IBIMA Publishing
Communications of International Proceedings
https://ibimapublishing.com/p-articles/46AI/2025/4646425/
Vol. 2025 (34), Article ID 4646425

Teaching AI to Read Consumer Law: Evaluating Modern Language Models for Semantic Search[1]*

Piotr POTIOPA

AGH University of Krakow
Faculty of Management, ul. Gramatyka 10, 30-067 Kraków, Poland
ORCID: 0000-0002-8176-7443

Correspondence should be addressed to: Piotr POTIOPA, ppotiopa@agh.edu.pl

* Presented at the 46th IBIMA International Conference, 26-27 November 2025, Ronda, Spain

Abstract

This paper compares four information retrieval methods: three semantic models (SBERT, E5-small, DistilRoBERTa-PL) and the classical lexical approach (BM25). The study is based on a corpus of Polish consumer law documents. Semantic models were used for embedding-based indexing and retrieval via FAISS, while the lexical approach used Elasticsearch with BM25. A question-answering (QA) system was built in two variants: (1) Semantic QA — retrieving legal text fragments based on meaning using sentence embeddings; (2) Lexical QA — traditional keyword-based retrieval. Both systems were evaluated on a custom set of 200 consumer law questions using standard IR metrics: precision@k, recall@k, Mean Reciprocal Rank (MRR), and Normalized Discounted Cumulative Gain (NDCG). Results show that semantic methods outperform BM25 in relevance (e.g., SBERT MRR@5 = 0.9194 vs BM25 = 0.5400), especially for queries requiring semantic understanding, with only a slight drop in absolute precision.

Keywords

Semantic search, legal information retrieval, language models, Sentence-BERT, BM25, consumer law, NLP, Polish law, FAISS, Elasticsearch

Introduction

Information retrieval in legal documents is essential to ensure access to applicable laws and regulations. Traditional keyword-based search systems (e.g., the BM25 model) match documents based on the exact occurrence of query terms in the text. While this approach is fast and proven, it has significant limitations — it does not "understand" word meanings and struggles with paraphrasing or synonyms. In the context of consumer law, this may mean that a colloquially formulated question is not matched to the appropriate legal provision if different terminology is used. For example, a query about a “product return” may fail to find the provision on withdrawal from a contract if there

[1] The APC was funded under subvention funds for the Faculty of Management and by program “Excellence Initiative - Research University” for the AGH University of Krakow

Cite this Article as: Piotr POTIOPA, Vol. 2025 (34) “Teaching AI to Read Consumer Law: Evaluating Modern Language Models for Semantic Search " Communications of International Proceedings, Vol. 2025 (34), Article ID 4646425, https://doi.org/10.5171/2025.4646425

is no direct keyword match. Recent advancements in machine learning — particularly transformer-based language models — enable semantic search, i.e., matching based on meaning rather than keywords. Models such as BERT and its derivatives can represent sentences as vectors (embeddings) in a semantic space, where semantically similar sentences lie close to each other, even if they do not share any common words (Devlin et al., 2019; Reimers & Gurevych, 2019). Reimers and Gurevych introduced Sentence-BERT (SBERT), a model that generates sentence embeddings that allow for fast and accurate semantic similarity comparison. This approach is successfully applied in web search engines, customer service chatbots, and legal document retrieval systems. In the legal domain in particular, language models have been shown to improve the relevance of legal document retrieval by understanding the context of legal queries (Sansone & Sperlí, 2022).

From a legal perspective, consumer law encompasses a set of rules regulating relations between businesses and consumers, particularly consumer protection, the right to information, the right of withdrawal from contracts, complaints, and alternative dispute resolution. This area has been largely shaped by European Union directives such as Directive~2011/83/EU on consumer rights, Directive~93/13/EEC on unfair terms in consumer contracts, and Directive~1999/44/EC on the sale of consumer goods and associated guarantees. In Poland, these provisions are implemented by, among others, the Act of 30 May 2014 on Consumer Rights and the Act of 23 September 2016 on Out-of-Court Consumer Dispute Resolution. These documents are publicly available on ISAP (https://isap.sejm.gov.pl/) and EUR-Lex (https://eur-lex.europa.eu).

It is also worth emphasizing the importance of the Polish Civil Code as a source of both general and detailed regulations applicable to consumer relations — in particular regarding warranty, contractual liability, forms of contract conclusion, and the entrepreneur's information obligations. These provisions, supplemented by case law and the practice of the Office of Competition and Consumer Protection (UOKiK), constitute the foundation of consumer legal protection in Poland (Federacja Konsumentów, 2018). Public legal databases such as ISAP and EUR-Lex provide access to up-to-date texts of statutes and directives that form the basis of this analysis.

Classical lexical methods such as BM25 still serve as a strong baseline in information retrieval tasks. However, recent studies (Wang et al., 2024b; Karpukhin et al., 2020) propose improved embedding models that surpass BM25. For example, the E5 model (Embeddings from bidirectional Encoder representations), trained contrastively on massive text corpora, was the first to outperform BM25 in zero-shot evaluation on the BEIR benchmark (Thakur et al., 2021). In open-domain QA tasks, dense representations (e.g., DPR) have also shown a significant advantage over BM25 in retrieving relevant text passages (Karpukhin et al., 2020). Simultaneously, models tailored for the Polish language have emerged — including Polish DistilRoBERTa, a distilled version of RoBERTa trained on a Polish corpus, which reduced the model size by half while preserving high embedding quality (Dadas, 2019).

In this study, four retrieval approaches were compared: three semantic models based on embeddings (SBERT (Reimers & Gurevych, 2019), E5-small (Wang et al., 2024b), and DistilRoBERTa-PL) and the classical lexical BM25 (Robertson & Zaragoza, 2009) implemented in Elasticsearch. The goal is to evaluate which approach performs best in the context of answering questions (QA) based on Polish legal texts related to consumer rights, particularly for queries requiring generalization or paraphrasing. This research contributes to the broader effort of modernizing access to legal information and improving its comprehensibility for citizens, as well as to the author's doctoral research on automating the semantic processing of legal texts.

Related Work

Natural Language Processing in the legal domain (Legal NLP) is a rapidly developing discipline focused on automating tasks such as information retrieval, contract analysis, court case outcome prediction, or data anonymization (Katz et al., 2023). Early systems relied mainly on rule-based approaches and statistical methods such as TF-IDF or BM25 (Robertson & Zaragoza, 2009), which, despite their simplicity and computational efficiency, have limitations in understanding the semantics of text (Potiopa et al., 2017).

The use of semantic representations for information retrieval is gaining increasing attention. SBERT (Sentence-BERT) has proven groundbreaking in sentence-level semantic matching tasks — it enables the generation of sentence embeddings that can be compared much more efficiently than raw BERT outputs, while maintaining high

semantic correlation. In the original SBERT paper, the model significantly improved performance in tasks such as information retrieval and textual similarity (Reimers & Gurevych, 2019). Follow-up studies expanded on this idea. Thakur et al. (Thakur et al., 2021) introduced the BEIR benchmark, which compared dozens of retrieval models in a zero-shot setup — models trained on large corpora (including various BERT/SBERT variants) often outperformed BM25 in tasks requiring semantic matching (Thakur et al., 2021). Wang et al. (Wang et al., 2024b) later proposed the E5 model family, trained contrastively on large-scale text pairs, achieving state-of-the-art results in unsupervised retrieval tasks — the E5-small model outperformed BM25 on many BEIR datasets (Wang et al., 2024b). E5 models are characterized by strong, universal sentence embeddings, making them promising for QA applications.

In the field of open-domain question answering (Open-Domain QA), the superiority of dense retrieval approaches over lexical ones has also been demonstrated. Karpukhin et al. introduced the DPR (Dense Passage Retrieval) model, in which a dual BERT encoder was trained to retrieve text passages matching the question — achieving a 9–19% higher top-20 answer retrieval performance compared to BM25 (Karpukhin et al., 2020). These results confirm that embedding-based models are capable of finding answers hidden behind paraphrased formulations, where keyword-based methods often fail. The legal domain has also seen attempts to apply transformers — for instance, in Polish, models such as HerBERT and Polish RoBERTa were trained on large corpora of legal and general texts. DistilRoBERTa-PL, developed by Dadas, is a compact version of the RoBERTa model (about 50% of the parameters), obtained through knowledge distillation (Dadas, 2019). Despite its smaller size, the model achieves performance comparable to the full-size RoBERTa on many NLP tasks (KLEJ benchmark (Rybak et al., 2020)), and its lower complexity enables faster retrieval. In this study, the SBERT, E5-small, and DistilRoBERTa-PL models were used to build vector-based semantic QA search engines. The classical BM25 model (Robertson & Zaragoza, 2009), implemented in the Elasticsearch engine and commonly used in search systems, served as the baseline. For vector indexing, the FAISS (Facebook AI Similarity Search) library was employed to enable fast retrieval from embedding collections (Reimers & Gurevych, 2019).

There are also domain-specific models designed for the legal field, such as LEGAL-BERT (Chalkidis et al., 2020), which are trained on corpora of legal texts, potentially improving their effectiveness in this domain. In the context of the Polish language, important work includes the development of Polish language models and benchmarks such as KLEJ (Rybak et al., 2020), as well as more recent initiatives like LEPISZCZE (Augustyniak et al., 2022) and PolQA (Rybak et al., 2022), which provide tools and datasets for evaluating NLP models in Polish. Noteworthy efforts also include the PL-MTEB benchmark (Poświata et al., 2024) and PIRB (Dadas et al., 2024), which assess embedding models for Polish, including in retrieval tasks and specifically within Polish legal texts. Valuable contributions also come from the development of Polish LLMs such as Bielik (Ociepa et al., 2024) and PLLuM (Kocoń et al., 2025), which the author also intends to evaluate in the context of legal text processing tasks. Although these models show strong potential, they were not included in the present study due to its focus on sentence-level embedding tasks, for which they are not yet optimized or readily available in a sentence encoding configuration. This makes direct comparison with established sentence embedding models challenging.

Information retrieval systems increasingly adopt hybrid techniques that combine the advantages of lexical search (e.g., BM25) with semantic approaches (vector-based models) (Li et al., 2024). Embedding vector indexing and search are efficiently handled using libraries such as FAISS (Johnson et al., 2021) or lexical search engines like Elasticsearch with Approximate Nearest Neighbor (ANN) search functionality (Elasticsearch, n.d.-a). Despite advancements, semantic search in specialized, narrow domains such as consumer law within a specific legal system (e.g., Polish law) remains an open research area, especially regarding the evaluation of off-the-shelf multilingual models compared to traditional methods, as well as the identification of best practices for their deployment.

Research Methodology

A. Data Collection and Preparation

The foundation of the study in terms of data was a corpus of Polish consumer law acts. A total of 11 statutes were used, including:

- The Act of 30 May 2014 on Consumer Rights (Ustawa o prawach konsumenta, 2014).

- The Act of 12 May 2011 on Consumer Credit (Ustawa o kredycie konsumenckim, 2011).
- The Act of 16 September 2011 on Timeshare (Ustawa o timeshare, 2011).
- The Act of 23 August 2007 on Counteracting Unfair Market Practices (Ustawa o przeciwdziałaniu nieuczciwym praktykom rynkowym, 2007).
- The Act of 16 February 2007 on Competition and Consumer Protection (Ustawa o ochronie konkurencji i konsumentów, 2007).

The texts of these legal acts were obtained from the ISAP (Internet System of Legal Acts). In addition, selected European Union legal acts related to consumer protection were included (a total of 12 EU directives), including Directive 2011/83/EU (on consumer rights) and Directive 2005/29/EC (on unfair commercial practices) (Directive 2011/83/EU, 2011; Directive 2005/29/EC, 2005). Due to the input sequence length limitation of the SBERT model (128 tokens for paraphrase-multilingual-MiniLM-L12-v2), the documents were split into smaller fragments (chunks). A splitting strategy based on articles was applied, and for longer articles — on sections or points. When such fragments still exceeded the token limit, further splitting into sentences with slight overlap (one sentence) was used to preserve context. In total, more than 1,000 fragments were prepared. Each fragment was assigned metadata identifying the source legal act and the article number — as a structural unit.

B. Modeling and Vector Representation

The semantic approach employed the SBERT model paraphrase-multilingual-MiniLM-L12-v2 from the Sentence Transformers library (Reimers & Gurevych, 2019; Sentence Transformers, n.d.). This model generates 384-dimensional embedding vectors for each legal text fragment. The choice of this particular model was motivated by its multilingual capabilities (important for potential extensibility to other EU languages) and solid performance in paraphrasing and semantic similarity tasks, while maintaining a relatively small model size. However, it should be noted that newer, specialized models for the Polish language or larger multilingual models may offer better performance, as indicated in the PL-MTEB benchmark (Poświata et al., 2024), where this model was not among the top performers. In this study, it serves as an accessible and widely adopted comparative baseline.

Additionally, the experiments included two other semantic models: the multilingual E5-small model (intfloat/multilingual-e5-small) and the Polish DistilRoBERTa-PL model (available, among others, as sdadas/st-polish-paraphrase on the Hugging Face platform). The E5 model (Wang et al., 2024a) is a state-of-the-art text embedding model trained contrastively on over a billion sentence pairs (in multiple languages, including Polish) using instruction tuning, which enables it to achieve outstanding results in information retrieval tasks. DistilRoBERTa-PL, on the other hand, is a distilled version of the RoBERTa model adapted to the Polish language and trained for tasks such as paraphrase detection (this model was evaluated, for instance, within the PIRB benchmark (Dadas et al., 2024)). The embeddings of legal text fragments generated by the above models were used for retrieval via the FAISS index, analogously to the SBERT setup.

To enable fast search over large document collections, the FAISS library (Facebook AI Similarity Search) (Johnson et al., 2021) was used, which is based on Approximate Nearest Neighbor (ANN) search methods. For comparison, the traditional BM25 retrieval method was also applied, defined by the following formula:

$$\mathrm{BM25}(q,d) = \sum_{i=1}^{n} \mathrm{IDF}(q_i) \frac{f(q_i,d)(k_1+1)}{f(q_i,d) + k_1\left(1 - b + b\frac{|d|}{\mathrm{avgdl}}\right)} \qquad (1)$$

where q is the query, d is the document, q_i is the i-th term in the query, IDF(q_i) is the inverse document frequency of term q_i, f(q_i,d) is the frequency of term q_i in document d, $|d|$ is the length of document d, $avgdl$ is the average document length in the collection, and k_1 and b are tuning parameters (typically $k_1 \in [1.2, 2.0]$, $b = 0.75$) (Robertson & Zaragoza, 2009; Elasticsearch, n.d.-b).

This formula allows for evaluating the quality of results in the classical lexical approach, while the vector-based method relies on similarity measures in vector space, such as cosine similarity. Embeddings of legal text fragments were indexed using FAISS by creating an *IndexFlatL2* index (exact nearest neighbor search using L2 distance, which is equivalent to maximizing cosine similarity for normalized vectors). For queries, they were also converted into embedding vectors using the same model and then compared with the vectors in the FAISS index to find the *k* most semantically similar fragments.

The simplified Python example in Listing 3.1 illustrates the process of building a SBERT+FAISS index and retrieving answers for a query.

```
from sentence_transformers import SentenceTransformer
import faiss

# Loading the SBERT model
model = SentenceTransformer('paraphrase-multilingual-MiniLM-L12-v2')

# Preparing embeddings for fragments (sample 'documents' list)
doc_embeddings = model.encode(documents, normalize_embeddings=True)

# Creating a vector index (L2 corresponds to normalized cos)
index = faiss.IndexFlatL2(doc_embeddings.shape[1])
index.add(doc_embeddings)

# Find the 5 closest chunks for an example question
query = "Czy mogę zwrócić produkt kupiony na odległość bez podania powodu?"
query_vec = model.encode([query], normalize_embeddings=True)
D, I = index.search(query_vec, k=5)
print("Most relevant acts:", I[0])
```

Listing 3.1 Simplified example of building an SBERT + FAISS index and retrieving answers

C. BM25 Baseline System

As a baseline system, BM25-based retrieval was implemented using the Elasticsearch engine (version 7.17). Legal text fragments were indexed in Elasticsearch, and standard match queries were used for retrieval with the default BM25 parameters provided by Elasticsearch (corresponding to $k_1 = 1.2$ and $b = 0.75$) (Elasticsearch, n.d.-b).

Results and Analysis

Dataset: A set of 200 questions was prepared to evaluate the systems. The questions were designed to require retrieving relevant articles from consumer law, including the Consumer Rights Act (2014) (Ustawa o prawach konsumenta, 2014), the Consumer Credit Act (2011) (Ustawa o kredycie konsumenckim, 2011), and the Timeshare Act (2011) (Ustawa o timeshare, 2011). For each question, fragments of legal acts containing the answer were manually identified (up to 5 articles per question), forming the ground truth for evaluation.

Evaluation methods: Each system was evaluated by comparing the top-5 results list against a set of reference answers (relevant fragments) for each question. Classic IR metrics were computed for the rankings: *precision@k* – the proportion of relevant results among the top-*k*, *recall@k* – the proportion of all relevant results retrieved within the top-k, *MRR* (Mean Reciprocal Rank) – the average reciprocal rank of the first correct answer, and *NDCG@k* (Normalized Discounted Cumulative Gain) – the normalized cumulative gain that accounts for the position of all relevant results up to rank *k* (Manning et al., 2008; Carnevali, 2023):

- Precision@k (P@k): the proportion of relevant documents among the top *k* retrieved results.

- Recall@k (R@k): the proportion of retrieved relevant documents out of all relevant documents for a given query, considering the top *k* results.
- Mean Reciprocal Rank (MRR): the average reciprocal rank of the first relevant result.

$$\mathrm{MRR} = \frac{1}{|Q|}\sum_{i=1}^{|Q|}\frac{1}{\mathrm{rank}_i} \tag{2}$$

where $|Q|$ is the number of queries, and $rank_i$ is the rank position of the first relevant document for the i-th query.

- Normalized Discounted Cumulative Gain (NDCG@k): a metric that accounts for both the relevance and ranking position of results.

$$\mathrm{NDCG@k} = \frac{\mathrm{DCG@k}}{\mathrm{IDCG@k}}, \quad \mathrm{DCG@k} = \sum_{j=1}^{k}\frac{\mathrm{rel}_j}{\log_2(j+1)} \tag{3}$$

where rel_j is the relevance score of the j-th result (binary in this study: 1 for relevant, 0 for non-relevant), and IDCG@k is the ideal DCG@k score for perfectly ranked results.

The comparison was performed for $k \in \{3, 5\}$.

Experiments and Results

Experimental results: All three semantic models clearly outperform BM25 in terms of retrieval quality. Table I presents the metric values obtained for each method (for k=3 and k=5). The SBERT model achieved the highest effectiveness — it reached the highest average precision and recall, as well as the top MRR and NDCG scores. The DistilRoBERTa-PL model ranked second, slightly behind SBERT, while the E5-small model placed third. The classic BM25 approach shows significantly lower effectiveness — its precision and recall are more than twice as low as those of the best semantic models, and the MRR score (0.54 in top-5) indicates that the correct answer rarely appears at the top of the BM25 ranking. The results also show that semantic methods are capable of retrieving a greater number of relevant fragments — for example, SBERT returned nearly 3 relevant answers in the top-5 on average (recall@5 = 0.7267), while BM25 returned only about 1.3 (recall@5 = 0.26).

The results for the proposed metrics are presented in Table I. Additionally, Fig. 1 provides a graphical comparison of the performance of all systems.

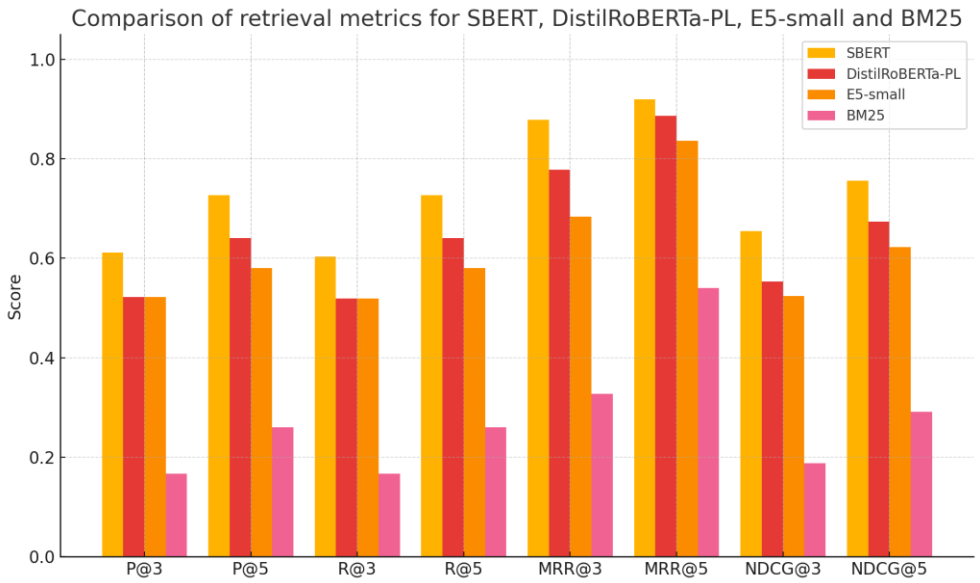

Fig. 1. Comparison of retrieval effectiveness metrics for four methods: semantic models (SBERT, E5-small, DistilRoBERTa-PL) and BM25. Metrics include precision@3, recall@3, MRR@3, NDCG@3 as well as precision@5, recall@5, MRR@5, NDCG@5 (higher values indicate better performance).

The chart in Fig. 1 illustrates the advantage of embedding-based methods over the lexical approach. For every quality metric, the SBERT, E5, and DistilRoBERTa-PL models achieve higher scores than BM25. The largest differences are observed in the MRR metric — SBERT reached a value of 0.92 (top-5), while BM25 only 0.54 — indicating that the correct answer was much more frequently placed at the top of the results list. The NDCG metric also performs significantly better for semantic models (e.g., NDCG@5 = 0.7560 for SBERT vs 0.2909 for BM25), highlighting more effective ranking of relevant fragments. A clear increase in precision is also visible — SBERT achieves a precision@5 of 0.7267, while BM25 reaches only 0.2600. This indicates a greater number of relevant answers among the top-ranked results for embedding-based models. Although they may occasionally return slightly less relevant documents, they provide substantially better overall coverage of correct responses.

When comparing the semantic models themselves, SBERT achieves the highest scores across all metrics, indicating better alignment with query language and higher semantic stability. The DistilRoBERTa-PL model scored slightly lower in precision and recall, but exhibits a high MRR, suggesting that it frequently places the correct answer very high in the ranking despite greater variance in results. E5-small shows generally good performance but noticeably lower NDCG, which may point to difficulty in properly ordering multiple relevant documents. These differences emphasize the influence of training approaches — SBERT, trained on paraphrase and NLI data, handles query meaning recognition more effectively, whereas contrastively trained models (like E5) may require additional fine-tuning for the legal domain. These findings suggest that the model choice should depend on the system's objective — if placing the correct answer at the very top is crucial, SBERT appears to be the most suitable option.

Table I. Comparison of the effectiveness of SBERT+FAISS, E5-small+FAISS, DistilRoBERTa-PL+FAISS, and BM25+Elasticsearch systems

Metrics	SBERT+FAISS	DistilRoBERTa-PL+FAISS	E5+FAISS	BM25+Elasticsearch
P@3	**0.6111**	0.5222	0.5222	0.1667
P@5	**0.7267**	0.6400	0.5800	0.2600

R@3	0.6028	0.5194	0.5194	0.1667
R@5	0.7267	0.6400	0.5800	0.2600
MRR@3	0.8778	0.7778	0.6833	0.3278
MRR@5	0.9194	0.8861	0.8361	0.5400
NDCG@3	0.6544	0.5527	0.5235	0.1881
NDCG@5	0.7560	0.6738	0.6221	0.2909

A qualitative analysis of the answers showed that the advantage of embedding-based models was particularly evident in the case of questions containing synonyms, hyponyms, hypernyms, or requiring an understanding of broader context that was not directly expressed through keywords. For example, in response to the question "*Czy mogę zwrócić produkt kupiony na odległość bez podania powodu?*", semantic models correctly identified fragments related to the right of withdrawal from a distance contract, even if the phrasing in the legal act was different. In such cases, BM25 often returned fragments containing the words "*product*"', "*kupić*", "*powód*", but not necessarily addressing the core of the question.

At the same time, it was observed that BM25 could be effective for highly precise questions containing unique legal terms that clearly identified the sought fragment. In some cases, SBERT could semantically "drift" if the query was very short or ambiguous.

Conclusions

The conducted study demonstrated that the use of language models such as Sentence-BERT (SBERT), DistilRoBERTa-PL, and E5 combined with FAISS vector indexing significantly improves the effectiveness of information retrieval in Polish consumer law documents compared to the classical BM25-based approach. The semantic system achieved higher values across all applied evaluation metrics, including P@k, R@k, MRR, and NDCG@k. The key advantage of the semantic approach lies in its ability to understand the meaning of queries and documents, allowing it to retrieve relevant information even when there is no exact keyword match. This is particularly valuable in the legal domain, which is characterized by specific terminology and complex formulations rarely used by laypersons searching for legal provisions.

Despite the promising results, some limitations should be acknowledged. The effectiveness of the SBERT model *paraphrase-multilingual-MiniLM-L12-v2*, although achieving the best results in the tests, may still be lower than that of newer, larger multilingual models or those specifically trained on Polish legal data. The strategy for splitting documents into fragments also has a significant impact on the results and requires further research and testing. Preparing a high-quality set of test questions and reference answers is time-consuming, but it is crucial for a reliable evaluation and should be a primary focus.

Future research directions should include testing more advanced language models (e.g., larger E5-family models or Polish models such as HerBERT (Mroczkowski et al., 2021) or *mmlw-roberta*), exploring hybrid techniques that combine semantic and lexical search (Li et al., 2024), and extending the system with capabilities for generating natural language answers (generative QA) instead of merely retrieving fragments of text. It is also worth investigating the impact of fine-tuning the models on a domain-specific corpus of Polish consumer law.

Acknowledgments

The author would like to thank the anonymous reviewers for their valuable comments, which helped improve the quality of this article. The author also acknowledges the contributions of developers and researchers who share their pretrained models with the community.

References

- Augustyniak, Ł., et al. (2022). This is the way: Designing and compiling LEPISZCZE, a comprehensive NLP benchmark for Polish. In *Advances in Neural Information Processing Systems 35 (NeurIPS 2022) Datasets and Benchmarks Track* (pp. 506–526). https://doi.org/10.48550/arXiv.2211.13112
- Carnevali, L. (2023, June). Evaluation measures in information retrieval. *Pinecone*. Retrieved May 20, 2025, from https://www.pinecone.io/learn/offline-evaluation/
- Chalkidis, I., Kouki, M., Passos, A., Tan, M., Boella, A. P. L., & Malik, S. (2020). LEGAL-BERT: The Muppets straight out of Law School. In *Findings of the Association for Computational Linguistics: EMNLP 2020* (pp. 2898–2904). https://doi.org/10.18653/v1/2020.findings-emnlp.261
- Dadas, S. (2019). *A repository of Polish NLP resources* [GitHub repository]. GitHub. https://github.com/sdadas/polish-nlp-resources
- Dadas, S., Perełkiewicz, M., & Poświata, R. (2024). PIRB: A comprehensive benchmark of Polish dense and hybrid text retrieval methods. *arXiv*. https://doi.org/10.48550/arXiv.2402.13350
- Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)* (Vol. 1, pp. 4171–4186). https://doi.org/10.18653/v1/N19-1423
- Directive 2005/29/EC of the European Parliament and of the Council of 11 May 2005 concerning unfair business-to-consumer commercial practices. (2005). *Official Journal of the European Union*, L 149/22. https://eur-lex.europa.eu/eli/dir/2005/29/oj/eng
- Directive 2011/83/EU of the European Parliament and of the Council of 25 October 2011 on consumer rights. (2011). *Official Journal of the European Union*, L 304/64. https://eur-lex.europa.eu/eli/dir/2011/83/oj/eng
- Elasticsearch. (n.d.-a). *K-nearest neighbor (kNN) search*. Retrieved May 20, 2025, from https://www.elastic.co/guide/en/elasticsearch/-reference/current/knn-search.html
- Elasticsearch. (n.d.-b). *BM25 similarity*. Retrieved May 20, 2025, from https://www.elastic.co/guide/en/elasticsearch/reference/current/index-modules-similarity.html#bm25
- Federacja Konsumentów. (2018). Odstąpienie od umowy – jak to zrobić? Retrieved May 20, 2025, from https://www.federacja-konsumentow.org.pl/n,45,1268,12,1,odstapienie-od-umowy%E2%80%93jak-to-zrobic.html
- Johnson, J., Douze, M., & Jégou, H. (2021). "Billion-Scale Similarity Search with GPUs," in *IEEE Transactions on Big Data*, vol. 7, no. 3, pp. 535-547, 1 July 2021, doi: https://doi.org/10.1109/TBDATA.2019.2921572.
- Karpukhin, V., et al. (2020). Dense passage retrieval for open-domain question answering. In *Proceedings of EMNLP 2020*. https://doi.org/10.18653/v1/2020.emnlp-main.550
- Katz, D. M., Hartung, D., Gerlach, L., Jana, A., & Bommarito II, M. J. (2023). Natural language processing in the legal domain. *arXiv*. https://doi.org/10.48550/arXiv.2302.12039
- Kocoń, J., Piasecki, M., Janz, A., Ferdinan, T., Radliński, Ł., Koptyra, B., Oleksy, M., Woźniak, S., Walkowiak, P., Wojtasik, K., Moska, J., Naskręt, T., Walkowiak, B., Gniewkowski, M., Szyc, K., Motyka, D., Banach, D., Dalasiński, J., Rudnicka, E., … Pęzik, P. (2025). *PLLuM: A family of Polish large language models*. arXiv. https://arxiv.org/abs/2511.03823.
- Li, X., Lipp, J., Shakir, A., Huang, R., & Li, J. (2024). BMX: Entropy-weighted similarity and semantic-enhanced lexical search. *arXiv*. https://doi.org/10.48550/arXiv.2408.06643
- Manning, C. D., Raghavan, P., & Schütze, H. (2008). Introduction to information retrieval. Cambridge University Press. https://doi.org/10.1017/CBO9780511809071
- Mroczkowski, R., Rybak, P., Wróblewska, A., & Gawlik, I. (2021). HerBERT: Efficiently pretrained transformer-based language model for Polish. In B. Babych, O. Kanishcheva, P. Nakov, J. Piskorski, L. Pivovarova, V. Starko, J. Steinberger, R. Yangarber, M. Marcińczuk, & S. Pollak (Eds.), *Proceedings of the 8th Workshop on Balto-Slavic Natural Language Processing* (pp. 1–10). Association for Computational Linguistics. https://aclanthology.org/2021.bsnlp-1.1/
- Ociepa, K., Flis, Ł., Wróbel, K., Gwoździej, A., & Kinas, R. (2024). Bielik 7B v0.1: A Polish language model – development, insights, and evaluation. *arXiv*. https://doi.org/10.48550/arXiv.2410.18565
- Potiopa, P., Karwatowski, M., Duda, J., Sasor, P., Wielgosz, M., & Muzykiewicz, B. (2017). Semantic search extension based on Polish WordNet relations in business document exploration. In *Proceedings of the International Conference on Internet of Things and Machine Learning*. https://doi.org/10.1145/3109761.3158401

- Poświata, R., Dadas, S., & Perełkiewicz, M. (2024). PL-MTEB: Polish massive text embedding benchmark. *arXiv*. https://doi.org/10.48550/arXiv.2405.10138
- Reimers, N., & Gurevych, I. (2019). 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. 3982–3992). https://doi.org/10.18653/v1/D19-1410
- 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
- Rybak, P., Mroczkowski, R., Tracz, J., & Gawlik, I. (2020). KLEJ: Comprehensive Benchmark for Polish Language Understanding. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (pp. 1191–1201, Online). Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.acl-main.111
- Rybak, P., Przybyła, P., Ogrodniczuk, M., et al. (2022). PolQA: A Polish question answering dataset. *arXiv*. https://doi.org/10.48550/arXiv.2212.08897
- Sansone, C., & Sperlí, G. (2022). Legal information retrieval systems: State-of-the-art and open issues. *Information Systems, 106*, 101967. https://doi.org/10.1016/j.is.2021.101967
- Sentence Transformers. (n.d.). *Pretrained models*. Retrieved May 20, 2025, from https://www.sbert.net/docs/pretrainedmodels.html
- Thakur, N., et al. (2021). BEIR: A heterogeneous benchmark for zero-shot evaluation of information retrieval models. *NeurIPS Datasets and Benchmarks Track*. https://doi.org/10.48550/arXiv.2104.08663
- Ustawa z dnia 16 lutego 2007 r. o ochronie konkurencji i konsumentów. (2007). Dz.U. 2007 Nr 50 poz. 331. https://isap.sejm.gov.pl/isap.nsf/DocDetails.xsp?id=wdu20070500331
- Ustawa z dnia 12 maja 2011 r. o kredycie konsumenckim. (2011). Dz.U. 2011 Nr 126 poz. 715. https://isap.sejm.gov.pl/isap.nsf/DocDetails.xsp?id=WDU20111260715
- Ustawa z dnia 23 sierpnia 2007 r. o przeciwdziałaniu nieuczciwym praktykom rynkowym. (2007). Dz.U. 2007 Nr 171 poz. 1206. https://isap.sejm.gov.pl/isap.nsf/DocDetails.xsp?id=wdu20071711206
- Ustawa z dnia 30 maja 2014 r. o prawach konsumenta. (2014). Dz.U. 2014 poz. 827. https://isap.sejm.gov.pl/isap.nsf/-DocDetails.xsp?id=wdu20140000827
- Ustawa z dnia 16 września 2011 r. o timeshare. (2011). Dz.U. 2011 Nr 230 poz. 1370. https://isap.sejm.gov.pl/isap.nsf/DocDetails.xsp?id=WDU20112301370
- Wang, L., Yang, N., Huang, X., Yang, L., Majumder, R., & Wei, F. (2024a). Multilingual E5 text embeddings: A technical report. *arXiv*. https://doi.org/10.48550/arXiv.2402.05672
- Wang, L., et al. (2024b). Text embeddings by weakly-supervised contrastive pretraining. https://doi.org/10.48550/arXiv.2212.03533