Evaluating Embedding Models and Aggregation Strategies for Court Decision Representation in a Heterogeneous Legal Graph

Samuel Baran[1,*], Martin Bača[1] and Šimon Horvát[1]

[1]Institute of Computer Science, Faculty of Science, Pavol Jozef Šafárik University in Košice, Jesenná 5, 040 01 Košice, Slovakia

Abstract

Constructing effective and comparable document representations from the raw textual data of court decisions remains a challenging task, particularly because of the length of the documents and also the varying importance of the information. The court decisions can be represented as a heterogeneous graph that captures relationships between decisions and associated legal concepts such as keywords and law references. This paper investigates the impact of transformer-based embedding models and aggregation strategies on the construction of court decision representations. The resulting representations are evaluated in a downstream classification task, complemented by intrinsic analysis. The results show that modern multilingual embedding models achieve highly similar performance in the downstream classification task, while the domain-specific model underperforms. Simple aggregation strategies provide comparable results, whereas weighted variants consistently yield lower performance and concatenation-based approaches offer only marginal differences. Intrinsic evaluation further indicates that different aggregation choices affect the structure of the embedding space, although these improvements do not consistently translate into downstream performance gains. Overall, the study highlights the robustness of modern embedding models and the limited impact of more complex aggregation strategies in the evaluated setting.

Keywords

Legal NLP, court decision representation, heterogeneous graph, document embeddings, legal references, aggregation strategies

1. Introduction

The growing volume of court decisions has increased the demand for automated methods for legal document analysis and categorization. A key factor influencing the effectiveness of such methods is the quality of document representations used in downstream tasks. Constructing informative representations of court decisions is, however, challenging due to the length of legal documents and the heterogeneous distribution of relevant information throughout the text.

One approach to addressing this challenge is to represent court decisions through selected document attributes rather than their complete textual content. In the adopted tripartite graph structure, court decisions (CD) are connected to legal references (LR) and keywords (KW), enabling document representations to be derived from their local graph neighbourhood. Such a representation provides a structured view of the legal corpus while reducing the amount of information that must be processed directly from the original documents.

Despite the simplicity of this approach, several important design choices remain open. In particular, the quality of court decision representations may depend on the selected embedding model, the strategy used to aggregate neighbouring node representations, and the incorporation of edge weights during aggregation. Understanding the impact of these factors is essential for constructing effective graph-based representations of legal documents.

Code and supplementary materials are available at: https://gitlab.science.upjs.sk/martin_baca/cd_3part_graph_embeddings.

ITAT'26: Information Technologies – Applications and Theory, September 25–29, 2026, Biele Karpaty, Slovakia

*Corresponding author.

samuel.baran@student.upjs.sk (S. Baran); martin.baca@student.upjs.sk (M. Bača); simon.horvat@upjs.sk (Š. Horvát)

0009-0005-8559-5422 (S. Baran); 0009-0005-8120-0948 (M. Bača); 0000-0002-3191-8469 (Š. Horvát)

© 2026 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

CEUR Workshop Proceedings ceur-ws.org ISSN 1613-0073

published 2026-11-09

CEUR-WS.org/Vol-4323/paper23.pdf

This paper investigates the influence of embedding models, aggregation strategies, and edge weighting on the quality of court decision representations derived from a tripartite graph of court decisions, legal references, and keywords. The resulting representations are evaluated using both intrinsic and downstream evaluation tasks, providing insights into the effectiveness of different design choices.

2. Related Work

Legal natural language processing (Legal NLP) has become an increasingly active research area, covering tasks such as legal document summarization, named entity recognition, question answering, argument mining, text classification, semantic search, and judgment prediction. Recent surveys emphasize that legal texts pose specific challenges for NLP systems, including long document length, complex domain-specific language, jurisdiction-dependent terminology, and the limited availability of open legal datasets [1]. These characteristics make the construction of effective legal document representations a non-trivial problem, especially in settings where documents need to be compared, retrieved, or classified at scale.

A substantial part of previous work has focused on adapting transformer-based models to legal document classification. BERT-style models have achieved strong results in many NLP tasks, but their direct application to legal documents is limited by the maximum input length of standard transformer architectures. Since legal documents often exceed this limit, simple truncation may remove important information and reduce classification performance. Limsopatham showed that BERT needs to be explicitly adapted for long legal documents and that domain-specific pre-training can improve performance in legal classification tasks [2]. Similar motivation appears in work on long-document legal encoders and legal-domain language models, where the main goal is to better capture relevant information distributed across lengthy legal texts.

Another line of research studies sentence and document embeddings for legal semantic representation. Sentence-BERT introduced an efficient way to obtain semantically meaningful sentence embeddings suitable for similarity search, clustering, and retrieval [3]. In the legal domain, this direction has been further explored through domain-adapted models such as Legal-BERT and Legal-SBERT, which aim to capture the specific meaning of legal terminology and improve downstream legal NLP tasks [4, 5]. Recent work also shows the practical relevance of multilingual embedding models in Central European legal settings. Csányi et al. evaluated multiple embedding models for semantic search over Hungarian court decisions and identified models such as *bge-m3* and *multilingual-e5-large* among the strongest candidates [6]. These findings are particularly relevant for low-resource and morphologically rich languages, where multilingual models may outperform or complement monolingual and domain-specific alternatives. At the same time, neural embeddings do not universally outperform traditional text representations. For example, Costa et al. reported that TF-IDF outperformed BERT-based embeddings on a Brazilian legal text classification task, which motivates empirical comparison rather than assuming the superiority of any single representation family [7].

Graph-based representations form another relevant direction for modeling textual and legal information. Earlier work on graph-based text classification represented documents as graphs of terms and showed that graph-derived features can outperform frequency-based representations in some classification settings [8]. In the legal domain, graph structures naturally arise from citations, references, argumentative relations, and links between legal entities. Recent work on legal argument mining has represented legal documents as graph structures and applied graph neural networks, graph augmentation, and collective classification to improve the identification and classification of argumentative text segments [9]. More generally, graph representation learning methods such as Graph Convolutional Networks and GraphSAGE provide a foundation for constructing node representations by combining node features with information from their neighborhoods [10, 11]. These methods are closely related to the idea of representing a target node through the aggregation of neighboring node information.

Our work is positioned at the intersection of these research directions. Unlike approaches that embed entire court decisions directly, we construct court decision representations from semantically enriched attributes extracted from an existing heterogeneous legal graph. The graph connects court decisions

to legal references and keywords, which are first embedded using multilingual, Slovak-specific, and legal-domain language models. Court decision embeddings are then obtained through simple and weighted pooling, concatenation of legal reference and keyword information, and principal component removal. The latter is related to post-processing methods that remove dominant common directions from embedding spaces, such as the all-but-the-top approach [12]. However, we do not apply the full SIF sentence embedding pipeline; instead, principal component removal is treated as a separate post-processing variant.

While previous work has studied legal document classification, long-document transformers, legal sentence embeddings, and graph-based legal NLP, less attention has been paid to the systematic evaluation of simple aggregation strategies for constructing court decision embeddings from attributes in an existing heterogeneous legal graph. This paper addresses this gap by comparing multiple embedding models and aggregation strategies in a supervised classification task over Slovak court decisions.

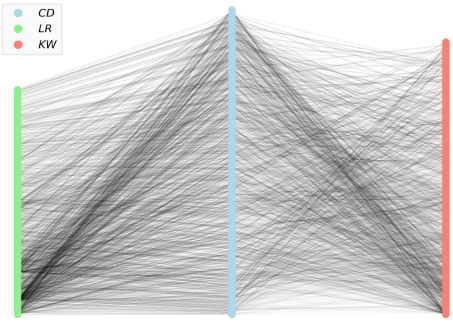
3. Methodology

This section outlines the methodology for constructing court decision representations and evaluating them through the downstream classification task described in Section 4.

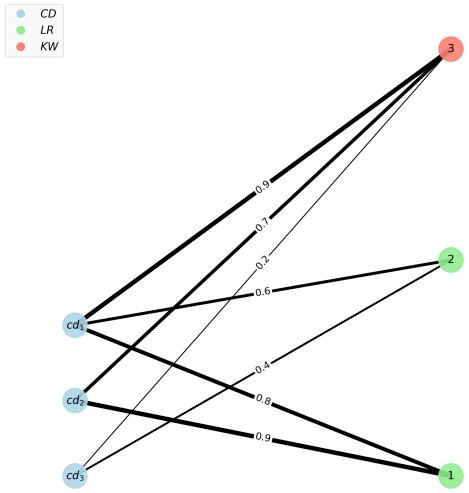
The methodology is based on a heterogeneous graph representation of court decisions and consists of three main components: dataset construction, node representation using embedding models, and aggregation of neighboring nodes embeddings into court decision representations.

3.1. Dataset

The experiments were conducted on a dataset of Slovak court decisions represented within a tripartite graph structure (see Figure 1.a) adopted from the work of Horvát et al. [13] The graph consists of three node types: court decisions (CD), legal references (LR), and keywords (KW). Edges connect court decisions to legal references and keywords, while no direct connections exist between nodes of the same type. Edge weights were computed according to the methodology proposed in the original work and were retained without modification.



(a) Overview of the tripartite graph structure with nodes representing court decisions (CD), legal references (LR), and keywords (KW), and weighted edges capturing connection strength. Reproduced from Horvát et al. [13].



(b) Illustrative subgraph of three court decisions (CD) and their neighborhood (LR and KW nodes) connected via weighted edges.

Figure 1: Tripartite graph representation of the court decision dataset. The left panel depicts the full graph structure, while the right panel illustrates a local neighborhood with weighted connections.

Table 1
Overview of embedding models used for LR and KW node representation.

Model	Type	Dim	Description
BGE-M3 [17]	Transformer embedding model	1024	General-purpose multilingual retrieval model.
EmbeddingGemma-300M [18]	Transformer embedding model	768	Lightweight multilingual embedding model.
multilingual-e5-large [19]	Transformer embedding model	1024	Sentence embedding model optimized for retrieval tasks.
SlovakBERT [20]	BERT-based encoder	768	Monolingual Slovak language model.
SlovakBERT-STS-STSB [20]	Sentence embedding model	768	Fine-tuned for semantic textual similarity.
Legal-Slovak-RoBERTa [16]	RoBERTa-based model	768	Domain-specific model trained on legal Slovak texts.

Legal references were extracted from court decisions using the approach proposed by Varga et al. [14], which combines rule-based extraction with similarity-based matching techniques. The textual content associated with legal reference nodes was obtained from the Collection of Laws of the Slovak Republic. Keywords were extracted using the TF-IDF-based method proposed by Szoplák et al. [15]

For the downstream classification task, a balanced subset of 20,000 court decisions was constructed from the criminal law domain. The objective is to classify court decisions into four criminal-law subcategories: Property, Life and Health, Family and Youth, and General Threats and to the Environment, with 5,000 randomly selected decisions per category.

The adopted graph contains textual information associated with legal reference (LR) and keyword (KW) nodes. It consists of 19,490 court decision (CD) nodes, 19,065 legal reference (LR) nodes, and 689 keyword (KW) nodes connected by 169,701 edges. The difference between the number of court decisions in the classification dataset and the number of CD nodes results from the graph construction procedure[13], in which decisions without any retained connections are not included in the graph

The adopted graph contains textual information associated with legal reference (LR) and keyword (KW) nodes. The resulting graph consists of 19,490 court decision (CD) nodes, 19,065 legal reference (LR) nodes, and 689 keyword (KW) nodes connected by 169,701 edges. Legal reference nodes are linked to the corresponding parts of the law text, while keyword nodes correspond to key phrases extracted from court decisions. These textual attributes are used to generate node embeddings, which subsequently serve as the basis for construcruction of court decision representations described in the following subsection.

For each of the six transformer-based language models listed in Table 1, embeddings were independently generated for LR and KW nodes. The selected models cover three groups of representation models: general-purpose multilingual embedding models, Slovak-specific models, and legal-domain models. This selection allows us to compare whether multilingual retrieval-oriented models, monolingual Slovak models, or a legal-domain Slovak model provide more suitable representations for the downstream court decision classification task. Legal-Slovak-RoBERTa is included as a Slovak legal-domain RoBERTa-based model from the MultiLegalPile/LegalLMs model family [16]. For sentence-transformer models, the corresponding model-specific encoding interface was used, while encoder-only transformer models were represented using mean pooling over token embeddings.

3.2. Court Decision Embedding Construction

Court decision representations are constructed by aggregating embeddings of neighboring legal reference and keyword nodes in the adopted graph illustrated in Figure 1.b. Simple pooling is included as a baseline aggregation strategy. Weighted pooling methods are evaluated to assess whether incorporating graph edge weights improves representation quality. Concatenation-based aggregation is used to

examine whether explicitly separating legal reference and keyword information is beneficial. Finally, principal component removal is considered as an optional post-processing step to investigate potential improvements in embedding quality.

For each court decision CD_i, let $N(CD_i)$ denote the set of neighboring nodes connected to CD_i. Each node $v \in N(CD_i)$ is represented by an embedding vector $e_v \in \mathbb{R}^d$, obtained using a selected embedding model.

3.2.1. Simple Pooling Aggregation

Let $\{e_1, e_2, \dots, e_n\}$ denote embeddings of neighboring nodes of a given court decision CD_i.

Mean pooling is defined as:

$$e_{CD_i} = \frac{1}{n} \sum_{j=1}^{n} e_j$$

Sum pooling is defined as:

$$e_{CD_i} = \sum_{j=1}^{n} e_j$$

Max pooling is applied element-wise across embedding dimensions:

$$e_{CD_i} = \max(e_1, e_2, \dots, e_n)$$

3.2.2. Weighted Pooling Aggregation

Let w_j denote the weight associated with the edge connecting CD_i and node j.

Weighted mean pooling is defined as:

$$e_{CD_i} = \frac{\sum_{j=1}^{n} w_j e_j}{\sum_{j=1}^{n} w_j}$$

Weighted sum pooling is defined as:

$$e_{CD_i} = \sum_{j=1}^{n} w_j e_j$$

3.2.3. Concatenation

Let $N_{LR}(CD_i)$ and $N_{KW}(CD_i)$ denote the sets of legal reference and keyword neighbors, respectively. Let $\mathrm{Agg}(\cdot)$ denote either mean pooling or sum pooling, as defined above.

Aggregation is performed independently for each set:

$$e_{LR} = \mathrm{Agg}\left(N_{LR}(CD_i)\right), \quad e_{KW} = \mathrm{Agg}\left(N_{KW}(CD_i)\right)$$

The final representation is obtained by concatenation:

$$e_{CD_i} = e_{LR} \,\|\, e_{KW}$$

where $\|$ denotes vector concatenation.

3.2.4. Principal Component Removal

Neural language model embeddings can exhibit anisotropy, with vectors concentrated along dominant direction in the embedding space [21, 22]. Suppressing this direction may reduce the influence of common structure and improve class discrimination.

As an additional experimental setting, first principal component removal (PCR) is applied to selected configurations of the aggregated embedding as an exploratory analysis, aiming to assess whether removing the dominant direction improves downstream classification performance.

Given a set of raw court decision embeddings $\{e_{CD_1}, \ldots, e_{CD_m}\}$, principal component analysis is performed to obtain the first principal component vector u.

Each embedding is then transformed as:

$$e_{CD_i}^{\text{PCR}} = e_{CD_i} - (u^\top e_{CD_i})u$$

This transformation removes the projection of the embeddings onto the dominant global direction in the embedding space.

The resulting court decision embeddings are used as input features for supervised classification models evaluated on a labeled dataset of court decisions.

4. Experimental Setup

This section describes the evaluation protocol used to assess the generated court decision embeddings. We evaluate the embeddings from two complementary perspectives. First, we measure their downstream usefulness in a supervised classification task. Second, we perform an intrinsic evaluation based on cosine similarity to analyze whether the embedding space preserves the expected class structure independently of a classifier.

4.1. Supervised Classification Evaluation

For each embedding model and aggregation strategy, document-level embeddings of court decisions were constructed and evaluated on a supervised classification task. The dataset was divided into training and test subsets using stratified sampling to preserve the original class distribution. The test set was reserved exclusively for final evaluation.

The quality of the generated embeddings was assessed using five classification algorithms: Logistic Regression, Linear Support Vector Machine (Linear SVM), Support Vector Machine with a radial basis function kernel (RBF SVM), Random Forest (RF), and Multilayer Perceptron (MLP).

For each combination of embedding model, aggregation strategy and classification algorithm, hyperparameter optimization was performed on the training set using grid search with five-fold cross-validation. The best hyperparameter configuration was selected according to the mean cross-validation score and subsequently evaluated on the held-out test set.

Performance was measured using mean cross-validation score, test accuracy, and macro-averaged F1 score. Since the evaluation dataset contains an equal number of instances in each category, accuracy was used as the primary evaluation metric, while macro F1 score was reported as a complementary measure of class-balanced performance.

4.2. Intrinsic Evaluation of Embedding Space

In addition to the supervised classification evaluation, we performed an intrinsic evaluation of the generated court decision embeddings. This evaluation assesses whether the embedding space preserves the expected class structure independently of a downstream classifier. All intrinsic metrics were computed using cosine similarity.

Let $D = \{d_1, d_2, \ldots, d_n\}$ denote the set of court decisions, $E = \{e_1, e_2, \ldots, e_n\}$ the corresponding embeddings, and y_i the class label of decision d_i. Cosine similarity between two decision embeddings is defined as:

$$\text{sim}(d_i, d_j) = \frac{e_i^\top e_j}{\|e_i\|\|e_j\|}.$$

We used four complementary intrinsic evaluation perspectives. First, the cluster margin metric compares the average similarity of a decision to decisions from the same class with its average similarity to decisions from different classes. For each decision d_i, the margin is computed as:

$$m(d_i) = s_{\text{in}}(d_i) - s_{\text{out}}(d_i),$$

where $s_{\text{in}}(d_i)$ denotes the average cosine similarity to decisions with the same class label and $s_{\text{out}}(d_i)$ denotes the average cosine similarity to decisions from other classes. The reported mean margin is then:

$$\text{Margin}_{\text{mean}}(E) = \frac{1}{n} \sum_{i=1}^{n} m(d_i).$$

Second, the cluster ranking evaluation treats each decision as a query and ranks all other decisions by decreasing cosine similarity. A retrieved decision is considered relevant if it belongs to the same class as the query. The ranking quality is measured using mean average precision:

$$MAP(E) = \frac{1}{n} \sum_{i=1}^{n} AP(d_i),$$

where $AP(d_i)$ is the average precision of the ranking produced for query decision d_i.

Third, the pairwise evaluation treats cosine similarity as a score for distinguishing same-class decision pairs from different-class pairs. All evaluated pairs are ranked by decreasing cosine similarity. Pairwise average precision is computed as:

$$AP_{\text{pairwise}}(E) = \frac{1}{P} \sum_{k=1}^{M} \text{Prec}(k) \cdot \text{rel}(k),$$

where M is the number of evaluated pairs, P is the number of same-class pairs, $\text{rel}(k)$ equals 1 if the pair at rank k is a same-class pair and 0 otherwise, and $\text{Prec}(k)$ is the precision among the top k ranked pairs. Higher AP_{pairwise} indicates that same-class pairs tend to receive higher similarity scores than different-class pairs.

Finally, the centroid-based evaluation compares each decision to class centroids. For each class c, the centroid is computed as the mean embedding of all decisions belonging to that class:

$$\mu_c = \frac{1}{|\{i \,:\, y_i = c\}|} \sum_{i \,:\, y_i = c} e_i.$$

Each decision is assigned to the class with the most similar centroid:

$$\hat{y}_i = \arg\max_{c} \text{sim}(e_i, \mu_c).$$

Centroid accuracy measures the proportion of decisions whose nearest class centroid matches their true class label:

$$\text{Acc}_{\text{centroid}}(E) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{I}(\hat{y}_i = y_i),$$

where $\mathbb{I}(\cdot)$ is the indicator function.

Cosine similarity ranges from -1 to 1, where 1 indicates maximum similarity. Consequently, the mean cluster margin theoretically ranges from -2 to 2, with higher values indicating stronger separation between same-class and different-class decisions, and 2 representing the theoretical optimum. MAP, pairwise AP, and centroid accuracy range from 0 to 1, with 1 representing perfect performance in each case.

Table 2
Impact of Embedding Models in Downstream Classification (test accuracy averaged over aggregation strategies)

Embedding model	Mean (%)	Std (%)	Min (%)	Max (%)
multilingual-e5-large	83.4	1.2	82.0	84.8
SlovakBERT-STS-STSB	83.3	1.0	81.8	84.2
BGE-M3	83.2	1.1	81.6	84.3
SlovakBERT	83.1	1.0	81.5	83.9
EmbeddingGemma-300M	82.9	1.0	81.4	83.7
Legal-Slovak-RoBERTa	79.4	1.4	77.5	81.5

5. Results

This section presents the experimental results of our study. We organize the evaluation into two complementary parts: extrinsic (downstream) and intrinsic analysis. The extrinsic evaluation is based on a multi-class classification task over 20,000 court decisions and uses test set accuracy as the primary metric. The intrinsic evaluation examines the structural properties of the embedding space independently of the downstream task.

Downstream Evaluation Results

We present the results of the extrinsic evaluation, focusing on classification performance under different representation strategies. The analysis is structured according to the three research questions addressing

(i) The impact of embedding models

(ii) The impact of aggregation strategies

(iii) The effect of weighting schemes in document construction

5.1. Impact of Embedding Models

This research question investigates how the choice of underlying embedding model influences the quality of document representations in the downstream classification task (Section 4.1).

The reported results, summarized in Table 2, are obtained by averaging performance across all aggregation strategies for each embedding model mentioned in Section **??**, so that results are not tied to a single aggregation strategy.

The results show that modern multilingual embedding models such as `multilingual-e5-large` and `bge-m3`, and SlovakBERT-based variants achieve highly similar average performance, with differences below one percentage point. This indicates that within this group, the choice of embedding model has a limited impact on downstream classification performance.

In contrast, the legal-domain-specific model (Legal-Slovak-RoBERTa) consistently underperforms all other models. Despite being trained on legal-domain data, it achieved the lowest average classification performance among the evaluated embedding models.

Overall, the results in Table 2 suggest that the evaluated multilingual and Slovak-specific models perform comparably in our experimental setting, while Legal-Slovak-RoBERTa consistently underperforms the remaining models.

5.2. Impact of Aggregation Strategy

We evaluate the effect of different aggregation strategies on downstream classification performance. We compare multiple pooling methods for constructing document representations, as described in Section 3.2.

Table 3
Impact of Aggregation Strategies on Classification Performance (test accuracy averaged over embedding models)

Aggregation type	Mean (%)	Std (%)	Min (%)	Max (%)
max pooling	83.7	1.2	81.5	84.8
concat_sum	83.3	1.6	80.0	84.3
mean pooling	83.2	1.8	79.6	84.3
sum pooling	83.1	1.6	79.9	83.9
concat_mean	82.4	1.5	79.4	83.2
weighted_mean pooling	81.0	1.7	77.6	82.2
weighted_sum pooling	81.0	1.7	77.5	82.0

Table 4
Effect of Principal Component Removal (PCR) on test accuracy of downstream task

Embedding model	Aggregation	Before (%)	After (%)	Δ (p.p.)
slovakbert-sts-stsb	concat_sum	84.2	84.1	-0.10
multilingual-e5-large	sum pooling	83.9	83.9	-0.08
slovakbert-sts-stsb	sum pooling	83.9	83.8	-0.10
bge-m3	concat_sum	83.9	83.5	-0.46
embeddinggemma-300m	concat_mean	83.2	83.2	-0.03
slovakbert-sts-stsb	concat_mean	83.1	83.0	-0.10
slovakbert-sts-stsb	weighted_mean pooling	81.8	81.8	-0.05
legal-slovak-roberta-base	weighted_mean pooling	77.6	77.5	-0.03

The results, summarized in Table 3, are obtained by averaging classification accuracy across all embedding models for each aggregation strategy.

We observe that simple pooling strategies, namely max, mean, sum, and concat-sum, achieve very similar average performance, ranging from 83.1% to 83.7%. These methods form a group of high-performing aggregation strategies.

When comparing sum pooling and its concatenation-based variant, concat-sum shows slightly higher average performance; however, the difference remains small. In contrast, concatenation-based mean pooling achieves lower average performance compared to standard mean pooling.

A consistent performance difference is observed between weighted and non-weighted variants. For both mean and sum pooling, the weighted versions achieve lower average accuracy compared to their counterparts, with a gap of approximately two percentage points.

Overall, the results indicate that performance differences are most pronounced between weighted and non-weighted aggregation strategies, while simple pooling methods without weighting exhibit highly similar performance.

5.3. Exploratory Analysis of Principal Component Removal

Table 4 summarizes the effect of principal component removal (PCR) on selected embedding and aggregation configurations. Across all evaluated configurations, PCR results in a decrease in classification accuracy. The observed reductions are generally small, typically around 0.1 percentage points, with the largest decrease of 0.46 percentage points observed for the combination of BGE-M3 and concat-sum aggregation.

5.4. Intrinsic Evaluation

The intrinsic evaluation provides a complementary view of the generated embedding spaces by measuring whether the embeddings preserve the expected class structure under cosine similarity. Table 5 summarizes the best-performing configurations according to the main intrinsic criteria.

Table 5
Best configurations according to intrinsic evaluation metrics.

Criterion	Configuration	$\text{Margin}_{\text{mean}}$	MAP	AP_{pairwise}	$\text{Acc}_{\text{centroid}}$
Best average rank	BGE-M3 + PCR concat mean	0.0609	0.3885	0.6264	0.7227
Best centroid accuracy	multilingual-e5-large + PCR concat mean	0.0598	0.3810	0.6181	0.7271
Best mean margin	BGE-M3 + PCR sum	0.0745	0.3827	0.6174	0.6955

The results show that no single configuration dominates all intrinsic metrics. `BGE-M3` with PCR-enhanced sum pooling achieves the strongest global class separation, reaching the highest mean margin. However, this configuration is not the most robust across the remaining criteria.

The most consistent configuration is `BGE-M3` with PCR-enhanced concatenated mean pooling. It achieves the best average rank across the evaluated intrinsic metrics and also obtains the highest MAP and pairwise average precision. This suggests that separately aggregating legal references and keywords and then concatenating their representations provides a better balance between global class separation and local neighbourhood structure.

The strongest alternative is `multilingual-e5-large` with PCR-enhanced concatenated mean pooling, which achieves the highest centroid accuracy. Overall, the intrinsic evaluation supports the use of concatenation-based aggregation, especially in combination with strong multilingual embedding models, as a stable strategy for constructing court decision representations.

6. Discussion

This section discusses the main findings of the experimental evaluation and their implications for document representation in legal classification tasks.

6.1. Embedding Models

The results indicate that modern multilingual embedding models achieve highly similar performance on the evaluated classification task. Differences between the best-performing models, including multilingual-e5-large, bge-m3, and SlovakBERT variants, are relatively small. In contrast, the domain-specific legal-Slovak-RoBERTa model consistently underperforms compared to the other evaluated embeddings.

These observations suggest that, for the considered dataset and task, modern sentence embedding models already provide a relatively saturated level of semantic representation quality, where further differences between architectures or training strategies have limited impact on downstream classification performance.

6.2. Aggregation Strategies

The evaluation of aggregation strategies shows that simple pooling methods, such as mean, sum, and max pooling, achieve highly comparable performance. Concatenation-based variants provide only marginal differences in accuracy, without a consistent improvement over simpler approaches. Weighted pooling methods, in contrast, consistently yield lower performance compared to their non-weighted counterparts.

Overall, these results indicate that simple aggregation strategies are sufficient for constructing effective document representations in this setting, while introducing weighting schemes does not lead to performance gains and may negatively affect classification accuracy.

6.3. Principal Component Removal and Intrinsic Analysis

As a supplementary exploratory analysis, principal component removal (PCR) was evaluated on a subset of embedding configurations. Across all tested cases, PCR does not improve downstream classification performance and results in a slight decrease in accuracy.

In parallel, intrinsic evaluation of the embedding spaces reveals a different pattern. Representations enhanced with PCR achieve strong performance in terms of cluster-based metrics, including margin, ranking-based measures, and centroid accuracy. However, no single configuration consistently dominates across all intrinsic metrics, and the best-performing setups vary depending on the evaluation criterion.

This discrepancy suggests a partial decoupling between intrinsic embedding space structure and downstream task performance in the evaluated setting. Due to the limited scope of the PCR experiments, these findings should be interpreted as preliminary and indicative of a potential direction for further investigation.

6.4. Summary of Findings

Overall, the results show that modern embedding models provide robust and largely comparable performance for legal document classification, simple aggregation strategies are sufficient and generally outperform more complex weighted variants, and intrinsic improvements in embedding space structure do not necessarily translate into improved downstream classification performance.

7. Conclusion

This paper presented a comparative study of embedding models and aggregation strategies for representing legal court decisions in a multi-class classification setting. We evaluated multiple state-of-the-art multilingual embedding models combined with various pooling-based aggregation methods, and further analyzed their performance using an additional post-processing step based on principal component removal (PCR) as well as intrinsic evaluation metrics.

The results show that modern multilingual embedding models achieve highly similar performance on the downstream classification task, while the domain-specific model consistently underperforms. Regarding aggregation strategies, simple pooling methods such as mean, sum, and max pooling perform comparably, with weighted variants consistently yielding lower accuracy. Concatenation-based aggregation provides only marginal differences in performance without clear and consistent improvements.

An additional analysis using principal component removal indicates a slight decrease in downstream classification performance across all evaluated configurations. In contrast, intrinsic evaluation suggests that PCR-based representations may improve certain structural properties of the embedding space, although no single configuration dominates across all intrinsic metrics.

Future work may focus on a more comprehensive study of principal component removal across a wider range of embedding configurations, as well as a deeper investigation into the relationship between intrinsic embedding space properties and downstream task performance. Further extensions may also include evaluation on additional datasets and exploration of learned aggregation mechanisms.

Acknowledgments

This work was supported by the Slovak Research and Development Agency under contract No. VV-MVP-24-0038, ALIT – Analysis of Liability for Internet Torts Using Machine Learning Methods.

Declaration on Generative AI

During the preparation of this work, the authors used ChatGPT (OpenAI) in order to: Grammar and spelling check. After using this tool, the authors reviewed and edited the content as needed and take

full responsibility for the publication's content.

References

[1] F. Ariai, J. Mackenzie, G. Demartini, Natural language processing for the legal domain: A survey of tasks, datasets, models, and challenges, ACM Comput. Surv. 58 (2025). URL: https://doi.org/10.1145/3777009. doi:10.1145/3777009.

[2] N. Limsopatham, Effectively leveraging BERT for legal document classification, in: N. Aletras, I. Androutsopoulos, L. Barrett, C. Goanta, D. Preotiuc-Pietro (Eds.), Proceedings of the Natural Legal Language Processing Workshop 2021, Association for Computational Linguistics, Punta Cana, Dominican Republic, 2021, pp. 210–216. URL: https://aclanthology.org/2021.nllp-1.22/. doi:10.18653/v1/2021.nllp-1.22.

[3] N. Reimers, I. Gurevych, Sentence-BERT: Sentence embeddings using Siamese BERT-networks, in: K. Inui, J. Jiang, V. Ng, X. Wan (Eds.), Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), Association for Computational Linguistics, Hong Kong, China, 2019, pp. 3982–3992. URL: https://aclanthology.org/D19-1410/. doi:10.18653/v1/D19-1410.

[4] I. Chalkidis, M. Fergadiotis, P. Malakasiotis, N. Aletras, I. Androutsopoulos, LEGAL-BERT: The muppets straight out of law school, in: T. Cohn, Y. He, Y. Liu (Eds.), Findings of the Association for Computational Linguistics: EMNLP 2020, Association for Computational Linguistics, Online, 2020, pp. 2898–2904. URL: https://aclanthology.org/2020.findings-emnlp.261/. doi:10.18653/v1/2020.findings-emnlp.261.

[5] J. Chauhan, Legal-sbert: Creating a sentence tranformer for the legal domain and generating data, 2023. URL: https://api.semanticscholar.org/CorpusID:259267090.

[6] G. M. Csányi, D. Lakatos, I. Üveges, A. Megyeri , J. P. Vadász, D. Nagy, R. Vági, From fact drafts to operational systems: Semantic search in legal decisions using fact drafts, Big Data and Cognitive Computing 8 (2024). URL: https://www.mdpi.com/2504-2289/8/12/185. doi:10.3390/bdcc8120185.

[7] J. A. F. Costa, N. C. D. Dantas, E. D. S. A. Silva, Evaluating text classification in the legal domain using bert embeddings, in: P. Quaresma, D. Camacho, H. Yin, T. Gonçalves, V. Julian, A. J. Tallón-Ballesteros (Eds.), Intelligent Data Engineering and Automated Learning – IDEAL 2023, Springer Nature Switzerland, Cham, 2023, pp. 51–63.

[8] K. Valle, Graph-based representations for text classification, 2011. URL: https://api.semanticscholar.org/CorpusID:59366266.

[9] G. Zhang, P. Nulty, D. Lillis, Argument mining with graph representation learning, in: Proceedings of the Nineteenth International Conference on Artificial Intelligence and Law, ICAIL '23, Association for Computing Machinery, New York, NY, USA, 2023, p. 371–380. URL: https://doi.org/10.1145/3594536.3595152. doi:10.1145/3594536.3595152.

[10] T. N. Kipf, M. Welling, Semi-supervised classification with graph convolutional networks (2016). doi:10.48550/arXiv.1609.02907.

[11] W. L. Hamilton, R. Ying, J. Leskovec, Inductive representation learning on large graphs, 2018. URL: https://arxiv.org/abs/1706.02216. arXiv:1706.02216.

[12] J. Mu, S. Bhat, P. Viswanath, All-but-the-top: Simple and effective postprocessing for word representations, ArXiv abs/1702.01417 (2017). URL: https://api.semanticscholar.org/CorpusID:3986339.

[13] Š. Horvát, D. Varga, Z. Szoplák, P. Gurský, S. Krajči, Graph-based categorization of court decisions by legal domain, in: Proceedings of the Twentieth International Conference on Artificial Intelligence and Law, ICAIL '25, Association for Computing Machinery, New York, NY, USA, 2026, p. 399–403. URL: https://doi.org/10.1145/3769126.3769251. doi:10.1145/3769126.3769251.

[14] D. Varga, M. Gojdič, Z. Szoplák, P. Gurský, Š. Horvát, S. Krajči, L. Antoni, Extraction of legal references from court decisions, in: Conference on Theory and Practice of Information Technologies, 2023. URL: https://api.semanticscholar.org/CorpusID:264545654.

[15] Z. Szoplák, P. Gurský, D. Varga, Optimizing Keyphrase Extraction for Court Decisions Using Legal References, 2024. doi:10.3233/FAIA241265.

[16] J. Niklaus, V. Matoshi, M. Stürmer, I. Chalkidis, D. Ho, MultiLegalPile: A 689GB multilingual legal corpus, in: L.-W. Ku, A. Martins, V. Srikumar (Eds.), Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, Bangkok, Thailand, 2024, pp. 15077–15094. URL: https://aclanthology.org/2024.acl-long.805/. doi:10.18653/v1/2024.acl-long.805.

[17] J. Chen, S. Xiao, P. Zhang, K. Luo, D. Lian, Z. Liu, M3-embedding: Multi-linguality, multi-functionality, multi-granularity text embeddings through self-knowledge distillation, in: L.-W. Ku, A. Martins, V. Srikumar (Eds.), Findings of the Association for Computational Linguistics: ACL 2024, Association for Computational Linguistics, Bangkok, Thailand, 2024, pp. 2318–2335. URL: https://aclanthology.org/2024.findings-acl.137/. doi:10.18653/v1/2024.findings-acl.137.

[18] H. S. Vera, S. Dua, B. Zhang, D. Salz, R. Mullins, S. R. Panyam, S. Smoot, I. Naim, J. Zou, F. Chen, D. Cer, A. Lisak, M. Choi, L. Gonzalez, O. Sanseviero, G. Cameron, I. Ballantyne, K. Black, K. Chen, W. Wang, Z. Li, G. Martins, J. Lee, M. Sherwood, J. Ji, R. Wu, J. Zheng, J. Singh, A. Sharma, D. Sreepathihalli, A. Jain, A. Elarabawy, A. Co, A. Doumanoglou, B. Samari, B. Hora, B. Potetz, D. Kim, E. Alfonseca, F. Moiseev, F. Han, F. P. Gomez, G. H. Ábrego, H. Zhang, H. Hui, J. Han, K. Gill, K. Chen, K. Chen, M. Shanbhogue, M. Boratko, P. Suganthan, S. M. K. Duddu, S. Mariserla, S. Ariafar, S. Zhang, S. Zhang, S. Baumgartner, S. Goenka, S. Qiu, T. Dabral, T. Walker, V. Rao, W. Khawaja, W. Zhou, X. Ren, Y. Xia, Y. Chen, Y.-T. Chen, Z. Dong, Z. Ding, F. Visin, G. Liu, J. Zhang, K. Kenealy, M. Casbon, R. Kumar, T. Mesnard, Z. Gleicher, C. Brick, O. Lacombe, A. Roberts, Q. Yin, Y. Sung, R. Hoffmann, T. Warkentin, A. Joulin, T. Duerig, M. Seyedhosseini, Embeddinggemma: Powerful and lightweight text representations, 2025. URL: https://arxiv.org/abs/2509.20354. arXiv:2509.20354.

[19] L. Wang, N. Yang, X. Huang, L. Yang, R. Majumder, F. Wei, Multilingual e5 text embeddings: A technical report, ArXiv abs/2402.05672 (2024). URL: https://api.semanticscholar.org/CorpusID:267547683.

[20] M. Pikuliak, Š. Grivalský, M. Konôpka, M. Blšták, M. Tamajka, V. Bachratý, M. Simko, P. Balážik, M. Trnka, F. Uhlárik, SlovakBERT: Slovak masked language model, in: Y. Goldberg, Z. Kozareva, Y. Zhang (Eds.), Findings of the Association for Computational Linguistics: EMNLP 2022, Association for Computational Linguistics, Abu Dhabi, United Arab Emirates, 2022, pp. 7156–7168. URL: https://aclanthology.org/2022.findings-emnlp.530/. doi:10.18653/v1/2022.findings-emnlp.530.

[21] J. Mu, P. Viswanath, All-but-the-top: Simple and effective post-processing for word representations, 2018. Publisher Copyright: © Learning Representations, ICLR 2018 - Conference Track Proceedings.All right reserved.; 6th International Conference on Learning Representations, ICLR 2018 ; Conference date: 30-04-2018 Through 03-05-2018.

[22] H. Sajjad, F. Alam, F. Dalvi, N. Durrani, Effect of post-processing on contextualized word representations, in: N. Calzolari, C.-R. Huang, H. Kim, J. Pustejovsky, L. Wanner, K.-S. Choi, P.-M. Ryu, H.-H. Chen, L. Donatelli, H. Ji, S. Kurohashi, P. Paggio, N. Xue, S. Kim, Y. Hahm, Z. He, T. K. Lee, E. Santus, F. Bond, S.-H. Na (Eds.), Proceedings of the 29th International Conference on Computational Linguistics, International Committee on Computational Linguistics, Gyeongju, Republic of Korea, 2022, pp. 3127–3142. URL: https://aclanthology.org/2022.coling-1.277/.