

Fine-tuning Language Model with Legal Text

Team members: Qingyang Bao, Youjia Li, Jiezhong Wu, Yifan Zhang

Introduction

In modern societies, law plays a vital role in shaping social interactions and maintaining order. With advances in natural language processing (NLP) and deep learning, there has been increasing interest in developing language models in the legal field.

Our project aims to develop a legal language model to assist in legal-related tasks. We chose to focus on language models because we needed to improve access to legal information and services.

To achieve our goal, we used pre-trained models and fine-tune models on a specialized legal dataset "Pile of Law". This dataset contains a large amount of legal text, including court decisions, statutes, and decrees, and will be used to train our model and improve its accuracy. Since it takes a lot of time to train the model using the complete dataset, we selected a part of it for our model training.

We choose BERT and GPT as the framework for our legal language models. By continuing to train the pre-trained BERT, LegalBERT and GPT on the Pile of Law dataset, the model can be further fine-tuned to complete the learning and training of the language model in the legal field.

We evaluated the performance of the models in two aspects: the LexGLUE benchmark score and text generation tasks given real-world law-related prompts or questions. The Lexical General Evaluation of Language Understanding (LexGLUE) benchmark is a widely used benchmark for evaluating the performance of NLP models, to ensure their accuracy and identify directions for improvement.

Goals and Discussion

In our effort to accomplish essential goals, there is a series of foundation work we finished, which serve as the critical groundwork towards these goals. We will firstly discuss these ground work below.

Groundwork

Text Data Pre-processing

Our models are trained on the Pile of Law dataset, a comprehensive collection of legal and administrative text primarily written in English. The dataset consists of approximately 256GB of

data sourced from 35 different legal and administrative data providers. These sources include court legal analyses, casebooks, US codes, government publications, statutes, and regulations.

For each dataset source, the data representation varies depending on the nature of the source. Since it takes a lot of time to train the model using the complete dataset, we selected 13 subsets of it for our model training.

First, we excluded the datasets not related to the United States from the "Pile of Law". Then, we specifically chose the "uscode", "cfr", "state_codes" datasets, which consists of The United States Code, U.S. Code of Federal Regulations and U.S. State Codes. Additionally, we selected 10 smaller public computing datasets from the remaining collection. A brief overview of the selected sub-datasets and their data representations is in Appendix 1.

We utilized the datasets library to load the Pile of Law dataset and consolidated the training sets of the 13 subsets into a unified dataset for training purposes. Each individual entry in the dataset corresponds to a single file, with entries potentially containing more than 10 thousand characters. In Figure 1, we present the initial text entry of our newly merged dataset, displaying the first 400 characters due to the length limitation.

```
text[0][:400]
```

```
'Administrative Conference of the United States\n\nSSA DISABILITY BENEFITS PROGRAM  
S:\nTHE DUTY OF CANDOR AND SUBMISSION OF ALL EVIDENCE\n\nFinal Report: Oct. 15, 20  
12\n\nThis report was prepared by the Office of the Chairman of the Administrative  
Conference of the\nUnited States. The views expressed do not necessarily reflect t  
hose of the Council, the members\nof the Conference, or its committees.\n\n\x0cCON  
TRIB'
```

Figure 1. The first 400 characters of the first entry in the merged dataset

We segmented the text into individual sentences by detecting the period symbol ('.') as a delimiter. The resulting collection of sentences is displayed in Figure 2, representing a bag of sentences.

```
bag[:3]
```

```
['Administrative Conference of the United States SSA DISABILITY BENEFITS PROGRAM  
S: THE DUTY OF CANDOR AND SUBMISSION OF ALL EVIDENCE Final Report: Oct',  
 ' 15, 2012 This report was prepared by the Office of the Chairman of the Adminis  
trative Conference of the United States',  
 ' The views expressed do not necessarily reflect those of the Council, the member  
s of the Conference, or its committees']
```

Figure 2. The splitted sentences of the first entry

The as-processed dataset can be directly used in the training for GPT and the mask task of BERT. But for the next sentence prediction(NSP) task, we need to create sentence pairs and label them as to whether they are consecutive pairs — or random pairs. As shown in Figure 3, we can see in the output of the final cell two pairs that we've created. The first is labeled as 0

and shows a IsNextSentence pair, the second is labeled as 1 and shows a NotNextSentence pair [3].

```
for i in range(2):
    print(label[i])
    print(sentence_a[i] + '\n---')
    print(sentence_b[i] + '\n')
```

0
192Bell, supra note 191, at 28-41 (summarizing commenters' opposition to mandatory disclosure requirements in proposed Rule 26(a))

193MODEL CODE OF PROF' L RESPONSIBILITY CANON 7 (1983)

1
Except for an agency's own rules, non-attorneys' conduct is largely unregulated

269 36 Moreover, in the non-attorney representative context, agencies like SSA do not have to address the potential conflict that may exist between state bar rules and its regulations

Figure 3. Sentence pairs and label for NSP

Fine-tuning Pre-trained Model

With the text data processed using the above techniques, we are able to perform fine-tuning tasks on pre-training language models. More specifically, we further trained the BERT, LegalBert and GPT2 using the same methods applied at their pre-training stage but with legal text data from the Pile of Law.

For the case of BERT and LegalBERT, we used “bert-base-uncased” and “nlpaulieb/legal-bert-small-uncased” version respectively. Original BERT[2] representation language model is based on two unsupervised tasks - masked LM and next sentence prediction. Therefore, the same set of tasks are performed on these two models. The building of the pipeline is inherited by the tutorial online [5] but we add much more ordinary codes to accommodate for our dataset. For the case of GPT2, we also applied the same unsupervised task - predicting the next word in a sentence given the context of previous words.

Limited by computational resources and training time, we are eventually able to fine-tune a total of 7 model instances. These instances encompass different variants of pre-trained models as well as variations in the amount of training data used. The time constraint forced us to select a small number of epochs, so we also trained the variations in epoch number to evaluate its impact. All model instances are summarized in the following table.

Index	Pretrained Model	Data Used% (total 256G)	Data Sources	Hyperparameter
1	bert-base-uncased	1.40%	'acus_reports', 'cc_casebooks', 'cfpb_creditcard_contracts', 'ed_policy_guidance', 'eoir', 'frcp', 'fre', 'ftc_advisory_opinions', 'hhs_alj_opinions', 'uscode'	Epoch=2
2	legal-bert-small-uncased			Epoch=2
3	bert-base-uncased			Epoch=5
4	legal-bert-small-uncased			Epoch=5
5	gpt2			Epoch=2
6	bert-base-uncased	3.72%	All of above + "cfr", "scotus_oral_arguments", "state_codes"	Epoch=2
7	legal-bert-small-uncased			Epoch=2

Table 1. Fine-tuned model instance summary

Essential goals

1. We will test our trained language models on a benchmark dataset: LexGLUE

LexGLUE is a benchmark dataset that contains several tasks, including the European Court of Human Rights (ECtHR), the US Supreme Court (SCOTUS), the European Union legislation (EUR-LEX), Labeled EDGAR (LEDGAR) which includes approx. 850k contract provisions, the UNFAIR-ToS dataset which contains 50 Terms of Service (ToS), and the Case Holdings on Legal Decisions (CaseHOLD). ECrHR, EUR-LEX, and UNFAIR-ToS are associated with Multi-label Classification tasks. SCOTUS and LEDGAR are associated with multi-class classification. CaseHOLD is associated with Multiple choice QA.[1]

LexGLUE requires a lot of computation power to run, especially with large models like GPT2 and BERT. Considering the resources that we have and the fact that we want to focus on US law rather than that of other countries and regions, we chose SCOTUS and CaseHOLD to be our tasks to test on. Also, for the original LexGLUE, every task is run with 5 random seeds to ensure robustness, we did not have time and computation power to do that so we only did once for each task.

We tested all of the models on those two tasks and below are the results.

Task	Model	Eval_Loss	Predict_Loss	Eval_Macro_F1	Eval_Micro_F1	Pred_Macro_F1	Pred_Micro_F1
SCOTUS	1	nan	nan	0.4622	0.5357	0.3614	0.4471

SCOTUS	2	0.9555	1.1809	0.7102	0.7857	0.6415	0.7435
SCOTUS	3	0.8880	1.0981	0.6499	0.7507	0.5869	0.7121
SCOTUS	4	0.9555	1.1809	0.7102	0.7857	0.6416	0.7435
SCOTUS	5	2.2806	2.1366	0.0462	0.2657	0.0457	0.2642
SCOTUS	6	0.8873	1.1112	0.6622	0.7636	0.5960	0.7057
SCOTUS	7	1.1981	1.6288	0.7112	0.7879	0.6453	0.7436
CaseHOLD	1	0.6994	0.7322	0.7222	0.7223	0.7103	0.7103
CaseHOLD	2	0.6895	0.7253	0.7398	0.74	0.7292	0.7292
CaseHOLD	3	0.6975	0.7420	0.7261	0.7262	0.7058	0.7058
CaseHOLD	4	0.6895	0.7253	0.7398	0.74	0.7292	0.7292
CaseHOLD	5	1.3320	1.3670	0.4343	0.4346	0.4300	0.4300
CaseHOLD	6	0.7821	0.8644	0.7224	0.7226	0.6975	0.6975
CaseHOLD	7	0.6472	0.6766	0.7398	0.7397	0.7239	0.7239

Table 2.LexGLUE evaluation

The experiments on BERT models are colored yellow, the experiments on LegalBert models are colored red, and the experiments on GPT2 models are colored blue.

From the results we can see that for BERT models, with our fine-tuning with increased epochs and data, the F1 increases for both evaluation and prediction in SCOTUS task. For the CaseHOLD task, however, the F1 values did not change too much. With the finetuning, the prediction F1 values actually dropped. The reason for that could be that SCOTUS is multi-class classification while CseHOLD is multiple-choice QA. For multiple-choice QA, the context is given, so more training on similar topics would actually lower the accuracy.

For the LegalBert model, the change in epochs did not change the accuracy at all, while the inclusion of more data used in fine-tuning slightly increased the accuracy and lowered the loss. This is different from that of BERT possibly because we used LegalBert-small which is such a small model that it's highly possible that we are facing the criteria of overfitting, which explains there's no change in accuracy when the number of epochs is changed and the fact that in the task of CaseHOLD, with the inclusion of more data, the losses decreased significantly while the F1 values stay relatively the same or even dropped a bit.

For GPT2, because the fine-tuning time is too long, we only did one training for it and the results seem worse than both the BERT and LegalBert one. We suspect that it is because both multi-class classification and multiple choice QA are tasks more suitable for encoder models - so generative/decoder models like GPT2 may not excel in them.

Compared to the original paper, the F1 they had for BERT in CaseHOLD is 70.8, and that in SCOTUS is 68.3(micro) and 58.3(macro). We can see that our approach successfully increased the accuracy in SCOTUS but not in CaseHOLD, which we suspect is the sensitivity to context and content for multiple-choice QA tasks.

For LegalBert, when comparing to the original paper, we have to keep in mind that we are using the smaller version of it. The F1 value they had for CaseHOLD is 75.3 and that for SCOTUS are 76.4(micro) and 66.5(macro). We have to admit that our approach did not improve the accuracy in LegalBert but it's highly possible that it's only due to the difference in model size.

2. We will also use prompts from Reddit subreddit LegalAdvice to demonstrate open question answering for our model.

Due to the distinctive nature and capabilities of GPT2-based and BERT-based models, we treated them differently to evaluate their ability to respond to real-world law advising prompts and questions.

GPT2-based model

For GPT2-based models, we employed the fine-tuned model (Model 5 in Table2) to generate text based on open-ended sentences as prompts collected from LegalAdvice Reddit. A subset of sample responses are collected in Table 3.

Prompts	Generated Response
<i>"You can file a police report in regards to"</i>	<i>"You can file a police report in regards to the incident first-responder."</i>
<i>"Unless it was a written check, I'm guessing it was an electronic transaction which means it is probably covered by EFTA. If so, the burden is on the bank to"</i>	<i>"Unless it was a written check, I'm guessing it was an electronic transaction which means it is probably covered by EFTA. If so, the burden is on the bank to prove that the check was made by means of electronic means"</i>
<i>"The burden of decision in a criminal case lies with"</i>	<i>"The burden of decision in a criminal case lies with the jury. "</i>
<i>"In a criminal trial, the prosecution must prove the defendant's guilt"</i>	<i>"In a criminal trial, the prosecution must prove the defendant's guilt beyond a reasonable doubt."</i>
<i>"Lawncare company provided service without expressed consent or prior payment, so I am"</i>	<i>"Lawncare company provided service without expressed consent or prior payment, so I am not responsible for any damages resulting from the use of this service."</i>

<i>"Is Amazon or other E-commerce companies required by US law to refund a product that was lost in transit regardless of how long the request was made?"</i>	<i>"Is Amazon or other E-commerce companies required by US law to refund a product that was lost in transit regardless of how long the request was made? If so, the seller must return the item item is returned to the seller for a refund."</i>
---	---

Table 3. Fine-tuned GPT2 model response with given prompts

From the generated responses, we can see that the fine-tuned GPT2 is capable of completing open-ended sentences in a law-related context. However, the model generally performs poorly when being asked a question. In most cases, the model will only repeat the question.

Bert-based model (legalbert_small variant)

For BERT-based models, we trained our legalbert_small model on the Stanford Question Answering Dataset (SQuAD) and got a variant called legalbert_QA model. In addition to the base legalbert model, it includes a specific head on top for question answering. This model outputs start and end scores for each item in the input sequence, which are used to predict the answer span in the input. Given a question and a context (such as a paragraph from a document), the model is trained to identify the span of text in the context that answers the question. The output layer is thus a simple linear layer that transforms the legalbert hidden states to start and end scores. We next present some contexts that are relevant to laws as well as answers the bert_QA model gives to the prompts (questions with hidden answers) in the following Table 4.

Prompts	Generated response
<p>Context: <i>"The Supreme Court of the United States (SCOTUS) is the highest court in the federal judiciary of the United States of America. It has ultimate and largely discretionary appellate jurisdiction over all federal and state court cases"</i></p> <p>Question: <i>"Is SCOTUS a federal or state court?"</i></p>	federal and state court
<p>Context: <i>"Intellectual property law grants owners exclusive rights to the products of human intellect, including patents, trademarks, and copyrights. A patent grants an inventor the right to exclude others from making, using, selling, offering to sell, or importing an invention for a certain number of years."</i></p>	human intellect ,

Prompts	Generated response
Question: "What rights are granted by intellectual property law?"	
<p>Context: "Employment law is a broad area that encompasses all areas of the employer/employee relationship. It includes not only collective labor legislation, which relates to the rights and duties of unions, employees, and employers, but also individual labor law, which concerns employees' rights at work and through the contract for work. Employment laws were first enacted to counter the harsh conditions that had become commonplace in factories and mines and to ensure fair pay and reasonable hours."</p> <p>Question: "What does employment law encompass?"</p>	<p>collective labor legislation, which relates to the rights and duties of unions, employees, and employers, but also individual labor law</p>

Table 4. legalbert_QA model response with given prompts

We emphasize that the generated results are not perfect, and oftentimes there are mistakes and even empty responses. This is highly likely due to the fact that we changed the weights of the base model of the BertForQuestionAnswering model without very thorough consideration due to time limit, such that some weights of BertForQuestionAnswering were not initialized from the model checkpoint and are newly initialized. This is something worth delving deeper after the project deadline to fully understand the mechanism. Before moving to the desired goals, we claim that we used the online resource [4] to equip the legalbert with the question-answering capability, while the prompting pipeline is purely done by Jiezhong.

Desired goals

1. We will change the percentage of Pile of Law Data used in fine-tuning and explore its effects

We completed this desired goal by setting two different levels of data usage as a percentage of the original full dataset (Pile of Law). The original dataset classifies collected legal text into multiple data sources (US codes, State codes, etc). For each data usage level we set, it maps to different subsets of data sources as shown in Table 2.

2. We will change the technique used for fine-tuning to layer freezing and explore its effects on the model performance.

We completed this desired goal by initializing a LegalBertSmall model with weights updated with state dict from fine-tuned legalbert model on “pile of law”. Then, we freeze all the layers before the classification layer, and create a new model for sequence classification with 14 classes for the purpose of testing on the SCOTUS dataset of LexGLUE. We replace the bert part of the model with the frozen layers, and change the classifier layer to fit with the new number of classes. Note that the SCOTUS dataset is a multi-class classification task with 14 classes. The training of this new legalbert_freezeLayer model is significantly faster since all layers apart from the last one are frozen. We compare the F1-score of standard legalbert and this new variant legalbert_freezeLayer

Model	Eval_Loss	Predict_Loss	Eval_Macro_F1	Eval_Micro_F1	Pred_Macro_F1	Pred_Micro_F1
Legalbert	0.9555	1.1809	0.7102	0.7857	0.6416	0.7435
legalbert_freezeLayer	1.1777	1.4765	0.7034	0.7836	0.6265	0.7435

Table 5. Layer freezing results

We see that if we use layer freezing, and impose a classification layer for the targeted tasks, all the evaluation metrics get slightly worse, but not much, which is an encouraging sign and indicates that we can use far less computational resources to achieve very similar model performances.

3. We will change to use the following alternative pre-trained models for fine-tuning - BERT, GPT, CaseLegalBERT and explore their performance.

We completed this desired goal by fine-tuning the BERT, GPT2 and LegalBERT pretrained model and compared their evaluation score in LexGlue as shown above.

Stretch goals

1. We will do a chatbot API for our best trained model and perform a survey on the usage of chatbot for legal consultation using Amazon Mechanical Turk.

We don't have time to make progress towards this goal.

2. We will improve our chatbot to be more sentimentally supportive.

We don't have time to make progress towards this goal.

Code and Documentation

We submitted the following notebook files.

Functionality	Notebook Name	Description
Fine-tuning	BERT_fine_tune.ipynb	Fine-tune BERT model with legal text
	LegalBERT_fine_tune.ipynb	Fine-tune LegalBERT model with legal text
	GPT2_fine_tune.ipynb	Fine-tune GPT2 model with legal text
Text Generation Tasks	GPT2_generation.ipynb	Employ fine-tuned GPT2 model to generate text with provided prompts
	QA_training.ipynb	In QA_training, we use our previous legalbert small model as the base model and equip it with the question-answering capability.
	Bert_QA_prompts.ipynb	In Bert_QA_prompts, we feed context and questions as prompts to the model for answers
LexGLUE Evaluation	LexGlue_SCOTUS.ipynb	Forked the original LexGLUE from https://github.com/coastalcph/lex-glue and run the LexGLUE pipeline on Colab.
	LexGLUE_CaseHOLD.ipynb	
	Many other LexGLUE jupyter notebooks to run parallel to save time	

Table 6. Code documentation

Reflections

What was interesting?

The concepts of Transformer and ChatGPT from the lectures were particularly relevant to our project. The specific aspects we learned from this project, such as applying and fine-tuning pretrained models on new datasets, were not covered extensively in other parts of the class. We learned through hands-on coding and conducting a comprehensive literature review on the fundamental principles of GPT, BERT, and their fine-tuning methodologies, as well as exploring various downstream techniques associated with these models.

What was difficult?

During our attempts to execute long-duration model training, we encountered a technical obstacle. The GPUs provided by Colab Pro+ that were accessible to us exhibited significant instability, often terminating the session unexpectedly after running for a period of 6-7 hours.

Also, training large language model is generally a very challenging task and it took us great efforts to debug, analyze and understand the mechanism and details of applying the pretrained models for downstreaming tasks.

What's left to do?

If given unlimited time and computation power, we would expand the fine-tuning process to a higher data usage percentage of "Pile of Law" and compare the evaluation against the amount of legal text data used.

Another area that deserves additional time and resource investment is the continued fine-tuning of our model using Question and Answer datasets. This effort aims to further enhance its ability to effectively respond to various types of questions. Pursuing this line of research will bring us closer to achieving our goal of deploying the model as a chatbot specialized in law advising.

Moreover, we would like to pursue the stretch goals that we proposed if possible. Given the experience we have gained from this quarter, we believe we are more capable of trying challenging tasks such as building a chatbot API and make chatbot more sentimentally supportive

References

- [1] Chalkidis, Ilias, Abhik Jana, Dirk Hartung, Michael Bommarito, Ion Androutsopoulos, Daniel Martin Katz, and Nikolaos Aletras. 2021. "LexGLUE: A Benchmark Dataset for Legal Language Understanding in English." *arXiv [cs.CL]*. arXiv. <http://arxiv.org/abs/2110.00976>.
- [2] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding." arXiv, May 24, 2019. Accessed: May 16, 2023. [Online]. Available: <http://arxiv.org/abs/1810.04805>
- [3] Briggs, J. (2021, September 2). How to fine-tune Bert with NSP. Medium. <https://towardsdatascience.com/how-to-fine-tune-bert-with-nsp-8b5615468e12>
- [4] Briggs, J. (2021, October 8). How to Train Bert For Q&A in Any Language <https://towardsdatascience.com/how-to-train-bert-for-q-a-in-any-language-63b62c780014>
- [5] Briggs, J, "Training BERT #5 - Training With BertForPretraining," YouTube video, February 12, 2022. [Online]. Available: <https://www.youtube.com/watch?v=IC9FaVPKIYc>. [Accessed: June 5, 2023].

Appendix

1. Overview of the selected sub-datasets

Subset	Representation	Train size
acus_reports	Reports from the Administrative Conference of the United States from 2010-2022	181
cc_casebooks	Educational Casebooks released under open CC licenses	59
cfpb_creditcard_contracts	Credit Card Contracts compiled by the U.S. Consumer Finance Protection Bureau	2023
ed_policy_guidance	Policy guidance documents from the U.S. Department of Education (2001-2022)	507
eoir	U.S. Executive Office for Immigration Review Immigration and Nationality Precedential Decisions	1671
frcp	U.S. Federal Rules of Civil Procedure	69
fre	U.S. Federal Rules of Evidence	51
ftc_advisory_opinions	Advisory opinions by the Federal Trade Commission	112
hhs_alj_opinions	Opinions from the U.S. Department of Health and Human Services Administrative Law Judges from 1985-2019	2083
uscode	The United States Code (laws)	43
cfr	U.S. Code of Federal Regulations	112
state_codes	: U.S. State Codes	43
scotus_oral_arguments	U.S. Supreme Court Oral Arguments	2083