Transferring Legal Natural Language Inference Model from a US State to Another: What Makes It So Hard?

Alice Saebom Kwak[1], **Gaetano Vincent Forte**[2], **Derek E. Bambauer**[2], **Mihai Surdeanu**[3]

[1]Department of Linguistics, The University of Arizona
[2]James E. Rogers College of Law, The University of Arizona
[3]Department of Computer Science, The University of Arizona
{alicekwak, fortegv, derekbambauer, msurdeanu}@arizona.edu

Abstract

This study investigates whether a legal natural language inference (NLI) model trained on the data from one US state can be transferred to another state. We fine-tuned a pre-trained model on the task of evaluating the validity of legal will statements, once with the dataset containing the Tennessee wills and once with the dataset containing the Idaho wills. Each model's performance on the in-domain setting and the out-of-domain setting are compared to see if the models can across the states. We found that the model trained on one US state can be mostly transferred to another state. However, it is clear that the model's performance drops in the out-of-domain setting. The F1 scores of the Tennessee model and the Idaho model are 96.41 and 92.03 when predicting the data from the same state, but they drop to 66.32 and 81.60 when predicting the data from another state. Subsequent error analysis revealed that there are two major sources of errors. First, the model fails to recognize equivalent laws across states when there are stylistic differences between laws. Second, difference in statutory section numbering system between the states makes it difficult for the model to locate laws relevant to the cases being predicted on. This analysis provides insights on how the future NLI system can be improved. Also, our findings offer empirical support to legal experts advocating the standardization of legal documents.

1 Introduction

This study investigates whether a legal natural language inference model trained on the data from one US state can be transferred to different US states. Natural Language Inference (NLI) is a textual reasoning task determining whether a premise entails, contradicts, or is neutral to a hypothesis. Kwak et al. (2022) suggests that the validity assessment of legal documents (e.g, wills) can be framed into a NLI task. It adapts the traditional NLI approach to fit in the legal domain. Unlike the general setting where a model gets two inputs (i.e., premise and hypothesis), the legal NLI model requires three inputs: a legal document of which validity would be assessed, a condition (a circumstance relevant to the validity assessment of the legal document), and a law. This is an intriguing adaptation, but it also introduces the potential of overfitting on the law texts.

Developing legal NLI models that can evaluate the validity of legal documents provides much benefit to legal professionals and anyone involved in writing such legal documents. Assessing the legal documents' validity through legal NLI models can reduce the time and resource required for review. It can also increase the validity of legal documents by preventing any errors at creation time. In addition, legal NLI models can serve as foundations for downstream tasks such as legal document review automation, electronic will system, and smart contract.

As US states have different legal systems, there is no guarantee that a legal NLI model trained on one state would work in another state. However, training a model with data from all US states is a time and labor intensive task. Further, it is often the case that we have access to data from only a few states. Given these practical difficulties, transferring a model trained on one US state to other states is a great alternative to training a model on every state from scratch. To this end, we explore whether a legal NLI model trained on one US state is transferable to another state. We also conduct error analysis to identify any pitfalls in transferring a model trained on one state to another.

We focus on a specific task in this study to evaluate a validity of a legal will statement. We chose to focus on this task for two reasons. First, probate code is one of the fields where US states are divided into two groups: states which have adopted Uniform Probate Code (UPC) in full vs. states which

Proceedings of the Natural Legal Language Processing Workshop 2023, pages 215–222

have not adopted the full UPC. This clear division makes a good testing case for our study. We tested if a model trained on the data from a state belonging to the former group (i.e., Idaho) works for the data from a state belonging to the latter group (i.e., Tennessee), and vice versa. Another reason is that this task is highly practical. Wills are important legal documents that allow people to maintain control over their assets. Unlike most legal documents, wills are commonly written by/for people without legal training. However, will execution/probate is not always straightforward, and there is a risk that mistakes in will writing or execution procedure invalidate parts of the will. Developing a model that can evaluate the validity of a will statements can help preventing such mistakes. In addition, as previously mentioned, the model can serve as a foundation for developing electronic will system. Given all the benefits that the model could provide, we decided to conduct our study on this task.

The main contribution of our study is as follows:

- We investigate domain transfer between two US states for a language model fine-tuned for legal NLI. We found out that such a legal NLI model trained on one state can be mostly transferred to another state. However, it is clear that the model's performance drops in the cross-state setting. The F1 scores of the Tennessee model and the Idaho model are 96.41 and 92.03 when predicting the data from the same state, but they drop to 66.32 and 81.60 when predicting the data from another state (i.e., Tennessee model predict Idaho data and vice versa).

- We conducted an error analysis on the model's cross-state predictions and identified two sources of error. We found out that stylistic differences between state laws (e.g., terms, formats, capitalization) and differences in statutory section numbering formats can be obstacles to transferring a model trained on one state to another.

2 Related Works

2.1 NLI in the legal domain

NLI in the legal domain is gaining attention in the recent years. Koreeda and Manning (2021) presents a dataset for document level natural language inference for contracts. Bruno and Roth (2022) introduces LawngNLI constructed from US legal opinions. It is a long-premise benchmark for in-domain generalization from short to long contexts and for implication-based retrieval. Mathur et al. (2022) presents a new document-level natural language inference model using optimal evidence selection. The study also proposes a new dataset called CaseHoldNLI on the task of legal judicial reasoning, which is used for the model testing. Kwak et al. (2022) introduces a legal NLI dataset for the validity assessment of legal will statements. Despite this increased interest in legal NLI tasks, there is no prior work that attempted cross-domain transfer in the legal NLI.

2.2 Domain transfer in legal natural language processing

There are several studies that have investigated domain transfer for legal natural language processing tasks. Salaka et al. (2018) proposes a transfer learning approach to build an automatic sentiment annotator for legal domain using the manually annotated data on movie reviews. Chalkidis et al. (2021) introduces a multi-lingual and multi-label legal document classification dataset for zero-shot cross-lingual transfer. Bannihatti Kumar et al. (2022) investigates the cross-domain transferability of text generation models for legal text. Niklaus et al. (2022) explores transfer learning techniques on Legal Judgment Prediction in various settings, including cross-lingual transfer, cross-domain transfer, cross-regional transfer, and cross-jurisdiction transfer. T.y.s.s et al. (2023) investigates domain adaptation from Legal Judgment Prediction on European court of Human Rights cases into an article-aware classification task.

Existing works mostly focus on either cross-domain (Salaka et al. 2018; Bannihatti Kumar et al. 2022; T.y.s.s et al. 2023) or cross-lingual (Chalkidis et al. 2021) transfer. Our study, on the other hand, explores the possibility of cross-jurisdiction transfer for the legal NLI task between the US states. Niklaus et al. (2022) explores the possibility of cross-jurisdiction transfer, but it is between different countries (from Indian to Swiss cases), not between the states within the same country. Also, as previously stated, there was no prior attempt made for exploring domain transfer in the legal NLI. Our study aims to fill in this gap.

Types of inputs	Example	Label
Statement	I, [Person-1], residing in the County of Cassia, State of Idaho, being of full legal age and being of sound and disposing mind and not acting under duress, menace, fraud, or undue influence of any person whomsoever, do make, publish and declare this my Last Will and Testament and expressly revoke all other and former Wills and Codicils to Wills made by me.	**support**
Condition	The testator was over 18 years old and of sound mind when executing this will.	
Law	15-2-501. WHO MAY MAKE A WILL. Any emancipated minor or any person eighteen (18) or more years of age who is of sound mind may make a will. A married woman may dispose of her property, whether separate or community, in the same manner as any other person subject to the restrictions imposed by this code.	

Table 1: The legal NLI datasets used in this study contain three inputs: statement, condition, and law. Statement is a sentence or a short paragraph excerpted from a legal will. Condition is a circumstance related to the evaluation of a will statement's validity. For example, testator's age or mental capacity at the time of will execution is a condition relevant to the evaluation of validity of the will execution. Law is a state provision on which the validity of a legal will statement is evaluated. The inputs are classified as either "support", "refute", or "unrelated", depending on whether the condition and the law support, refute or are unrelated with the given will statement.

3 Methods

To figure out if a legal NLI model trained on a single US state can be transferred to other states, we compared a model's performance on the in-domain data (i.e., the data which is from the same state as the model is trained on) and its performance on the out-of-domain data (i.e., the data which is from a state other than the one the model is trained on). We fine-tuned a pre-trained model on a task of evaluating the validity of legal will statements, once with the dataset containing the Tennessee wills and once with the dataset containing the Idaho wills.

The format of the datasets is distinct from the traditional NLI datasets in that it contains three types of inputs (statement, condition, and law) rather than two (premise and hypothesis). *Statement* is a sentence or a short paragraph excerpted from a legal will. Its legal validity is assessed based on the other two input types. *Condition* is a circumstance relevant to the evaluation of validity of a will statement. Lastly, *Law* is a state provision that the will statement's validity is assessed on. These inputs are labeled as either *support*, *refute*, or *unrelated*, depending on whether the condition and the law support, refute, or are unrelated to the will statement given. The table 1 shows the datasets are formatted with an example for each input type. For further details on the datasets, see the section 3.1. The model we fine-tuned with our dataset is roberta-large-mnli (Liu et al., 2019). For the implementation of the fine-tuning, see the section 3.2.

We used the fine-tuned models to predict both the in-domain testing data (i.e., Tennessee model predicts Tennessee data and Idaho model predicts Idaho data) and the out-of-domain data (i.e., Tennessee model predicts Idaho data and Idaho model predicts Tennessee data). To better understand the models' behaviors, we generated confusion matrices with the models' predictions and conducted error analysis by using Local interpretable model-agnostic explanations (LIME; Ribeiro et al. 2016).

3.1 Datasets

We use two datasets in our study: a dataset containing legal wills from Tennessee (Kwak et al., 2022) and a dataset containing legal wills from Idaho. The Tennessee dataset was introduced in Kwak et al. (2022). The dataset contains 23 wills from Tennessee, splitted into 1,014 data points. Each data point consists of a legal will statement (usually a sentence excerpted from a will) accompanied with hypothetical conditions at the time of will execution and/or probate and state laws relevant (or irrelevant) to the evaluation of the legal will statement's validity. The dataset is annotated with three

Model - Data	Precision	Recall	F1	Accuracy
TN - TN (in-domain)	96.67	96.25	96.41	96.86
ID - ID (in-domain)	91.06	93.34	92.03	93.82
TN - ID (out-of-domain)	80.59	62.62	66.32	76.23
ID - TN (out-of-domain)	80.70	83.42	81.60	82.94

Table 2: This table shows the performances of the Tennessee (TN) model and the Idaho (ID) model in two test settings: in-domain and out-of-domain. In-domain means that the model predicted on the testing partition from the same state, while out-of-domain means that the model predicted on the data from a different state. Both models show decent performance on the in-domain test setting. The F1 scores (weighted average) for Tennessee model and Idaho model in the in-domain setting are 96.41 and 92.03, respectively. Both models' F1 scores (weighted average) are considerably lower in the out-of-domain setting, but Tennessee model suffer more dramatically (Tennessee model: 66.32, Idaho model: 81.60)

labels: support, refute, and unrelated. ("support": condition & law supports the will statement; "refute": condition & law refutes the will statement, "unrelated": the law is unrelated to the validity assessment of the given legal will statement.) See table 1 for the further details on the format of the dataset.

We created a new dataset containing legal wills from Idaho, following the format of the dataset introduced in Kwak et al. (2022). We collected 14 Idaho wills from the U.S. Wills and Probates datasets in Ancestry.[1] The collected wills were manually anonymized by replacing any personal information into special tokens (e.g., names into [Person-n] and addresses into [Address-n]) and so on), as suggested by Suntwal et al. (2019). The wills were splitted into 1,039 statements. We added hypothetical conditions and state laws relevant (or irrelevant) to the evaluation of the statement's validity, as Kwak et al. (2022) did. The dataset was also annotated with three labels: support, refute, and unrelated. The annotation was done by two annotators: one law student and one non-law graduate student. They were given a clear annotation guideline. They also had ample discussions during the annotation process to ensure the consistency of the dataset. The kappa agreement score between the two annotators is 0.89. The dataset creation and annotation process was supervised and reviewed by a law professor.

3.2 Model fine-tuning

We fine-tuned a transformer model with the two datasets introduced above. We chose to fine-tune roberta-large-mnli (Liu et al., 2019) as it was shown to be the best performing model in the given task setting by Kwak et al. (2022). We went through two separate fine-tuning processes: once with the Tennessee dataset to get a model working on Tennessee legal wills ("Tennessee model" from now on) and once with the Idaho dataset to get a model working on Idaho legal wills ("Idaho model" from now on). The two fine-tuning processes were done independently of each other, meaning that the result of these fine-tuning processes was two separate models (Tennessee model and Idaho model) each learned Tennessee data and Idaho data respectively, not one model that learned both. The fine-tuning was done on PyTorch 1.11.0 with Cuda 11.3 using the HuggingFace Trainer class, and hyperparameters were tuned on the development partition.

3.3 Error Analysis

We generated confusion matrices with the models' predictions in the cross-state setting (i.e., Tennessee model predicting Idaho data and Idaho model predicting Tennessee data). With the generated matrices, we identified the most salient error types. NLP experts and law experts in our team worked together to analyze the potential causes for these error types. We employed Local interpretable model-agnostic explanations (LIME; Ribeiro et al. 2016) to analyze the errors. LIME provides human interpretable explanations on individual predictions of any machine learning classifiers (or models) by highlighting the words that had impact on the model's prediction. With the LIME explanations, we were able to find potential causes for the major error types.

4 Results

Table 2 shows the performances of Tennessee model and Idaho model in two different settings: in-domain (predicting the testing partitions from

[1] Court documents such as probated wills are in the public domain in the US.

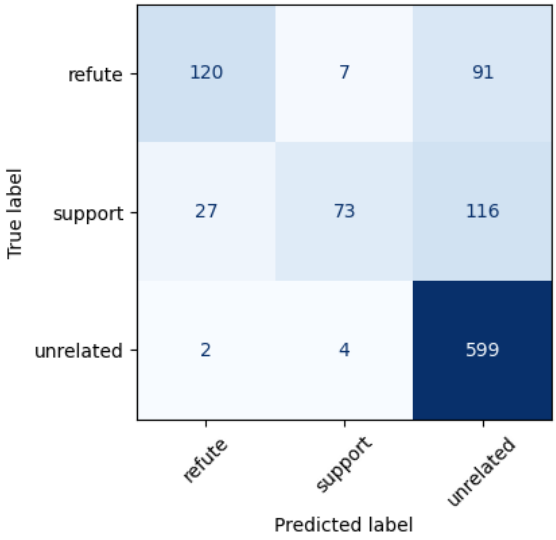

Figure 1: A confusion matrix generated with the Tennessee model's predictions on Idaho data. It is noticeable that the majority of the model's errors (207/247) is from incorrectly predicting support (116/247) or refute cases (91/247) as unrelated cases.

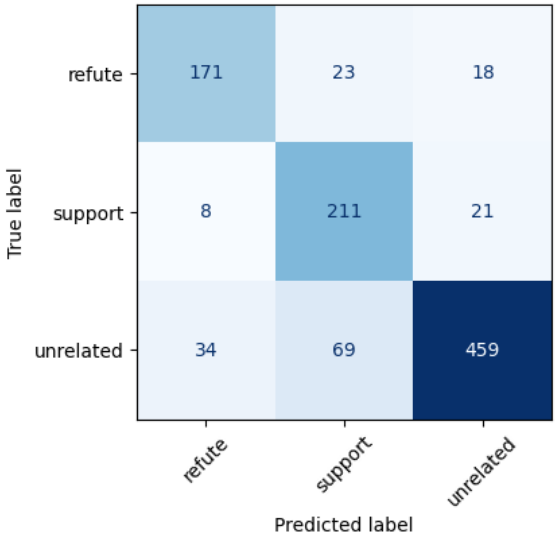

Figure 2: A confusion matrix generated with the Idaho model's predictions on Tennessee data. It is worth noting that more than half of the model's errors (103/173) is from incorrectly predicting unrelated cases as support (69/173) or refute cases (34/173).

the same state) and out-of-domain (predicting the data from a different state). Both Tennessee model and Idaho model show decent performance in the in-domain setting, as suggested by high F1 scores (Tennessee model: 96.41, Idaho model: 92.03). However, in the out-of-domain setting, the performance of both models drops considerably. Both models' F1 scores are considerably lower in the out-of-domain setting, but Tennessee model suffers more dramatically (Tennessee model: 66.32, Idaho model: 81.60).

To figure out the cause of the performance drop, we generated confusion matrices with both models' predictions in the out-of-domain setting. Figure 1 shows the confusion matrix generated with Tennessee models' predictions on Idaho data. It is noticeable that the majority of the model's errors (207/247) is from incorrectly predicting support (116/247) or refute cases (91/247) as unrelated cases. Figure 2 presents the confusion matrix generated with Idaho models' predictions on Tennessee data. This time, more than half of the errors (103/173) originated from incorrectly predicting unrelated cases as support (69/173) or refute cases (34/173).

5 Error Analysis

We identified two major error types by observing the confusion matrices. The first is that the Tennessee model wrongly predicts support or refute cases as unrelated when predicting Idaho data (207/247). The second is that the Idaho model wrongly predicts unrelated cases as support or refute cases (103/173). We focus on analyzing these two error types in this section.

5.1 Failure to Recognize Relevant Laws

The major issue found from the Tennessee model's performance in the out-of-domain setting is that it fails to identify relevant laws from Idaho state code. The model confuses support or refute cases with unrelated cases, meaning that it does not recognize the laws that are relevant to will execution or probate. There are two potential causes for this error. First, it can be that the Idaho dataset contains many laws that are not in the Tennessee dataset. Second, it can be that the stylistic differences in law (e.g., terms, formats, capitalization) made it difficult for the model to recognize similar laws. To find out what portion of errors are attributable to the differences in the state codes, we compared the laws contained in the Tennessee dataset to the ones contained in the Idaho dataset.[2]

As a result of the comparison, we found that only 79 out of 207 errors are attributable to the

[2]As the model would have learned only the laws that are contained in the dataset, we restricted the range of comparison only to the laws that are contained in the datasets used for the model training. Therefore, our analysis on the difference of Tennessee and Idaho statutes may not hold outside the context of this study.

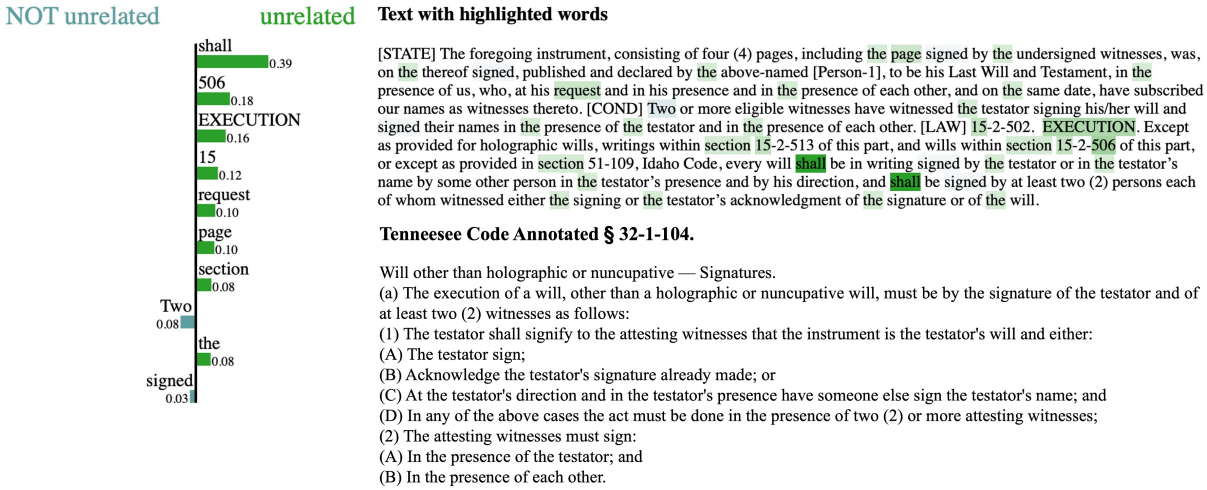

Figure 3: The chart on the left and the figure on the upper right present the LIME explanation on a case where Tennessee model failed to recognize the law relevant to the will execution (Idaho Statute 15-2-502) and made wrong prediction on the Idaho data(a "support" case as an "unrelated" case). The explanation shows that the model made prediction based on terms and numbers (e.g., 506, 15, request, section) that are not present in the equivalent law in Tennessee (Tennessee Code Annotated, 32-1-104; presented in the lower right side).

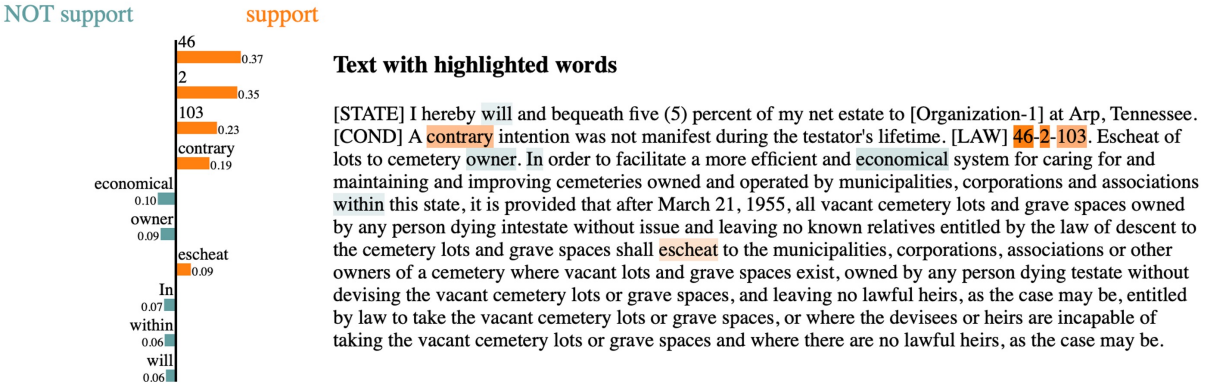

Figure 4: The figure presents the LIME explanation on a case where Idaho model makes wrong prediction on the Tennessee data based on the statutory section number. The model predicted an "unrelated" case as a "support" case, based on the statutory section number (i.e., 46, 2, 103).

differences between the laws contained in the Tennessee dataset and the ones contained in the Idaho dataset. These error cases contained Idaho statutes of which equivalence could not be found from the statutes in the Tennessee dataset. One such example is Idaho Statute 15-3-914 ("DISPOSITION OF UNCLAIMED ASSETS"). This law states that the personal representative shall distribute the share of any missing heir, devisee or claimant to their trustee, and if no trustee has been appointed for them, the personal representative shall file the report of abandoned property required by section 14-517. However, the equivalence of such statute cannot be found from the Tennessee dataset. Therefore, it is not surprising that Tennessee model failed to recognize this statute as a relevant one.

However, the rest of the errors (128/207) are not attributable to the differences in the state codes as the provisions contained in these cases have the equivalence in the Tennessee state code.[3] In these cases, it is likely that the model failed to see the connections between the similar provisions due to the stylistic differences in law. For example, Tennessee model failed to identify the Idaho provision pertaining to will execution (Idaho Statute 15-2-502), even though there is a similar provision in Tennessee (Tennessee Code Annotated 32-1-104). The model made incorrect predictions on every data

[3]The criterion we used to determine the equivalence of two laws is whether one law can replace the other in evaluating the will statements' validity without altering the result.

point containing this law (39 errors in total) and dictated it as unrelated despite the law's relevance to the will execution. Figure 3 presents a LIME explanation for one of such cases. The explanation shows that the model relied on terms/numbers that are not present in the equivalent law in Tennessee (Tennessee Code Annotated 32-1-104), such as *section*, *506*, and *15*. This shows that the model failed to see the connection between the two laws, even though the model was already trained on the Tennessee law (Tennessee Code Annotated 32-1-104).

It is also worth noting that the model relied on the word *EXECUTION*, which is highly relevant to the given will statement, when dictating the law as "unrelated". The model (i.e., roberta-large-mnli) is case-sensitive, so it distinguishes between the word in lower case (*execution*) and the word in upper case (*EXECUTION*). In Tennessee state code, it is very uncommon that a word is fully capitalized. However, in Idaho state code, all titles are presented in full capitalization (e.g., Idaho Statute 15-2-502. EXECUTION). As the model did not see the word in full capitalization during the training, the model fails to recognize the relevance of the word to the given will statement. This illustrates how a minor format difference, such as capitalization, can interfere with the model's performance.

5.2 Difference in Statutory Section Numbering System

The Idaho model makes incorrect predictions on cases that should be labeled as "unrelated" in the out-of-domain setting (103/173). The LIME explanations revealed that the model primarily relies on statutory section numbers when making these wrong predictions. Figure 4 shows one such case. The model depended on *46*, *2*, and *103* when predicting the case as "support."

This error is attributable to the difference in statutory section numbering system. While majority of the Idaho laws (which are mostly unrelated to will execution and/or probate) use Title-Chapter+Section numbering (e.g., Title: 14, Chapter: 1, Section: 1 = 14-101), a small portion of laws (e.g., Uniform Probate Code) use Title-Chapter-Section numbering (e.g., Title: 15, Chapter: 2, Section: 502 = 15-2-502). Most of the laws in the dataset that have Title-Chapter-Section numbering are relevant to the will execution and/or probate.[4] The Idaho model uses this numbering format difference to distinguish between laws relevant to will execution and/or probate and those that are not. However, this pattern does not exist in Tennessee data as all Tennessee laws have the same numbering format (i.e., Title-Chapter-Section). As Tennessee laws use the same numbering format as the Uniform Probate Code, the Idaho model would have predicted a large portion of "unrelated" cases in Tennessee data as "support" or "refute" based on the statutory section number format. This shows how a seemingly unrelated factor, such as statutory section numbering format, can have an impact on the model's performance.

6 Conclusion

This study found that legal natural language inference model trained on the data from one US state can be transferred to another state, but the performance deteriorates considerably. The F1 scores of the Tennessee model and the Idaho model are 96.41 and 92.03 when predicting the data from the same state, but they drop to 66.32 and 81.60 when predicting the data from another state (i.e., Tennessee model predict Idaho data and vice versa).

Our error analysis found two major causes for the performance drop. First, we found that the model struggles with identifying the equivalent laws across states when there are stylistic differences (e.g., terms, formats, capitalization, and etc.) between laws. We also found that difference in statutory section numbering system between the states makes it harder for the model to locate laws relevant to the cases that they are predicting on.

This analysis gives hints for the design of the future legal NLI systems. For instance, it is expected that including a knowledge base with equivalent laws between states would enhance the model's performance on predicting the cross-state data. The knowledge base should help with the normalization of statutory section numbers across the states. Normalization of the law text formats (e.g., capitalization) would also improve the model's performance on the data from across the states.

The analysis also provides empirical support to legal experts for the argument that legal documents such as wills should be standardized. As illustrated by our analysis, differences in legal documents' texts or formats between the states can easily interfere with the model's performance.

[4] Uniform Commercial Code (UCC) laws in Idaho Statute also have Title-Chapter-Section numbering, but the number of UCC laws included in our dataset is very few (8 out of 1039 data points.

Standardization of legal documents would facilitate the development of legal NLI system that can work across the states. Our open-access dataset and source code are publicly available at: https://github.com/ml4ai/nli4wills-corpus

Acknowledgements

We thank the reviewers for their thoughtful comments and suggestions. This work was partially supported by the National Science Foundation (NSF) under grant #2217215, and by University of Arizona's Provost Investment Fund. Mihai Surdeanu declares a financial interest in lum.ai. This interest has been properly disclosed to the University of Arizona Institutional Review Committee and is managed in accordance with its conflict of interest policies.

References

Vinayshekhar Bannihatti Kumar, Kasturi Bhattacharjee, and Rashmi Gangadharaiah. 2022. Towards cross-domain transferability of text generation models for legal text. In *Proceedings of the Natural Legal Language Processing Workshop 2022*, pages 111–118, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

William Bruno and Dan Roth. 2022. LawngNLI: A long-premise benchmark for in-domain generalization from short to long contexts and for implication-based retrieval. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 5019–5043, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Ilias Chalkidis, Manos Fergadiotis, and Ion Androutsopoulos. 2021. MultiEURLEX - a multi-lingual and multi-label legal document classification dataset for zero-shot cross-lingual transfer. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6974–6996, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Yuta Koreeda and Christopher Manning. 2021. ContractNLI: A dataset for document-level natural language inference for contracts. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 1907–1919, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Alice Kwak, Jacob Israelsen, Clayton Morrison, Derek Bambauer, and Mihai Surdeanu. 2022. Validity assessment of legal will statements as natural language inference. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 6047–6056, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Puneet Mathur, Gautam Kunapuli, Riyaz Bhat, Manish Shrivastava, Dinesh Manocha, and Maneesh Singh. 2022. DocInfer: Document-level natural language inference using optimal evidence selection. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 809–824, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Joel Niklaus, Matthias Stürmer, and Ilias Chalkidis. 2022. An empirical study on cross-X transfer for legal judgment prediction. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 32–46, Online only. Association for Computational Linguistics.

Marco Túlio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. "why should I trust you?": Explaining the predictions of any classifier. *CoRR*, abs/1602.04938.

Viraj Salaka, Menuka Warushavithana, Nisansa de Silva, Amal Shehan Perera, Gathika Ratnayaka, and Thejan Rupasinghe. 2018. Fast approach to build an automatic sentiment annotator for legal domain using transfer learning. In *Proceedings of the 9th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 260–265, Brussels, Belgium. Association for Computational Linguistics.

Sandeep Suntwal, Mithun Paul, Rebecca Sharp, and Mihai Surdeanu. 2019. On the importance of delexicalization for fact verification. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3413–3418, Hong Kong, China. Association for Computational Linguistics.

Santosh T.y.s.s, Oana Ichim, and Matthias Grabmair. 2023. Zero-shot transfer of article-aware legal outcome classification for European court of human rights cases. In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 605–617, Dubrovnik, Croatia. Association for Computational Linguistics.