STRUCTURE OF KEY PROCESSES IN UZBEK PARAPHRASING SYSTEM

Akhmedova Khusniya

Department of Info communication Engineering, TUIT, Tashkent, Uzbekistan

ABSTRACT

This article examines the conceptual model, architecture, and structure of the information system for paraphrasing texts in the Uzbek language. The conceptual model represents all processes in the system. In addition, the processes of paraphrasing sentences in the Uzbek language are also separately expressed. The purposes of use of the information system users are separately indicated. The developed information system also has the ability to identify paraphrased sentences in documents, which makes it easier to check publications in editorial offices, identify and isolate duplicate sentences in scientific articles and dissertations. The general architecture and user interface of the system have been developed. The roles of user, specialist, and programmer have been introduced in the system, and separate functions have been defined and analyzed for each role.

KEYWORDS

Paraphrase, Information System, Conceptual Model, Structure, Uzbek Language

1. INTRODUCTION

Due to the inherent complexity of natural language, a wide range of research directions has been actively developing worldwide within the field of natural language processing. These directions include automatic text analysis, automatic text synthesis, the creation and maintenance of automated lexical resources, the development of information retrieval systems, machine translation, the design of automated language learning systems, and the development of software tools for addressing problems in applied linguistics.

Information systems have existed since the emergence of society, because at different stages of development society required systematized, pre-prepared information for its management. This is especially true for production processes - processes related to the production of material and cultural goods. Because they are of vital importance for the development of society. An information system is a set of technologies, communication systems, information resources and specialists that provide workers in various fields with the transfer and processing of information about an object to perform management functions. It is used to store, process and search information in automated information systems, as well as to perform operations related to the collection, preparation and transmission of information on computers, as well as to provide information to the consumer. These systems have wide functional capabilities and are capable of storing and processing very large volumes of information.

2. RELATED WORKS

In the field of computational linguistics worldwide, a number of scholars have conducted extensive research on semantic analysis and analyzers. Notable contributions in this area include

the studies of A.V. Tuzov, M.V. Mozgovoy, A.V. Sokirko, N.A. Schlayefer, A.V. Mochalova, A.R. Gatiatullin, B. Yergesh, A. Sharipbay, G. Bekmanova, S. Lipniski, Y.A. Kanevskiy, Hamroyeva Sh.[3] and K.K. Boyarskiy. Issues related to the development of morphological analyzers have also been widely investigated in both global and Turkic linguistics. In particular, the works of P.S. Bakasova, M.G. Malkovskiy, A.S. Starostin, Ye.A. Kanevskiy, N.V. Kolpakova, S. Bird, E. Klein, and O.A. Mitrofanova provide detailed insights into this area. The studies conducted by numerous international researchers are significant in that they address a range of fundamental problems in the field of Natural Language Processing (NLP). Research efforts aimed at processing the Uzbek language are reflected in the works of M. Musayev, O. Xamdamov, B. Elov, I. Bakayev [4], B. Akmuradov, M. Abdullayeva, and X. Axmedova, among others [5-7]. Based on the linguistic models proposed by these researchers, initial steps are currently being taken toward the development of semantic analyzers for the Uzbek language [8].

Numerous paraphrase corpora and systems have been developed worldwide, and a comparative analysis of these systems is presented in Table 16. As can be seen from Table 1, modern paraphrasing systems developed at the global level provide the ability to paraphrase input sentences across different styles or domains [2]. However, this functionality has not been addressed in the proposed information system for paraphrasing Uzbek-language sentences.

Table 1. Comparison of global and Uzbek paraphrase corpora

Paraphrasing System	**Multilingual**	**Number of characters**	**Extraction of paraphrased sentences from a file**	**Style-based Paraphrasing**	**Adapted for Uzbek-language Texts**
QuillBot	26	125	-	+	-
Myessaywriter.ai	26	250	-	+	-
Ahrefs' Paraphrasing	42	2048	-	+	-
WriteHuman	28	200	-	+	-
Grammarly	30+	500	-	+	-
CleverSpinner		200	-	+	-
Spin Rewriter		200	-	+	-
Perefrazr.io	1+	300	-	+	-
Prepostseo Paraphrasing Tool	5+	500	-	+	-
Muharrir AI	1	125	-	-	-
ZeroGPT	+	300	-	+	+
Uzbek Paraphrase	1+	1000	+	-/+	+

3. INFORMATION SYSTEM STRUCTURE AND ARCHITECTURE

Based on the conducted comparative analyses, it was identified that several systems for paraphrasing sentences in the Uzbek language have been developed. However, these systems do not provide functionality for automatically detecting and extracting paraphrased sentences from file-based texts.

In addition, the ChatGPT system is available, which supports not only sentence-level paraphrasing but also the detection of paraphrases within files. This system was also tested; however, the results showed that its accuracy in extracting paraphrases was relatively low.

Modern paraphrasing information systems also offer the capability to paraphrase texts in different styles (e.g., formal, literary, and scientific). At present, this functionality has not yet been implemented in the developed information system and is planned to be addressed in the subsequent stages of the research.

The information system for paraphrasing sentences in the Uzbek language is an automated information system that processes information and provides information to the user. That is, both the input and output values of the system are information. Like other automated information systems, the developed information system has its own structure, architecture, and conceptual model, which are described below. Figure 1 presents the structure of the developed information system, according to which the system is built on the basis of the following components [1].

Software - this component represents the software part of the information system and includes both frontend and backend modules. The software is implemented in the Python programming language, as Python provides modern technologies and numerous functionalities for natural language processing. It is highly convenient for performing operations on text and offers specialized libraries. For instance, libraries such as NLTK and NLP are well-suited for working with language constructs. Additionally, Python facilitates easy integration with contemporary machine learning models.

Mathematical Support includes the GloVe algorithm, version 3 of the Gemma model [5], the cross-entropy loss function, the multilingual-E5-text-embedding model, and cosine similarity algorithms. The specific applications and usage contexts of these algorithms have been described above.

Information Support consists of tables containing sentences and words in the Uzbek language, as well as text files extracted from various official web pages and articles. Words, sentences, and phrases in tabular form are used to identify paraphrased sentences in Uzbek, study their usage in different styles, and adapt them to various domains. Text-based data was employed to develop a model for paraphrasing sentences in the Uzbek language. The volume of this information support can be dynamically expanded; that is, users can add new sentences along with their paraphrases.

Linguistic Support forms the core of the information system. This support includes root words in the Uzbek language along with their synonyms, phrases and their synonyms, as well as sentences containing these phrases. Additionally, text data used to develop the sentence paraphrasing model for Uzbek sentences also constitutes a part of the linguistic support.

Organizational and Methodological Support refers to the methods employed during the development of the information system. This includes evaluation techniques, model creation procedures, and data handling methods. Additionally, the approach for comparing the developed information system with its other equivalents also constitutes a part of this support.

Technical Support refers to the technical resources required both during the development of the information system and for making it accessible to a wider audience. In this context, computers are essential for writing and testing the software components of the system, while storage devices are necessary for storing large volumes of text data. Servers play a crucial role in training large-scale models and evaluating the generated models. Furthermore, hosting services and domain management are important for deploying the developed information system to the public.

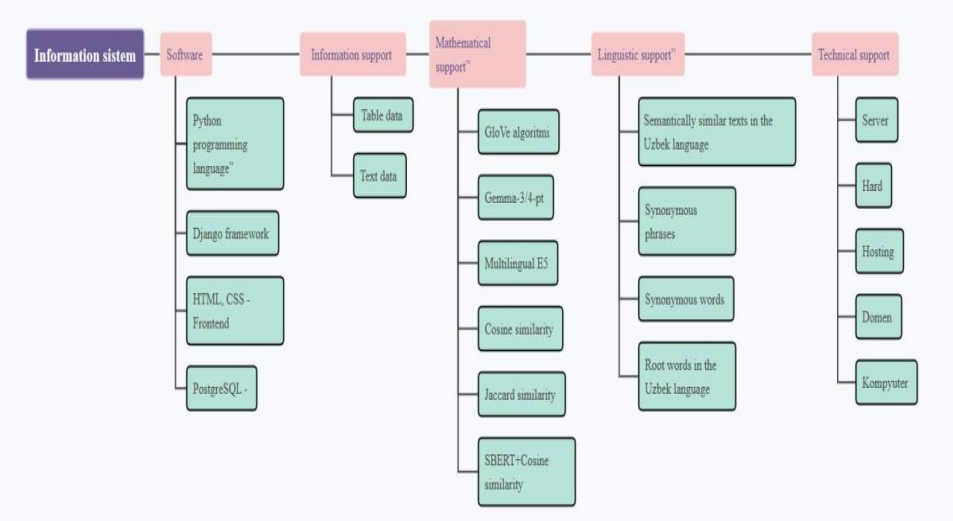

Figure 1. Structure of the Information System

The aforementioned support components collectively form the information system for paraphrasing sentences in the Uzbek language. The architecture of the information system is presented in Figure 2 and is based on the Model-View-Template (MVT) architecture.

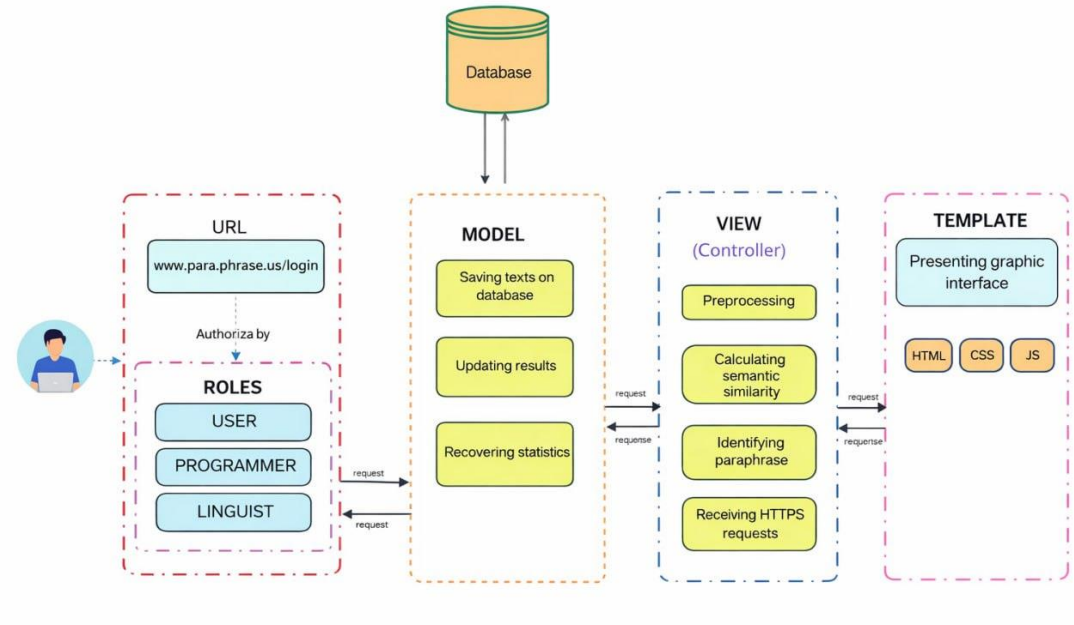

Figure 2. Architecture of the Information System

As illustrated in Figure 2, the **MODEL** component of the information system consists of .py files that implement operations such as creating the database, inserting data, and reading data. The **VIEW** component functions as the controller of the system, serving as the management layer between the database and the presentation of information to the user. The **TEMPLATE** component is responsible for forming the frontend part of the information system.

4. PURPOSES OF INFORMATION SYSTEM

The model reflecting the tasks involved in the development of the information system is referred to as the conceptual model, which illustrates the system's functions and the main steps for accomplishing them. As shown in Figure 3, the information system for paraphrasing sentences in the Uzbek language comprises two primary processes: the paraphrase identification process and the paraphrase generation process.

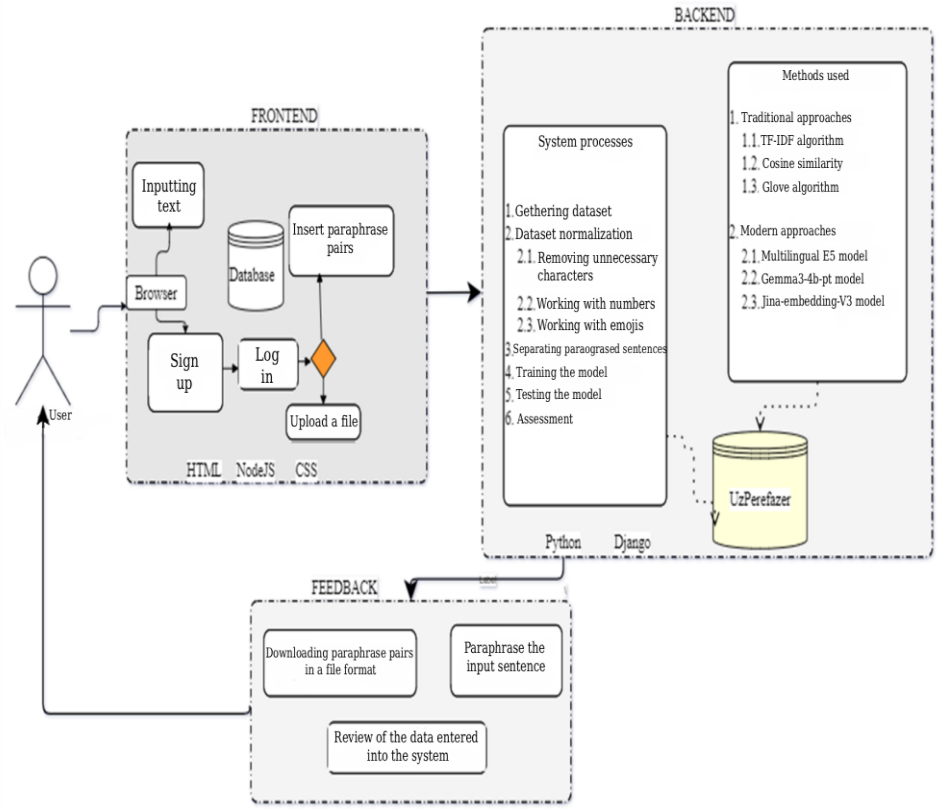

Figure 3. Conceptual model of the information system

As illustrated, each process consists of several stages. In the paraphrase identification process, the user is required to upload a file, and the system identifies paraphrases among the sentences contained in the uploaded file. The resulting output is displayed on the user interface, and users also have the option to download these results in .txt format. At this point, it is also important to consider the users and their access rights to the system.

The system includes users with three different roles: linguist, programmer, and regular user. Regular users can utilize the information system for the purposes illustrated in Figure 4. The roles of the linguist and the programmer are very important in the system. The linguist selects texts for training and testing and labels sentence pairs as paraphrase or non-paraphrase. The programmer, in turn, is responsible for testing and optimizing the system

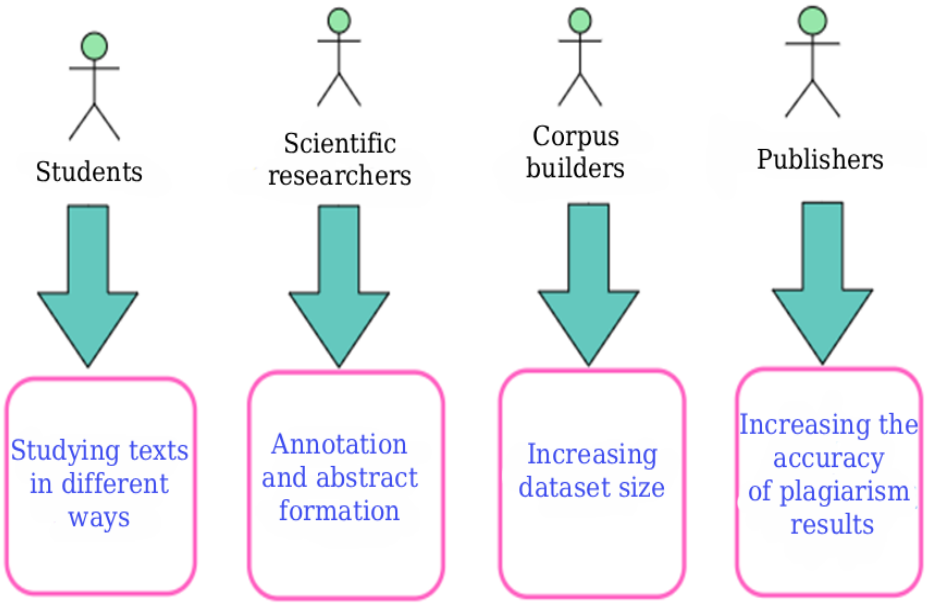

Figure 4. Purposes of using the information system by users

By users we refer to the users illustrated in Figure 4. Students can use the system to practice expressing texts in different styles. Currently, the capability to paraphrase Uzbek texts in various styles is not available; implementation of this functionality is planned for the subsequent stages of the research.

Scientific researchers can generate annotations for their scientific articles and dissertations using the information system. This can be achieved through the paraphrase detection module of the system. Specifically, the annotation is formed by aggregating sequences of sentences whose degree of paraphrasing falls within a specified range.

Corpus developers can increase the size of their datasets within the information system by generating paraphrases. As is well known, large-scale datasets can be effectively used to solve various NLP tasks.

Publishing houses can use the system during the pre-publication process to identify sentences with a high degree of paraphrasing and revise them accordingly. As a result, the revised texts are likely to achieve better results when subjected to plagiarism detection.

As shown in Figure 5, registered users of the system are provided with the ability to input paraphrased sentences, detect paraphrased sentences within uploaded files, and download the results. Users who are not registered can only use the sentence paraphrasing functionality. The main page of the information system provides a sentence paraphrasing feature that is accessible to all users. In addition, the system offers extended functionalities that can be accessed by users after registering in the system.

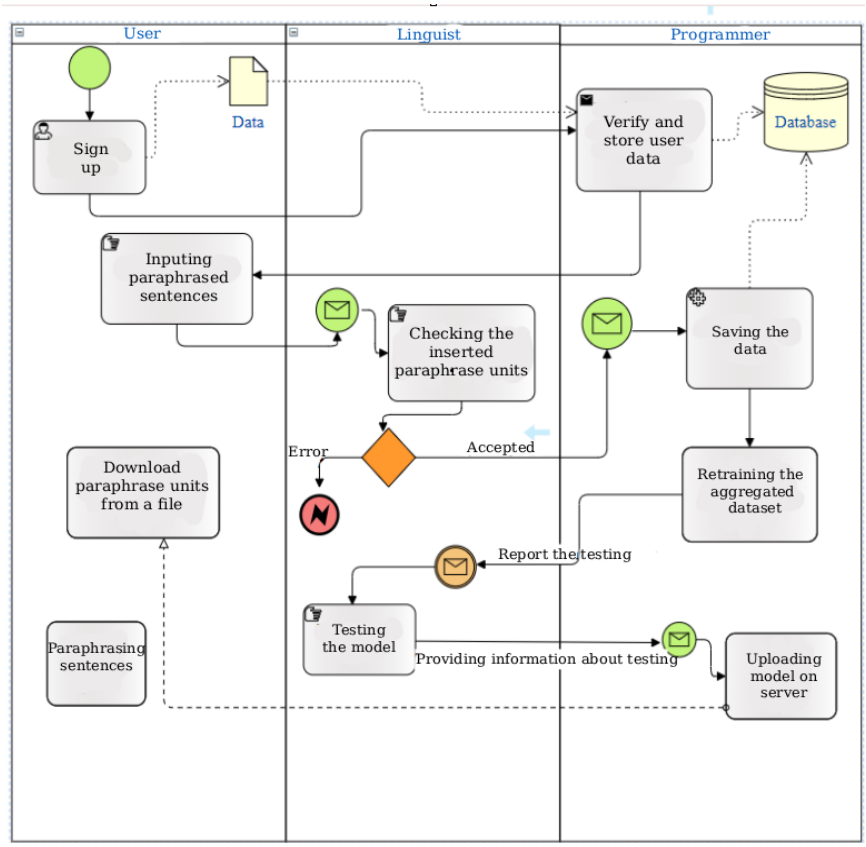

Figure 5. Roles implemented in the system and their associated privileges

These tasks were carried out using different approaches and a range of models built on modern transformer architectures. In developing the information system, the formation of the dataset - considered the core component of the system - was identified as a primary task. For model development, the Gemma3-4B-PT model was utilized. The model was trained on a 194 MB dataset consisting of 375K triplet sentences. Fine-tuning the model required four RTX 4090 GPUs. The training process was conducted over 11 days for a total of five epochs. As a result, the *gemma_text_to_paraphraser/uz* model was obtained. The system accepts user-uploaded files in *.txt*, *.doc*, and *.docx* formats to extract paraphrased sentences. As an output, sentences with a semantic similarity score higher than 85% are identified and the results are provided to the user in *.txt* format. To address this task, the *multilingual-E5-large-instruct* model was employed. The model generates semantic vector representations of sentences, and cosine similarity was used to compute similarity scores between them. As a result, the degree of semantic similarity between sentences was accurately determined.

The *Gemma-text-to-paraphrased_V2* model, obtained as a result of fine-tuning the *Gemma3-4B-PT* model, was evaluated, and the following results were achieved.

For evaluation purposes, the test sentences were divided into three groups:

1 sentences containing 11–15 tokens;
2 sentences containing 5–10 tokens;
3 inputs consisting of 1–4 tokens.

According to the results, sentences with a higher number of tokens preserved semantic meaning more accurately in the paraphrasing outputs. The test results are presented below.

Table 2. Results obtained based on the number of tokens

№	Types of sentences	Tests		Percentage indicator
		Number of original sentences	Number of accurately paraphrased sentences	
1	1–4 token	180	109	60%
2	5-10 token	250	231	92%
3	11-15 token	500	496	99%
			Average:	84%

The analysis results were calculated and evaluated using evaluation methods. The most commonly used metrics for evaluating system results are Precision, Recall, and the F1-score.

Table 3. Results obtained from the model

Model	google/gemma3-large
Precision	**0.954**
Recall	**0.955**
Accuracy	**0.913**
F1 score	**0.954**

The calculations presented are the results obtained from testing the *Gemma-text-to-paraphrased_V2* model, which was produced by fine-tuning the *Gemma3-4B-PT* model.

5. CONCLUSIONS

This article discusses the design and implementation of an information system aimed at paraphrasing sentences in the Uzbek language. During the study, the system's structure, architecture, and conceptual model were developed, and its main functional capabilities were analyzed in detail. The system was built based on modern NLP technologies, particularly the third version of the Gemma model, multilingual embedding models, and cosine similarity methods.

The developed information system enables users to paraphrase sentences in Uzbek, identify paraphrased sentences within texts, and access the results in a convenient format.

As a result, the developed information system has proven to be an effective tool for addressing paraphrasing tasks in the Uzbek language and can be widely applied in scientific and practical research, as well as in text processing and plagiarism reduction processes. Expanding the system's functions in the future, including the introduction of paraphrasing capabilities tailored to different styles and domains, is considered one of the promising directions for further research.

REFERENCES

[1] A. Thyagarajan, Siamese Recurrent Architectures for Learning Sentence Similarity, Artificial Intelligence 30(1) (2016) 2786-2792. DOI: 10.1609/aaai.v30i1.10350.

[2] Lin, Z., & Wan, X. (2021). Pushing Paraphrase Away from Original Sentence: A Multi-Round Paraphrase Generation Approach. Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021, 1548-1557-p. https://doi.org/10.18653/v1/2021.findings-acl.135
[3] Hamroyeva Sh. O‘zbek tili mualliflik korpusini tuzishning lingvistik asoslari: Filol.fan. bo‘yicha falsafa doktori (PhD). – Qarshi, 2018.–250 p.
[4] Bakayev I.I. «O‘zbek tili so‘z shakllarini morfologik tahlil qilish modellari va algoritmlari»: texnika fan. fan. bo‘yicha falsafa doktori (PhD)…disser.avtoref. – Toshkent, 2021. –12-22 p.
[5] Abdullayeva M. Nutq texnologiyalari asosida eshitish qobiliyati cheklangan bemorlarni reabilitatsiya qilish axborot tizimini ishlab chiqish: texnika fan. fan. bo‘yicha falsafa doktori (PhD) – Toshkent, 2023. –10-20 p.
[6] Axmedova X. O‘zbek tilidagi gaplarni semantik tahlil qilishning modeli, algoritmlari va axborot tizimini ishlab chiqish: texnika fan. fan. bo‘yicha falsafa doktori (PhD) – Toshkent, 2023.–5-23 p.
[7] Akmuradov B., Khamdamov U., Elov J., Sultanov D., Narzullayev I., “Organization of Initial Text Processing in the Uzbek Language Synthesizer”, 2021 International Conference on Information Science and Communications Technologies (ICISCT), 2021, pp. 1-5, doi: 10.1109/ICISCT52966.2021.9670048
[8] Thomas Mesnard, “Gemma 3 Technical Report,” arXiv:2503.19786 (Gemma family technical report). https://arxiv.org/abs/2503.19786
[9] Dj. Sultanov, X. Axmedova. Tabiiy tilni qayta ishlashda qo‘llaniladigan nlp modellari tahlili. Raqamli texnologiyalar va iqtisodiyot: zamonaviy muammolari va istiqbollari mavzusida Respublika ilmiy-amaliy konferenciya 2024-yil 24-25-aprel 22-25 p.

AUTHOR

Akhmedova Khusniya was born on May 15, 1986, in Tashkent Republik of Uzbekistan. She serves as a lecture of Department of Info communication Engineering, Tashkent University of Information Technologies named after Muhammad al-Khwarizmi. Her research interests include natural language processing (NLP), deep learning methods and language paraphrasing systems.