# Semantic retrieval of Uzbek seismic safety regulations

**Farrukh Ishkobilov[1], Abdulatif Meyliev[2], Mironshoh Ortikov[3], Javlon Ibragimov[4], Umida Turaeva[5], Elmurod Jumaev[6]**
University of Information Technologies and Management, Karshi, 180212, Kurgan str, Uzbekistan
[2]Corresponding author
**E-mail:** [1]*farruxishqobilov@gmail.com*, [2]*abdulatif1909@gmail.com*, [3]*mironshoh5500@gmail.com*, [4]*boburdusyorov@gmail.com*, [5]*turayevaumida775@gmail.com*, [6]*elmurodjumayev272@gmail.com*

*Received 7 March 2026; accepted 11 April 2026; published online 16 July 2026*
*DOI https://doi.org/10.21595/vp.2026.26510*

Check for updates

*77th International Conference on Vibroengineering in Almaty, Kazakhstan, June 11-12, 2026*



**Abstract.** This study presents a semantic retrieval approach for Uzbek seismic safety regulations and engineering documents. The proposed system combines paragraph-level indexing with two retrieval models, a classical TF-IDF baseline and a FastText-based subword embedding model, to improve access to relevant regulatory and technical text units in a morphologically rich language environment. A specialized corpus was compiled from seismic engineering and safety documentation and segmented into paragraphs for fine-grained semantic search. The retrieval performance was evaluated on 100 domain-specific queries using Precision@5, Recall@5, Mean Average Precision (MAP), and Mean Reciprocal Rank (MRR). The experimental results show that the FastText model outperformed TF-IDF across all major metrics, achieving Precision@5 = 0.7444 and Recall@5 = 0.6875, compared with 0.6017 and 0.5418, respectively. The observed improvement was statistically significant according to paired t-test analysis ($t = 15.1372$, $p < 0.001$). The findings indicate that subword-based semantic modeling improves retrieval quality for Uzbek seismic safety documentation and can support faster access to relevant engineering regulations and compliance information.

**Keywords:** semantic retrieval, seismic safety regulations, FastText, TF-IDF, Uzbek language, engineering documentation.

## 1. Introduction

Recent advances in engineering digitalization have increased the volume of seismic safety regulations, construction standards, and technical compliance documents that must be accessed during design, assessment, and risk analysis procedures. In such environments, efficient retrieval of relevant normative provisions is essential for timely engineering decision-making and regulatory compliance. Classical information retrieval methods remain an important foundation for document search, but engineering practice increasingly requires more accurate retrieval of semantically related text fragments rather than simple keyword overlap [1], [2]. Seismic engineering studies commonly focus on structural safety, risk reduction, and resilience of infrastructure in seismic conditions, which further highlights the need for effective access to technical and regulatory documentation [7-9].

However, conventional keyword-based retrieval is often insufficient for Uzbek-language technical texts because Uzbek is a morphologically rich agglutinative language. Lexical variation caused by suffixation increases token dispersion and reduces the effectiveness of purely frequency-based matching. Although distributed word representations have substantially improved semantic modeling in information retrieval and text similarity tasks, the application of such methods to Uzbek seismic safety documentation remains limited [3], [5]. Among available embedding approaches, FastText is particularly suitable for agglutinative languages because it incorporates subword information and can preserve semantic similarity across morphologically related forms [4]. More advanced multilingual contextual models have also demonstrated strong potential for future semantic retrieval development [6], [10].

This study therefore proposes a semantic retrieval approach for Uzbek seismic safety regulations and engineering documents based on paragraph-level indexing and semantic similarity estimation. The framework compares a classical TF-IDF baseline with a FastText-based subword embedding model in order to evaluate whether subword-aware semantic representation improves retrieval quality in a domain-specific engineering corpus. The system is assessed on specialized seismic safety documentation using standard retrieval metrics and statistical validation, with the aim of supporting faster and more accurate access to relevant engineering regulations and compliance information [1], [2], [4], [5].

## 2. Methods

### 2.1. System overview

The proposed retrieval framework was designed as a modular semantic search system for Uzbek seismic safety documentation. The processing pipeline includes document collection, text cleaning, paragraph segmentation, vector representation, semantic similarity estimation, and ranking of the most relevant text units. This structure enables fine-grained retrieval of regulatory and technical passages rather than whole-document matching. The overall workflow of the proposed system is illustrated in Fig. 1.

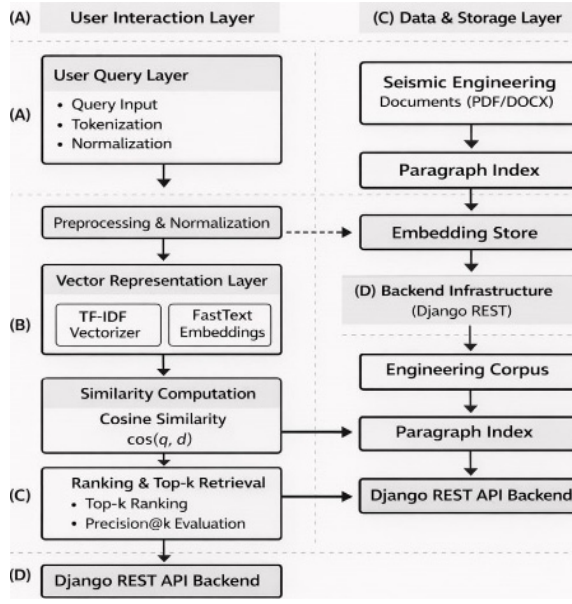


**Fig. 1.** General workflow of the semantic retrieval system

### 2.2. Corpus construction

A specialized domain corpus was compiled from seismic safety regulations, construction compliance documents, seismic risk mitigation guidelines, and related engineering publications. During preprocessing, technical documents were cleaned and segmented at the paragraph level to preserve coherent normative and technical units. Paragraph-level indexing was selected in order to reduce document-level noise and improve retrieval precision for engineering search scenarios. The main corpus characteristics are summarized in Table 1.

The paragraph-level structure of the corpus supports fine-grained retrieval of normative and technical text units and provides a practical basis for evaluating semantic ranking methods in Uzbek seismic documentation.

**Table 1.** Main characteristics of the corpus

| Parameter | Value |
|---|---|
| Number of documents | 120 |
| Indexed paragraphs | 8,450 |
| Total tokens | 777,400 |
| Vocabulary size | 18,000 |

### 2.3. Retrieval models

Two retrieval models were implemented and compared. The first model was a classical TF-IDF baseline representing paragraphs in a statistical vector space. The second model used FastText embeddings with subword information, which is particularly suitable for agglutinative languages because morphologically related forms can share representational components [3], [4]. For each paragraph, the final vector representation was computed as the arithmetic mean of the token vectors, as shown in Eq. (1):

$$V_p = \frac{1}{n}\sum_{i=1}^{n} V(w_i), \tag{1}$$

where $V(w_i)$ denotes the embedding of the $i$-th token and $n$ is the number of tokens in the paragraph.

### 2.4. Evaluation procedure

The system was evaluated on 100 domain-specific queries derived from seismic engineering and regulatory documentation. Retrieval effectiveness was assessed using Precision@5, Recall@5, Mean Average Precision (MAP), Mean Reciprocal Rank (MRR), and F1-score. Since both models were tested on the same query set, paired statistical testing was used to determine whether the observed performance differences were significant.

## 3. Results and discussion

The comparative retrieval results are presented in Table 2. The FastText model outperformed the TF-IDF baseline across all evaluation metrics. Precision@5 increased from 0.6017 to 0.7444, while Recall@5 improved from 0.5418 to 0.6875. Similar gains were observed for MAP and MRR, indicating more accurate and more stable ranking of relevant seismic safety paragraphs. These results confirm that subword-based semantic representation is more effective than purely lexical matching for Uzbek technical texts with high morphological variation [4], [5].

The statistical significance of the observed improvement was assessed using a paired $t$-test on 100 domain-specific queries. The obtained result, $t = 15.1372$ and $p < 0.001$, indicates that the performance difference between TF-IDF and FastText is statistically significant and cannot be explained by random variation alone. Therefore, the superiority of the FastText-based retrieval model is not only numerical but also statistically reliable.

**Table 2.** Comparison of TF-IDF and FastText retrieval performance

| Model | Precision@5 | Recall@5 | MAP | MRR | F1-score |
|---|---|---|---|---|---|
| TF-IDF | 0.6017 | 0.5418 | 0.573 | 0.596 | 0.570 |
| FastText | 0.7444 | 0.6875 | 0.712 | 0.731 | 0.714 |

From a practical engineering perspective, the results show that paragraph-level semantic retrieval can improve access to seismic safety regulations, compliance-related clauses, and technical requirements needed during engineering assessment and decision-making. In Uzbek seismic documentation, lexical variation caused by agglutinative word formation reduces the

effectiveness of surface-level matching, whereas FastText preserves semantic proximity across related forms through subword modeling [4]. This makes the retrieval process more robust for regulatory and engineering texts, where relevant information is often expressed through terminological variation rather than identical wording. In this respect, the proposed approach provides a useful intermediate solution between classical sparse retrieval methods [1], [2] and more advanced multilingual contextual models that may be considered in future work [6], [10].

## 4. Conclusions

This study presented a semantic retrieval approach for Uzbek seismic safety regulations and engineering documents based on paragraph-level indexing, TF-IDF modeling, and FastText subword embeddings. The proposed framework was developed to address the limitations of conventional keyword-based retrieval in morphologically rich technical texts, where lexical variation often reduces the effectiveness of surface-level matching. The experimental evaluation on 100 domain-specific queries demonstrated that the FastText model consistently outperformed the TF-IDF baseline across all major retrieval metrics, including Precision@5, Recall@5, MAP, MRR, and F1-score. The observed improvement was confirmed by paired $t$-test analysis ($t = 15.1372$, $p < 0.001$), indicating that the performance gain was statistically significant rather than incidental.

Beyond the quantitative improvement, the findings show that subword-aware semantic modeling is especially suitable for Uzbek seismic documentation, where relevant regulatory and technical information is frequently expressed through morphologically variable forms. The study also confirms the importance of paragraph-level indexing for engineering regulations, since such texts are naturally organized into semantically coherent normative units. From a practical perspective, the proposed system can support faster retrieval of safety clauses, compliance requirements, and technical provisions needed in seismic engineering workflows. Thus, the framework may serve as a useful decision-support component for digital engineering environments dealing with regulatory and safety documentation. Future work should focus on expanding the corpus, improving annotation robustness, and integrating contextual multilingual models such as XLM-R and mBERT to further enhance semantic retrieval quality.

## Acknowledgements

The authors have not disclosed any funding.

## Data availability

The datasets generated during and/or analyzed during the current study are available from the corresponding author on reasonable request.

## Conflict of interest

The authors declare that they have no conflict of interest.

## References

**[1]** G. Salton and C. Buckley, “Term-weighting approaches in automatic text retrieval,” *Information Processing and Management*, Vol. 24, No. 5, pp. 513–523, 1988, https://doi.org/10.1016/0306-4573(88)90021-0

**[2]** C. D. Manning, P. Raghavan, and H. Schütze, *Introduction to Information Retrieval*. Cambridge: Cambridge University Press, 2008, https://doi.org/10.1017/cbo9780511809071

**[3]** T. Mikolov, K. Chen, G. Corrado, and J. Dean, “Efficient estimation of word representations in vector space,” in *International Conference on Learning Representations*, 2013.

**[4]** P. Bojanowski, E. Grave, A. Joulin, and T. Mikolov, "Enriching word vectors with subword information," *Transactions of the Association for Computational Linguistics*, Vol. 5, pp. 135–146, Dec. 2017, https://doi.org/10.1162/tacl_a_00051

**[5]** T. Kenter and M. de Rijke, "Short text similarity with word embeddings," in *Proceedings of the 24th ACM International Conference on Information and Knowledge Management*, pp. 1411–1420, 2015, https://doi.org/10.1145/2806416.2806475

**[6]** A. Conneau et al., "Unsupervised cross-lingual representation learning at scale," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 8440–8451, Jan. 2020, https://doi.org/10.18653/v1/2020.acl-main.747

**[7]** B. Harsono, F. M. Mahadiva, Tavio, T. M. Mirkadirovich, and H. Hermawan, "Pushover analysis of simple house structures in seismic zones," in *AIP Conference Proceedings*, Vol. 3480, No. 1, p. 050035, Jan. 2026, https://doi.org/10.1063/5.0319164

**[8]** S. S. Sulaymanov, K. M. Nurmatov, and N. B. Chetkova, "Strengthening the capacity of infrastructure facilities of the Angren-Pap railway line to reduce seismic risk," in *AIP Conference Proceedings*, Vol. 2943, No. 1, p. 030056, 2023, https://doi.org/10.1063/5.0141972

**[9]** E. Shipacheva, S. Shaumarov, A. Gulamov, and M. M. Talipov, "Modeling of interaction of external enclosing structures of buildings with the internal and external environments," in *ICTEA: International Conference on Thermal Engineering*, Vol. 1, No. 1, 2024.

**[10]** N. Reimers and I. Gurevych, "Sentence-BERT: sentence embeddings using Siamese BERT-networks," in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 3980–3990, Jan. 2019, https://doi.org/10.18653/v1/d19-1410