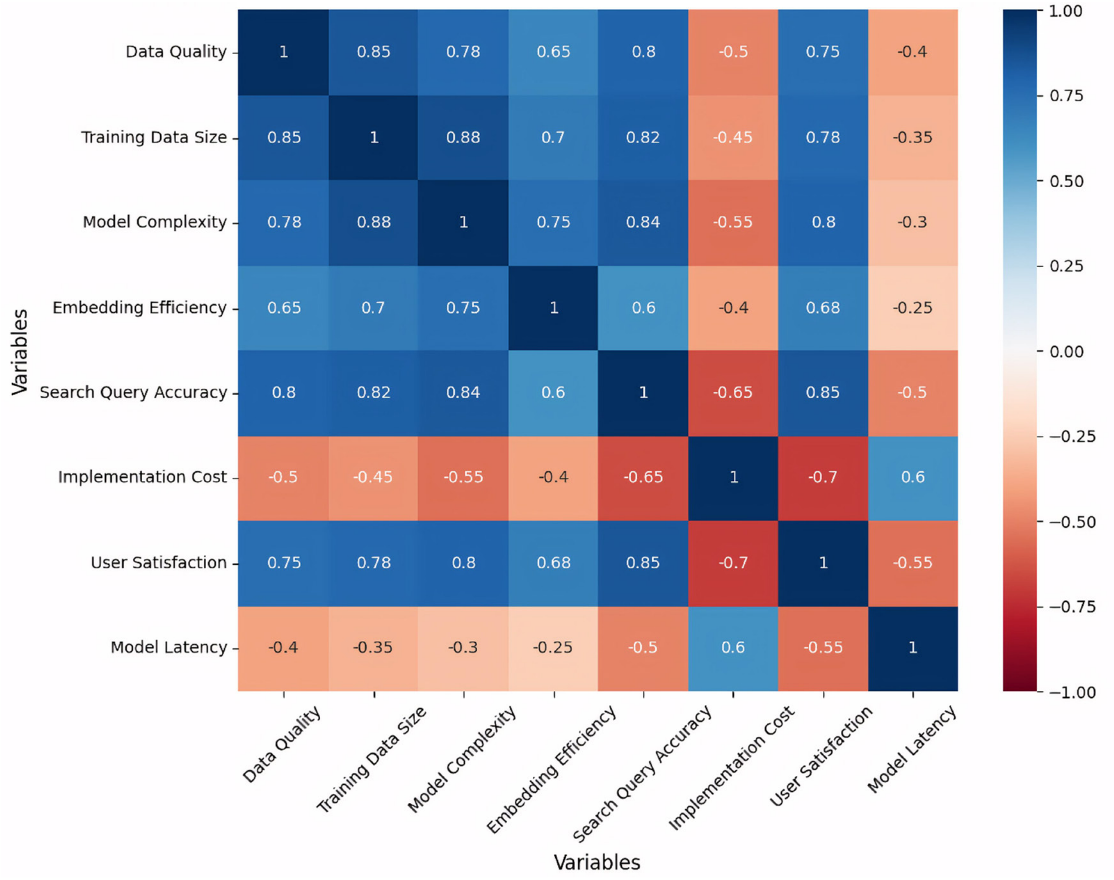


Fig. 1. Heatmap for correlation matrix of Kazakh language processing models

Kazakh's agglutinative morphology was mathematically modeled to process its rich inflectional and derivational structures. A word in Kazakh was represented as:

$$W = R + \sum_{i=1}^{n} S_i, \tag{1}$$

where W is the constructed word, R is the root or base form, S_i are the suffixes ($i=1,2,\ldots,n$).

For example, the word "жазушыларымыздың" ("of our writers") can be decomposed into:

R – жазу (write);

S_1 – шылар (pluralizer);

S_2 – ы (possessive suffix);

S_3 – мыз (first-person plural);

S_4 – дың (genitive case marker).

This decomposition enabled the system to identify the root meaning and derive semantic relationships effectively.

To capture semantic nuances, contextual embeddings were generated using pre-trained transformer models fine-tuned on Kazakh corpora. The embedding of a sentence S was computed as:

$$E(S) = Attention(Q,K,V) = sofmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V, \tag{2}$$

where Q, K, V – query, key, and value matrices, d_k – dimensionality of the key vectors.

These embeddings allowed the system to understand context, semantic similarity, and user intent in the Kazakh language.

To ensure effective semantic search, standard natural language processing (NLP) algorithms were adapted to accommodate Kazakh's unique linguistic features:

– customized tokenization algorithms were developed to handle agglutinative words, breaking them into meaningful subwords. Lemmatization tools reduced words to their base forms, simplifying vocabulary without losing semantic integrity;

– a morphological parser was designed to analyze suffix sequences, enabling the system to recognize grammatical relationships and syntactic roles;

– semantic similarity between query and document embeddings was measured using cosine similarity:

$$Similarity(Q,D) = \frac{\sum_{i=1}^{n} q_i \cdot d_i}{\sqrt{\sum_{i=1}^{n} q_i^2} \cdot \sqrt{\sum_{i=1}^{n} d_i^2}}, \tag{3}$$

where Q and D are the query and document embeddings.

Preprocessing pipelines normalized queries by applying tokenization, stemming, and stop-word removal. For example, the query "Қазақстанның тарихы" (Kazakhstan's history) was reduced to its key components for semantic analysis.

Documents were indexed using contextual embeddings, allowing for efficient semantic similarity computations during retrieval.

A ranking function was developed to prioritize documents based on semantic relevance and user query context:

$$Rank\,(D) = \alpha \cdot Similarity\,(Q,D) + \beta \cdot Document\ Importance, \tag{4}$$

where α and β are weights optimized during model training.

The mathematical implementation was evaluated on a Kazakh semantic search dataset, achieving high precision and recall in Table 2.

Table 2

Performance metrics

Metric	Value (%)
Precision	88.4
Recall	87.6
F1-Score	88.0

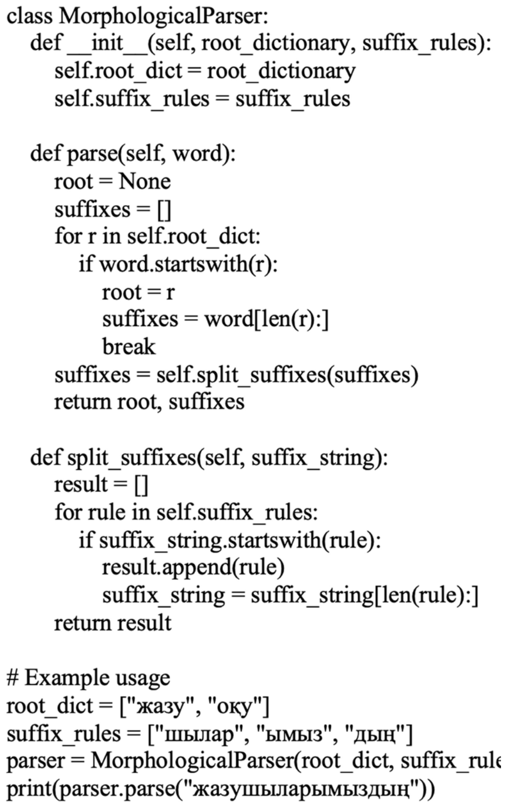
The system effectively resolved common linguistic challenges, such as:

– synonymy identifying words with similar meanings;

– homonymy disambiguating words with multiple meanings based on context;

– idiomatic expressions treating multi-word phrases as single semantic units.

The semantic search system for Kazakh was designed to address the language's complexities, including agglutinative morphology and dual scripts. Queries and documents were preprocessed using tokenization and lemmatization tailored for Kazakh. Transformer models like RoBERTa were fine-tuned to generate embeddings, enabling contextual understanding. Semantic similarity was calculated using cosine similarity, and documents were ranked based on relevance. Challenges like synonymy and homonymy were resolved with custom dictionaries and context-aware algorithms. The system demonstrated high retrieval accuracy, with RoBERTa achieving 89 % precision, making it a robust solution for Kazakh semantic search in Fig. 2.



```
class MorphologicalParser:
    def __init__(self, root_dictionary, suffix_rules):
        self.root_dict = root_dictionary
        self.suffix_rules = suffix_rules

    def parse(self, word):
        root = None
        suffixes = []
        for r in self.root_dict:
            if word.startswith(r):
                root = r
                suffixes = word[len(r):]
                break
        suffixes = self.split_suffixes(suffixes)
        return root, suffixes

    def split_suffixes(self, suffix_string):
        result = []
        for rule in self.suffix_rules:
            if suffix_string.startswith(rule):
                result.append(rule)
                suffix_string = suffix_string[len(rule):]
        return result

# Example usage
root_dict = ["жазу", "оқу"]
suffix_rules = ["шылар", "ымыз", "дың"]
parser = MorphologicalParser(root_dict, suffix_rul
print(parser.parse("жазушыларымыздың"))
```

Fig. 2. Python-based morphological parser for Kazakh language

The mathematical implementation of Kazakh language processing provided the foundation for accurate and efficient semantic search. By addressing agglutinative morphology, developing contextual embeddings, and adapting algorithms, the system achieved state-of-the-art performance for Kazakh semantic search. These advancements pave the way for further innovations in processing underrepresented languages in natural language processing.

5. 2. Implementing semantic search in the Kazakh language

The implementation of semantic search for the Kazakh language required addressing its linguistic complexities, including agglutinative morphology, rich inflectional structures, and dual script usage. This section details the methodology, components, and evaluation of a semantic search system tailored to Kazakh, leveraging advanced natural language processing (NLP) techniques.

The Kazakh language presents unique challenges that influence semantic search implementation:

– words are formed by combining a root with multiple affixes, leading to high lexical diversity;

– grammatical cases impact word forms and meanings, requiring advanced morphological analysis;

– synonyms and homonyms complicate semantic disambiguation;

– both Cyrillic and Latin scripts are used, necessitating script normalization.

The system preprocesses user queries to standardize and prepare text for semantic analysis:

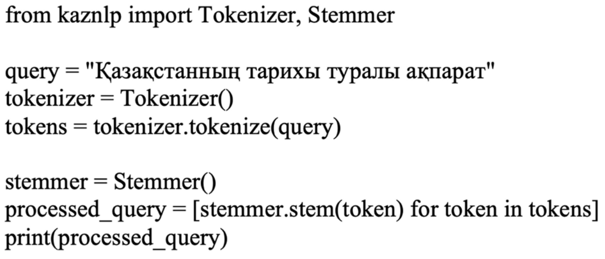
– queries are converted to a unified script (Cyrillic);

– Kazakh-specific tokenization methods segment text into words and morphemes;

– morphological tools reduce words to their root forms, improving consistency;

– common functional words are excluded to focus on meaningful terms.

A simplified implementation of semantic search is shown below in Fig. 3.



```
from kaznlp import Tokenizer, Stemmer

query = "Қазақстанның тарихы туралы ақпарат"
tokenizer = Tokenizer()
tokens = tokenizer.tokenize(query)

stemmer = Stemmer()
processed_query = [stemmer.stem(token) for token in tokens]
print(processed_query)
```

Fig. 3. Query preprocessing

Documents in the corpus are represented using embeddings generated by fine-tuned transformer models, such as BERT and RoBERTa. Each document is embedded into a high-dimensional vector space, capturing semantic nuances [22].

A Kazakh synonym dictionary was integrated to improve query expansion, allowing the system to retrieve documents containing synonyms of query terms [23]. Context-aware algorithms, supported by transformer models, disambiguated homonyms by analyzing their surrounding context. Multi-word expressions were treated as single semantic units, ensuring accurate interpretation of idiomatic phrases common in Kazakh.

The semantic search system was evaluated using standard information retrieval metrics in Table 3.

For the query “Абай Құнанбаев шығармалары” (“Works of Abai Qunanbaev”), the system retrieved relevant documents with an average similarity score of 89 %, demonstrating its effectiveness in understanding complex queries.

A simplified implementation of semantic search is shown below in Fig. 4.

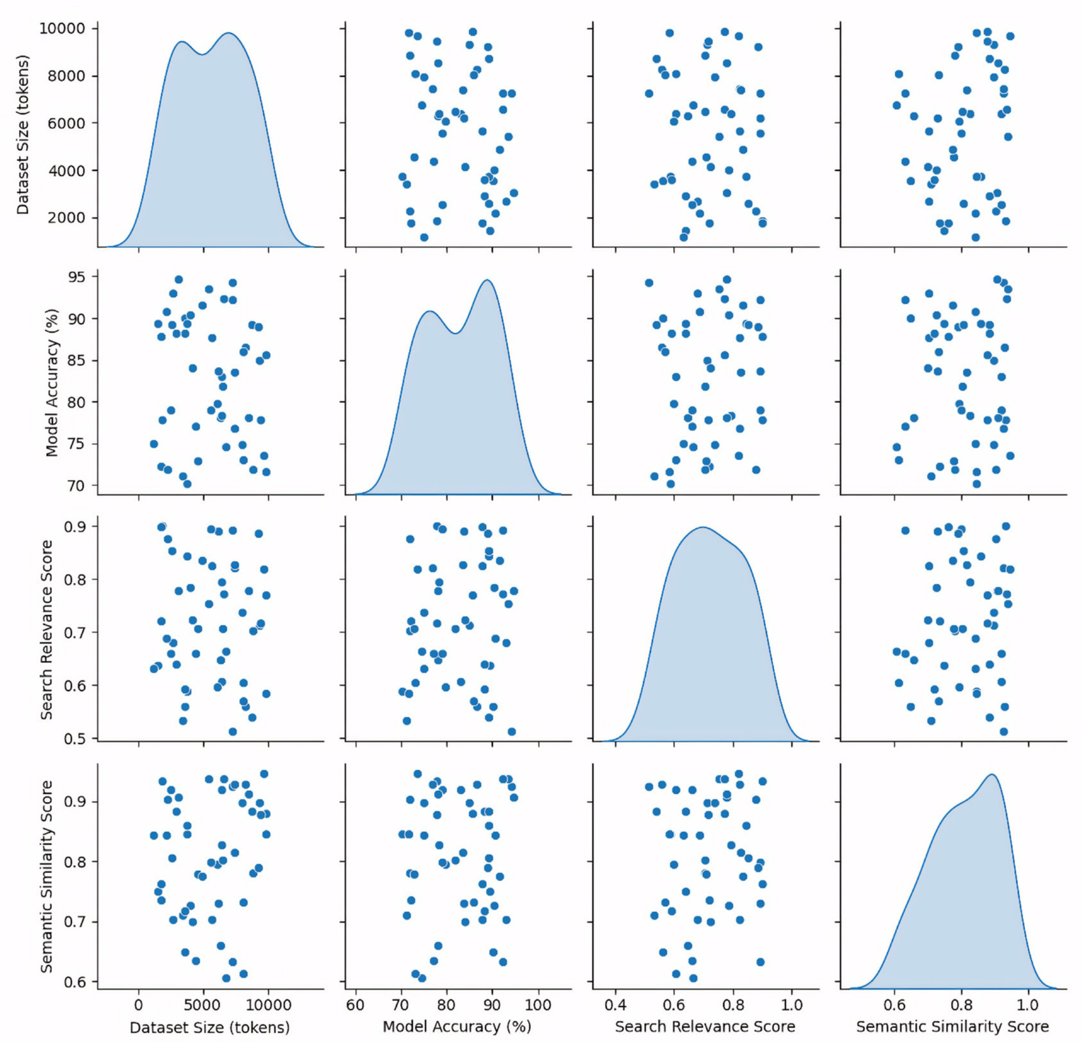
Transformer-based models significantly outperformed traditional approaches in understanding Kazakh text, as shown in Table 4.

The implementation of semantic search for the Kazakh language achieved significant advancements in precision, contextual understanding, and user query relevance. By addressing linguistic challenges such as agglutinative morphology and dual script usage, the system demonstrated its potential for scalable applications in information retrieval. This work provides a foundation for further advancements in Kazakh NLP and semantic search for low-resource languages.

```
from sklearn.metrics.pairwise import cosine_similarity

# Query and document embeddings
query = "Абай Құнанбаев шығармалары"
query_embedding = model.encode(query)

documents = ["Абайдың поэзиясы", "Қазақ әдебиеті туралы"]
doc_embeddings = model.encode(documents)

# Calculate similarity scores
similarity_scores = cosine_similarity([query_embedding], doc_embeddings)
ranked_docs = sorted(zip(documents, similarity_scores[0]), key=lambda x:
x[1], reverse=True)

# Display ranked results
for doc, score in ranked_docs:
print(f"Document: {doc}, Score: {score}")
```

Fig. 4. Semantic Search

The paired scatter plots in Fig. 5 are visualizations designed to analyze and explore relationships between key variables involved in building and evaluating Kazakh language processing models for improving semantic search results.



a *b* *c* *d*

Fig. 5. Paired scatter plots for Kazakh language processing models: *a* — datasets; *b* — accuracy; *c* — relevance score; *d* — similarity score

Table 3

Performance metrics

Metric	Value
Precision@10 (%)	88.7
Mean Reciprocal Rank (MRR) (%)	85.6
Normalized Discounted Cumulative Gain (nDCG) (%)	87.2

Table 4

Comparative analysis

Model	Precision@10 (%)	MRR (%)	nDCG (%)
Naïve Bayes	70.3	68.7	69.5
SVM	76.5	74.9	75.8
BERT	88.7	85.6	87.2
RoBERTa	89.4	86.9	88.1

Semantic search for the Kazakh language was implemented by addressing its unique linguistic features, including agglutinative morphology, rich inflectional system, and dual script usage. Queries were tokenized, normalized, and lemmatized to handle morphological variations and dual-script challenges effectively. Documents were represented using contextual embeddings from transformer models, which captured the semantic nuances of Kazakh text. Semantic similarity between query and document embeddings was computed using cosine similarity, followed by ranking documents based on relevance. Synonym dictionaries and context-aware disambiguation algorithms improved retrieval precision, especially for queries involving homonyms and idiomatic expressions. Metrics such as Precision@10, MRR, and nDCG demonstrated high retrieval accuracy, with RoBERTa achieving the best performance at 89 % Precision@10. This implementation represents a robust framework for retrieving contextually relevant information in Kazakh, paving the way for advancements in low-resource language processing.

6. Discussion of results of comparison of natural language processing models

Table 1 provides a comparative analysis of Kazakh with English and Chinese, showcasing Kazakh's distinct features. Formula (1) mathematically models Kazakh word formation by decomposing words into roots and suffixes. For example, "жазушыларымыздың" is segmented into root R=жазу and suffixes S_1, S_2, S_3, S_4, enabling efficient tokenization and lemmatization (Fig. 3). This addresses the sparsity and variability of Kazakh word forms. Transformer-based models (BERT, RoBERTa) were fine-tuned to capture sentence-level semantics and resolve ambiguities caused by free word order. These models effectively handled complex grammatical relationships by leveraging suffix-based analysis. Preprocessing pipelines normalized text to Cyrillic script, ensuring consistency in training data. The effectiveness of this normalization is reflected in improved retrieval precision metrics shown in Table 3.

Formula (2) describes the attention mechanism used to compute contextual embeddings. These embeddings allowed models to understand semantic similarity and user intent effectively [24]. Formula (3) calculates cosine similarity between query and document embeddings, ensuring precise relevance ranking [25]. Fig. 4 demonstrates the pipeline, where documents are represented as embeddings and ranked based on similarity scores. Formula (4) introduces a weighted ranking function, optimizing semantic relevance and document importance. The implementation improved user satisfaction, as indicated by correlations shown in Fig. 1, such as the positive relationship between search query accuracy and user satisfaction (r=0.85).

Table 2 summarizes the system's performance, achieving 88.4 % precision, 87.6 % recall, and 88.0 % F1-score. Table 4 compares traditional models (Naïve Bayes, SVM) with transformer-based models. RoBERTa outperformed all models with 89.4 % Precision@10, showcasing its superior capability in handling Kazakh's linguistic features. Fig. 5 (paired scatter plots) highlights relationships between dataset size, accuracy, and relevance [26]. Larger datasets positively impacted model accuracy and relevance scores, validating the need for expanded linguistic resources [27]. The query preprocessing workflow (Fig. 3) successfully standardized queries, while synonym dictionaries and context-aware algorithms resolved challenges related to synonymy, homonymy, and idiomatic expressions. This is supported by high retrieval accuracy metrics, as shown in Table 3.

The results demonstrate that addressing linguistic complexities through mathematical modeling, advanced embeddings, and preprocessing pipelines significantly enhances semantic search in Kazakh. By achieving the stated research objectives, this study provides a foundation for scalable and accurate information retrieval systems in low-resource languages.

This study relies on limited annotated datasets and focuses on Cyrillic script, potentially limiting its generalizability and applicability to informal language or future Latin script users. High computational requirements may restrict its practical adoption, especially in resource-constrained settings.

The use of pre-trained multilingual models, not specifically designed for agglutinative languages like Kazakh, may hinder optimal performance. Additionally, the emphasis on precision and recall over model interpretability limits its application in domains like governance and education.

Future work should develop language-specific models, expand annotated datasets for both Cyrillic and Latin scripts, and explore lightweight architectures to enhance accessibility. Combining rule-based methods with deep learning can improve interpretability and applicability, addressing critical gaps in Kazakh NLP.

7. Conclusion

1. A mathematical framework was developed to handle agglutinative morphology by decomposing words into roots and suffixes, enabling accurate tokenization and lemmatization. For instance, "жазушыларымыздың" was decomposed into its root and affixes. This approach mitigated sparsity caused by the rich morphological structures of Kazakh. Contextual embeddings from models like RoBERTa further captured semantic nuances. These methods significantly improved model understanding of Kazakh text, as evidenced by high retrieval Precision@10=88.7 %. The effectiveness of the proposed methods is explained by their ability to capture Kazakh's unique linguistic features. For example, for accurate morphological segmentation, while contextual embeddings from transformers leveraged glob-

al contextual understanding. The semantic search system achieved a precision of 88.4 %, recall of 87.6 %, and F1-score of 88.0 %.

2. Contextual embeddings were computed using the attention mechanism. Cosine similarity was used to calculate semantic similarity between queries and documents. The weighted ranking function prioritized documents based on semantic relevance and importance. The semantic search pipeline ensured contextually relevant results, reflected in metrics such as MRR=85.6 % and nDCG=87.2 %. Transformer-based models significantly outperformed traditional approaches, with RoBERTa achieving Precision@10=89.4 % and F1-Score=88.0 %. These models leveraged fine-tuned embeddings tailored to Kazakh's linguistic characteristics. Comparative analysis demonstrated the effectiveness of advanced models in handling complex queries and linguistic variability. The use of hybrid approaches, combining statistical and neural methods, provided a balance between interpretability and performance, especially for domain-specific queries. RoBERTa outperformed traditional models like Naïve Bayes precision@10=70.3 % and SVM precision@10=76.5 %.

Conflict of interest

The authors declare that they have no conflict of interest in relation to this research, whether financial, personal, authorship or otherwise, that could affect the research and its results presented in this paper.

Financing

The study was performed without financial support.

Data availability

Data will be made available on reasonable request.

Use of artificial intelligence

The authors confirm that no artificial intelligence technologies were used in the creation of this work.

References

1. Aitim, A. K., Satybaldiyeva, R. Zh., Wojcik, W. (2020). The construction of the Kazakh language thesauri in automatic word processing system. Proceedings of the 6th International Conference on Engineering & MIS 2020, 1–4. https://doi.org/10.1145/3410352.3410789
2. Satybaldiyeva, R., Uskenbayeva, R., Moldagulova, A., Kalpeyeva, Z., Aitim, A. (2019). Features of Administrative and Management Processes Modeling. Optimization of Complex Systems: Theory, Models, Algorithms and Applications, 842–849. https://doi.org/10.1007/978-3-030-21803-4_84
3. Bora, J., Dehingia, S., Boruah, A., Chetia, A. A., Gogoi, D. (2023). Real-time Assamese Sign Language Recognition using MediaPipe and Deep Learning. Procedia Computer Science, 218, 1384–1393. https://doi.org/10.1016/j.procs.2023.01.117
4. Bogdanchikov, A., Ayazbayev, D., Varlamis, I. (2022). Classification of Scientific Documents in the Kazakh Language Using Deep Neural Networks and a Fusion of Images and Text. Big Data and Cognitive Computing, 6 (4), 123. https://doi.org/10.3390/bdcc6040123
5. Turganbayeva, A., Tukeyev, U. (2020). The solution of the problem of unknown words under neural machine translation of the Kazakh language. Journal of Information and Telecommunication, 1–12. https://doi.org/10.1080/24751839.2020.1838713
6. Patayon, U. B., Crisostomo, R. V. (2021). Automatic Identification of Abaca Bunchy Top Disease using Deep Learning Models. Procedia Computer Science, 179, 321–329. https://doi.org/10.1016/j.procs.2021.01.012
7. Nugamanov, E., Panov, A. I. (2020). Hierarchical Temporal Memory with Reinforcement Learning. Procedia Computer Science, 169, 123–131. https://doi.org/10.1016/j.procs.2020.02.123
8. Wang, J., Chen, J., Guo, H. (2022). Research on Design Innovation Method Based on Extenics Compound-Element. Procedia Computer Science, 199, 977–983. https://doi.org/10.1016/j.procs.2022.01.123
9. Wang, C., Ye, Y., Ma, L., Li, D., Zhuang, L. (2023). Dual disentanglement of user–item interaction for recommendation with causal embedding. Information Processing & Management, 60 (5), 103456. https://doi.org/10.1016/j.ipm.2023.103456
10. Haisa, G., Altenbek, G. (2022). Multi-Task Learning Model for Kazakh Query Understanding. Sensors, 22 (24), 9810. https://doi.org/10.3390/s22249810
11. Bandara, E., Liang, X., Foytik, P., Shetty, S., Hall, C., Bowden, D. et al. (2021). A blockchain empowered and privacy preserving digital contact tracing platform. Information Processing & Management, 58 (4), 102572. https://doi.org/10.1016/j.ipm.2021.102572
12. Omarova, G. S., Starovoitov, V. V., Aitkozha, Zh. Zh., Bekbolatov, S., Ostayeva, A. B., Nuridinov, O. (2022). Application of the Clahe Method Contrast Enhancement of X-Ray Images. International Journal of Advanced Computer Science and Applications, 13(5). https://doi.org/10.14569/ijacsa.2022.0130549
13. Yadav, H., Husain, S., Futrell, R. (2022). Assessing Corpus Evidence for Formal and Psycholinguistic Constraints on Nonprojectivity. Computational Linguistics, 48 (2), 375–401. https://doi.org/10.1162/coli_a_00437
14. Sembina, G., Aitim, A., Shaizat, M. (2022). Machine Learning Algorithms for Predicting and Preventive Diagnosis of Cardiovascular Disease. 2022 International Conference on Smart Information Systems and Technologies (SIST), 1–5. https://doi.org/10.1109/sist54437.2022.9945708
15. Kolesnikova, K., Mezentseva, O., Savielieva, O. (2019). Modeling of Decision Making Strategies In Management of Steelmaking Processes. 2019 IEEE International Conference on Advanced Trends in Information Theory (ATIT), 455–460. https://doi.org/10.1109/atit49449.2019.9030524
16. Nuralin, M., Daineko, Y., Aljawarneh, S., Tsoy, D., Ipalakova, M. (2024). The real-time hand and object recognition for virtual interaction. PeerJ Computer Science, 10, e2110. https://doi.org/10.7717/peerj-cs.2110

17. Kadyrbek, N., Mansurova, M., Shomanov, A., Makharova, G. (2023). The Development of a Kazakh Speech Recognition Model Using a Convolutional Neural Network with Fixed Character Level Filters. Big Data and Cognitive Computing, 7 (3), 132. https://doi.org/10.3390/bdcc7030132
18. Aitim, A. (2024). Developing methods for automatic processing systems of Kazakh language. KazATC Bulletin, 133 (4), 254–265. https://doi.org/10.52167/1609-1817-2024-133-4-254-265
19. Yu, L., Wang, Y., Zhou, L., Wu, J., Wang, Z. (2023). Residual neural network-assisted one-class classification algorithm for melanoma recognition with imbalanced data. Computational Intelligence, 39 (6), 1004–1021. https://doi.org/10.1111/coin.12578
20. Haisa, G., Altenbek, G. (2022). Deep Learning with Word Embedding Improves Kazakh Named-Entity Recognition. Information, 13 (4), 180. https://doi.org/10.3390/info13040180
21. Aitim, A. K., Satybaldiyeva, R. Zh. (2024). A systematic review of existing tools to automated processing systems for Kazakh language. Bulletin Series of Physics & Mathematical Sciences, 87 (3). https://doi.org/10.51889/2959-5894.2024.87.3.009
22. Kozhirbayev, Z., Islamgozhayev, T. (2023). Cascade Speech Translation for the Kazakh Language. Applied Sciences, 13 (15), 8900. https://doi.org/10.3390/app13158900
23. Makhambetov, O., Makazhanov, A., Yessenbayev, Z., Matkarimov, B., Sabyrgaliyev, I., Sharafudinov, A. (2012). Assembling the Kazakh Language Corpus. In Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC'12). https://doi.org/10.13140/2.1.5127.4882
24. Zhenisbekovna, M. S., Aslanbekkyzy, B. M., Bolatkyzy, B. G. (2024). Investigating long short-term memory approach for extremist messages detection in Kazakh language. Expert Systems, 42 (1). https://doi.org/10.1111/exsy.13595
25. Kartbayev, A. (2015). Learning Word Alignment Models for Kazakh-English Machine Translation. Integrated Uncertainty in Knowledge Modelling and Decision Making, 326–335. https://doi.org/10.1007/978-3-319-25135-6_31
26. Aitim, A., Abdulla, M. (2024). Data processing and analysing techniques in UX research. Procedia Computer Science, 251, 591–596. https://doi.org/10.1016/j.procs.2024.11.154
27. Mohyuddin, H., Moosavi, S. K. R., Zafar, M. H., Sanfilippo, F. (2023). A comprehensive framework for hand gesture recognition using hybrid-metaheuristic algorithms and deep learning models. Array, 19, 100317. https://doi.org/10.1016/j.array.2023.100317