The Gap between Deep Learning and Law: Predicting Employment Notice

Jason T. Lam
jason.lam@queensu.com
Queen's University
Kingston, Ontario

David Liang
david.liang@queensu.com
Faculty of Law, Queen's University
Kingston, Ontario

Samuel Dahan
samuel.dahan@queensu.ca
Faculty of Law, Queen's University
Kingston, Ontario
Cornell Law School, Cornell University
Ithaca, New York

Farhana Zulkernine
farhana@cs.queensu.com
School of Computing, Queen's University
Kingston, Ontario

ABSTRACT

This study aims to determine whether Natural Language Processing with deep learning models can shed new light on the Canadian calculation system for employment notice. In particular, we investigate whether deep learning can enhance the predictability of notice period, that is, whether it is possible to predict notice period with high accuracy. A major challenge with the classification of reasonable notice is the inconsistency of the case law. As argued by the Ontario Court of Appeal, the process of determining reasonable notice is "more art than science". In a previous study, we assessed the predictability of reasonable notice periods by applying statistical machine learning to a hand-annotated dataset of 850 cases. Building on this past study, this paper utilizes state-of-the-art deep learning models on a free-text summary of cases. We further experiment with a variety of domain adaptations of state-of-the-art pretrained BERT-esque models. Our results appear to show that the domain adaptations of BERT-esque models negatively affected performance. Our best performing model was an out-of-the-box RoBERTa base model which achieved a 69% accuracy using a +/-2 prediction window.

CCS CONCEPTS

• **Applied computing** → **Law**; • **Computing methodologies** → **Artificial intelligence**; **Natural language processing**.

KEYWORDS

Deep Learning - Employment Law - Reasonable Notice - Employment Termination - Legal Analytics - Predictive Analytics - Consistency

ACM Reference Format:
Jason T. Lam, David Liang, Samuel Dahan, and Farhana Zulkernine. 2020. The Gap between Deep Learning and Law: Predicting Employment Notice. In *Proceedings of the 2020 Natural Legal Language Processing (NLLP) Workshop, 24 August 2020, San Diego, US*. ACM, New York, NY, USA, 5 pages.

1 INTRODUCTION

The system of law that governs work in Canada (outside Quebec) consists of three overlapping regimes: the common law regime, the regulatory regime and the collective bargaining regime (also called labour law or the law of unionized workers). This paper focuses on the common law regime and in particular the employment law principles that apply to notice of termination, one of the most litigated issues in Canada. One peculiarity of the Canadian system is that while each province has specific regulatory standards, common law of employment, in principle, applies in the western provinces of Canada [10]. Common law is usually defined as a system of judge-made rules that uses a precedent-based approach to case law. Earlier decisions pertaining to similar facts or legal issues guide later decisions in an attempt to create legal predictability. In an effort to ensure legal certainty and predictability, judges must follow the reasoning in earlier cases that address the same legal issues and similar facts. That said, common law rules and their interpretations evolve with societal values, and therefore, the interpretation and application of common law principles can sometime be inconsistent and unpredictable, including when it comes to termination notice [16].

Upon termination, if the employment relationship is governed by an indefinite contract and if there is no termination provision limiting the employee's rights, the employer has the obligation to provide either notice or pay in lieu of notice[1]. Should the employer fail to comply with this obligation, courts attempt to determine what compensation the employee would have received during that period if adequate notice had been provided, as well as damages for that loss, less any mitigation income. Courts typically begin their analysis of what constitutes "reasonable notice" by looking at the so-called "Bardal factors", described in the landmark case Bardal v. Globe & Mail Ltd: 1) age of the employee, 2) length of service, 3) character of employment and 4) availability of similar employment[2] [1].

[1]There is no obligation to provide reasonable notice when there is a lawful termination provision (see Machtinger v. HOJ Industries, SCC) or where there is a fixed-term contract [7] [10]
[2]Most employment contracts are of indefinite length, and the law implies a term that employers must provide employees with reasonable notice that the relationship is ending. See, e.g., Machtinger v. HOJ Industries Ltd, [1992] 1 SCR 986, 7 OR (3d) 480

While the Bardal test has been designed as an objective calculation system, there is no clear indication of how much weight should be given to each factor, nor of how the factors should be utilized [1]. Accordingly, the case law on employment notice has been noted to be inherently inconsistent and subjective, and there does not seem to be a "right" figure for reasonable notice. As Justice Dunphy wrote of calculation of reasonable notice periods, "[it] is more art than science but must be one that is fair in all of the circumstances" [2]. There have also been additional layers of complication as judges have considered factors beyond those explicitly mentioned in Bardal, such as inducement, in which an employer's act in bad faith results in aggravated damages (Wallace Damages) [3].

In this paper we investigate whether deep learning models can enhance the predictability of termination notice. Using advances in pretrained models such as BERT [9], we investigate the effectiveness of domain adaptations as well as benchmark our results against deep learning models that have shown success across multiple domains and in law. We build on previous work in Dahan et al. [8], in which we analyzed the prediction of reasonable notice using statistical machine learning on a hand-annotated tabular dataset and demonstrated a lack of consistency amongst the judgements.

In the balance of this paper we present related deep learning research for legal text analytics and a description of the problem, followed by a discussion of the data, models and methods utilized in this research. We end with our results and discussions followed by a conclusion and recommendations for future work.

2 RELATED WORK

In literature, the application of NLP to legal analytics is still new and there exist very few implementations in predicting court decisions or classification of legal data [8]. Soh et al. [13] explored multiple statistical machine learning approaches, out-of-the-box deep learning models such as BERT and a shallow convolutional neural network to classify 6,277 Singapore Supreme Court Judgements into their 31 different legal areas. Their results showed that the statistical model performed best with a micro-F1 of 63.2 and a macro-F1 of 73.3 [13]. Our problem statement differs from Soh et al. [13] as they classified the area of law utilizing the entirety of a case while we wish to predict the outcome. While in machine learning, 3-4 instances of a sample is considered to be too few, in law, 3-4 samples of precedence are considered to be plenty. Few-shot predictors attempt to generalize classes with few training samples. Luo et al. [20] proposed a model for predicting criminal charges leveraging related law articles. They used a hierarchical attention mechanism to create a document representation and a different stack of attention components was trained to select the best supporting legislative statutes for a given case. Their model was trained on 50,000 case documents extracted from China Judgments Online[3], and only predicted criminal charges that had at least 80 cases. For the sake of simplicity, the authors only considered cases in which there was a single defendant. Luo et al. [20] reported an F1 micro/macro of 90.21/80.48 and compared the performance of their model to other baselines they had built. Hu et al. [14] performed three experiments that trained multiple baselines including the model presented by Luo et al. [20], on three datasets comprising 61,589/153,521/306,900 factual case summaries generated from China Judgments Online. They reported better results than Luo et al. [20] macro-F1 scores of 64.0/67.1/73.1 on their small/medium/large datasets, respectively. We note in Hu et al. [14] they segment the maximum document length to 500 and China Judgements Online appear to be a database of fact descriptions not full cases. Chalkidis et al. [6], used a dataset of European Court of Human Rights(ECHR) cases totalling 11,500. They reported the results of a variety of deep learning architectures on three tasks: binary classification, multi-class classification and case importance prediction. They further introduced a Hierarchical-BERT, which first produced fact embeddings which are used with a self-attention mechanism to formulate the document embedding. Their Hierarchical-BERT was their performing model in both their binary and multi-label classification tasks with an F1 of 82 and 60.8. In their case importance task, their majority-class classifier achieved the lowest mean-squared error of 0.369, with Hierarchical-BERT achieving the best Spearman's ρ of .527. Although we implement similar models to Chaldkidis et al. [6], we note that our task of predicting an outcome differs from the classification of an entire document.

We note that as of this writing, there does not appear to be existing literature on using deep learning for the prediction of reasonable notice from free text.

3 PROBLEM DESCRIPTION

The aim of this project is to predict how judges determine 'reasonable notice' in employment termination cases. As argued earlier, if the employment relationship is governed by an indefinite contract and the employer wishes to terminate the employment relationship for any reason, the employer has the obligation to provide notice — usually denoted in months — or pay in lieu of notice, calculated according to the Bardal factors[4].

While the primary goal of this research is to assess the predictive power of deep learning models when it comes to notice of termination, it is worth noting that this research is drawn from a larger project aiming at developing an open-source system for small-claims disputes, including employment disputes: MyOpenCourt. This system aims to promote access to small-claims justice and provide legal help to self-represented litigants by democratizing legal analytics technology. Like many AI legal tools, MyOpenCourt provides legal information that requires users to fill out a multiple-choice questionnaire and outputs a single numerical prediction along with a list of relevant precedents.

Instead, in this research we propose a system that outputs a prediction of reasonable notice based on a free-text summary of the case law. We used deep learning in conjunction with Natural Language Processing (NLP) to calculate the period of reasonable notice from manually typed summaries of adjudicated cases. The summaries are unstructured text data written in plain English (i.e. not "legalese"), collected from WestLaw's Quantum service[5]. These summaries contain sufficient information for a trained legal professional to approximately determine the notice award without looking at the outcome of the case. We decided to use summaries

[3]http://wenshu.court.gov.cn/

[4]Payment in lieu of notice' is immediate compensation at an amount equal to that an employee would have earned as salary or wages by working through the whole notice period

[5]https://www.westlawnextcanada.com/quantums/

instead of entire cases because of the the "fussy" nature of common law text [11]. The lack of styling in Canadian cases made it difficult to automatically parse the extraction of the legal facts. As our goal is to predict the outcome of a case, we require the inputs to not contain any mention of the legal analysis or the outcome. The full legal cases are long, often more than ten pages and averaging 5,000 words, and introduce ambiguity through an abundance of information that must be carefully filtered to extract useful information. Furthermore, legal cases are decided by a multitude of judges, which leads to many idiosyncrasies in writing styles and case structuring.

4 DATA

The WestLaw Quantum Service[5] provides a brief synopsis of the case, often including the judgement on the reasonable notice period. Since the goal of this research is to predict reasonable notice period, any mention of the judgement regarding the notice period was manually removed, leaving only factual descriptions of the plaintiff and the characteristics of the case. We prepended the input with the year of the judgement, occupation category, age, salary, job title and duration of employment of the plaintiff extracted from the summaries. If the information could not be found, nothing was prepended to the summary. The summaries appeared to be written in plain English by human writers. Each case outcome is considered to be its own class, with outcomes of 25 months or greater being grouped together. Our classification task had a total of 25 classes.

5 MODELS AND METHODS

Our dataset totaled 1,695 cases for training and 409 for testing. We did not utilize a development set and evaluated the final models on the testing set. All experiments were completed on an IBM Power8 server with 512 GB of memory, 64 cores (hyper-threaded to 128), four Nvidia K80 GPUs, Red Hat Enterprise Linux Server release 7.6 and ppc64le architecture.

5.1 Hierarchical Attention Network

Extracting deep semantic and contextual understanding from text data is essential for every NLP task. Bahdanau et al. [4] first proposed attention mechanisms for machine translation by learning how to align the original text with the translated words. Rather than the traditional approach of attempting to distill an entire document into a vector, Yang et al. [23] introduced the Hierarchical Attention Network (HAN) to encode smaller chunks of text that are then used to inform future encodings. The HAN first learned the importance of each word which informed its sentence embedding, and a separate attention mechanism learned the importance of each sentence to inform the final document representation. In our HAN we used SpaCy sentence bound detection and tokenization [12].

We utilized pretrained 200-dimension GloVe vectors with a LSTM containing hidden dimensions of 75 and an attention dimension of 50. We optimized with a stochastic gradient descent optimizer with a learning rate of 0.06 , a batch size of 32, a momentum of 0.9 and dropout of 0.5. The learning rate was reduced by a factor of 0.95 when our performance stopped improving. Each epoch took approximately six minutes to execute.

5.2 Few-shot Learning

While in machine learning 3-4 instances of a sample is considered to be too few, in law, 3-4 samples are considered to be plenty. Few-shot models often utilize a method which only requires a few instances of a training sample to be able to generalize. Given our sparse dataset and the success Hu et al. [14] had demonstrated in criminal law, we implemented a modified version for the prediction of reasonable notice. We followed the implementation details presented in Hu et al. [14] except we adopted the sentence embeddings presented by Lin et al.[18] and generated $\boldsymbol{r}$ number of attention vectors for each attribute, as the original model yielded poor results. The architecture presented by Hu et al. [14] created a document representation by combining an attribute-aware embedding with one that is attribute-free. The attribute-aware embedding resulted from an average pooling of the sentence embeddings from four stacked self-attention mechanisms. For a single mechanism we predict an attribute (e.g. a person's age or duration) by self-attending to multiple parts of the text simultaneously. Four labels were used to train the attribute-aware mechanisms to predict the length of employment, age of employee, character of employment, and availability of similar employment. These labels were hand-annotated by a team of Queen's Law students. The attribute-free embedding comprised a max-pooling of the hidden states generated by the encoder.

We utilized pretrained 300-dimension GloVe vectors that were fine-tuned, a hidden dimension of 300, and dropout of 0.5. An attention $\boldsymbol{r}$ of 30 was used. We used an Adam optimizer with a learning rate of 0.001 and reduced the learning rate when a metric has stopped improving by a factor of 0.95. Alpha, the scaling on the attribute-aware loss, had a value of 0.3 and each epoch took approximately 8 minutes to execute.

5.3 BERT-esque and Domain adaptation

The Bidirectional Encoder Representations from Transformers (BERT) from Devlin et al. [9] has recently laid the foundation of pretraining models by leveraging language modelling, transfer learning, and fine-tuning on downstream tasks. As one of the main strengths of BERT is its generalized understanding of the English language, and pretrained BERT models have been publicly released[6], we leveraged RoBERTa for predicting reasonable notice. We further experimented with Robustly Optimized BERT Pretraining Approach (RoBERTa), which held the top spot in the GLUE benchmarks [19] during the course of our research. Architecturally, RoBERTa did not differ from the original BERT model; instead, Liu et al. [19] utilized an additional 160GBs of data and further fine-tuned hyperparameters. Through experimentation of learning rates and batch sizes, the authors determined that the next sentence prediction only offered marginal to no performance improvements. Using only hyperparameter fine-tuning and additional data, RoBERTa achieved an almost 7% improvement over the original BERT model on the GLUE benchmarks.

5.3.1 Domain Adaptations. Following common practice for improving pretrained model performance, we experimented by domain-adapting our BERT-esque models on full reasonable notice case

[6] https://github.com/google-research/bert

texts as well as the Harvard case law dataset [21]. In our BERT implementations, we utilized the $BERT_{base}$ model from HuggingFace[7]. We further experimented with domain adapting a $RoBERTa_{base}$ model using Facebook AI's implementation[8]. In addition, we further pretrained both models using only the masked language modelling (MLM) criterion, consistent with the results from Liu et al. [19]. Five epochs were used for all MLM pretraining. For our end results, pretrained language models were further trained for text classification with ten epochs and fine-tuned on 409 remaining cases. Classification training was performed using a batch size of 16, using the default losses and optimizers. The classification head of BERT took 20 minutes to train while MLM took 2.5 hours. For BERT, we fine-tuned on the full case text that correspond to each of the respective 1,695 reasonable notice cases in our training set.

In our RoBERTa implementation, we fine-tuned on approximately four million cases from Harvard's case law project. We determined cases before 1960 to be linguistically different from present-day legal documents and thus these cases were removed. We note that the Harvard case law project only includes cases from the United States. To create an accurate comparison with our BERT implementation, we domain-adapted $RoBERTa_{base}$ to the same set of full case texts. The classification head of RoBERTa took 30 minutes to fine-tune, while MLM took 3 hours. Our batch size was 256 and peak learning rate was 0.0001 in accordance with Liu et al. [19].

6 RESULTS AND DISCUSSION

Approach	Acc. (+/-2)
HAN	67%
Few-shot w/ Self-attention	51%
BERT+base	61%
BERT+full cases	49%
RoBERTa+full cases	63%
RoBERTa+Harvard	65%
RoBERTa+base	**69%**

Table 1: Summary of results for predicting the number of months awarded for reasonable notice using case summaries.

The output of our system was classified as correct if it was within +/-2 months of the ground truth label to account for situational variability (i.e.Eq 1). We refer to this as the output window.

$$groundtruth - 2 \leq prediction \leq groundtruth + 2 \quad (1)$$

Our $RoBERTa_{base}$ model had the highest accuracy of 69%. A summary of the results can be seen in Table 1.

Interestingly, our best-performing model was $RoBERTa_{base}$ out-of-the-box and not a domain-adapted version. Our HAN out-performed the majority of our pretrained models and was only marginally worse than $RoBERTa_{base}$. The performance of HAN may be attributable to architectural structuring, as each sentence in our case summaries roughly contains a statement of fact, allowing our HAN to learn the importance of each sentence (fact) and weigh it accordingly prior to creating a document representation. This may be replicating the thought process of the judiciary. Furthermore, we believe our mixed results from our few-shot model may be attributable to the size of our dataset. In Hu et al. [14], their smallest training dataset contained over 61,000 training samples, compared to our 1,695. The issue of data in our few-shot model appear to be supported with the performance of our BERT-esque models in which the majority of models performed better. BERT-esque models are pretrained to have a generalized understanding of language out of the box, making it easier to fine-tune on a specific classification task. Furthermore, the task of predicting reasonable notice requires additional knowledge beyond an understanding of the natural language. In fact, it requires knowledge on judicial bias and dispute settlement - something that seasoned employment lawyers have built throughout their interactions with judges and colleagues and cannot be learned from case law. This may partially explain our mixed results insofar as the majority of employment disputes are resolved through negotiation. Thus, considering that the case law constitutes only a small piece of the data, it may be argued that our model could have performed better had it be trained on settlement agreements.

In our experiments we utilized a classification approach over a regression as we believe this best replicates the decision-making process of a judge. In deciding the amount of reasonable notice a plaintiff should receive one could presume a judge would locate similar past cases and adjust their ruling based on differences of fact. We utilize classification to anchor the number of months of reasonable notice and broaden the output window in our metric to account for the differences of fact. In addition, our findings in Dahan et al. [8], in which we used regression along with a hand-labeled tabular dataset, indicated regression to be a poor predictor of reasonable notice.

A key finding of this research is that our domain adaptations did not yield significant improvements when compared to the out-of-the-box pretrained models, as reported in Rietzler et al. [21] and commonly noted as a promising avenue for other domains (e.g. SciBERT [5] and BioBERT [17]). Instead, domain adaptations appeared to negatively affect the performance of both our *RoBERTa* and *BERT* models. Despite the language of the Canadian reasonable notice cases being more similar to our case summaries than the Harvard Case Law dataset, our results from $RoBERTa_{base}$, domain-adapted on the Harvard dataset, performed slightly better than the ones trained on full Canadian cases. This finding was in line with one of the conclusions set out by Liu et al. [19], which emphasized the importance of volume for datasets. Furthermore, as Liu et al. [19] performed extensive hyperparameter experimentation we opted to utilize their recommended learning rates and batch sizes. It appears current deep learning solutions may not be able to accurately grasp the many unique characters of each legal dispute. Unfortunately, our performance suggests that in spite of recent advances, deep learning may not be able to accurately predict reasonable notice from free text. That said, our mixed results may also be explained by the fact that judges are inherently unpredictable when it comes to the application of the Bardal test, and thus predicting notice is an almost impossible task for deep learning models, but also for experienced lawyers.

[7] huggingface.co
[8] https://github.com/pytorch/fairseq

7 CONCLUSION AND FUTURE WORK

In this paper we applied multiple deep-learning solutions to human-written case summaries to classify the reasonable notice period a plaintiff should be awarded. Our best performing model was $RoBERTa_{base}$ which achieved a 69% accuracy with a +/-2 month window. Domain adaptations negatively affected our performance in the case of RoBERTa, and marginally improved performance in the case of BERT. Given the significant successes of deep learning in various domains, the relatively poor efficiency found in predicting notice periods may come as a surprise. In fact, results from this paper appear to be consistent with our previous findings in Dahan et al. [8], suggesting the inconsistencies in the case law make it difficult to predict reasonable notice accurately. Thus, it can be argued that our results have very little to do with the quality of the model, and that it may not be possible to reduce our prediction error, mainly because of the inherently inconsistent nature of the dataset.

At the time of our research, RoBERTa was the top-performing model on the GLUE benchmarks [22], where it ranks 10^{th} as of this writing. Utilizing the top performing model from the benchmarks may yield better results. While our domain adaptation used 1,695 full cases, and the superior performance of domain adapting on an adjacent domain of American law, collecting a larger dataset of Canadian cases may be an interesting avenue to explore for further experimentation. While deep learning and BERT-esque models have proven successful in numerous domains, it appears a gap still exists for deep learning applications in the legal field. That said, other areas — such as art, which has been considered too artistic to be impacted by AI — have shown some recent successes. In particular, deep learning models have successfully been trained to paint like humans [15], which leads us to be optimistic about future applications in the legal field. We believe that further advances in the field of NLP and deep learning will be able to perform well on this task in the future.

ACKNOWLEDGMENTS

We would like to thank the team at the Conflict Analytics Lab, the Scotiabank Centre for Customer Analytics and the Center for Law in the Contemporary Workplace at Queen's University (Jonathan Touboul, Maxime Cohen, Kevin Banks, Simon Townsend, William Quiglietta, Zach Berg, Mackenzie Anderson, Brandon Loehle, Max Saunders, Sean Gulrajani, Shane Liquornik, Yuri Levin, Mikhail Nediak, Stephen Thomas) as well as our colleagues Joshua Karton and Bill Flanagan for supporting the project as well as contributing to the development of the idea.

REFERENCES

[1] [n.d.]. Bardal v. Globe & Mail Ltd. (1960), 24 D.L.R. (2d).
[2] [n.d.]. Fraser v. Canerector Inc., 2015 CarswellOnt 4796, 2015 ONSC 2138 at para 32.
[3] [n.d.]. Wallace v. United Grain Growers Ltd., 1997 CanLII 332 (SCC), (1997) 3 S.C.R. 701.
[4] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473* (2014).
[5] Iz Beltagy, Arman Cohan, and Kyle Lo. 2019. Scibert: Pretrained contextualized embeddings for scientific text. *arXiv preprint arXiv:1903.10676* (2019).
[6] Ilias Chalkidis and Dimitrios Kampas. 2019. Deep learning in law: early adaptation and legal word embeddings trained on large corpora. *Artificial Intelligence and Law* 27, 2 (2019), 171–198.
[7] Bruce Curran and Sara Slinn. 2016. Just Notice Reform: Enhanced Statutory Termination Provisions for the 99%. *Osgoode Legal Studies Research Paper* 61 (2016), 2017.
[8] Samuel Dahan, Jonathan Touboul, Jason Lam, and Dan Sfedj. 2020. *Predicting Employment Notice with Machine Learning: Promises and Limitation.*
[9] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805* (2018).
[10] David J Doorey. 2020. The Law of Work: Common Law and the Regulation of Work. *Regulation* 285 (2020).
[11] James Holland and Julian Webb. 2013. *Learning legal rules: a students' guide to legal method and reasoning.* Oxford university press.
[12] Matthew Honnibal and Mark Johnson. 2015. An Improved Non-monotonic Transition System for Dependency Parsing. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Lisbon, Portugal, 1373–1378. https://aclweb.org/anthology/D/D15/D15-1162
[13] Jerrold Soh Tsin Howe, Lim How Khang, and Ian Ernst Chai. 2019. Legal Area Classification: A Comparative Study of Text Classifiers on Singapore Supreme Court Judgments. *arXiv preprint arXiv:1904.06470* (2019).
[14] Zikun Hu, Xiang Li, Cunchao Tu, Zhiyuan Liu, and Maosong Sun. 2018. Few-shot charge prediction with discriminative legal attributes. In *Proceedings of the 27th International Conference on Computational Linguistics*. 487–498.
[15] Zhewei Huang, Wen Heng, and Shuchang Zhou. 2019. Learning to paint with model-based deep reinforcement learning. In *Proceedings of the IEEE International Conference on Computer Vision*. 8709–8718.
[16] Grant Lamond. 2006. Precedent and analogy in legal reasoning. (2006).
[17] Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2020. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics* 36, 4 (2020), 1234–1240.
[18] Zhouhan Lin, Minwei Feng, Cicero Nogueira dos Santos, Mo Yu, Bing Xiang, Bowen Zhou, and Yoshua Bengio. 2017. A structured self-attentive sentence embedding. *arXiv preprint arXiv:1703.03130* (2017).
[19] 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* (2019).
[20] Bingfeng Luo, Yansong Feng, Jianbo Xu, Xiang Zhang, and Dongyan Zhao. 2017. Learning to Predict Charges for Criminal Cases with Legal Basis. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. 2727–2736.
[21] Alexander Rietzler, Sebastian Stabinger, Paul Opitz, and Stefan Engl. 2019. Adapt or get left behind: Domain adaptation through bert language model finetuning for aspect-target sentiment classification. *arXiv preprint arXiv:1908.11860* (2019).
[22] Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R Bowman. 2018. Glue: A multi-task benchmark and analysis platform for natural language understanding. *arXiv preprint arXiv:1804.07461* (2018).
[23] Zichao Yang, Diyi Yang, Chris Dyer, Xiaodong He, Alex Smola, and Eduard Hovy. 2016. Hierarchical attention networks for document classification. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. 1480–1489.