УДК 004.94(045)

ENHANCING INFORMATION RETRIEVAL WITH DJANGO-BERT INTEGRATION: A PATH TO QUESTION-ANSWER SYSTEMS

Kazangapova B.A., Zhappas Z.A., Suleimenov D.M., Suienaly A.A.

Almaty Technological University, Almaty, Kazakhstan

Today, in our information explosion age, data is constantly increasing in amount, increasing the likelihood we might be overwhelmed and drowned in the flow. Trying to find specific information on the internet, or getting the right answer for any query, is now turning into an actual problem, particularly in the scenario of an expanding information sphere. Individuals invest much energy and effort in searching for the specific information they require from any source, hindering decision making and work performance.

In such an instance, tools that enable such an undertaking to be sped up and made easier come into focus. Artificial intelligence (AI) and quality control systems are among such tools. Using these tools is possible to locate the information needed and give the right response much quicker than an individual can. Such systems can analyze texts, recognize patterns, and extract really meaningful information using machine learning algorithms, along with NLP-based technology.

Quality control systems process user requests as an input and, applying sophisticated analysis, determine the best and most meaningful answers. Here, the BERT model plays an especially significant role. Developed on the transformer framework and trained on massive quantities of text, it has an excellent language comprehension ability in the style of humans. For this reason, BERT is an extremely valuable assistant for tasks in information retrieval and comprehension.

A practical application has been the joining together of BERT and the Django framework. It brings powerful web development together with deep text analysis. It allows the development of web applications capable of handling user queries and providing the right answers through semantic text analysis. It presents wide

possibilities for the development of smart systems usable in the education, healthcare, science, and other sectors.

This method is marked by simplicity of usage owing to the convenient development framework of Django and BERT's high performance in the production of excellent answers. It guarantees high text-processing accuracy owing to the model being able to capture meaningful information and come up with meaningful as well as accurate responses. Furthermore, the system is flexible and can be easily adapted for customized tasks and scaled when needed.

Overall, combining Django and BERT allows building powerful smart systems able to respond with high accuracy and speed to any type of user query.

Methods and Materials.

The goal of the research is the realization of an intelligent information retrieval platform, embedding the BERT deep learning model into the Django web framework. The suggested solution should provide high accuracy in semantic text query processing, high response rates, and adaptive scalability, in order for the platform to be applicable in any fields of science and in multilingual environments. In addition, the platform should outperform conventional methods for information retrieval in parameters such as answer accuracy, query processing rate, and stability in high-load scenarios, combining thoroughly high-end natural language handling tools with modern-day web development tools. Part of the aim during the research is the universality and practicality of the resulting platform, implying its applicability in educational establishments, scientific institutions, and the business world alike. The resulting system should not only be able to give the right answers for complicated user questions, but it should be able to learn and refine itself constantly based on new specialist data through self-training and self-tuning. Great attention will be paid to performance, especially efficient use of resources and minimization of response latency. The developed solution should act as the transition point between the scientific community and practical application scenarios for smart search technologies, realizing the application of artificial intelligence advances in everyday life directly.

The system architecture integrates Django as the web framework and BERT via Hugging Face Transformers. It supports RESTful APIs, Redis caching, and Celery-based asynchronous processing to ensure scalability and responsiveness. Text preprocessing includes cleaning, normalization, lemmatization, and tokenization. These steps improve data consistency and ensure compatibility with BERT's input requirements for multilingual texts. The REST API built on Django REST Framework handles user requests and responses while embedding BERT for semantic analysis. This enables efficient integration with external systems. Redis and Celery are used for caching and asynchronous task management, reducing latency and improving stability under high load. The developed platform was compared with TF-IDF and BM25 in terms of accuracy, response time, and stability, showing clear advantages in semantic relevance and scalability. The system was validated on datasets from education, HR, and healthcare in English, Russian, and Kazakh languages, confirming its adaptability.

Results.

Implementing and Testing the Django-BERT Architecture.

In the early stage, the BERT model was integrated with the Django web framework via the Hugging Face Transformers library. The architecture built allowed for modular interactions among all the components of the system, including the REST API for accepting and handling user requests, orchestrating the BERT module for semantic text analysis, and the generation and provision of the response back to the users. Special care was taken while ensuring smooth coordination among the back-end operations such that data transferred between the API endpoints and the NLP module was efficient. Optimization in disk I/O was also achieved through applying Redis for cache storage and async task handling through the implementation of Celery. All the optimizations were explicitly included in code for the stability and response of the application while in high demand and multiple users concurrently access it. The system was thereby able to demonstrate its performance in supporting scalable and reliable operation,

establishing a sound foundation for further practical application and experimental analysis in other application scenarios.

Evaluation of Text Preprocessing Effectiveness.

The second phase of the study zeroed in on a detailed comparison of text preprocessing strategies—cleaning, normalization, lemmatization, and tokenization—across English, Russian, and Kazakh. Given the linguistic complexity and morphological variation across English, Russian, and Kazakh, the preprocessing workflows were carefully adapted—drawing on both standard NLP libraries and bespoke in-house code when off-the-shelf tools fell short. The experiments made one thing clear: preprocessing isn’t a background step; it plays a front-line role in enabling meaningful information extraction. Enhanced normalization and lemmatization routines helped preserve the semantic structure of texts, especially when handling domain-specific or multilingual input. The cleaner, more consistent data that resulted fed smoothly into the BERT model, making the subsequent semantic analysis and response generation markedly more effective. This phase underscored a crucial takeaway—robust, context-aware preprocessing is not just a performance booster; it's a prerequisite for scalable, intelligent retrieval systems that hold up in real-world conditions.

Experimental Comparison with Traditional Methods.

To evaluate how the system holds up against established techniques, a series of direct comparisons were conducted between the BERT-Django platform and traditional retrieval models like TF-IDF and BM25. The testing framework assessed not only the relevance of returned results, but also system responsiveness and performance consistency under growing user demand. The results spoke for themselves: the semantic depth offered by the BERT model allowed for a far more accurate interpretation of user queries, consistently yielding answers that aligned better with user intent. But it wasn’t just about smarter results—the architecture’s integration of caching mechanisms and asynchronous task execution proved equally valuable, enabling the platform to maintain low-latency responses and steady throughput even under stress. Collectively, these outcomes reinforced the

system’s readiness for deployment in environments where both precision and scalability are non-negotiable.

Scalability and Stress Testing.

At this stage, the system was evaluated under high-load conditions, with simultaneous processing of up to 500 requests per minute. Special attention was given to assessing the platform’s ability to maintain operational stability and responsiveness during periods of intensive user activity. The use of Celery for asynchronous job runs and Redis for caching data in an efficient manner played an important part in keeping the response times consistent and avoiding system crashes. The tests validated the maturity of the platform for industrial usage and mass implementation. The design provided an excellent ability to support multiple queries in parallel with no adverse impact on performance, hence confirming its viability for real-time application in scenarios with high user interactions and stringent reliability needs.

To validate the platform’s effectiveness in real-world scenarios, tests were conducted using actual datasets from education, healthcare, and human resource management domains.

The multilingual capabilities were put to the test with queries in Russian, English, and Kazakh, each chosen to reflect domain-specific language and context. Particular attention was paid to how well the system handled specialized terminology and the unique linguistic features present in each field. Following targeted fine-tuning on relevant language corpora, the platform consistently demonstrated high accuracy across all three languages. More importantly, it showed a strong ability to adjust to the semantic and contextual subtleties of different knowledge areas. These results underscored the system’s adaptability, positioning it as a viable, cross-domain solution for intelligent information retrieval in both single-language and multilingual settings.

References:

1. Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. arXiv preprint arXiv:1810.04805. https://arxiv.org/abs/1810.04805
2. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. Advances in Neural Information Processing Systems, 30, 5998–6008. https://papers.nips.cc/paper/7181-attention-is-all-you-need.pdf
3. Robertson, S., & Zaragoza, H. (2009). The Probabilistic Relevance Framework: BM25 and Beyond. Foundations and Trends in Information Retrieval, 3(4), 333–389. https://doi.org/10.1561/1500000019
4. StartXLabs. (2023). Mastering Django and Machine Learning: Building AI-Powered Web Apps. Medium. https://medium.com/@StartXLabs/mastering-django-and-machine-learning-building-ai-powered-web-apps-9038246fe23c
5. Skvortsov, N. (2023). Improving Information Retrieval with Django-BERT. Habr. https://habr.com/en/articles/769168/