Towards Semantic Comparison of Examination Regulations: A Prototype for Cross-Institutional Paragraph Analysis

Douglas Blank[1,*], Stefan Conrad[1]

[1]Heinrich Heine University, Universitätsstraße 1, 40225 Düsseldorf, Germany

Abstract

Examination regulations define key formal rules in higher education but are difficult to compare across institutions due to inconsistent formatting and heterogeneous structure. This paper presents a prototype pipeline for extracting and semantically analyzing paragraphs from German bachelor-level examination regulations. We apply regular-expression-based segmentation to isolate legal-style paragraphs and compare their content using TF-IDF vectors and sentence embeddings. Our initial results show that while surface-level representations suffice for intra-institutional comparisons, semantic embeddings, especially when fine-tuned, are necessary for reliable cross-institutional similarity. The proposed method lays the groundwork for large-scale regulatory analyses, enabling structured comparisons of academic policies across universities.

Keywords

Examination Regulations, Paragraph Extraction, Semantic Text Similarity, Sentence Embeddings, Legal Document Processing, Cross-Institutional Analysis

1. Introduction

Examination regulations play a central role in defining the formal and legal framework of academic programs. They regulate study structures, examination types, credit rules, and degree requirements. In large-scale educational analyses, such as efforts to understand causes of high dropout rates, differences between these regulations across universities can be highly relevant.

However, these documents are difficult to analyze automatically. They are typically published as PDFs with inconsistent formatting, contain dense legal language, and lack a standardized structure. This makes it challenging to extract and compare individual rules across institutions.

This paper presents a first prototype for extracting and representing individual paragraphs from German bachelor-level examination regulations. We combine regular expression-based text segmentation with vector-based similarity techniques and sentence embeddings to enable semantic comparison of paragraphs from different sources.

The goal is to lay the groundwork for future large-scale analyses of regulatory similarities in higher education. Although still in an early stage, the approach demonstrates promising directions for bridging structural and semantic heterogeneity in legal-educational texts.

2. Related Work

Text processing of legal documents is a highly challenging task in natural language processing (NLP) due to their complex structure and domain-specific terminology. Moreover, legal texts are often used in high-stakes contexts, such as by courts or legal professionals, where misinterpretations or errors

36th GI-Workshop on Foundations of Databases (Grundlagen von Datenbanken), September 29 - October 01, 2025, Regensburg, Germany

*Corresponding author.

douglas.blank@hhu.de (D. Blank); stefan.conrad@hhu.de (S. Conrad)

https://dbs.cs.uni-duesseldorf.de/mitarbeiter.php?id=blank (D. Blank);
https://dbs.cs.uni-duesseldorf.de/mitarbeiter.php?id=conrad (S. Conrad)

0009-0005-2158-0862 (D. Blank); 0000-0003-2788-3854 (S. Conrad)

can lead to serious consequences, including the misapplication of laws or financial harm. Even in less critical settings, individuals may rely on automatically processed legal content to make personal or educational decisions, which further underscores the importance of robust and accurate methods.

Although the type of legal documents addressed in this work, which is German bachelor-level examination regulations, arguably belongs to a lower-risk category, they still demand careful processing. These documents occupy a small and underexplored niche within the broader domain of legal texts. To the best of our knowledge, there exists no prior work that specifically targets this document type, particularly in the German language.

In recent years, research on legal NLP has grown substantially. While much of the focus has been on text classification, the next most commonly addressed tasks are information extraction and information retrieval [1]. The rise of word embeddings and especially transformer-based models has further accelerated progress in this domain. Among these, BERT [2] and its legal-domain adaptation LEGAL-BERT [3] are widely used for various downstream tasks such as classification, summarization, and question answering on legal corpora.

More recently, the use of large language models (LLMs) in legal NLP has attracted increased attention [4, 5]. However, their capabilities in legal reasoning, factual correctness, and explainability still require further systematic investigation.

Despite the availability of such pre-trained legal models, none are directly applicable to our setting. For example, LEGAL-BERT is trained exclusively on English legal corpora, such as EU and US legislation or court rulings. Its structure, vocabulary and training context differ significantly from German examination regulations. Currently there exists no analogous model for German legal texts that supports semantic similarity tasks at the paragraph level.

The most notable German-language contribution is the Legal Entity Recognition (LER) dataset [6], which provides high-quality annotations for named entity recognition in legal documents. Based on this dataset, German BERT models have been fine-tuned for Legal NER [7], achieving strong results. However, these models are tailored to entity recognition rather than capturing paragraph-level semantics.

Our work therefore addresses a gap in the intersection of German legal NLP and semantic similarity modeling. In contrast to prior work, we focus on comparing semantically related paragraphs across structurally inconsistent legal documents using contrastive learning and fine-tuned sentence embeddings.

3. Data and Problem Setting

This work originates from a larger research project that investigates examination regulations of German bachelor programs in the context of identifying factors associated with high dropout rates. A comprehensive analysis requires access to a large and diverse collection of such regulations from universities across Germany.

However, there is no centralized repository for these documents, and not all universities publish them openly. While we plan to issue formal data access requests to all German universities as part of the broader project, this step has not yet been taken. For this initial study, we therefore rely on a small sample of examination regulations manually collected from publicly available university websites.

These regulations are typically available as PDF files, which must first be processed to extract their textual content. Tools such as PyMuPDF [8] are suitable for this task and work well for most documents. However, in some cases these tools fail, either due to unusual formatting or because the PDFs are actually scanned images without embedded text. From our experience, this issue occurs particularly with older documents. In such cases, we fall back on optical character recognition (OCR) using tools like Tesseract [9] to recover the textual content.

While there is no standardized format for how examination regulations are structured, certain recurring elements can be observed. Typically, a document may begin with a cover page, followed by a short introductory section that places the regulation in a legal context, indicating, for example, which

overarching laws it builds upon or refines. Some documents also include a table of contents listing all defined paragraphs.

The main body of the document consists of the actual legal paragraphs, which appear sequentially and define the specific rules for the respective bachelor's program. These may cover topics such as admission to examinations, required and elective modules, grading procedures, or the roles and responsibilities of examiners. At the end, some documents include additional content that is not legally binding, such as general information about the university or acknowledgments of staff members.

For illustration, Figure 1 shows a typical excerpt from a regulation paragraph, covering rules related to module examination outcomes. Note the legal style, nested structure, and the mixture of definitions and conditions, all of which pose challenges for both text extraction and semantic analysis.

§ 13 Modulprüfungen: Bestehen und Nichtbestehen
(1) Eine Prüfungsleistung ist mit Erfolg erbracht und die Modulprüfung somit bestanden, wenn sie mindestens mit „ausreichend" (kleiner oder gleich 4,0) bewertet wurde.
(2) Eine Modulprüfung wird als nicht bestanden bewertet, wenn sie mit der Note „nicht ausreichend" (5,0) bewertet wurde.
(3) Die kumulative Modulprüfung zu einem Modul ist bestanden, wenn alle geforderten Prüfungsleistungen mit „ausreichend" oder besser bewertet und alle geforderten Studienleistungen erbracht wurden. Andernfalls wird die kumulative Modulprüfung mit der Note „nicht ausreichend" (5,0) bewertet.
(4) Mit dem Bestehen der Modulprüfung sind alle gemäß Anhang auf das betreffende Modul entfallenden Leistungspunkte erworben.

Figure 1: Excerpt from a German examination regulation regarding the passing and failing of module exams.

3.1. Paragraph Extraction

Assuming that all documents share a common structural core, consisting of a sequence of legal paragraphs formatted as shown in Figure 1, we can apply regular expressions to extract these units. Specifically, we assume that while the beginning and end of each document may vary (e.g., cover pages, metadata, or appendices), the central portion follows a consistent paragraph-based layout.

To extract individual paragraphs, we use a regular expression that identifies lines starting with the paragraph symbol (§) followed by a paragraph number, and captures all subsequent text until the beginning of the next paragraph or the end of the document:

```
(^§\s*\d+.*?)(?=^§|\Z)
```

The regular expression is applied in multiline mode, so that "^" matches the beginning of any line rather than only the beginning of the entire document. In addition, dot-all mode is enabled, allowing the wildcard operator "." to match newline characters as well. This configuration ensures that multi-line paragraphs are captured correctly.

While this regular expression provides a powerful baseline for extracting paragraphs from examination regulations, it is far from perfect. One notable issue is that the final matched paragraph includes not only the actual paragraph content, but also all remaining text until the end of the document. In some cases, this includes metadata or university-related appendices that are not relevant to our analysis.

A second type of artifact results from the preprocessing step in which the PDF is converted into plain text. Since all visible content is extracted, including headers, footers, and page numbers, these elements can appear inside the extracted paragraph blocks. This introduces noise into the resulting text segments and may affect downstream processing.

We attempt to reduce such noise through basic text refactoring operations. These include removing lines that consist only of isolated numbers (e.g., page numbers), collapsing multiple empty lines, merging orphaned single-character lines with their successors, and restoring hyphenated words that were split

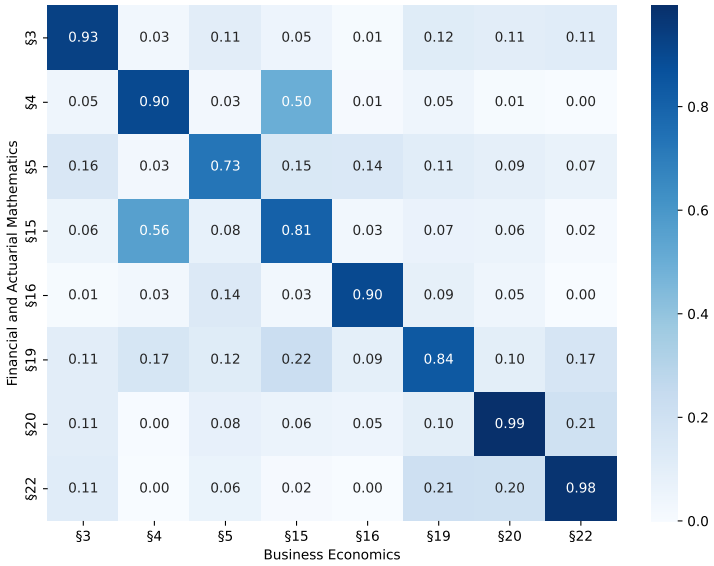

Figure 2: Pairwise comparison of TF-IDF representations for selected paragraphs from the examination regulations of the bachelor programs Business Economics and Financial and Actuarial Mathematics. For a brief description of the individual paragraphs, see Table 1

across line breaks. While these steps improve the structure and readability of the extracted content, they do not yet address document-specific artifacts such as recurring footers or university metadata.

We acknowledge these limitations and plan to refine our extraction pipeline in the future, for example by incorporating layout-aware filtering or manually curated rules. For the purposes of this study, however, we proceed with the current method and assume that remaining artifacts do not significantly impact our preliminary similarity analysis.

4. Approach

Now that we have a method for extracting paragraphs from the source documents, the next step is to transform these text segments into a representation suitable for analysis. As a first approach, we use the term frequency-inverse document frequency (TF-IDF) [10] representation, implemented via the scikit-learn library [11].

TF-IDF produces a sparse vector representation of a document based on the frequency of its words, scaled by how often those words appear across the entire corpus. While this approach does not capture word order or context, it offers a robust and interpretable baseline for comparing textual content on a statistical level. In our case, each extracted paragraph is treated as an individual document, resulting in one TF-IDF vector per paragraph.

These vector representations can now be compared to one another, for example by computing the cosine similarity between pairs of paragraphs. As a first experiment, we apply this technique to paragraphs extracted from examination regulations of bachelor programs at our institution, *Heinrich Heine University Düsseldorf*.

To explore the similarity structure visually, we compute pairwise cosine similarities between all paragraph representations and display the resulting matrix as a heatmap. This allows us to identify clusters of paragraphs with potentially overlapping regulatory content. In Figure 2, we present a similarity matrix based on a sample of TF-IDF vectors derived from paragraphs taken from two different examination regulations: *Business Economics* and *Financial and Actuarial Mathematics*. A brief description of the content of the respective paragraphs is provided in Table 1.

As shown in the heatmap, one paragraph from the first document exhibits a high similarity to a paragraph in the second document that covers the same regulatory topic. Among all other paragraph

Table 1
Description of the paragraphs in Figures 2, 4 and 5

Paragraph	Description
§ 3	Start of Studies and Admission Requirements
§ 4	Standard Period and Scope of Study
§ 5	Examinations / Deadlines / Examination Dates
§ 15	Academic Requirements
§ 16	Type / Scope of the Bachelor's Examination
§ 19	Passing the Bachelor Examination
§ 20	Voluntary Supplementary Modules
§ 22	Invalidity of the Bachelor Examination

Table 2
Description of the paragraphs in Figures 3 and 6

Paragraph	Description
§ 1	Study Program Objectives and Purpose
§ 8	Examination Committee
§ 13	Repetition of Examinations
§ 18	Bachelor's Thesis
§ 23	Access to Examination Records
§ 24	Entry into Force and Publication

pairs, similarity scores remain comparatively low, with the possible exception of §4 and §15, which show slightly elevated similarity. This can be explained by overlapping content. Several rules stated in §4 are reiterated or extended in §15.

Overall, the similarity values align well with the thematic relationships between paragraphs, suggesting that TF-IDF representations may already provide a reasonable baseline for intra-institutional comparisons. However, this no longer holds when comparing documents from different institutions.

In Figure 3, we show similarity scores between paragraphs from the *Business Administration* regulation and another regulation from a different university, anonymized here as *Study Program X*. While some topic-related paragraph pairs exhibit slightly higher similarity than unrelated pairs, the absolute scores remain low and ambiguous. In other words, even thematically related paragraphs may not be clearly distinguishable from unrelated ones based on TF-IDF alone.

A likely explanation is that institution-specific stylistic and structural conventions influence word distributions. Documents from the same university are often written by the same authors or follow consistent editorial guidelines, which benefits TF-IDF but fails to generalize. Since TF-IDF is based purely on term frequency, it captures surface-level lexical similarity but ignores semantic meaning.

To address this limitation, we explore sentence-level embeddings generated by pre-trained Sentence-Transformers [12]. These models map entire sentences or paragraphs into dense vector representations that aim to reflect semantic content more robustly and contextually.

In our experiments, we evaluated several pre-trained Sentence-Transformer models, including the multilingual *cross-en-de-roberta-sentence-transformer* [13]. However, these models did not yield satisfactory results for our task. As shown in Figure 4, which replicates the previous comparison from Figure 2, most paragraphs appear overly similar to one another, making meaningful distinctions difficult.

We hypothesize that this is due to the general-purpose nature of these models, which are trained on diverse and broad datasets. As a result, they tend to encode high-level contextual similarities, such as the fact that all input texts are part of examination regulations, while underestimating finer-grained differences in content.

We therefore propose training a specialized model on domain-specific data. To the best of our knowledge, no sentence embedding model currently exists that has been trained on German examination

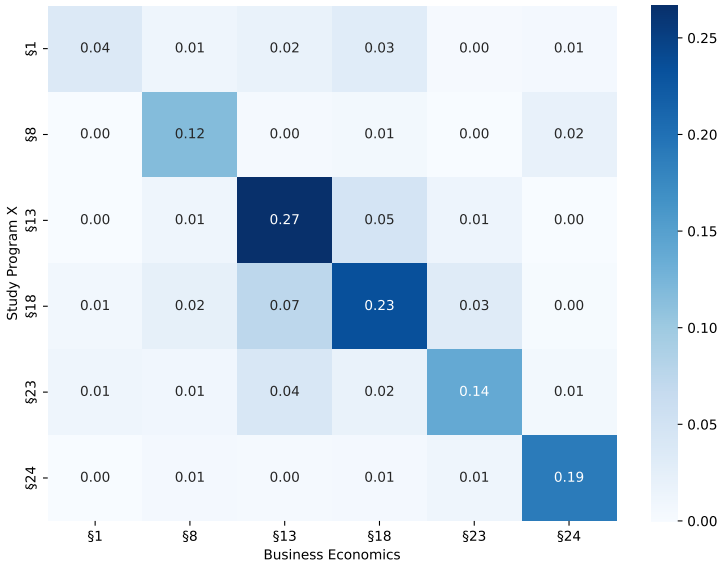

Figure 3: Pairwise comparison of TF-IDF representations for selected paragraphs from the examination regulations of the bachelor programs Business Economics and an anonymized program from another university. For a brief description of the individual paragraphs, see Table 2

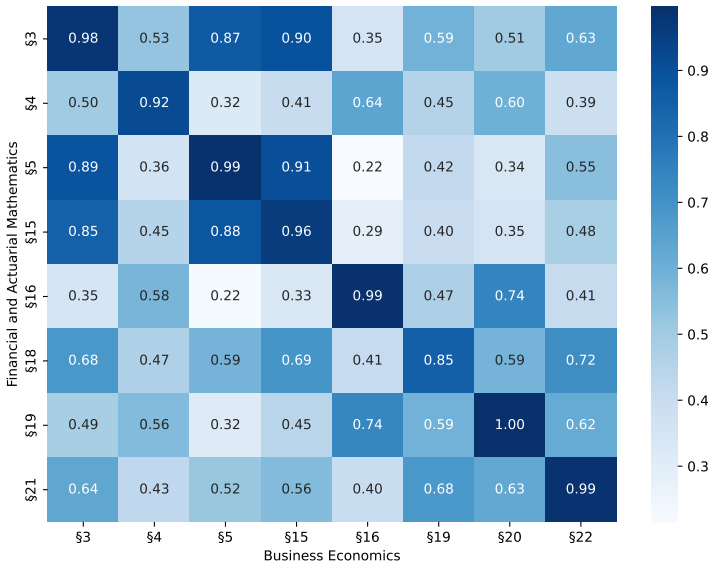

Figure 4: Pairwise cosine similarity of sentence-transformer representations (cross-en-de-roberta-sentence-transformer, pre-trained) for the same paragraph selection as in Figure 2.

regulations. As a next step, we aim to construct such a model and assess whether it can better capture paragraph-level distinctions relevant to our analysis.

For this purpose, we are currently building a small, manually curated dataset of examination regulations. As outlined earlier, we intend to submit formal data access requests to all German universities, asking for bachelor-level examination regulations from the past ten years. While the eventual response rates remain uncertain, we hope this effort will enable the creation of a sufficiently large and representative corpus to support domain-specific model training.

We train our model using a contrastive learning approach inspired by self-supervised methods [14], based on cross-document paragraph similarity. Let $\mathscr{D} = \{D_1, D_2, \ldots, D_n\}$ be the set of documents, where each document D_i consists of a set of paragraphs.

For every paragraph $p \in D_i$, we construct one positive pair per other document D_j with $j \neq i$.

Specifically, we define the paragraph $q^* \in D_j$ that is most similar to p as its positive counterpart:

$$(p, q^*) \text{ is a positive pair } \iff q^* = \arg\max_{q \in D_j} \text{sim}(p, q)$$

In addition, all other paragraphs $r \in D_j \setminus \{q^*\}$ and all paragraphs $r \in D_i \setminus \{p\}$ are used to generate negative pairs:

$$(p, r) \text{ is a negative pair } \iff r \in (D_j \setminus \{q^*\}) \cup (D_i \setminus \{p\})$$

This strategy results in $n - 1$ positive pairs and many more negative pairs per paragraph. We employ a contrastive loss to bring semantically aligned content (positive pairs) closer together in the embedding space, while pushing apart unrelated or less relevant paragraphs (negative pairs). The contrastive loss is defined as

$$\mathscr{L} = y \cdot \|u - v\|^2 + (1 - y) \cdot \max(0, m - \|u - v\|)^2,$$

where u and v are the embeddings of two paragraphs, and $y \in \{0, 1\}$ is a binary label indicating whether the respective paragraphs form a positive pair ($y = 1$) or a negative pair ($y = 0$). The margin parameter $m > 0$ defines the minimum distance that negative pairs should be apart in the embedding space. If the embeddings of a negative pair are closer than this threshold, the loss increases, encouraging the model to push them further apart.

While we currently do not possess a dataset large enough to train a sentence embedding model from scratch, we keep this option open and may revisit it once a sufficiently large collection of examination regulations becomes available through our formal data access requests.

In the meantime, we fine-tune the pre-trained cross-en-de-roberta-sentence-transformer on our initial dataset. If this adapted model shows improved performance on our similarity task, we intend to further refine and scale this approach as additional data becomes available.

At the current stage of our work, we use a dataset consisting of only 18 examination regulation documents, handpicked from publicly accessible university websites across several German federal states. As mentioned before, documents originating from the same institution tend to follow similar stylistic and structural patterns, either because they are authored by the same administrative offices or because newer documents intentionally replicate earlier formats for the sake of consistency.

To avoid overfitting to such layout- or style-specific patterns, we deliberately include documents from a diverse range of institutions. This variety is intended to encourage the model to focus on the semantic content of paragraphs rather than superficial formatting similarities.

From the extracted paragraphs, we construct a training dataset consisting of 217 manually defined positive pairs and 3,244 derived negative pairs, following the contrastive learning scheme described earlier. At this stage, we rely on small supervised samples for the positive examples, as we have not yet identified a suitable heuristic or learning signal to automate the matching of semantically equivalent paragraph pairs.

One approach we are considering for future iterations is to use the fine-tuned model itself as a weak supervision signal. Specifically, we could identify, for each paragraph, its most similar counterpart in other documents according to the current model, and use these matches as positive training pairs. A model trained on this automatically generated data could then be used to produce the next generation of labels, allowing for a process of iterative refinement.

However, if the generated pairs are of insufficient quality, the model may be guided in the wrong direction, potentially reinforcing noise or drifting away from meaningful semantic distinctions. We leave the exploration of appropriate strategies for such self-supervised label generation to future work.

Fine-tuning is performed using the AdamW optimizer on an NVIDIA RTX 5080 GPU with a batch size of 32 for 500 training steps. We use a learning rate of 2×10^{-5} and a warmup ratio of 0.1. These settings were selected as reasonable defaults and not subjected to extensive hyperparameter tuning, as optimization of training dynamics is beyond the scope of this work.

The training procedure is implemented using the SentenceTransformers [12] framework, including its ContrastiveLoss module with a margin of $m = 0.5$.

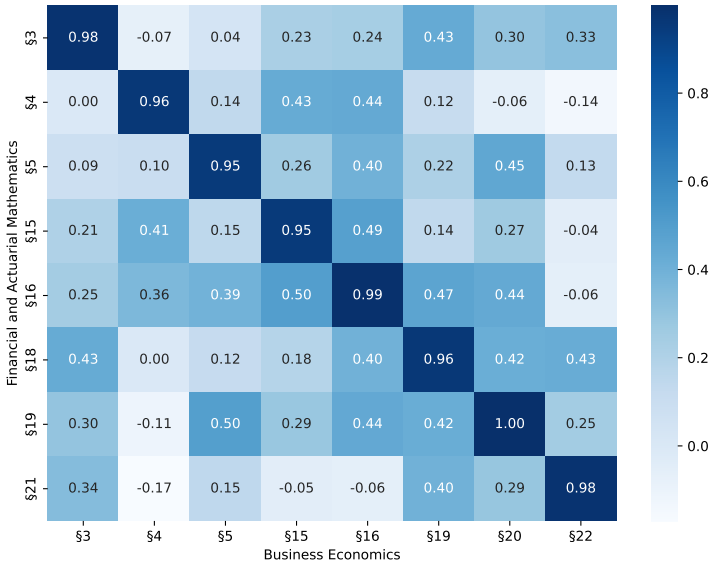

Figure 5: Pairwise cosine similarity of sentence-transformer representations (cross-en-de-roberta-sentence-transformer, after our fine-tuning) for the same paragraphs as in Figure 4.

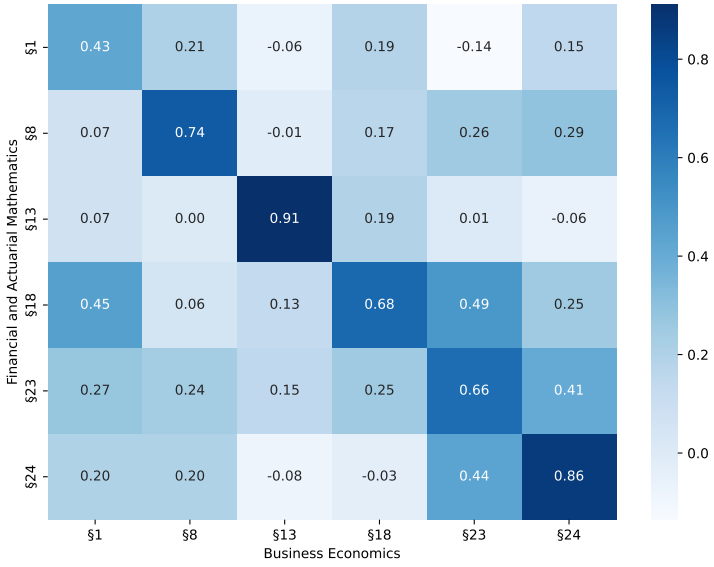

Figure 6: Pairwise cosine similarity of sentence-transformer representations (cross-en-de-roberta-sentence-transformer, after our fine-tuning) for the same paragraphs as in Figure 3.

The resulting model produces more consistent and interpretable similarity scores between paragraphs. As illustrated in the comparison between Figure 4 and Figure 5, semantically related paragraphs are more accurately identified after fine-tuning. Additionally, the fine-tuned model enables meaningful comparisons between paragraphs from different documents (Figure 6), a task in which TF-IDF representations had previously failed (cf. Figure 3).

These results demonstrate the feasibility of our approach: we successfully conceptualized and implemented a pipeline that extracts paragraphs from examination regulations, transforms them into vector-based representations, and enables semantic comparison across documents. While this study represents an early proof-of-concept, the observed improvements suggest that our methodology is promising, particularly if the acknowledged limitations are addressed and the system is scaled to a broader dataset.

5. Conclusion and Outlook

This paper presented an initial step toward a broader, long-term analysis of examination regulations in German higher education. The work is part of an interdisciplinary project that investigates institutional rules and responsibilities as potential factors contributing to high dropout rates or prolonged study durations.

To support this research, we plan to develop a system that allows for cross-university comparisons of regulatory structures. One envisioned feature is the ability to identify and highlight differences in specific regulatory aspects, such as exam admission criteria, between institutions with significantly different dropout rates. Users would be able to select a given paragraph from one university's regulation and retrieve semantically similar paragraphs from others.

With this larger goal in mind, the present paper first examined the structure of examination regulation documents and proposed a simple method to extract individual paragraphs using regular expressions. Although this method is effective as a baseline, it occasionally captures irrelevant text fragments, such as page numbers or footers introduced during PDF preprocessing. We plan to refine this step in future iterations to improve extraction quality and structural segmentation.

In the second part, we explored how extracted paragraphs can be represented in ways that reflect their underlying semantic meaning, enabling pairwise comparison. While TF-IDF vectors offer a simple and interpretable representation, they are often insufficient when comparing documents from different institutions, due to stylistic and structural inconsistencies. To address this, we fine-tuned a pre-trained sentence transformer model using a contrastive learning approach and demonstrated that the resulting embeddings yield more meaningful semantic similarity scores across institutional boundaries.

Although we currently rely on a small set of manually defined positive pairs, the proposed training method shows promising results even on limited data. As a next step, we aim to expand this training process by developing a fully self-supervised method for generating high-quality positive pairs. We believe that scaling the model with a larger and more diverse dataset, that is currently under construction, will significantly improve its semantic understanding.

In addition to the limitations of the training data, it is also important to consider structural assumptions inherent in our current approach. Specifically, we assume that the examination regulations being compared follow a structurally similar format, that is, each paragraph in one document has a clearly corresponding counterpart in the other. However, this is not always the case. For example, a concept that is expressed within a single large paragraph in one document may be distributed across multiple smaller paragraphs in another. Conversely, a given paragraph may not have any meaningful counterpart at all.

While such discrepancies could potentially be mitigated by allowing flexible similarity thresholds or by aggregating multiple paragraphs for comparison, another complication arises in this specific data setting. Some universities define general examination regulations that apply across all programs, and supplement these with program-specific rules. In such cases, a complete representation of the regulation requires the combination of multiple documents, which further complicates the alignment task.

Beyond fine-tuning, we also consider training a model entirely from scratch in the future. In addition, further models could be developed for related tasks such as named entity recognition and question answering within the context of examination regulations, depending on future project requirements.

Acknowledgments

This work was supported by the Federal Ministry of Research, Technology and Space (BMFTR, formerly BMBF) under grant number 16FG001B as part of the project ”RegelWerk”.

Declaration on Generative AI

The author(s) have not employed any Generative AI tools.

References

[1] D. M. Katz, D. Hartung, L. Gerlach, A. Jana, M. J. B. II, Natural language processing in the legal domain, 2023. URL: https://arxiv.org/abs/2302.12039.

[2] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding, 2019. URL: https://arxiv.org/abs/1810.04805.

[3] I. Chalkidis, M. Fergadiotis, P. Malakasiotis, N. Aletras, I. Androutsopoulos, Legal-bert: The muppets straight out of law school, 2020. URL: https://arxiv.org/abs/2010.02559.

[4] J. Lai, W. Gan, J. Wu, Z. Qi, P. S. Yu, Large language models in law: A survey, 2023. URL: https://arxiv.org/abs/2312.03718.

[5] D. H. Anh, D.-T. Do, V. Tran, N. L. Minh, The impact of large language modeling on natural language processing in legal texts: A comprehensive survey, in: 2023 15th International Conference on Knowledge and Systems Engineering (KSE), 2023, pp. 1–7. doi:10.1109/KSE59128.2023.10299488.

[6] E. Leitner, G. Rehm, J. Moreno-Schneider, Fine-grained Named Entity Recognition in Legal Documents, in: M. Acosta, P. Cudré-Mauroux, M. Maleshkova, T. Pellegrini, H. Sack, Y. Sure-Vetter (Eds.), Semantic Systems. The Power of AI and Knowledge Graphs. Proceedings of the 15th International Conference (SEMANTiCS 2019), number 11702 in Lecture Notes in Computer Science, Springer, Karlsruhe, Germany, 2019, pp. 272–287. 10/11 September 2019.

[7] H. Darji, J. Mitrović, M. Granitzer, German bert model for legal named entity recognition, in: Proceedings of the 15th International Conference on Agents and Artificial Intelligence, SCITEPRESS - Science and Technology Publications, 2023, p. 723–728. URL: http://dx.doi.org/10.5220/0011749400003393.

[8] b. o. c. b. J. X. M. Artifex Software, Inc., R. Liu, PyMuPDF: Python bindings for mupdf, https://pymupdf.readthedocs.io/, 2025. Version 1.26.3.

[9] R. Smith, An overview of the tesseract ocr engine, in: Ninth International Conference on Document Analysis and Recognition (ICDAR 2007), volume 2, 2007, pp. 629–633. doi:10.1109/ICDAR.2007.4376991.

[10] C. D. Manning, P. Raghavan, H. Schütze, Introduction to Information Retrieval, Cambridge University Press, Cambridge, UK, 2008. URL: http://nlp.stanford.edu/IR-book/information-retrieval-book.html.

[11] L. Buitinck, G. Louppe, M. Blondel, F. Pedregosa, A. Mueller, O. Grisel, V. Niculae, P. Prettenhofer, A. Gramfort, J. Grobler, R. Layton, J. VanderPlas, A. Joly, B. Holt, G. Varoquaux, API design for machine learning software: experiences from the scikit-learn project, in: ECML PKDD Workshop: Languages for Data Mining and Machine Learning, 2013, pp. 108–122.

[12] N. Reimers, I. Gurevych, Sentence-bert: Sentence embeddings using siamese bert-networks, 2019. URL: https://arxiv.org/abs/1908.10084.

[13] P. May, cross-en-de-roberta-sentence-transformer, https://huggingface.co/T-Systems-onsite/cross-en-de-roberta-sentence-transformer, 2020. Licensed under the MIT License. Copyright (c) 2020 Philip May, T-Systems on site services GmbH.

[14] A. Jaiswal, A. R. Babu, M. Z. Zadeh, D. Banerjee, F. Makedon, A survey on contrastive self-supervised learning, 2021. URL: https://arxiv.org/abs/2011.00362.