UDC 004.891.2 https://doi.org/10.54251/2616-6429.2026.01.0014nu

Zh.D. Iztayev[1], D.K. Seitkulov[1], B.E. Aidaulet[1]*, S.D. Kurakbayeva[1], Sh. Ganbari[2]
[1]Candidate of Pedagogical Sciences, Associate Professor, M. Auezov SKU, Shymkent, Kazakhstan
[1]Master's student, M. Auezov SKU, Shymkent, Kazakhstan
[1]Master of Science, Lecturer, M. Auezov SKU, Shymkent, Kazakhstan
[1]Candidate of Technical Sciences, Professor, M. Auezov SKU, Shymkent, Kazakhstan
[2]PhD, Associate Professor, Islamic Azad University, Ashtian Branch, Iran
***Corresponding author's email:** bekarys.aidaulet@mail.ru

DEVELOPMENT OF AN INTELLIGENT QUESTION ANSWERING SYSTEM FOR PROCESSING LEGISLATIVE TEXTS USING TRANSFORMER MODELS

Abstract

This paper presents the development of an intelligent question answering system for processing legislative texts using transformer models. The proposed approach is based on natural language processing techniques and semantic similarity computation between user queries and normative legal documents. Pre-trained transformer-based language models, including KazBERT and XLM-RoBERTa, are employed to generate vector representations of textual data. Semantic similarity between questions and legislative articles is calculated using the cosine similarity metric, and the most relevant text fragments are selected using an extractive method. The system is implemented as a modular web-based software solution, enabling scalability and further model integration. Experimental evaluation conducted on a corpus of legislative acts of the Republic of Kazakhstan demonstrates satisfactory performance in terms of Accuracy, F1-score, and response time, confirming the applicability of the proposed system for automated intelligent legal information systems.

Keywords: semantic analysis, intelligent system, question and answer system, natural language processing, model, artificial intelligence.

Introduction

At the present stage, the rapid development of information technologies has a significant impact on all spheres of society, including the legal system. The process of digitalization necessitates expanding access to legislative information and establishing effective mechanisms for its processing and use. However, the complex structure and formal language of legislative texts make information comprehension difficult for ordinary users. In this regard, the application of intelligent information systems in the legal domain has become one of the pressing issues [1].

In recent years, question-answering systems based on Natural Language Processing (NLP) and artificial intelligence technologies have been developing intensively. Such systems enable the automatic retrieval and presentation of semantically relevant information by processing user queries formulated in natural language. Nevertheless, most studies in this field are focused on English or Russian, while the number of solutions designed to work with Kazakh-language legislative texts remains limited. The agglutinative nature of the Kazakh language, the scarcity of linguistic resources, and the lack of sufficiently annotated datasets significantly complicate the development of intelligent systems for this domain [2].

Materials and Methods

Overview of Intelligent Systems for Processing Legislative Texts

In recent years, intelligent question-answering systems designed for the automated processing of legislative information have been developing rapidly and have become an essential component of the digitalization of legal services. The primary objective of such systems is to promptly provide relevant legal norms, regulatory acts, or interpretations in response to user queries expressed in natural language. At present, a number of analogous systems are in use at both global and regional

levels.

At the international level, some of the most widely used legal information systems include the LexisNexis and Westlaw platforms. These systems offer intelligent search mechanisms based on extensive databases of legislative materials and judicial practice. However, the aforementioned platforms operate on a subscription-based model and provide limited support for the Kazakh language [3].

Systems such as Harvey AI and Casetext (CoCounsel) employ transformer-based large language models to analyze legal documents, answer legal queries, and process contracts. The main advantages of these systems lie in their high contextual relevance and accuracy of responses. Nevertheless, they are predominantly designed to operate in English [4].

In the Republic of Kazakhstan, the adilet.zan.kz portal provides official texts of legislative acts. However, this portal relies on traditional keyword-based search and does not fully support intelligent question-answering functionality. A comparative analysis of existing legislative information systems is presented in Table 1.

Table 1 – Comparative Analysis of Analogous Systems for Legislative Information Processing

No.	System Name	Application Domain	Supported Languages	Intelligent QA	Adaptation to National Legislation	Accessibility
1	LexisNexis	International law, case law	English and other languages	Yes	No	Paid
2	Westlaw	Legal search and analytics	English	Yes	No	Paid
3	Harvey AI	Document analysis, legal advisory services	English	Yes	No	Paid
4	Casetext (CoCounsel)	Legal analytics	English	Yes	No	Paid
5	adilet.zan.kz	Legislative acts of the Republic of Kazakhstan	Kazakh, Russian	No	Yes	Free

Most existing systems are primarily designed for professional lawyers and operate in foreign languages. From this perspective, the development of a domestic intelligent question-answering system for processing Kazakh-language legislative information represents a relevant and necessary scientific and practical task [5].

Description of Data Processing Methods

During the study, the dataset consisted of the current normative legal acts of the Republic of Kazakhstan. The legal texts were obtained from the adilet.zan.kz portal as the primary data source.

The dataset included key codes and laws in the fields of labor law, civil law, and administrative law. Legal texts were structured by articles and clauses, with each article considered as a separate textual unit. Overall, the dataset comprised several thousand legal articles. During the preprocessing stage, the texts were cleaned of extraneous characters and standardized into a uniform format.

In the development of the intelligent question-answering system, contemporary Natural Language Processing (NLP) and machine learning methods were applied, taking into account the specific features of legislative information processing, as summarized in Table 2. The proposed system is designed to semantically analyze user queries and automatically retrieve the relevant

information from the legislative database [6].

Table 2 – Data Processing Stages and Methods

No.	Stage Name	Method / Algorithm	Brief Description
1	Text Preprocessing	Tokenization, Normalization	Standardizing the format of queries and legal texts, reducing linguistic noise
2	Semantic Vectorization	KazBERT, XLM-RoBERTa	Creating semantic vector representations of the texts
3	Similarity Calculation	Cosine Similarity	Determining semantic closeness between the query and legal texts
4	Answer Generation	Extractive Method	Directly retrieving the relevant article or excerpt from the legal text
5	Evaluation of Results	Accuracy, F1-score, Time	Assessing the quality of responses and system performance

Figure 1 illustrates the overall workflow of the intelligent question-answering system. First, the user query is received by the system and undergoes the preprocessing stage, during which the text is cleaned and transformed into a format suitable for further processing.

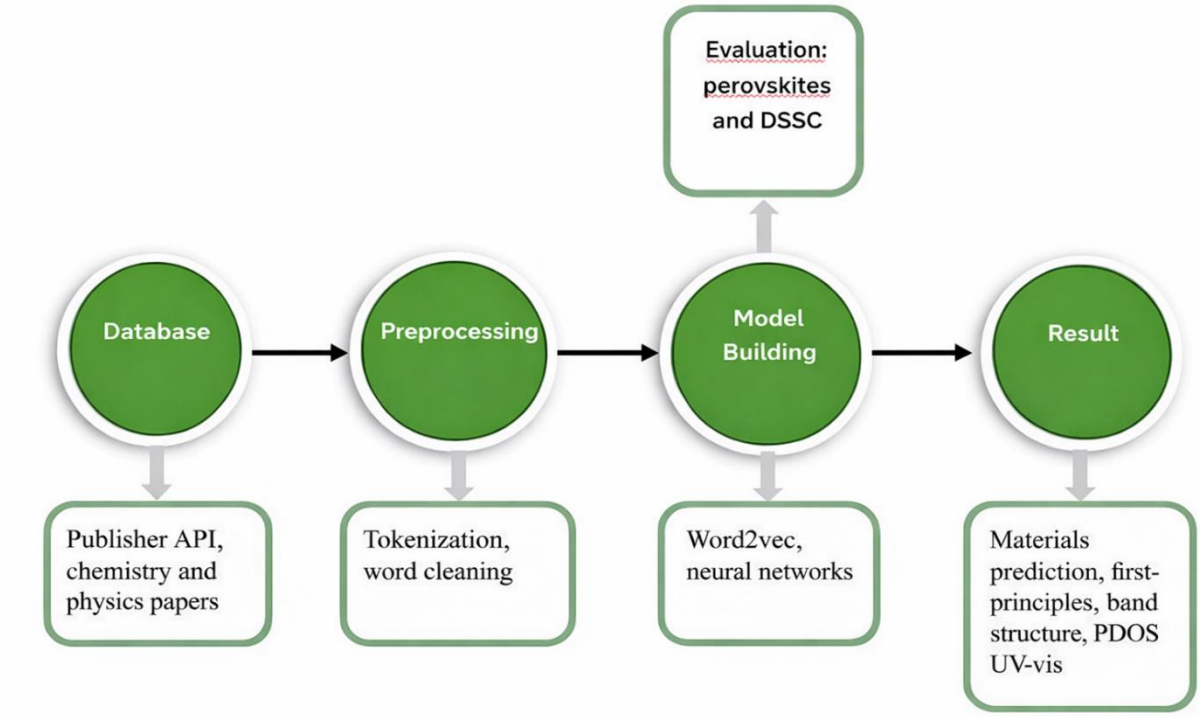

Figure 1 – Operation algorithm of the intelligent question answer system

In the second stage, the user query and legislative texts are transformed into vector representations using semantic models. The semantic similarity between the resulting vectors is calculated using the cosine similarity metric. Legal articles with the highest similarity scores are selected for answer generation.

At the final stage, the system presents the user with the relevant legal article or a concise excerpt extracted from the legislative text. Thus, the proposed scheme fully illustrates the data processing and response generation workflow of the intelligent question-answering system. The system's operational

algorithm begins with storing data in the database, followed by a preprocessing stage. Based on the processed data, a model is constructed, and the resulting outputs are subsequently stored in the database.

Additionally, the proposed approach ensures a scalable architecture that can be extended to other branches of law by expanding the legislative dataset. The use of semantic models improves the robustness of the system when handling linguistically complex and context-dependent legal queries. Overall, the developed algorithm enhances the accessibility of legislative information for non-expert users while maintaining acceptable accuracy and response time [7].

Architecture of the Intelligent Question-Answering System

The architecture of the proposed intelligent question-answering system is based on modular and multi-layered principles. According to the stages presented in Table 3, the architecture covers all key processes, from receiving a user query to generating a precise response based on legislative information.

Table 3 – Description of the System Architecture and Its Modules

No.	Layer Name	Module Name	Brief Description
1	User Interaction Layer	Web/Chat Interface	User queries are accepted in natural language. The interface is user-friendly and does not require legal expertise.
2	Preprocessing Layer	Text Preprocessing and Linguistic Analysis	Text cleaning, tokenization, lemmatization, and removal of stop words are performed.
3	Semantic Analysis Layer	Semantic Matching	Semantic similarity between the user query and legislative texts is determined.
4	Data Storage Layer	Legislative Database	Normative legal acts of the Republic of Kazakhstan are stored. Legal texts are structured by articles and clauses.
5	Result Presentation Layer	Answer Generation Module	A response is generated and presented to the user based on the selected legal norms.

The proposed architectural solution enables future system scalability, supports the integration of new legislative data, and allows for the incorporation of more advanced artificial intelligence models, thereby enhancing the overall functionality and adaptability of the system. The layered architecture ensures clear separation of responsibilities between modules, which improves system maintainability, facilitates debugging, and supports efficient updates as legal regulations and AI technologies evolve [8].

Results and discussion

To assess the performance of the proposed intelligent question-answering system, legal queries formulated in the Kazakh language were used, and the system’s ability to provide correct answers was evaluated. For the experiment, 50 questions related to common legal issues encountered in everyday life were selected. For each question, the correct answer in the form of a relevant legal article or its content was determined in advance.

The evaluation results show that the system demonstrated stable and reliable performance. The accuracy level reached 82%, indicating that the system was able to correctly answer the majority of the legislative questions. The F1-score value of 0.81 confirms that the balance between the precision and completeness of the answers is at a satisfactory level. The average response time ranged from 2 to 3 seconds, which makes the system convenient for practical use.

Overall, the experimental results confirm the effectiveness of the proposed intelligent question-answering system in processing legislative information. The system was able to identify relevant legal provisions based on semantic meaning even when the user queries did not contain exact legal terminology. This is particularly important for non-expert users, as many citizens are not familiar with formal legal language, and therefore require intuitive and accessible tools for obtaining legal information.

In addition, the obtained results indicate that the use of semantic models significantly improves the system's robustness when handling linguistically diverse and context-dependent queries. The system demonstrates the potential to reduce the time and effort required to search for relevant legal information compared to traditional keyword-based search tools. These findings suggest that the proposed approach can serve as a practical foundation for developing user-oriented legal information services in the Kazakh language and can be further enhanced by expanding the dataset and incorporating more advanced language models.

Conclusion

This study examined the technology for developing an intelligent question-answering system designed to process legislative information in the state language. The primary objective of the system is to automatically identify and present relevant legal articles in response to legislative queries formulated in the Kazakh language.

The experimental results demonstrated that the system is capable of correctly answering the majority of the submitted queries. Compared to traditional keyword-based search approaches, the intelligent method proved to be more effective, as it takes into account the semantic meaning of the user's question and enables the retrieval of contextually relevant legal norms.

The proposed intelligent question-answering system facilitates access to legislative information in the state language and can contribute to improving users' legal awareness. In future work, the system's capabilities may be further enhanced by expanding the dataset and integrating generative response modules, thereby increasing both the coverage and the expressiveness of the provided answers.

References

1. Rajchandar K., Gupta P., Suthar G., Sidhu K. S., Sarkar R., Satyanarayana P. Natural Language Processing for AI-Powered Legal Document Analysis // 2025 International Conference on Computing Technologies & Data Communication (ICCTDC), 2025. P. 1-5.
2. Usen E.B. Application of neural network models for semantic analysis of publications in social mediasocial networks // International Scientific Journal "BULLETIN OF SCIENCE", 2025, vol. 4. No. 2(83), pp. 480-485.
3. Baegizova A. S., Myrzabekova G. E., Alimagambetova A. Z., Mukhamedrakhimova G. I., Kassim M. analysis of short texts using intelligent clustering methods // International Journal of information and communication technologies, 2025, Vol.6. No. 2, p.23-36.
4. Kaibasova D. zh., Makhanova B. M., Suleimen A. E. methods of Natural Language Processing (NLP) in determining the text Stilin in the Kazakh language // university proceedings, 2021, No. 2, p. 172-176.
5. Esengabylov I., Aldabergenova A., Orazbayeva A., Adambekova A., Zhapsarbayev G. Analysis of the practical application of neural networks for handwriting recognition // Bulletin of KazATK, No.136(1), pp.155-161.
6. Nithya M., Harini S., Kavyadharshini S., Srinidhi K. AI-Driven Legal Automation to Enhance Legal Processes with Natural Language Processing // International Conference on IoT Based Control Networks and Intelligent Systems (ICICNIS), 2024, №1, P. 1246-1253.

7. Akanov B. B., Muratovich D. modern technologies for creating a chat bot // Bulletin of the Kazakh-American Free University, 2021, No. 3, p. 233-238.
8. Zakharova O. I. Semantic analysis and synthesis of textual data // VSU Bulletin. Series: System Analysis and Information Technologies, 2024, No. 4, pp. 182-208.

Ж.Д. Изтаев[1], Д.К. Сейткулов[1], Б.Е. Айдәулет[1]*, С.Д. Куракбаева[1], Ш. Ганбари[2]
[1]п.ғ.к., қауымдастырылған профессор, zhalgasbek71@mail.ru, М.Әуезов атындағы ОҚУ, Шымкент, Қазақстан
[1]магистрант, seitkulov.daulet80@gmail.com, М.Әуезов атындағы ОҚУ, Шымкент, Қазақстан
[1]магистр, оқытушы, bekarys.aidaulet@mail.ru, М.Әуезов атындағы ОҚУ, Шымкент, Қазақстан
[1]т.ғ.к., профессор, sevam@mail.ru, М.Әуезов атындағы ОҚУ, Шымкент, Қазақстан
[2]PhD, қауымдастырылған профессор, myrshg@gmail.com, Ислам Азад университеті, Аштиян, Иран

ТРАНСФОРМЕРЛІК МОДЕЛЬДЕРДІ ҚОЛДАНУ АРҚЫЛЫ ЗАҢНАМАЛЫҚ МӘТІНДЕРДІ ӨҢДЕУГЕ АРНАЛҒАН ИНТЕЛЛЕКТУАЛДЫ СҰРАҚ-ЖАУАП ЖҮЙЕСІН ӘЗІРЛЕУ

Түйін

Бұл мақалада трансформерлік модельдерді қолдану арқылы заңнамалық мәтіндерді өңдеуге арналған интеллектуалды сұрақ-жауап жүйесін әзірлеу қарастырылады. Ұсынылған тәсіл пайдаланушы сұрағы мен нормативтік-құқықтық актілер арасындағы семантикалық ұқсастықты анықтауға негізделген табиғи тілдерді өңдеу әдістеріне сүйенеді. Мәтіндердің векторлық көріністерін қалыптастыру үшін алдын ала оқытылған KazBERT және XLM-RoBERTa трансформерлік тілдік модельдері пайдаланылды. Семантикалық ұқсастық cosine similarity метрикасы арқылы есептеліп, ең сәйкес заңнамалық мәтін фрагменттері экстрактивті әдіс негізінде таңдалып алынады. Жүйе модульдік архитектураға ие веб-негізделген бағдарламалық шешім ретінде іске асырылды. Қазақстан Республикасының заңнамалық актілері корпусында жүргізілген эксперименттік бағалау нәтижелері жүйенің дәлдік (Accuracy), F1-score және жауап беру уақыты көрсеткіштері бойынша қанағаттанарлық нәтижелер көрсететінін дәлелдейді және оны интеллектуалды құқықтық ақпараттық жүйелерді автоматтандыруда қолдануға болатынын көрсетеді.

Кілттік сөздер: семантикалық талдау, интеллектуалды жүйе, сұрақ-жауап жүйесі, табиғи тілдерді өңдеу, модель, жасанды интеллект.

Ж.Д. Изтаев[1], Д.К. Сейткулов[1], Б.Е. Айдәулет[1]*, С.Д. Куракбаева[1], Ш. Ганбари[2]
[1]к.п.н., ассоциированный профессор, zhalgasbek71@mail.ru, ЮКУ им. М. Ауэзова, Шымкент, Казахстан
[1]магистрант, seitkulov.daulet80@gmail.com, ЮКУ им. М. Ауэзова, Шымкент, Казахстан
[1]магистр, преподаватель, bekarys.aidaulet@mail.ru, ЮКУ им. М. Ауэзова, Шымкент, Казахстан
[1]к.т.н., профессор, sevam@mail.ru, ЮКУ им. М. Ауэзова, Шымкент, Казахстан
[2]PhD, ассоциированный профессор, myrshg@gmail.com, Исламский университет Азад, Аштиян, Иран

РАЗРАБОТКА ИНТЕЛЛЕКТУАЛЬНОЙ ВОПРОСНО-ОТВЕТНОЙ СИСТЕМЫ ДЛЯ ОБРАБОТКИ ЗАКОНОДАТЕЛЬНЫХ ТЕКСТОВ С ИСПОЛЬЗОВАНИЕМ ТРАНСФОРМЕРНЫХ МОДЕЛЕЙ

Аннотация

В статье рассматривается разработка интеллектуальной вопросно-ответной системы для обработки законодательных текстов с использованием трансформерных моделей. Предлагаемый подход основан на методах обработки естественного языка и вычисления семантического сходства между пользовательским запросом и нормативно-правовыми актами. Для получения векторных представлений текстов применяются предобученные трансформерные языковые модели KazBERT и

XLM-RoBERTa. Оценка семантической близости осуществляется с использованием метрики cosine similarity, после чего релевантные фрагменты законодательных текстов извлекаются экстрактивным методом. Архитектура системы реализована в виде модульного веб-ориентированного программного решения, обеспечивающего масштабируемость и возможность интеграции дополнительных моделей. Экспериментальная проверка на корпусе законодательных актов Республики Казахстан показала, что разработанная система обеспечивает удовлетворительные показатели точности (Accuracy), F1-score и времени отклика, что подтверждает ее применимость для автоматизированных интеллектуальных правовых информационных систем.

Ключевые слова: семантический анализ, интеллектуальная система, система вопросов и ответов, обработка естественного языка, модель, искусственный интеллект.