Chair 17 - Information Systems and Business Process Management
TUM School of Computation, Information and Technology
Technical University of Munich

Automating Legal Text Classification in Business Process Management: A Machine Learning Approach

Internship Report - Approaching Information System Challenges with Natural Language Processing

Patrick Ahrend ✉

TUM School of Computation, Information and Technology, Technical University of Munich

✉ patrick.ahrend@tum.de

January 30st, 2024

Abstract — This report presents a comprehensive exploration of machine learning and natural language processing techniques for the classification of legal texts for Business Process Management (BPM). It focuses on developing and evaluating models that aim to identify relevant legal documents for business processes, addressing key challenges in business process compliance. Utilizing diverse embedding techniques like TF-IDF, Word2Vec, Glove, Fasttext, Bert, and GPT-3, the study contrasts traditional machine learning models with advanced approaches, including fine-tuned large language models (LLM) and rule-based methods. The research highlights the efficacy of machine learning in managing the complexities of legal texts and offers insights into potential practical applications, particularly in reducing manual workload and aiding in regulatory compliance. However, the study acknowledges the limitations posed by the dataset's size and domain-specific nature. Future research with more extensive and varied data is essential to validate and generalize these findings.

1 Introduction

At the core of every organization lie its business processes, which are often the focus of dedicated departments such as business solutions or process excellence. These departments strive to quantify business processes with key performance indicators, constantly adapting them to meet objectives in a timely, cost-effective, and qualitative manner—an overarching goal that presents both challenges and opportunities [20]. A pervasive challenge organizations face is to identify which legal documents are relevant to their processes. Typically, the task of ensuring that business practices are in compliance with relevant laws and regulations falls to compliance departments [6, 18]. Within the field of BPM, this task is referred to as business process compliance, which is often manual and labor-intensive [7]. With an increase in development of natural language processing (NLP), this project is a machine learning approach to classify legal text by its relevant to business processes. Further, it shows an outlook on fine-tuning and rule-based approaches to this problem and compares the results.

2 Related Work

As the approach is creating an automated machine learning approach to classify legal text. One can argue that the related literature can be divided into two sections.

2.1 NLP within business process management and compliance

Han van der Aa et al. [20]. highlights the potential of NLP in the field of Business Process Management (BPM). Particularly interesting was the mention of NLP techniques within the field of Regulation. Examples of such are conducted by the following: Han van der Aa et al. [19] suggested the concept of a behavioral space, which comprises all different interpretations of a textual process description within businesses. They further evaluated the usefulness of the behavioral space for compliance checking. Sapkota et al. [18] introduced a framework to extract regulation entities into a machine-interpretable format. The results were close to manual annotation in a case study of the Pharmaceutical industry. El Hamdani et al.[6] proposed a combination of a rule-based and machine-learning approach for automated compliance checking of the General Data Protection Regulation (GDPR). This is particularly interesting as this study also used the privacy policies of the GDPR in the approach, but created business processes out of the articles and used the article as relevant regulatory text.

Arguably, the most related work to this one was done by Cai et al. [17]. They study how assess-

ing legal regularity text for business processes can be assisted by manual and (semi-)automated approaches. The semi-manual approach is a crowd study with Amazon MTurk, where untrained individuals annotate regulation text based on a provided process description and BPMN. The automated approaches are a state-of-the-art natural language processing legal information retrieval (SOTA NLP LIR) with S-bert [15] and using the large language model GPT-4 with zero-shot approach and specific prompt engineering. Further, the approaches vary on scales of automation, transparency, and reproducibility. For comparison, a expert analysis was conducted to have a gold standard to evaluate the methods against in a quantitative manner. Evaluated on precision, recall, and accuracy, GPT-4 was able to outperform the SOTA NLP LIR and crowd study approach. The authors reasoned this by GPT-4 ability to use a lot of contextual information as input to determine the relevance of annotation of the legal text.

2.2 Machine Learning Text Classification on Legal Text

Outside the field of BPM, there has been extensive work on classifying legal text to support legal matters. Christian JH Mahoney et al. [11] worked on making text classification of legal text more explainable. They proposed a framework that summarized the positively labeled documents into 50-word snippets, to quickly determine false positives. Boella et al. [1] presented a topic-based classification of Italian legal text, emphasizing the importance of the support vector machine model. They achieved 97.50 percent accuracy on single paragraphs and 76.23 percent on article level. Chen et al. [3] hinted that although different techniques are being used for text classification of legal text, there is still room for improvement. They showed promising results of 84.49 percent for a domain-specific text classification with pre-trained word embeddings deep learning algorithms and a Random Forest Classifier.

This report expands upon the work conducted in the course Master Internship: Approaching Information System Challenges with Natural Language Processing from Catherine Sai. Thus, this study worked on the same problem of classifying regulatory text for business processes, but with a machine learning approach.

3 Approach

In Figure 1, the proposed machine learning pipeline is displayed. The following sections go into detail for each step.

3.1 Data

The data for this study were sourced from three distinct domain areas: 1) GDPR, 2) Austrian Smart Meter, and 3) Australian legal documents [5]. These sources provided a range of processes: seven processes from GDPR, three business processes from Australian data, and six processes from the Smart Meter data. The labeling methodology included three distinct groups for the input. Group A comprised relevant passages from relevant documents, Group B included irrelevant passages from relevant documents, and Group C contained irrelevant passages from irrelevant documents. The distribution across these groups was approximately 10 percent from Group A, and 45 percent each from Groups B and C, resulting in a relevance to irrelevance ratio of 1:9. For the representation of the processes, either existing process descriptions were utilized, or new descriptions were generated based on the Business Process Model and Notation (BPMN) of the processes. Consequently, the dataset encompassed 1860 data points, each consisting of a Process Description, Legal Text, and a Label denoting relevance (Relevant or Irrelevant). This dataset is accessible in the /data/raw/ directory of the project repository [16].

Initially, the project focused on applying deep learning to this problem and thus needed a lot more data. For this, GPT-3.5 was fine-tuned to assist in automating data collection. However, this approach was subsequently revised due to the model's tendency to overlabel passages as relevant and a high level of confidence in its outputs. As these newly relevant passages could not be confidently added to the gold standard due to the lack of legal knowledge, this approach was aborted. This lack of large volumes of data is a common problem for BPM tasks, as it requires legal expertise and domain knowledge [2, 20]. It is also noteworthy that the lengths of the data sources varied. The majority of the Australian data consisted of passages, while GDPR and Smart Meter data predominantly featured sentence-level information. Passages were not broken down into sentences as they often comprised lists or enumerations, which could potentially introduce redundancy rather than provide additional value to the analysis.

3.2 Embeddings

Prior to building embeddings, the data was preprocessed to convert into lowercase and remove punctuation. The reason for this was that some embedding methods are case-sensitive, and to further limit the vocabulary size. For vectorization of the text corpus into vectors of weighted features, which models can be trained on, a diverse range of embedding techniques were used. These techniques, ranging from simple to complex, included TF-IDF, Word2Vec, GloVe, FastText, Bert, and GPT-3. The selection of both static and contextual embeddings from transformers aimed to evaluate their efficacy in capturing information from the text corpus for the model to learn on. This goes to the point that generally speaking, there is a claim that simpler embeddings like TF-IDF are not capable of building high-quality feature classification of legal text [3]. Thus, also a relatively large vector size of 1000 features was chosen for TF-IDF, compared to other simpler approaches like Glove and Word2Vec with 300 dimensions, to see if vector dimensions matter in feature efficiency. The specific settings for these embeddings are detailed in the /src/features/build word embeddings.py file in the project repository [16].

Two distinct types of embeddings were created for this study: 1. **Combined**: This method involved concatenating the process description and legal text before embedding them together.
2. **Separate**: In this approach, the process description and legal text were embedded separately and then concatenated.

For further feature enrichment, additional features were added to the word embeddings which made sense to the use case. These included the cosine similarity between the process descriptions and legal texts, and the frequency of the top 100 most common words. This was done with the aim of enhancing the models' ability to understand the relationship between process descriptions and legal texts. To evaluate the significance of these added features, a feature importance analysis was conducted using a Random Forest Classifier. Due to the large size of the feature set (101 dimensions) and the extensive dimensions of the embeddings, Principal Component Analysis (PCA) with a 95 percent variance threshold was applied to reduce dimensionality. However, the added features demonstrated minimal importance, leading to not being used further in the analysis. Detailed results of the feature importance analysis are available in the /references/feature importance directory of the project repository [16].

3.3 Explanatory Data Analysis

As word embeddings were often a black box, an explanatory data analysis was conducted using Uniform Manifold Approximation and Projection (UMAP) [12]. It is a non-linear dimension reduction technique that can be quite good at preserving structure and relations in high-dimensional data. The primary focus of this analysis was to explore the correlation between process descriptions and legal texts as represented in the embeddings.

One has to be careful to not over-interpret the results of UMAP as it reduces dimensions and thus does not fully reflect the actual data. It's possible that other patterns might emerge or vary within the reduced dimension. In this analysis, four different UMAP configurations were employed: one for each embedding type (Combined and Separate), one exclusively for legal text, and one for separate embeddings visualized in multiple UMAPs for each embedding type.

The analysis of different embedding approaches using UMAP revealed distinct patterns. In the UMAP dedicated to legal text, no clear clusters were observed across any embeddings. However, certain groupings of data points were noted, possibly indicating texts from the same documents, especially considering the overlap in documents within labeling Group C. The multiple UMAPs displayed a pattern where legal texts circle a central cluster of process descriptions. This could be attributed to the similarity in writing style across process descriptions, which is quite different from the legal text corpus. Interestingly, clear clusters can be seen for each process in both the separate and combined embeddings, suggesting that each process may have a distinct vocabulary.

An analysis of the placement of data points from different label classes indicated that relevant classes consistently formed a subset within the irrelevant classes. This observation aligns with the hypothesis that the broader classification of irrelevant texts (Group C) creates larger clustering in UMAP.

Comparing various embeddings, TF-IDF and GPT appeared most effective in clustering different processes. GloVe, FastText, Word2Vec, and Bert all had some form of overlap in the Smart Meter or GDPR processes. This was somewhat unexpected, especially in the case of TF-IDF outperforming BERT. One possible explanation is that the word frequency information captured by TF-IDF is sufficiently informative, whereas the additional context provided by transformer-based embeddings like BERT may introduce noise. This hypothesis is supported by the ob-

servation that TF-IDF and GPT, which have larger vector dimensions, showed more distinct clustering patterns. An experiment to increase vector sizes in TF-IDF, Word2Vec, and FastText did not result in significant changes in clustering, reinforcing the initial findings.

3.4 Modelling

For traditional machine learning, this study incorporated nine models from the Scikit-learn library. The Logistic Regression model was selected as a baseline due to its class weight parameter. The class weight parameter is an essential feature for handling the imbalanced dataset, which can otherwise lead to models only predicting the dominant class, and is quite relevant to this use case.

Additionally, a Decision Tree Classifier was employed as a standard tree model. Tree-based ensemble models such as Random Forest and Gradient Boosting were also utilized. The rationale for including these models stemmed from their proficiency in feature identification within embeddings, and there are less prone to overfitting [3]. The study further included two variations of Naive Bayes classifiers: Gaussian and Bernoulli. These models are known for their effectiveness in scenarios with smaller datasets and limited vocabularies [3].

A Support Vector Machine model was chosen as it appears to perform quite well on legal text classification [3, 8]. It also has the class weight parameter. Additionally, a perceptron model as it a basic form of a neural network, aligning with the study's intention to use deeper neural network techniques in advanced approaches.

Lastly, a Stochastic Gradient Descent (SGD) model was used as well due to its efficacy in managing large-scale and sparse data. The initial training phase for these models was conducted using default parameters, followed by a phase of hyperparameter tuning. The tuning process utilized grid search and 5-fold cross-validation to optimize model performance.

3.5 Advanced Techniques

For the advanced approach, a Recurrent Neural Network (RNN) was implemented using PyTorch, featuring Tanh activation functions and binary cross-entropy as the loss function. The process of fine-tuning involved determining optimal values for hidden size, number of layers, learning rate, and batch size for each embedding type. This fine-tuning was accomplished using the Ray Tune library [10]. Detailed results of the hyperparameter optimization are documented in the /data/processed/hyperparameters directory of the project repository [16].

Subsequently to RNN, the study progressed to fine-tune pretrained BERT Models for the classification task. Specifically, This involved working with both the 'base-uncased' and 'large-uncased' versions of BERT. Unlike the standard BertForSequenceClassification, a customized implementation was adopted for a more granular control and understanding of the network's final layer. This custom approach utilized the Bert Model class from the Hugging Face transformers library, complemented by a sigmoid activation function applied to the pooled output. Binary cross-entropy was again employed as the loss function for these models. Comprehensive details of both implementations are available in the /notebooks/advanced approach directory in the project repository [16].

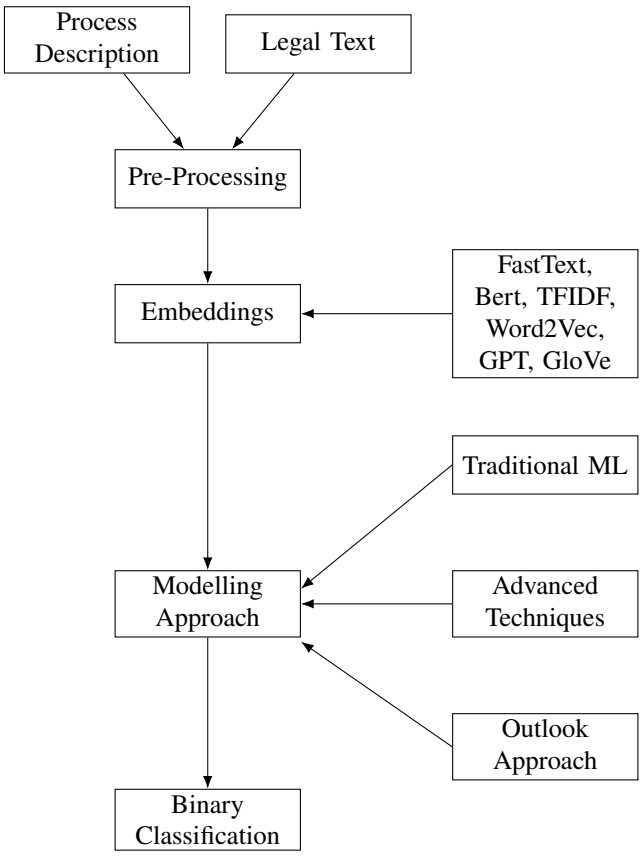

Figure 1 The legal text classification pipeline with preprocessing, embeddings, modeling and binary classification stage.

4 Evaluation

The evaluation of the models was conducted in two parts: a quantitative analysis focusing on key metrics, and a qualitative analysis based on the interpretation of data points. This was carried out on a gold standard comprising 266 observations, where the relevant class represented 11 percent of the dataset, similarly to the original data set. The processes selected for this analysis were those with the highest volume of data points, drawn from the three data domain sources.

4.1 Quantitative Analysis

For the quantitative evaluation of text classification, recall, precision, and accuracy are the most frequently used metrics [9]. This study specifically focused on recall due to the critical nature of identifying all relevant passages within the legal documents. Given the context of the use case, the cost of overlooking relevant regulatory requirements outweighs having more false positives, which might lead to increased manual review but ensures regulatory compliance.

The recall results for various fine-tuned models, corresponding to their best-performing embedding types, are presented in Table 1.

Table 1 Model Performance - Recall Results

Model	Embedding Type	Recall
Random Forest	All Similar	0.67
Decision Tree	FastText	0.73
Gradient Boosting	FastText	0.6
Logistic Regression	Glove	0.83
Bert-Base	Bert	0.5
Bert-Large	Bert	0.2
SGD	All Similar	0.0
SVC	TF-IDF	0.57
RNN	All Similar	0.0
Perceptron	GPT	1
Bernoulli NB	TF-IDF	0.67
Gaussian NB	TF-IDF	0.73

The analysis revealed that there was no variance in recall between the combined and separate embeddings. This observation counters the initial hypothesis that the increased dimensionality of separate embeddings might adversely affect recall scores, potentially due to the 'curse of dimensionality'. UMAP visualizations supported this finding, showing similar clustering patterns for both types of embeddings.

For the tree models, Gradient Boosting exhibited the least effective performance, in contrast to the Random Forest, which consistently excelled across various embeddings. The Decision Tree Classifier demonstrated a notable high recall score when utilizing fastText embeddings, yet its performance was suboptimal with other embeddings, especially when compared to Random Forest. Logistic Regression displayed robust performance with Bert, GPT, TF-IDF, and particularly with GloVe embeddings. However, its recall score were zero for both fasttext and word2vec embeddings, indicating that its predictive capacity in this context is heavily reliant on the type of embedding employed.

It is important to note that the Perceptron model achieved a recall score of 1 exclusively with GPT embeddings, while invariably scoring 0 with other embeddings. This lack of consistency suggests that the model consistently predicts a single class, a hypothesis that will be further examined in the subsequent qualitative analysis.

RNN, SGD, and SVC models were unable to effectively learn from the data provided. An unexpected finding was the lower recall score of the larger BERT variant compared to its base model. Additionally, it was observed that traditional machine learning models, like Random Forest and Logistic Regression, surpassed transformer-based models in terms of recall scores.

Moreover, it is crucial to highlight that inaccuracies in the prediction of even a single data point results in a lower recall score in the magnitude of 3 percent. This is due to the proportion of the true class within the 266 gold standard instances constitutes merely 11 percent. Visualization of the recall scores in a heat map and model performances are in the /references/model results directory of the project repository [16].

4.2 Qualitative Analysis

The qualitative examination of model outputs revealed consistent patterns in misclassification across various models, which are noteworthy for understanding the models' behavior beyond the quantitative scores. A recurring observation was that all models invariably misclassified the same nine data points from the Australian 'travel insurance claim process', labeling them as irrelevant (0) when they should have been marked relevant (1). Furthermore, a manual examination of the nine data points was conducted. Which, without legal knowledge, failed to identify these data points as relevant. This observation suggest limitations in the gold standard.

The Decision Tree model demonstrated the best performance, with its misclassifications being exclusively

confined to the travel insurance process. The Random Forest model had a similar pattern, with only one additional misclassified data point from the same process. The Gradient Boosting model slightly deviated by misclassifying one further data point from the travel insurance process and one from the 'Know Your Customer' process, also from the Australia data.

Logistic Regression faced similar challenges with the travel insurance data points as the tree-based models, but also was able to predict 2 data points of these correctly on the GloVe embeddings exclusively. BERT-based models, both 'base' and 'large', show quite different misclassification behaviors, indicating a disparate approach to learning from the dataset. This make sense as they are transformer based models.

For the SGD, SVC and RNN models uniformly predicted all data points as irrelevant, demonstrating an inability to learn from the data in order to classify legal text correctly. The Percheron model did the same for all models except GPT, where it oppositely predicted all data points of the test set as relevant, leading to a recall of 1. In contrast, the Bernoulli Naive Bayes model incorrectly labeled several GDPR and smart meter data points as relevant, showcasing a lot of false positive predictions. Gaussian Naive Bayes shared this tendency, while also incorrectly classifying some Australian data points. For a granular view of the misclassification made by each model, one can use the frontend provided in the project's GitHub [16].

5 Comparison of Approaches

As this was a project within the seminar where we had different approaches to the same problems. This study also wanted to compare the models of my machine learning approach with results of large language model and rule-based approaches.

5.1 Large Language Model

For a comparison with the LLM approach, GPT-3.5 was fine-tuned for this study. The selection of GPT-3.5 over alternatives like LLAMA2 was influenced by familiarity with GPT and the technical limitations encountered with LLAMA2, particularly its incompatibility with the M1 chip of my laptop. This choice was also driven by a desire to achieve results comparable to those documented in existing literature [17].

The fine-tuning process involved adapting GPT-3.5 to enhance its proficiency in understanding and analyzing the nuances of legal language and process-specific terminologies. The training consists of a series of dialogs with 3 elements, similar to this work [17]. Firstly, it starts with an explanation of the task to be performed, then includes a process description, the legal text to classify and a binary classification label indicating the relevance of the legal text to the process description (0 for 'Irrelevant', 1 for 'Relevant'). Only in the training data, the completion of the dialog which is the actual label of the data point is included. For testing, the maximum token parameter was set to 1 to focus solely on binary classification, and a conservative temperature setting of 0.2 was employed to minimize response randomness.

While only prompt engineering, as suggested in [2], was considered, it was beyond the scope of a comparison approach and fine-tuning lead faster to comparable results. To ensure the robustness of the findings, GPT-3.5's responses were validated through multiple queries, consistently yielding identical results.

5.2 Rule-Based

In addition to the LLM approach, this study implemented two rule-based methodologies for comparative analysis. The first approach uses the cosine similarity to classify legal texts based on their relevance to given process descriptions. The core idea was to calculate the mean cosine similarity between the process description and the legal text for each label within each process. This mean value then served as a threshold for classifying the labels. Specifically, a legal text was classified as relevant if its cosine similarity with the process description was greater than the mean similarity for label 0 for that process, and vice versa. Notably, only one process from the Australian data had a larger mean similarity for label 0 than label 1.

Building on the rule-based methodology, the second approach was inspired by the UMAP clustering and K-Nearest Neighbors (KNN) algorithms. This approach first involved clustering the data by process and label, then calculating the mean centroid for each cluster, therefore having 2 clusters per process for label 0 and label 1. The classification of a legal text was determined by calculating the distance between these mean centroids of label 0 and label 1 for its respective process. The label corresponding to the nearest centroid was then assigned to the data point. One also has to mention that the rule-based approaches are rather simple. A far more complex one was conduced in [17].

5.3 Outlook Evaluation

Quantitative results can be seen in Table 2. The fine-tuned GPT-3.5 shows the similar results to the Decision Tree, which shows that is it quite capable of understanding nuances in the data and performing the text classification task. The rule-based mean centroid approach has the highest recall of 0.93 on TF-IDF and GPT embeddings. This aligns with the insights gathered from the UMAP visualization, as these embeddings, created the best clustering. On the other embeddings, the results ranged from 0.77 to 0.87. The cosine similarity approach however performed quite similar based with 0.87 recall being the best and on other embeddings spanning from 0.7 to 0.87. It appears that the cosine similarity is informative as a feature for the text classification task, contradictory to the feature importance. However, the sustainability of these rule-based methods with different datasets remains questionable, as their effectiveness heavily relies on the quality of clustering.

Table 2 Outlook Approaches - Recall Results

Model	Embedding Type	Recall
GPT 3.5	GPT	0.73
Mean Centroid	TF-IDF/GPT	0.93
Cosine Similarity	GPT	0.87

Qualitatively, GPT-3.5 demonstrated difficulties with some, but not all, of the problematic nine data points from the Australian travel insurance claim process. It also wrongly predicted relevance in two instances within the 'Know Your Customer' process. The mean centroid method encountered a completely different set of misclassifications. This could be explained by the method being purely a numerical analysis of vectors. Similarly, the cosine similarity approach tended to overestimate relevance on some of the data points, which contributing to its highly lower recall score.

6 Discussion

The findings underscore the complexity of applying various machine learning and NLP techniques to the classification of legal texts for business process compliance. One key insight was that there is a clear difference in disparity in learning capabilities among the models. Notably, models such as SVD, Perceptron, SGD and the RNN showed that they are not able to learn from the data. This could be attributed to the complexity of the legal text itself. Legal documents present a unique challenge due to their mix of legal and domain terminology. In [3, 14], the authors explained this on the example of cyber law. A legal text of cyber law would examine how legal concepts, such as obligations or rights, are relevant to domain terminology like Wi-Fi and passwords. Further, for this problem, there was also additional complexity with process vocabulary from different domains.

One also has to mention, that there are clear limitations of this study, particularly within the size of the data and its origin domains. The gold standard is limited, and there may not be generalizability of the findings to broader applications. Additionally, the dataset's construction, where the proportion of relevant data was set at 10 percent, based on findings within Australian datasets. This may not adequately represent of real-world scenarios, especially since the dataset was compiled by an individual lacking legal expertise. Such a constraint in dataset design could significantly influence the applicability and validity of the study's outcomes in practical settings.

In addition, unidentified patterns within the data could have also influenced the outcome. Another notable limitation comes from the exclusive use of the process descriptions to represent the business processes. The process description is subjective to my limited knowledge of the process and may not reflect the actual process well.

Nevertheless, this project has demonstrated that machine learning approaches can obtain results comparable to those achieved by advanced techniques like fine-tuning LLMs. This also has implications to real-world applications of these models. From a practical standpoint, the complexity and costs of these approaches differ vastly. Constructing a machine learning pipeline, I would argue, is more complex than fine-tuning a GPT model. Also, the fine-tuning costs of GPT-3.5 are approximately 8 euros, in contrast to embeddings from GPT costing approximately for this project. Another consideration is data quality, as GPT models, with their large input size and pre-training, may not require as extensive a dataset to produce reasonable results.

Future research should aim to employ a more varied and extensive dataset to validate and generalize the findings of this study. Diversifying the dataset will help in overcoming the current limitations and provide more robust and widely applicable insights.

7 Outlook

There are various advancements of this project, both on the level of embeddings and on model approaches. An advanced approach could be to use larger and more specific LLMs such as S-Bert [15], which is fine-tuned for sentence level embeddings or GPT-4 instead of GPT-3.5
For the deep learning approach, different architectures may show improved results. While RNN struggled to learn in this context, architecture such as Long Short-Term Memory (LSTM) or even autoencoders could be considered. The benefit of autoencoders is that they could first be trained to understand legal terminology before being applied to this classification task. Further, directed graph neural networks are able to indicate relationships between data. This is interesting for this problem since the legal text is classified in relation to the process descriptions.

Improvements on the embedding level can be made by using the newly released embedding models from OpenAI or selecting those that perform well on the Hugging Face leaderboard [4, 13].

Apart from the improvements, an interesting future topic could be to focus more on prompt engineering and compare Bert and GPT for this BPM task. In [2] there is a mention that there is a lack of a semantic comparison between different pre-trained language models, namely Bert and GPT for BPM tasks.

In conclusion, combining machine learning or fine-tuned approaches with the specialized knowledge of legal experts can potentially create a powerful tool for businesses. Such a tool could significantly decrease manual workload and semi-automate regulatory compliance efforts, although its development and implementation will require ongoing, nuanced research, particularly involving more diverse data sources for broader applicability and validation of the findings.

References

[1] Guido Boella, Luigi Di Caro, and Llio Humphreys. "Using classification to support legal knowledge engineers in the Eunomos legal document management system". In: *Fifth international workshop on Juris-informatics (JURISIN)*. 2011.

[2] Kiran Busch et al. "Just tell me: Prompt engineering in business process management". In: *International Conference on Business Process Modeling, Development and Support*. Springer. 2023, pp. 3–11.

[3] Haihua Chen et al. "A comparative study of automated legal text classification using random forests and deep learning". In: *Information Processing & Management* 59.2 (2022), p. 102798.

[4] Huggingface Leaderboard for Classification Embeddings. In: URL: https://huggingface.co/spaces/mteb/leaderboard.

[5] Natural Language Processing Business Process Management data. In: URL: https://github.com/EduardoBre/nlp-bpm-data/tree/main.

[6] Rajaa El Hamdani et al. "A combined rule-based and machine learning approach for automated GDPR compliance checking". In: *Proceedings of the Eighteenth International Conference on Artificial Intelligence and Law*. 2021, pp. 40–49.

[7] Mustafa Hashmi et al. "Are we done with business process compliance: state of the art and challenges ahead". In: *Knowledge and Information Systems* 57.1 (2018), pp. 79–133.

[8] Thorsten Joachims. "Text categorization with support vector machines: Learning with many relevant features". In: *European conference on machine learning*. Springer. 1998, pp. 137–142.

[9] Qian Li et al. "A survey on text classification: From shallow to deep learning". In: *arXiv preprint arXiv:2008.00364* (2020).

[10] Richard Liaw et al. "Tune: A Research Platform for Distributed Model Selection and Training". In: *arXiv preprint arXiv:1807.05118* (2018).

[11] Christian J Mahoney et al. "A framework for explainable text classification in legal document review". In: *2019 IEEE International Conference on Big Data (Big Data)*. IEEE. 2019, pp. 1858–1867.

[12] Leland McInnes et al. "UMAP: Uniform Manifold Approximation and Projection". In: *The Journal of Open Source Software* 3.29 (2018), p. 861.

[13] OpenAI New Embeddings Models. In: URL: https://openai.com/blog/new-embedding-models-and-api-updates.

[14] Adeline Nazarenko and Adam Wyner. "Legal NLP introduction". In: *Traitement automatique des langues* 58.2 (2017), pp. 7–19.

[15] Nils Reimers and Iryna Gurevych. "Sentence-bert: Sentence embeddings using siamese bert-networks". In: *arXiv preprint arXiv:1908.10084* (2019).

[16] Project Repository. In: URL: https://github.com/patrickahrend/TUM-NLP-Praktikum.

[17] Catherine Sai et al. "Identification of Regulatory Requirements Relevant to Business Processes: A Comparative Study on Generative AI, Embedding-based Ranking, Crowd and Expert-driven Methods". In: *arXiv preprint arXiv:2401.02986* (2024).

[18] Krishna Sapkota et al. "Extracting meaningful entities from regulatory text: Towards automating regulatory compliance". In: *2012 Fifth IEEE International Workshop on Requirements Engineering and Law (RELAW)*. IEEE. 2012, pp. 29–32.

[19] Han Van der Aa, Henrik Leopold, and Hajo A Reijers. "Checking process compliance against natural language specifications using behavioral spaces". In: *Information Systems* 78 (2018), pp. 83–95.

[20] Han Van der Aa et al. "Challenges and opportunities of applying natural language processing in business process management". In: *COLING 2018: The 27th International Conference on Computational Linguistics: Proceedings of the Conference: August 20-26, 2018 Santa Fe, New Mexico, USA*. Association for Computational Linguistics. 2018, pp. 2791–2801.