


How Fine Tuning Affects Contextual Embeddings: A Negative Result Explanation

Ha-Thanh Nguyen¹ ^a, Vu Tran², Minh-Phuong Nguyen³, Le-Minh Nguyen³ and Ken Satoh¹

¹National Institute of Informatics, Tokyo, Japan

²Institute of Statistical Mathematics, Tokyo, Japan

³Japan Advanced Institute of Science and Technology, Ishikawa, Japan

Keywords: Contextual Embedding, Fine-Tuning, Impact Explanation.

Abstract: Recently, deep learning models trained on large amounts of data have achieved impressive results in the field of legal document processing. However, being seen as black boxes, these models lack explainability. This paper aims to shed light on the inner behavior of legal learning models by analyzing the effect of fine-tuning on legal contextual embeddings. This paper provides pieces of evidence to explain the relationship between the moving of contextual embeddings and the effectiveness of a model when fine-tuned on legal tasks. It can help further explain the effect of finetuning on language models. To this end, we use multilingual transformer models, fine-tune them on the lawfulness classification task, and record the changes in the embeddings. The experimental results reveal interesting phenomena. The method in this paper can be used to confirm whether a deep learning model truly gains the knowledge in a legal problem to make the predictions or simply memorize the training examples, or worse, predict randomly.


1 INTRODUCTION

Deep learning models have achieved impressive results in several legal document processing tasks (Kano et al., 2017; Rabelo et al., 2019; Nguyen et al., 2020; Chalkidis et al., 2020; Shao et al., 2020; Nguyen et al., 2021b). However, the explainability of the results remains an important issue for deploying in legal practice. For example, the law often contains technical terms that have a specific meaning. If a model cannot identify the meaning of these terms, it is likely to fail in further processing the documents. In previous work (Nguyen et al., 2021a), Nguyen et al. propose explanation metrics and visualizations that can be used to explain the state-of-the-art legal pretrained models. These explanations can be used to understand their performance, improve them, or build new models. These explanations are based on the model's output at a static status, i.e. when the model has already been trained. Although it is an effective approach to unbox the model, it cannot explain what is happening to the model while it is being trained or finetuned.

Embeddings are essential for any deep learning

model. Word embeddings are vectors that represent the meanings of the words, this is the very first approach that allows presenting the semantic meaning of a word in a computer system (Mikolov et al., 2013) instead of only its lexical meaning, i.e. the sequence of symbols that represents that word. However, the words in a sentence do not always have the same meaning. The same word can have different meanings in different contexts. This problem is handled by contextual embeddings. Contextual embeddings are vectors that represent the meaning of the words in their context (Vaswani et al., 2017). Contextual embeddings are trained along with the model, so they can be considered as a part of the model. Transformers are a type of contextual embedding model. Transformer models are pre-trained on large amounts of data and then fine-tuned on the specific data. Transformer models have achieved impressive results in many NLP tasks including legal document processing (Devlin et al., 2018; Raffel et al., 2020; Brown et al., 2020; Lewis et al., 2019).

In this paper, we present a way to understand how legal fine-tuning affects contextual embeddings. To this end, we use multilingual transformer models and fine-tune them on the task of lawfulness classification. In this task, given a text, which can be a legal state-

^a <https://orcid.org/0000-0003-2794-7010>

ment, the model predicts whether the text is lawful or unlawful. We record the changes in the metrics on accuracy and embeddings. The experimental results show that there is a relationship between accuracy and the quality of the embeddings. Observations in this paper can be used to explain and improve pretrained language models for legal document processing. We also hope that this paper will encourage researchers to use more explainable and transparent methods for developing legal machine-learning models. This paper is a side result of our effort to create a powerful language model on the NMSP approach (Nguyen et al., 2021b). The model fails our expectations but it reveals interesting phenomena.

2 RELATED WORK

The legal domain is always considered a difficult domain for natural language processing tasks (Rabelo et al., 2020; Nguyen and Nguyen, 2021; Vuong et al., 2022). The legal text is different from the general text in some aspects such as the presence of technical terms, the specific language, and the specific structure. Recently, using transformer models has yielded promising results in the field of law. The standard way to evaluate the model performance is to use some evaluation metrics such as accuracy, F1 score, and ROC curve. However, these metrics cannot show the relationship between the model’s performance and the model’s understanding of the legal problem. For example, a model can predict the result of a case correctly, but this does not mean that the model has understood the legal problem. This is a big problem when building a model to assist lawyers. Lawyers need to know how the model works. If the model relies on a few unimportant details from the data to make the predictions, then, these predictions will be useless.

Nguyen et al. (Nguyen et al., 2021a) propose two evaluation metrics to test the quality of legal embeddings. These evaluation metrics are based on the number and the positions of legal terms in the embeddings. These evaluation metrics can be used to understand the model’s understanding of legal terms. They argue that if a model can locate the relative positions of legal terms in its embedding, they have more chance to model the high-abstract concepts in the legal domain, generalize better, and achieve high performance. However, the authors do not propose a way to evaluate the effect of fine-tuning legal documents on the model.

LVC stands for *Legal Vocabulary Coverage*, which is the percentage of legal terms in the embed-

ding. Simply put, if the embedding contains 25K legal terms out of 100K terms, then, the LVC is 25%. This metric is reasonable because legal terms are the most important part of legal documents and the most correct way to describe legal concepts. With subword representation, the model can try to understand an unknown term by its subparts. However, this representation is a workaround approximation and can lead to some problems. For example, a non-legal term can be split into some legal subparts, but it does not mean that this non-legal term represents the legal concept.

LECA stands for *Legal Embedding Centroid-based Assessment*, which can be used to understand the relative positions of legal terms in the embedding. LECA is based on the fact that the positions of the legal terms in the embedding space should be closer together. This metric first finds the average embedding vector of all legal terms as a centroid. Then, the Euclidean distance between the legal terms and the legal centroid is calculated. The average distance of all legal terms is considered the value of this assessment. If this value is small, the legal terms are close to the legal centroid and the legal concepts are likely to be represented properly in the embedding space.

LVC remains the same during finetuning, whereas the LECA changes because of backpropagation. We observe that the legal terms in the embedding space become closer and closer to each other in most of our experiments, which leads us to conduct experiments to see how this metric changes before and after finetuning. This information can reveal how better the model understands legal terms by being finetuning on a legal dataset.

3 EXPERIMENTS

In our experiment, we initially wanted to create XLM-Paralaw as a state-of-the-art model that outperforms existing contextual embeddings-based models. However, our expectations were subverted, and we conducted experiments to get additional clues to explain the model’s failure.

3.1 XLM-Paralaw

In the NMSP approach (Nguyen et al., 2021b), in order to improve the performance of language models, we train the model on a multilingual corpus. This approach is inspired by the idea that the translation process can be considered a form of knowledge transfer. Besides, in translation, we need to base on the context to find the most suitable translation. This is similar to the way a language model is trained. In the NMSP

approach, the model is trained to determine whether a sentence is a translation of the following sentence of a given sentence.

The previous model trained on this approach achieved state-of-the-art results on COLIEE 2021¹. The backbone of this model is BERT multilingual. This result brings us an assumption that using a larger model that pretrained on multilingual data can lead to better results. Hence, we choose XLM-RoBERTa (Conneau et al., 2019) as the main architecture of the model.

The data used to further pretrain XLM-RoBERTa to get XLM-Paralaw is the data set described in (Nguyen et al., 2021b). This data set has been collected from the legal translation corpora of Japanese law². From the original pairs of Japanese-English sentences, the training samples are generated as follows. First, we take a sentence, to make a positive sample, and we pair it with the following sentence (the original or the translation). To make a negative sample, we pick random sentences in the corpus. Now we have pairs of sentences with labels ('positive' or 'negative'). The final training set is a collection of 718,000 pairs with labels.

XLM-Paralaw is then further pretrained on this data set. We randomly sample 10% of the original corpus for validation. The pretraining process is stopped when we do not see improvement after one epoch. After pretraining, we finetune and test XLM-Paralaw on COLIEE 2021 with the lawfulness classification task. **The results show that the model's performance is far from what we expected.** To better understand this phenomenon, we conduct an extensive experiment to see how well the model represents legal concepts in its vocabulary and compare it with other models. The results are summarized and analyzed in Section 3.3.

3.2 Experimental Settings

We choose multilingual transformer models as the study objects. They share the same architecture of contextual embedding. Our strategy is to measure the LECA value before and after fine-tuning the lawfulness classification task to see how finetuning changes the model's performance as well as the contextual embeddings.

In terms of experimental resources, we have the legal bi-lingual dataset in English-Japanese and pre-trained language models constructed from the parallel corpus. We choose the following models for our experiments:

- BERT Multilingual Base Uncased (Devlin et al., 2018): The uncased version of BERT multilingual model trained on 104 languages.
- BERT Multilingual Base Cased (Devlin et al., 2018): The cased version of BERT multilingual model trained on 104 languages.
- Paralaw Nets - NFSP (Nguyen et al., 2021b): A model trained on the Paralaw dataset with the NFSP approach.
- Paralaw Nets - NMSP (Nguyen et al., 2021b): A model trained on the Paralaw dataset with the NMSP approach.
- XLM-RoBERTa Base (Conneau et al., 2019): A model trained on 100 languages.
- XLM-Paralaw³: A model further pretrained from XLM-RoBERTa Base on the Paralaw dataset with the NMSP approach, described in Section 3.1.

In terms of the finetuning task, we choose the lawfulness classification proposed by Nguyen et. al. (Nguyen et al., 2019). Given a legal statement, the task is to classify it into one of two classes: lawful or unlawful. This task is perfect for our purpose of checking whether the model gives the correct prediction based on the same features as human experts. We fine-tune the above models on the English version, Japanese version, and English-Japanese bi-lingual version of the task. For each language, we limit the number of legal terms \mathcal{L} to 1000 and the number of legal sample sentences \mathcal{D} to 5000. The purpose of this experiment is not about finding the best model for this task, but to explain the behavior of the transformer model finetuned in legal tasks. The data is provided by COLIEE competition, after data processing and argumentation, we have in total of 3,950 samples in English, 3,478 samples in Japanese, and 7,428 samples in bi-lingual version. The data is split into a train set and a test set at a ratio of 9:1. We use the following parameter: learning rate = $3e-5$, batch size = 16, sequence length = 512, dropout rate = 0.1, weight decay = 0.01, adam epsilon = $1e-8$. What we pay attention to is the difference of accuracy and LECA before and after finetuning.

3.3 Experimental Results

The experimental results of all embeddings and all datasets are reported in Tables 1, 2, and 3. There are two metrics for evaluation: accuracy (higher is better) and LECA (lower is better). In most cases, the accuracy of all embeddings increases after finetuning, and

¹<https://sites.ualberta.ca/~rabelo/COLIEE2021/>

²<https://www.japaneselawtranslation.go.jp/>

³<https://huggingface.co/nguyenthanhasia/XLM-Paralaw>

Table 1: Results on Japanese dataset in accuracy (higher is better) and LECA (lower is better).

Embedding	Before Finetuning		After Finetuning	
	Accuracy	LECA	Accuracy	LECA
BERT Multilingual Base Uncased	0.4899	0.6402	0.5187	0.6354
BERT Multilingual Base Cased	0.4784	0.4951	0.5591	0.4434
ParaLaw Nets - NFSP	0.5101	0.2987	0.5793	0.2812
ParaLaw Nets - NMSP	0.4986	0.1256	0.5908	0.1121
XLM-RoBERTa	0.4784	0.2277	0.5908	0.2132
XLM-Paralaw	0.5187	0.0000	0.5014	0.0000

Table 2: Results on English dataset in accuracy (higher is better) and LECA (lower is better).

Embedding	Before Finetuning		After Finetuning	
	Accuracy	LECA	Accuracy	LECA
BERT Multilingual Base Uncased	0.4886	0.5486	0.5646	0.5287
BERT Multilingual Base Cased	0.4937	0.4148	0.5468	0.4044
ParaLaw Nets - NFSP	0.4835	0.2233	0.5848	0.2252
ParaLaw Nets - NMSP	0.4734	0.0975	0.5443	0.0821
XLM-RoBERTa	0.4835	0.1971	0.5646	0.1755
XLM-Paralaw	0.5190	0.0000	0.5063	0.0000

Table 3: Results on bi-lingual dataset in accuracy (higher is better) and LECA (lower is better).

Embedding	Before Finetuning		After Finetuning	
	Accuracy	LECA	Accuracy	LECA
BERT Multilingual Base Uncased	0.5013	0.6014	0.5903	0.5532
BERT Multilingual Base Cased	0.5054	0.4592	0.5943	0.3922
ParaLaw Nets - NFSP	0.5148	0.2612	0.5984	0.2251
ParaLaw Nets - NMSP	0.4906	0.1068	0.5930	0.0723
XLM-RoBERTa	0.5135	0.2109	0.6024	0.1887
XLM-Paralaw	0.5013	0.0000	0.5202	0.0000

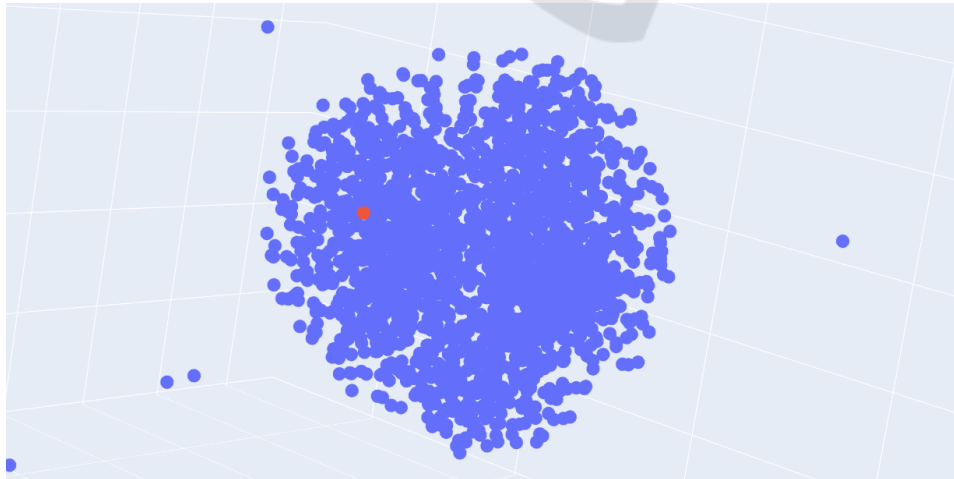


Figure 1: Visualization of XLM-Paralaw embedding using t-SNE. Red points represent legal terms, blue points represent non-legal terms.

the LECA value also decreases. We were surprised to see XLM-Paralaw achieve impressive LECA scores of 0.0000 (from 10^{-9} to 10^{-8}) on all datasets. However, we don't see the same results in accuracy.

From the experimental results, we can see that XLM-RoBERTa and ParaLaw Nets have significant improvements on all datasets. These results may come from the multilingual ability of XLM-RoBERTa and ParaLaw Nets. XLM-RoBERTa was trained in 100 different languages, while ParaLaw Nets were trained on English-Japanese bilingual data. BERT Multilingual's improvements are less significant and the XLM-Paralaw has no notable improvement on the dataset. Both the decrease of LECA and the increase in accuracy on this model is insignificant.

Looking at the increasing speed of the accuracy and the decreasing speed of LECA, we can see that the improvement in accuracy does not always mean a better representation of legal terms in the model's embedding. In the case of NFSP on the English dataset, the LECA value is increased but the accuracy also improves. From this observation, we can have a certain assumption that, in this case, the improvement in accuracy is associated with the model's performance on non-legal terms or by chance. As a result, solely using accuracy to evaluate the model's performance is not always an effective way. XLM-Paralaw's performance after one epoch is still close to randomness. We can see that the LECA score of this model is extremely small, which means that it locates vectors representing legal terms in a narrow space. Our hypothesis is that over-optimization on the legal terms caused the model difficulty in differentiating legal terms and understanding the relationship between them and non-legal terms.

For a better understanding of the XLM-Paralaw embedding, we reduce the dimension using t-distributed stochastic neighbor embedding algorithm and visualize it using a 3D scatter plot. The visualization is displayed in Figure 1. We observed that the XLM-Paralaw's embedding locates the vectors representing legal terms in a very narrow space (we can only see one red point in the figure). This visualization is consistent evidence with our hypothesis.

4 DISCUSSION

As an attempt to explain the poor performance of XLM-Paralaw, we conduct experiments to compare this model and other candidates. The experimental results bring us several observations.

First, the XLM-Paralaw model's performance is inferior and inconsistent with our initial expectations.

We present some pieces of evidence that this was caused by the over-optimization of the contextual embeddings with the pretraining strategy. This observation suggests that it is essential to find a balance point in pretraining to achieve the best model's performance.

Second, contextual embedding is different from word embedding. During the training process, the relative positions of the vectors move. As a result, the relative positions of the vectors representing the terms may be very different in the final embedding. In the case of XLM-Paralaw, when the legal terms are compressed in a very narrow space, the relationships between them are damaged, which leads to the poor performance of this model.

Third, we want to emphasize that to explain the behaviors of a model or confirm a hypothesis, multiple metrics should be considered. To evaluate the performance of a model, we should look at not only the accuracy but also the LECA score. Besides, visualization of the embedding (for example, t-SNE) is also a useful tool for understanding the model.

5 CONCLUSIONS

In this paper, we propose a way to evaluate pretrained embedding effectiveness on a legal task. We prove that using solely accuracy to evaluate model performance is not always help us to understand the nature of the model. We propose to use LECA (Legal Embedding Centroid-base Assessment) before and after finetuning to measure the effectiveness of the legal terms learned by the model. The experimental results show that in most cases, improving the ability to represent legal terms will improve the performance in the legal task, but not always. In the case of XLM-Paralaw, over-optimization on legal terms caused the opposite effect. This paper will be the starting point of a more in-depth investigation of legal term representation in the language model.

REFERENCES

- Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al. (2020). Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.
- Chalkidis, I., Fergadiotis, M., Malakasiotis, P., Aletras, N., and Androutsopoulos, I. (2020). Legal-bert: The muppets straight out of law school. *arXiv preprint arXiv:2010.02559*.

- Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, E., Ott, M., Zettlemoyer, L., and Stoyanov, V. (2019). Unsupervised cross-lingual representation learning at scale. *arXiv preprint arXiv:1911.02116*.
- Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.
- Kano, Y., Kim, M.-Y., Goebel, R., and Satoh, K. (2017). Overview of coliee 2017. In *COLIEE@ ICAIL*, pages 1–8.
- Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., Stoyanov, V., and Zettlemoyer, L. (2019). Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461*.
- Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., and Dean, J. (2013). Distributed representations of words and phrases and their compositionality. *Advances in neural information processing systems*, 26.
- Nguyen, H., Tran, V., and Nguyen, L. (2019). A deep learning approach for statute law entailment task in coliee-2019. *Proceedings of the 6th Competition on Legal Information Extraction/Entailment. COLIEE*.
- Nguyen, H. T., Binh, D. T., Quan, B. M., and Le Minh, N. (2021a). Evaluate and visualize legal embeddings for explanation purpose. In *2021 13th International Conference on Knowledge and Systems Engineering (KSE)*, pages 1–6. IEEE.
- Nguyen, H.-T. and Nguyen, L.-M. (2021). Sublanguage: A serious issue affects pretrained models in legal domain. *arXiv preprint arXiv:2104.07782*.
- Nguyen, H.-T., Tran, V., Nguyen, P. M., Vuong, T.-H.-Y., Bui, Q. M., Nguyen, C. M., Dang, B. T., Nguyen, M. L., and Satoh, K. (2021b). Paralaw nets—cross-lingual sentence-level pretraining for legal text processing. *arXiv preprint arXiv:2106.13403*.
- Nguyen, H.-T., Vuong, H.-Y. T., Nguyen, P. M., Dang, B. T., Bui, Q. M., Vu, S. T., Nguyen, C. M., Tran, V., Satoh, K., and Nguyen, M. L. (2020). Jnlp team: Deep learning for legal processing in coliee 2020. *arXiv preprint arXiv:2011.08071*.
- Rabelo, J., Kim, M.-Y., Goebel, R., Yoshioka, M., Kano, Y., and Satoh, K. (2019). A summary of the coliee 2019 competition. In *JSAI International Symposium on Artificial Intelligence*, pages 34–49. Springer.
- Rabelo, J., Kim, M.-Y., Goebel, R., Yoshioka, M., Kano, Y., and Satoh, K. (2020). Coliee 2020: methods for legal document retrieval and entailment. In *JSAI International Symposium on Artificial Intelligence*, pages 196–210. Springer.
- Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., Liu, P. J., et al. (2020). Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*, 21(140):1–67.
- Shao, Y., Mao, J., Liu, Y., Ma, W., Satoh, K., Zhang, M., and Ma, S. (2020). Bert-pli: Modeling paragraph-level interactions for legal case retrieval. In *IJCAI*, pages 3501–3507.
- Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems*, 30.
- Vuong, Y. T.-H., Bui, Q. M., Nguyen, H.-T., Nguyen, T.-T.-T., Tran, V., Phan, X.-H., Satoh, K., and Nguyen, L.-M. (2022). Sm-bert-cr: a deep learning approach for case law retrieval with supporting model. *Artificial Intelligence and Law*, pages 1–28.