Cross-Domain Generalization and Knowledge Transfer in Transformers Trained on Legal Data

Jaromír ŠAVELKA [a,1], Hannes WESTERMANN [b], and Karim BENYEKHLEF [b]

[a] *School of Computer Science, Carnegie Mellon University*

[b] *Cyberjustice Laboratory, Faculté de droit, Université de Montréal*

Abstract. In this paper we examine the ability of pre-trained language models to generalize beyond the legal domain and dataset they were trained on. We transform three separate datasets of annotated legal cases with different type systems to use a single type system (namely, determining whether a sentence describes facts or not). Then, we train SVM and RoBERTa models on the different source datasets, and measure their performance on the remaining datasets. We also create models trained on combined datasets, and measure their performance. Overall, we find that the performance of the models is surprisingly high, despite being trained and evaluated between datasets from different domains and jurisdictions. The RoBERTa model seems to be able to benefit from its pre-trained language representation to better capture the abstract idea behind the new type system. Further, the models benefit from training on multiple datasets. This potentially opens the door for models trained on separate datasets that could perform better in previously unseen domains, and to overcome cold-start problems when using active learning methods.

Keywords. Transformers, cross-domain generalization, case-law, rhetorical roles, sentence classification

1. Introduction

In this paper we examine the ability of pre-trained language models to generalize beyond the domain and dataset they were trained on in the legal domain. We explore if and how different datasets and type systems can be re-cast in such a way that a language model trained on one dataset could be successfully utilized on other datasets. As high-quality legal data sets are scarce, the efficient use of those that are available is of utmost importance in AI & Law research.

Written legal cases typically follow a certain pattern in their reasoning. For example, legal cases often contain the following sections:

1. Claims of the plaintiff and responses/counter-claims of the defendant

[1] E-mail: jsavelka@andrew.cmu.edu

Proceedings of the 2020 Workshop on Automated Semantic Analysis of Information in Legal Text (ASAIL)

2. A description of factual circumstances of the case as seen by the parties and as determined by the court
3. Legal rules applicable to the factual circumstances
4. Application of the rules to the factual circumstances
5. Court's conclusions and outcomes of the case

The sections play different roles in a decision, and could carry different meaning dependent on the context in which they are read. A judge, a law student, a party to a dispute, or a legal professional could all read the different sections with different focus depending on what they need to learn. It is an indispensable skill for any professional lawyer to quickly identify and understand the different sections of a case.

Due to its prominence the task has attracted abundant research in AI & Law (see e.g. [14,17,1]). Researchers have worked on automating the identification of the different sections, using sophisticated ML models, on a sentence, paragraph or other level. The automatic segmentation of cases could serve as a reading aid for actors in the legal world, as a way to improve legal search results, and as an important prerequisite for future research in AI & Law focused on specific sections of the cases.

Typically, research in prediction of rhetorical roles includes dataset creation and training of ML models on that dataset. The dataset usually focuses on one or several domains. The model is then tested against a test set, typically sampled randomly from the dataset. This shows the predictive capability of the model on test data from the same distribution as the training data.

However, it can be difficult to evaluate how well a model would perform if applied to other, previously unseen datasets that might have different distributions, e.g. whether the model is able to generalize beyond the annotated dataset. For example, it is possible that a model has only learned a specific characteristic of a single domain or jurisdiction to describe factual circumstances, rather than the abstract idea of what a sentence describing factual circumstances looks like. If the goal is to apply the model in a general fashion, on cases from different jurisdictions, domains and countries, it is therefore important to evaluate the performance of the system on datasets it has not seen at all during training.

Evaluating the models on multiple datasets, however, is not trivial. Firstly, publicly available datasets are scarce in AI & Law. Further, the datasets that are available typically use custom type systems, making cross-dataset evaluation challenging. In this paper, we identify an example meta type system, that generalizes across three datasets. We use the meta type system to train models on data from different domains and jurisdictions and assess performance of the models across the domains.

We compare the ability of different types of ML models to generalize from one dataset to another. First, as a baseline we use a Support Vector Machines classifier which learns the correlation between word and/or n-gram occurrences and a certain label. Then, we evaluate the more recently developed BERT model (bidirectional encoder representation from transformers). This model has been pre-trained on massive corpora of text, to learn a language model. It can then be further fine-tuned on a down-stream task, such as text classification in a specific domain, achieving strong results with little training data by leveraging the previously learned language representation.

2. Hypotheses

By conducting the experiments described in Section 4, we investigate the following hypotheses:

- *H1* - Machine learning models are able to generalize the knowledge learned from one dataset and apply it successfully on another dataset.
- *H2* - The BERT model is able to leverage its understanding of language to abstract beyond the specific domain vocabulary, thereby showing better performance in generalizing across the domains than the SVM model.
- *H3* - Training the models on pooled data from multiple datasets improves performance and results in more robust models.

Showing that the models generalize well among different domains could have important implications for the real-world applicability of the methods. Further, they could be employed in active or interactive learning settings, where annotators are presented with predictions by the model and confirm/correct these to create accurate models. This could drastically lower the number of expensive annotations that need to be performed to start the annotation of novel datasets.

3. Related Work

The identification of rhetorical rules of sentences in legal cases has been investigated by several researchers. The authors of [14] used Conditional Random Field models with custom features to predict the rhetorical role of sentences in three distinct domains. Walker et. al [17] trained a number of machine learning models to predict the rhetorical role of sentences in decision from the U.S. Board of Veteran Appeals and compared them to rule-based approaches. In [22] the researchers created an interface to easily create boolean search rules for sentence classification. Bhattacharya et al. [1] created a dataset of decisions from the Indian Supreme Court from 5 domains, and trained Hierarchical BiLSTM CRF models, achieving good results across domains. These papers trained and evaluated the models on the same datasets. The contribution in this paper is that we investigate the capability of models trained on one dataset to generalize to *other* datasets. Therefore, we re-cast three datasets (including the publicly available datasets from [17] and [1]) into a meta type system. Then we investigate the ability and respective performance of SVM and RoBERTa models trained on a single or multiple datasets to generalize to the other datasets. To our knowledge, training models for the classification of legal texts on one dataset and evaluating the performance on other datasets has not been extensively investigated previously, and is thus a novel contribution.

We further investigate whether BERT-based models are able to exploit their pre-trained language model to achieve higher performance in generalizing across datasets. The use of BERT on legal texts has numerous examples. In [3] BERT is evaluated on classification of claim acceptance given judges' arguments. A task of retrieving related case-law similar to a case decision a user provides is tackled in [11]. In [2] BERT is evaluated as one of the approaches to predict court decision outcome given the facts of a case. BERT has been successfully used for classification of legal areas of Supreme Court judgments. [8] The authors of [10] combine BERT with a simple similarity measure to tackle the challenging task of case law entailment. BERT was also used in learning-to-rank settings for retrieval of legal news [12] and case-law sentences interpreting statutory concepts [13]. In [20], the researchers investigate the additional fine-tuning of language models on related tasks to improve performance in the analysis of legal entailment.

We show that models pre-trained on one dataset can to some extent perform predictions on other datasets. This could be used to bootstrap new datasets in other domains, decreasing the effort of annotation. This follows a steady line of work in AI & Law on making annotation more effective. Westermann et al. [22] proposed and assessed a method for building strong, explainable classifiers in the form of Boolean search rules. Employing an intuitive interface, the user develops Boolean rules for matching instead of annotating the individual sentences. In [21] a method for using pre-trained language models to identify semantically similar sentences suitable for annotation is proposed. Šavelka and Ashley [16] evaluated the effectiveness of an approach where a user labels the documents by confirming (or correcting) the prediction of a ML algorithm (interactive approach). The application of active learning has further been explored in the context of classification of statutory provisions [19] and eDiscovery [4,5,7].

4. Experimental Design

In order to evaluate the ability of models to generalize beyond a single domain, we employ an experimental design consisting of several steps. First, we identify three datasets containing a categorization of sentences by rhetorical role (Section 4.1). Then, we identify a meta type system that we can transform all the type systems into to create a task shared among the datasets (Section 4.2). We then fine-tune and evaluate a pre-trained language model and a Support Vector Machine model (Section 4.4) on different combinations of these datasets (Sections 4.4 and 4.5).

4.1. Data

In this work we utilize three datasets. The first one comes from [17]. The authors analyzed 50 fact-finding decisions issued by the U.S. Board of Veterans' Appeals ("BVA") from 2013 through 2017, all arbitrarily selected cases dealing with claims by veterans for service-related post-traumatic stress disorder (PTSD). For each of the 50 BVA decisions in the PTSD dataset, the researchers extracted all sentences addressing the factual issues related to the claim for PTSD, or for a closely-related psychiatric disorder. These were tagged with the rhetorical roles [18] the sentences play in the decision. These were Finding, Reasoning, Evidence, Legal Rule, and Citation. [2]

The second dataset also focuses on the rhetorical roles of sentences. Bhattacharya et al. [1] analyzed 50 opinions of the Supreme Court of India ("ISC"). The cases were sampled from five different domains in proportion to their frequencies (criminal, land and

[2] Dataset available at `https://github.com/LLTLab/VetClaims-JSON`

Table 1. Label distributions of the datasets used in this work.

	BVA	CB	ISC	Total
Facts	2,420	4,182	2,219	8,821
Non-Facts	3,733	5,783	9,380	18,896
Total	6,153	9,965	11,599	27,717

property, constitutional, labor and industrial, and intellectual property). The decisions were split into 9,380 sentences and manually classified into one of the seven categories according to the rhetorical roles they play in a decision. These were Facts, Ruling (lower court), Argument, Ratio, Statute, Precedent, Ruling (present court).[3]

We also created a brand-new data set by scraping the case briefs from a publicly available Case Brief Summary database ("CB").[4] The case briefs were categorized in terms of the areas of regulation, such as administrative law, business law, or criminal law (11 categories in total). In total, we were able to obtain 715 unique case briefs. The case briefs are structured into a number of sections with headings. We extract the sections based on an extensive battery of regular expressions to segment the retrieved briefs into individual sections. While there were over 100 unique section heading names we were able to identify six main types to which we could map many of the different variations (e.g., all of Legal Issue, Issues, and Issue map to a single category). We applied a specialized legal case sentence boundary detection system [15] to segment the sections into sentences. This results in the dataset comprising 9,965 sentences, each with one of the six labels corresponding to the section of the brief where the sentence occurred. The possible sections are Facts, Issue, Conclusion, Procedural History, Reasoning, and Rule.

4.2. Task

The datasets described in 4.1 use different annotation type systems and stem from different domains and decision makers. However, parts of the type systems overlap, making it possible to map certain types to a new meta type system that allows the training of models and evaluation of the trained models between datasets.

In order to map the source type systems into a single type system, we identify the label of the source datasets that most closely corresponds to a description of factual circumstances in the cases. These sentences are tagged as "Facts". All other sentences are tagged as "Non-facts". The labels that we chose to represent facts in the different datasets are as follows:

1. BVA - "Evidence" - defined as a sentence that primarily states the content of the testimony of a witness, states the content of documents introduced into evidence, or describes other evidence.
2. CB - "Facts" - Sentences stemming from a section with the title "Facts", typically describing the factual background to a case.
3. ISC - "Facts" - refers to the chronology of events that led to filing the case, and how the case evolved over time in the legal system (e.g., First Information Report at a police station, filing an appeal to the Magistrate).

[3] Dataset available at `github.com/Law-AI/semantic-segmentation`

[4] `http://www.casebriefsummary.com/`; Note that as of 2020-11-29 the site is no longer available.

After transforming the datasets into this schema, the distributions of Facts vs Non-Facts are as described in Table 1 for each dataset. While the source type systems are different, and the definitions of the corresponding types are not the same, for the purposes of the training and evaluation we assume that the new type system describes the same task when applied across the three datasets (i.e., we assume that we have transformed three related tasks into a unified one).

4.3. Dataset splits

We randomly split all the data sets into training, validation, and test sets using ratios of 50-25-25. The splitting for all four datasets is performed at the level of documents, i.e., all the sentences from a single document are in the same fold. We did not take the size of a document into account, meaning that the number of sentences could vary slightly between datasets. The training split of the datasets is used for training the models, the validation split is used for model selection and hyperparameter optimization, while the test set is used for final evaluation.

4.4. Models

In this work, we use the RoBERTa (a robustly optimized BERT pretraining approach) described in [9] as the starting point for our experiments.[5] Out of the available models we chose to work with the smaller roberta.base model that has 125 million parameters, in order to be able to iterate the experiments faster than the larger models. RoBERTa is using the same architecture as BERT [6]. However, the authors of [9] conducted a replication study of BERT pre-training and found that BERT was significantly undertrained. They used the insights thus gained to propose a better pre-training procedure.

As baseline, we use a Support Vector Machine (SVM) classifier. SVM constructs a hyper-plane in a high dimensional space, which is used to separate the classes from each other. As an implementation of SVM we use the scikit-learn's Support Vector Classification module.[6] As features we use the bag of words of (1-3)-grams weighted by TF-IDF.

4.5. Experiments

We train the models on all the different possible pools of training data. The possible pools are BVA, CB, ISC, BVA+CB, BVA+ISC, CB+ISC, and BVA+CB+ISC. Both RoBERTa and SVM are trained on each of these pools separately, and the results are presented for prediction on the test split of each dataset (Section 5).

In all the experiments, we fine-tune the base RoBERTa model, with a linear classification layer on top of the pooled output, for 10 epochs on the training splits of the selected datasets. We use the batch size of 8 which is the maximum allowed by our hardware setup (1080Ti with 11GB) given we set the length of a sequence to 512 (maximum). As optimizer we use Adam with initial learning rate set to $4e^{-5}$. We store a model's checkpoint after the end of each training epoch. The checkpoints are evaluated on the

[5] `github.com/pytorch/fairseq/tree/master/examples/roberta`

[6] `scikit-learn.org/stable/modules/generated/sklearn.svm.LinearSVC.html`

corresponding validation set. The model with the highest F_1 on the validation set is then selected as the one to make predictions on the test sets. The performance on the test sets is what we report in the results section.

The SVM is optimized via gridsearch over the few most important hyperparameters (C, class weight, number of iterations). The various hyperparameter combinations are evaluated against the validation dataset. The model with the highest F_1 on the validation set is then selected as the one to make predictions on the test sets. The performance on the test sets is what we report in the results section.

For evaluation we report the F1-measure (F_1), i.e., the traditional information retrieval measures, to evaluate performance of the trained models. Since the task is binary, the application of the measure is straightforward.

$$P = \sum_{i=1}^{|S|} \frac{TP}{TP+FP} \qquad R = \sum_{i=1}^{|S|} \frac{TP}{TP+FN} \qquad F_1 = \frac{2PR}{P+R}$$

TP stands for true positives, FP for false positives, and FN for false negatives.

5. Results

5.1. H1 - Machine learning models are able to generalize the knowledge learned from one dataset and apply it successfully on another dataset.

In order to evaluate this hypothesis, we investigate the scores in Table 2. Of course, the models with the highest scores are the models where the training data stems from the same datasets as the testing data (cells colored in gray). Further, the RoBERTa model performs better than the SVM model across the board, which is expected as RoBERTa is a much more complex model (compared to SVM) pre-trained for language understanding.

To investigate H1 we assess the scores for models trained on one dataset when applied an another dataset. For the RoBERTa model, the maximum drop in performance is 0.27 in F1-score, when trained on BVA and evaluated against ISC. In general, the drop is smaller, with strong performances even when applied on different datasets. The fact that the RoBERTa model trained on the CB dataset achieves and F1-score of .84 on the BVA dataset, despite the different characteristics of the datasets, is quite impressive.

The SVM model seems to have problems learning the distributions of the ISC dataset, resulting in low performance when trained on this dataset. This might be due to the transformed ISC dataset being more imbalanced than the other two (see Table 1. Overall, the RoBERTA model trained on the CB dataset performs the strongest on average. This might be due to the large distribution of domains contained in this dataset.

5.2. H2 - The BERT model is able to leverage its understanding of language to abstract beyond the specific domain vocabulary, thereby generalizing better across the domains than the SVM model.

The RoBERTa model appears to transfer the knowledge learned on any one dataset to prediction on the other two much better than the baseline SVM model (Table 2). For

Table 2. F_1 scores for models trained on specific dataset training pools (rows) predicting on testing datasets (columns), compared between SVM and RoBERTa. Grey cells indicate that the training data includes the target dataset. values in parentheses indicate difference to the same model type trained on the training data matching the test data.

	BVA (test)		CB (test)		ISC (test)		Avg (test)	
	SVM	RoBERTa	SVM	RoBERTa	SVM	RoBERTa	SVM	RoBERTa
BVA (train)	.92	.94	.67 (-.11)	.71 (-.12)	.44 (+.03)	.44 (-.15)	.68	.70
CB (train)	.60 (-.22)	.84 (-.10)	.78	.83	.50 (+.09)	.57 (-0.02)	.63	.75
ISC (train)	.11 (-.81)	.67 (-.27)	.19 (-.59)	.65 (-.18)	.41	.59	.24	.64

Table 3. F1-scores for models trained on specific dataset training pools (rows) predicting on testing datasets (columns), compared between SVM and RoBERTa. First row ("Matching train") displays performance of models when trained and evaluated on same dataset. Parentheses show difference in F1-score to model trained and evaluated on the same dataset.

	BVA (test)		CB (test)		ISC (test)	
	SVM	RoBERTa	SVM	RoBERTa	SVM	RoBERTa
Matching train	.92	.94	.78	.83	.41	.59
BVA+CB	.91 (-.01)	.93 (-.01)	.78 (+.00)	.84 (+.01)	.41 (+.00)	.55 (-.04)
BVA+ISC	.91 (-.01)	.93 (-.01)	.46 (-.32)	.75 (-.09)	.46 (+.05)	.59 (+.00)
CB+ISC	.55 (-.37)	.85 (-.09)	.71 (-.07)	.82 (-.01)	.53 (+.12)	.60 (+.01)
BVA+CB+ISC	.90 (-.02)	.93 (-.01)	.68 (-.10)	.82 (-.01)	.50 (+.09)	.59 (+.00)

example, RoBERTa trained on CB performs at $F_1 = 0.84$ on BVA. This is a sizeable performance hit when compared to RoBERTa trained on BVA ($F_1 = 0.94$). However, the SVM model goes from $F_1 = 0.92$ to $F_1 = 0.60$. While it is customary for BERT models to perform better than traditional ML models, the sizeable difference of .24 in this instance might indicate that there is a qualitatively different understanding going on, as RoBERTa can understand the semantic meaning rather than the specific vocabulary used. The difference does not hold up for all datasets, however.

5.3. H3 - Training the models on pooled data from multiple datasets improves performance and results in more robust models.

Table 3 shows the results for models trained on the matching training data (e.g. the model trained on BVA evaluated on BVA) in the top row, and models trained on multiple pooled datasets and evaluated on a single test dataset below. Overall, there does not seem to be a strong improvement when adding another dataset to the training data.

Concerning robustness, the results suggest that RoBERTa is much better at handling data coming from different datasets, achieving impressive performances on both datasets. This is not the case for the SVM baseline. The RoBERTa model trained on all three datasets is especially impressive as it achieves within 0.01 of all the individual datasets simultaneously. Such a model would likely generalize better to other domains (not considered here) as compared to the model trained exclusively on one of the three datasets.

This robustness can further be illustrated by the fact that combining data from the two non-target datasets seems to yield higher performance on the target datasets than training just on either of the non-target datasets. For example, training the RoBERTa model on both BVA and ISC yields higher performance when predicting CB (.75) than using either BVA (.71) or ISC (.65) for training on their own.

6. Discussion

The results reveal several interesting properties of the evaluated models. First of all, they show that it is indeed possible to train a model on one dataset, and use it for prediction on another dataset with decent performance. This holds especially true for the powerful RoBERTa model. It should be noted that the models in section 2 have never seen any sample from the target dataset, except when the source dataset and the target dataset are the same, and still achieve strong results. This shows a clear path towards using a model trained on one dataset to bootstrap a new dataset for active or interactive learning purposes. It also shows that the tasks defined in terms of their respective type systems are related, despite being created in different contexts and jurisdictions. The relatedness enables ML models to be successfully applied across domains.

Further, several interesting observations can be made when comparing the performance of the SVM and RoBERTa model. In several instances, the drop between the model trained and evaluated on the same dataset and a model trained on another dataset is larger for the SVM model. This could indicate that RoBERTa leverages its language understanding to abstract beyond the specific vocabulary used.

Finally, the RoBERTa models trained on multiple datasets seem to retain high performance on the specific datasets they are trained on, while also improving performance on datasets they were not trained on. This is another promising result showing that the models can learn patterns even from datasets created in different contexts and jurisdictions, and use the additional data to create robust and strong classifiers.

7. Conclusion and Future Work

In this paper, we have provided an example of how datasets created in different contexts can be re-cast to support the same task, and utilized towards a common goal. The pre-trained language models seem to be able to understand the underlying idea behind a task to a larger extent than SVM models, resulting in higher performance when applied to different datasets than they were trained on. Further, training the models on several domains improved robustness and to some extent performance. Overall, these results suggest that pre-trained language models have several desirable properties when training on multiple datasets, making them an ideal tool for efficient utilization of the existing AI & Law datasets to support future research in novel tasks.

We observed that it is possible to train a model on one dataset, and use it for prediction on another one with decent performance. We plan to investigate whether this could be used to overcome the cold start problem to improve efficacy in an active learning context. Another angle of investigation is to experiment with the transformation of the datasets, and identifying other pertinent documents, potentially even across languages. Finally, using larger, state-of-the art models is warranted to investigate additional improvements in predictive performance.

References

[1] Bhattacharya, P., S. Paul, K. Ghosh, S. Ghosh, and A. Wyner. “Identification of Rhetorical Roles of Sentences in Indian Legal Judgments.” *JURIX 2019*, IOS Press.

[2] Chalkidis, I., I. Androutsopoulos, and N. Aletras. “Neural legal judgment prediction in english.” *Proceedings of the 57th Annual Meeting of the ACL*. 2019.

[3] Condevaux, C., S. Harispe, S. Mussard, and G. Zambrano. “Weakly Supervised One-Shot Classification Using Recurrent Neural Networks with Attention: Application to Claim Acceptance Detection.” *JURIX*. 2019.

[4] Cormack, G., and M. Grossman. “Scalability of continuous active learning for reliable high-recall text classification.” In *Proc. 25th ACM Int’l Conf. on Info. & Knowledge Management*. 2016.

[5] Cormack, G., and M. Grossman. “Autonomy and reliability of continuous active learning for technology-assisted review.” *arXiv preprint arXiv*:1504.06868 (2015).

[6] Devlin, J., M. Chang, K. Lee, and K. Toutanova. “Bert: Pre-training of deep bidirectional transformers for language understanding.” *NAACL*, 2019.

[7] Hogan, C., R. Bauer, and D. Brassil. “Human-aided computer cognition for e-discovery.” In *ICAIL*. 2009.

[8] Soh, J., H. K. Lim, and I. E. Chai. “Legal Area Classification: A Comparative Study of Text Classifiers on Singapore Supreme Court Judgments.” *NLLP Workshop*, 2019.

[9] Liu, Y., M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov. “Roberta: A robustly optimized bert pretraining approach.” arXiv preprint arXiv:1907.11692 (2019).

[10] Rabelo, J., M. Kim, and R. Goebel. “Combining similarity and transformer methods for case law entailment.” *ICAIL*. 2019.

[11] Rossi, J., and E. Kanoulas. “Legal Search in Case Law and Statute Law.” *JURIX*. 2019.

[12] Sanchez, L., J. He, J. Manotumruksa, D. Albakour, M. Martinez, and A. Lipani. “Easing Legal News Monitoring with Learning to Rank and BERT.” *European Conference on Information Retrieval*. Springer, Cham, 2020.

[13] Savelka, J. Discovering sentences for argumentation about the meaning of statutory terms. *Diss. University of Pittsburgh*, 2020.

[14] Savelka, J., and K. D. Ashley. “Using CRF to detect different functional types of content in decisions of united states courts with example application to sentence boundary detection.” *ASAIL 2017*.

[15] Savelka, J., Walker, V. R., Grabmair, M., & Ashley, K. D. (2017). Sentence boundary detection in adjudicatory decisions in the united states. *Traitement automatique des langues*, 58, 21.

[16] Savelka, J., G. Trivedi, and K. Ashley. “Applying an interactive machine learning approach to statutory analysis.” *JURIX*. 2015.

[17] Walker, V. R., K. Pillaipakkamnatt, A. M. Davidson, M. Linares, and D. J. Pesce. “Automatic Classification of Rhetorical Roles for Sentences: Comparing Rule-Based Scripts with Machine Learning.” *ASAIL 2019*.

[18] Walker, V. R., J. H. Han, X. Ni, and K. Yoseda. “Semantic types for computational legal reasoning: propositional connectives and sentence roles in the veterans’ claims dataset.” *ICAIL*. 2017.

[19] Waltl, B., J. Muhr, I. Glaser, G. Bonczek, E. Scepankova, and F. Matthes. “Classifying Legal Norms with Active Machine Learning.” *JURIX*. 2017.

[20] Westermann, H., J. Savelka, and K. Benyekhlef. “Paragraph Similarity Scoring and Fine-Tuned BERT for Legal Information Retrieval and Entailment.” *COLIEE*. 2020.

[21] Westermann, H., J. Savelka, V. Walker, K. Ashley, and K. Benyekhlef. “Sentence Embeddings and High-speed Similarity Search for Fast Computer Assisted Annotation of Legal Documents.” *JURIX*. 2020.

[22] Westermann, H., J. Savelka, V. Walker, K. Ashley, and K. Benyekhlef. “Computer-Assisted Creation of Boolean Search Rules for Text Classification in the Legal Domain.” *JURIX*. 2019.