Received: 15 August 2024 | Revised: 28 February 2025 | Accepted: 10 March 2025 | Published online: 8 April 2025

RESEARCH ARTICLE

Journal of Computational and Cognitive Engineering
2025, Vol. 4(3) 319–331
DOI: 10.47852/bonviewJCCE52024104

BON VIEW PUBLISHING

Legal Text Analytics for Reasonable Notice Period Prediction

Jason Lam[1], Yuhao Chen[1,*], Farhana Zulkernine[1] and Samuel Dahan[2]

[1] *School of Computing, Queen's University, Canada*

[2] *Department of Law, Queen's University, Canada*

Abstract: Applications of deep learning (DL) to generate text embeddings and natural language processing (NLP) have shown wide success in semantic interpretations of domain-specific text data when applied to downstream tasks such as predicting the next word, information extraction for classification, analyzing social media feeds, classifying text, and creating compressed representations. While DL and NLP have been widely applied across numerous domains, researchers have recently begun to apply these techniques to the field of law due to the challenges in processing legal case descriptions. Attention-based models have shown promising results in predicting criminal charges using unstructured text as an input, but little work has been done on data representing the Canadian legal system, especially employment law. The legal field poses many challenges, such as the amount of legal data publicly available in Canada, the verbosity of judgments, the legal jargon used in judgments, and the subjectivity of outcomes that pose many challenges in processing legal text data. Many of the state-of-the-art systems require expensive hand-annotated labels that are often unobtainable. In this study, we investigate the prediction of reasonable notice for termination of employment in the field of law. To address these challenges, we propose domain-adapted BERT variations specifically trained for legal texts. We assess the performance of various attention-based and pre-trained models using human-typed summaries of legal judgment and present a detailed analysis of the data and the results to provide insights for further exploration in this area. Our approaches also provide interesting insights about this specific type of legal case focusing on employment law, given the subjective nature of judges and the variability in outcomes.

Keywords: deep learning, reasonable notice prediction, employment law, legal analytics

1. Introduction

In Canada's common law system, employment law is derived from case law precedence and statutory legislation [1]. Case law precedence represents principles established in previous legal cases decided by a presiding judge. With precedence being one of our primary sources of law, much of the law is open to interpretation and subjectiveness. Judgment prediction, therefore, poses difficult challenges in processing the jargon of legal text and learning the contextual mappings for the classification of judgments regarding termination of employment.

According to legal precedence, reaffirmed by the Supreme Court of Canada, the amount of reasonable notice awarded should be based on the following four factors, known as the Bardal factors: (1) age of the employee, (2) length of service, (3) nature of employment, and (4) availability of similar employment [2]. While the law states that the judge should utilize the Bardal factors in wrongful dismissal cases, it does not indicate how much emphasis should be given to each factor and whether the factors must be utilized [2]. Furthermore, over the years, Canada's state of economy such as the recession, has contributed to additional variances in reasonable notice calculations and impacted the availability of similar employment. In past cases, the Federal Court of Appeal has repeatedly emphasized that "[the determination of reasonable notice] is not an exact science, but a discretionary power that must take into account all relevant circumstances" [3]. In addition, there were other complications in many cases, such as in the Hawkes Dismissal case, where the employer, Max Aicher, induced Mr. Hawkes to leave his job by making empty promises [4]. In the end, the judge awarded Mr. Hawkes higher compensation. Based on this incident, the judge needs to consider factors other than the Bardal factors when deciding to proceed with the trial.

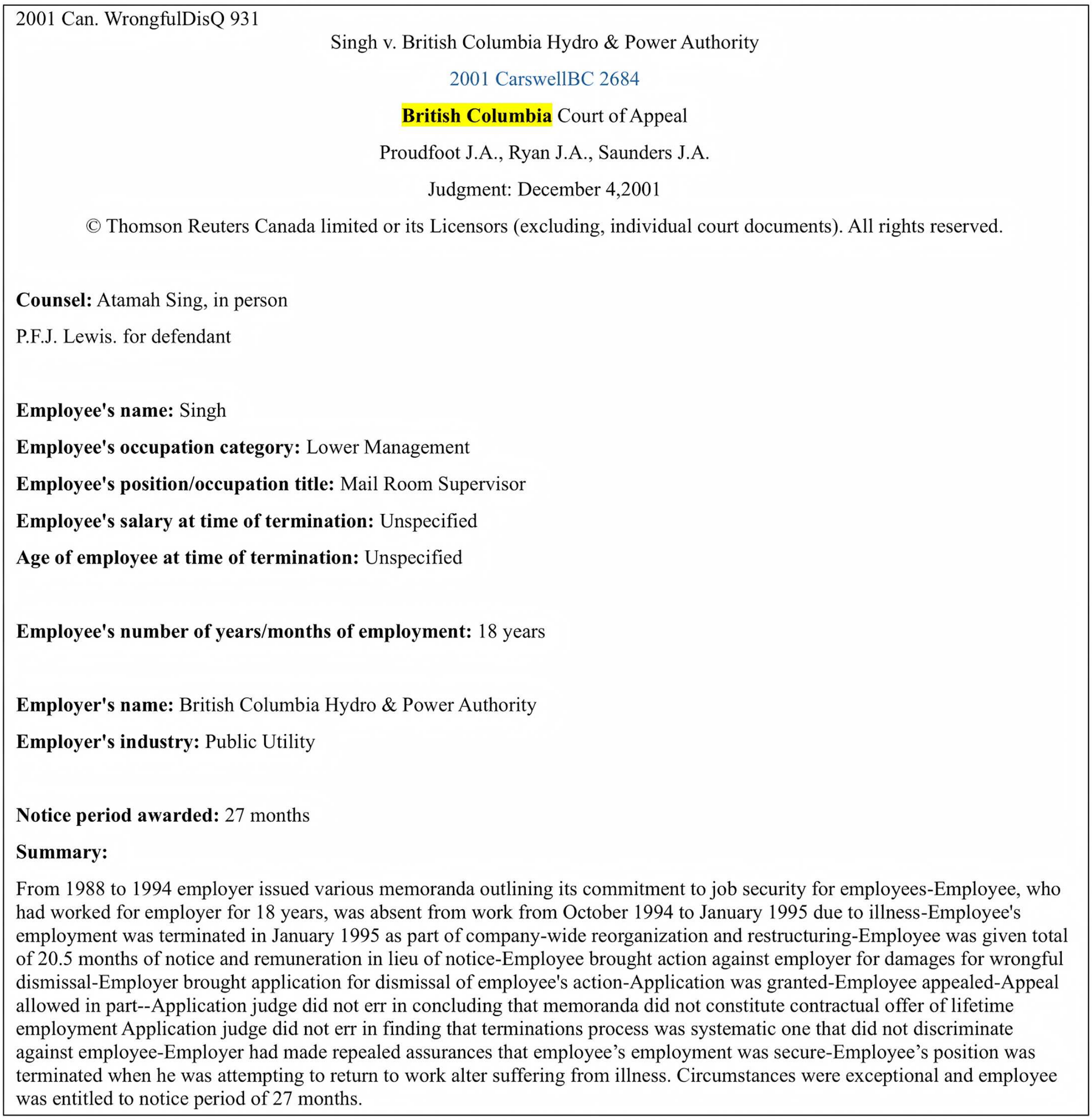
We propose a natural language processing (NLP) and deep learning (DL)-based approach for calculating the reasonable notice period using handwritten (typed) summaries of legal case descriptions as shown in Figure 1. The summaries are unstructured text data written in plain English (i.e., not legalized), which we collected from Westlaw's Quantum service.

Employees are entitled to numerous rights governed by Canada's labor law[1]. These laws regulate the relationship between an individual employee and the employer. In employment law, when employees are terminated, they have the right to be fairly compensated based on several factors. Specifically, under Canadian precedence, when employees are terminated, they are entitled to receive compensation in lieu of reasonable notice. In many cases, the reasonable notice period that an employee is entitled to becomes

***Corresponding author:** Yuhao Chen, School of Computing, Queen's University, Canada. Email: yuhao.chen@queensu.ca

[1] https://laws-lois.justice.gc.ca/eng/acts/L-2/

Figure 1
Example of data from Westlaw Quantum services



2001 Can. WrongfulDisQ 931

Singh v. British Columbia Hydro & Power Authority

2001 CarswellBC 2684

British Columbia Court of Appeal

Proudfoot J.A., Ryan J.A., Saunders J.A.

Judgment: December 4,2001

© Thomson Reuters Canada limited or its Licensors (excluding, individual court documents). All rights reserved.

Counsel: Atamah Sing, in person

P.F.J. Lewis. for defendant

Employee's name: Singh

Employee's occupation category: Lower Management

Employee's position/occupation title: Mail Room Supervisor

Employee's salary at time of termination: Unspecified

Age of employee at time of termination: Unspecified

Employee's number of years/months of employment: 18 years

Employer's name: British Columbia Hydro & Power Authority

Employer's industry: Public Utility

Notice period awarded: 27 months

Summary:

From 1988 to 1994 employer issued various memoranda outlining its commitment to job security for employees-Employee, who had worked for employer for 18 years, was absent from work from October 1994 to January 1995 due to illness-Employee's employment was terminated in January 1995 as part of company-wide reorganization and restructuring-Employee was given total of 20.5 months of notice and remuneration in lieu of notice-Employee brought action against employer for damages for wrongful dismissal-Employer brought application for dismissal of employee's action-Application was granted-Employee appealed-Appeal allowed in part--Application judge did not err in concluding that memoranda did not constitute contractual offer of lifetime employment Application judge did not err in finding that terminations process was systematic one that did not discriminate against employee-Employer had made repealed assurances that employee's employment was secure-Employee's position was terminated when he was attempting to return to work alter suffering from illness. Circumstances were exceptional and employee was entitled to notice period of 27 months.

a matter of primary contention between employees and employers. Determining this period is one of the most litigated issues in the Canadian judicial system [1, 3]. The reasonable notice period can vary for employees depending on the length of employment and availability of similar employment, among other reasons.

By developing a tool to predict the reasonable notice period, we can provide the general public with improved access to justice. As the cost of litigation is generally very high, the general public is often hesitant to pursue legal remedies. This research aims to provide the public with an estimation of reasonable notice to help them weigh the potential rewards against litigation costs. Furthermore, we believe that an artificial intelligence (AI) solution to reasonable notice can help employers and employees negotiate settlements by providing an anchoring point for discussing a severance package. The primary research questions addressed in this study are:

1) Given a *structured* input of Bardal factors, can we predict a reasonable notice period? The structured inputs must be extracted by the domain experts, which is expensive and time-consuming.
2) Given a *free text summary* of a case, including the plaintiff's description, how can we predict the reasonable notice period that a person should receive? This can eliminate the need to manually extract the structured input.

The broader goal of our collaborative research is to propose a system that (1) predicts the notice periods or severance rewards and (2) explains which parts of the input contribute to the predicted notice period.

1.1. Contribution

The main contributions of this work are as follows:

1) We combine multiple state-of-the-art (SOTA) DL models to design and implement our novel multi-attention DL model to predict a reasonable notice period focusing on Canada's employment law.
2) We implement an end-to-end pipeline to preprocess legal summaries, extract key features from the data using NLP techniques, and use the extracted features to train and validate our DL model. For the prediction of reasonable notice under employment law using DL, we believe we are one of the first researchers to address this issue.
3) We compare our model's performance with two top-performing pre-trained DL models at the time: Bidirectional Encoder Representations from Transformers (BERT) [5] and Robustly Optimized BERT multi-attention Approach (RoBERTa) [6] for predicting judgments.
4) In addition, we domain adapt BERT to Legal Adapted BERT (LA-BERTbase) and RoBERTa (LA-RoBERTabase) using approximately 4 million cases from Harvard's case law project. Due to their substantial length, often tens to hundreds of pages, legal documents present significant challenges in analysis and processing [7]. Pretraining models like RoBERTa are highly adaptable, allowing us to fine-tune them for the specifics of Canadian legal cases. This approach provides a practical solution for addressing Canada's lack of open-source datasets while still developing a model that is grounded in Canadian legal standards.

The rest of the paper is organized as follows. To contextualize our approach within the current landscape of DL and legal text analytics, we explore recent advancements in NLP applications within DL and legal contexts in Section 2. The methodology containing an overview of our approach is provided in Section 3. Section 4 presents the implementation details. Experimental results for validation are presented in Section 5. The paper concludes in Section 6 with a discussion of possible future work directions.

2. Background and Related Work

There has been an increase in providers of subscription-based AI-legal services (e.g., Blue J Legal[2]), where these services utilize machine learning (ML) to assist large legal firms. A client using these services must fill out pre-defined forms with drop-down menus to provide the necessary information for the system to output a prediction. Unfortunately, only allowing a person's case to be defined by preset options fails to capture the minute details of a case, resulting in the loss of important information. Recent research has emphasized these limitations, noting that AI systems that rely on structured inputs often miss fine details that are critical to legal judgments [8]. Alternatively, DL is an approach that learns the case features independently, automatically determines what is important, and can be trained on the unstructured input of raw text data. This flexibility makes DL particularly suited to legal tasks, as the complexity and variability of cases require more sophisticated models capable of handling unstructured data [8, 9].

2.1. Deep learning models

Attention-only models have been dominating NLP-based General Language Understanding Evaluation (GLUE) benchmarks [10, 11]. Vaswani et al. [12] introduced a novel non-recursive encoding architecture called Transformer that achieved SOTA performance in NLP benchmarks and was significantly faster to train. Previous attention mechanisms [2] with recurrent neural networks (RNNs) require an iterative process as new states are generated from previous states, resulting in a serialized pipeline. Instead, Transformer models apply a parallel computation architecture and calculate the attention weights with a single matrix multiplication, which are fed into the decoder at once. Therefore, we opted to explore Transformer-based models in this study.

Bahdanau et al. [13] were the pioneers in proposing attention mechanisms for machine translation, aiming to capture more intricate meanings. Yang et al. [14] developed the Hierarchical Attention Network (HAN), which encodes smaller text segments to inform subsequent encodings, moving beyond the traditional approach of compressing entire documents into single vectors.

Lin et al. [15] reported significant improvements in results using their self-attention-based sentence embedding method on three different challenges: determining the age of users from a dataset of 68,485 randomly selected Twitter tweets, performing sentiment analysis on 500,000 randomly selected Yelp reviews and classifying their ratings into five classes, and computing textual entailment by training on 570,000 samples from the stanford natural language inference (SNLI) corpus. In the last challenge, they achieved 84.4% accuracy compared to the SOTA performance of 84.6%. The authors mentioned that a major limitation of their work was that the model was heavily reliant on the end discriminative model, not allowing this framework to be easily trained using unsupervised learning.

The seminal methodology of Transformer has led to numerous architectures. One of the most well-known examples is BERT [5], which uses stacked Transformer encoders to train a language model. It achieved new SOTA results on multiple NLP benchmarks. RoBERTa [6] is a variation of BERT, which applies several approaches to improve and optimize BERT. In recent years, researchers have been exploring the capabilities of BERT variations for applications in specific domains, including the legal [16, 17], medical [18, 19], and social media [20, 21] domains. These studies aim to assess how effectively BERT models can handle domain-specific linguistic patterns and data requirements, potentially enhancing the performance of NLP tasks in these specialized areas.

2.2. NLP application in legal analytics

In academic literature, the application of NLP in legal data analytics is still new, and only a few implementations exist [22, 23] in predicting court decisions or the classification of legal data [3] that are used in practical and commercial systems. Dahan et al. [3] focused on the same problem and trained various statistical ML methods on manually extracted and curated structured data features from the Westlaw case data on the awarded notice period for termination of employment. The extracted data represented the Bardal factors. An analysis of the correlation of the various Bardal factors with the judgment is presented in the paper to validate the ML algorithms, which demonstrate a strong correlation of the Bardal factors with the judgment compared to the other features of the legal cases [3].

Howe et al. [24] explored several ML approaches, including pre-trained DL models such as BERT and a shallow convolutional

[2] https://www.bluejlegal.com/

neural network, to classify 6,277 Singapore Supreme Court judgments across 31 legal areas. Their findings demonstrated that the statistical model outperformed the others, achieving a micro-$F1$ score of 63.2 and a macro-$F1$ score of 73.3 [24]. Notably, while ML typically requires large sample sizes for robust model training, the legal domain often considers 3–4 case precedents as sufficient for establishing legal principles.

Luo et al. [25] introduced a predictive model for determining criminal charges using relevant legal articles. They applied a hierarchical attention mechanism to generate a comprehensive document representation. They trained a series of attention layers to identify the most relevant legislative statutes for each case. The model was built on 50,000 case records from China Judgments Online, focusing exclusively on criminal charges with at least 80 instances to ensure adequate data for training. To simplify their analysis, the study was limited to cases with a single defendant. The model achieved F1 scores of 90.21 (micro) and 80.48 (macro), outperforming several baseline models.

Hu et al. [26] conducted three experiments on automatic charge prediction, a task that aims to predict final charges based on descriptions of facts in criminal cases. In their study, they performed three experiments that trained multiple baselines including the model presented by Luo et al. [25], across three datasets containing 61,589, 153,521, and 306,900 factual case summaries from China Judgments Online. They outperformed Luo et al. [25] achieving macro-$F1$ scores of 64.0, 67.1, and 73.1 on their small/medium/large datasets, respectively.

Medvedeva et al. [27] investigated ML techniques for predicting decisions of the European Court of Human Rights (ECtHR).

Li et al. [28] employed parameter-efficient techniques that leveraged extensive legal data extraction from public legal forums to adapt DL models to the legal domain using unsupervised learning. The outcomes demonstrated that the approach could achieve calibration comparable to existing SOTA models across several tasks while utilizing a reduced number of training parameters. However, it is noteworthy that none of these explored DL techniques for analyzing legal text data.

Lam et al. [29] introduced an approach to enhance a pre-trained domain-adapted language model. They employed the model for text embedding by incorporating a classification layer. This augmentation aimed to facilitate the acquisition of significant legal attributes for the purpose of detecting similar legal cases of personal injury.

Before making a final decision for a legal judgment, people often analyze and compare similar legal cases or possible charges to help them decide [30]. Based on this idea, Zhang et al. [30] proposed a contrastive learning framework to capture the fine-grained differences between similar law cases for legal judgment prediction (LJP). Their framework achieved SOTA performance on real-world datasets.

Building on existing methodologies, our approach applies and extends domain-specific models to explore the unique aspects of Canadian employment law in predicting reasonable notice periods.

3. Methodology

This research aims at providing transparent legal services to the general public in terms of informed advice regarding the notice period an employee can expect as a decision by our judicial system using predictive analytics with DL models. Existing legal services based on AI techniques typically offer black-box legal advice by requiring the consumers to fill in a brief questionnaire with *yes* or *no* answers. The information is used to produce a single numerical value indicating the predicted notice period along with a generic explanation[3]. Such systems may be useful to legal professionals, but not to the general public, who need to know their chances of getting good compensation if they take their cases through the courts.

3.1. Overview of the approach

In employment law, reasonable notice is specified as the number of months for which an employee gets compensated after the termination of a job. We map the prediction of a reasonable notice period as a classification task and develop DL models to predict one of 25 classes, where classes 1 to 24 indicate 1 to 24 months of notice respectively, and all values greater than 24 are classified as class 25.

Based on the existing literature and the suitability of the SOTA models for the research problem, we selected several models to implement and then improved their performance. Due to a lack of existing literature on utilizing DL in legal text analytics and the data we used, we implemented a variety of DL models to compare our model performance for validation and developed the best use case. We observed that certain outcomes (e.g., 3, 6, 12, and 18 months of awarded notice period) have a significant number of cases in the training samples compared to the other outcomes (e.g., 7, 8 months). Unbalanced training data can result in bias in the trained models. Few-shot models are designed to work with unbalanced and small training datasets [31]. We, therefore, chose the few-shot approach as a starting point in our research, given the lack of data. Justifications for selecting the models we implemented are given below:

1) Few-shot model [26]: Due to the simplicity and ability to attend to and learn data features that influence the prediction outcome from a small set of training data using few-shot learning, we selected the few-shot model proposed by Hu et al. [26] to address our problem. We later attempted to augment our few-shot approach with the self-attention mechanism from Lin et al. [15].
2) Self-attend model [15]: We thought that a simpler model utilizing only the self-attention mechanism from Lin et al. [15] could lead to a more generalizable model. The model demonstrated its ability to train numerous layers of attention to focus on multiple topic areas. We hypothesized that each attention layer could attend to a specific Bardal factor.
3) Hierarchical Attention Network (HAN) [14]: Given that our case summaries had a natural hierarchical structure, where each fact of a case was roughly contained in a single sentence, we thought that the HAN would perform well [14]. The model can learn to attend to information within the input data that leads to the correct prediction. Thus, it has the potential to inform users about the influencing factors that lead to the judgment predicted by the model about the notice period (i.e., the facts that contribute the most toward the severance calculation). We implemented a HAN [14] utilizing case summaries to potentially allow better explainability in predicting reasonable notice awards.

We implemented the following models for validating our approach:

1) Baseline models: We trained and evaluated numerous statistical models and selected two best-performing ML models: namely, multi-linear regression (MLR) and gradient boosting regressor (GBR), as the baseline models to compare model performances. These models were trained on manually extracted data features

[3] https://www.bluejlegal.com/ca

instead of unstructured text summaries. The data features correspond to the factors defined explicitly in the Bardal case law, such as the age of the employee, character of employment, availability of similar jobs and length of employment. We wanted to compare the DL model performance with statistical ML models trained on handcrafted data, which is an expensive and few-shot process as done in earlier work [3, 29].

2) Time-consuming models: Since Transformer-based architectures have been dominating NLP benchmarks (e.g. GLUE [11]), we also implemented two popular models: RoBERTa [6] and BERT [5]. These models effectively capture contextual information into rich embeddings suitable for downstream classification tasks, while also minimizing the reliance on expensive hand-annotated labels. We domain adapted the general pre-trained RoBERTa [6] and BERT [5] and used the generated embeddings to feed into a classification layer to predict the class of notice period.

We explain the domain adaptation and data preprocessing approaches in the subsequent sections.

3.2. Domain adaptation

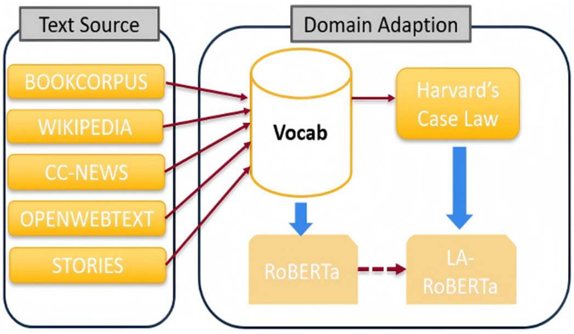
In recent years, deep pre-trained language models (PLM) such as RoBERTa [6] and BERT [5] have dominated a variety of NLP benchmarks. However, due to the heavy computational overhead and resource requirements, it is very difficult for individuals and small organizations to pre-train language models. Fortunately, large companies such as Google and Facebook provide the weights of PLMs to assist researchers who do not have access to high-performance computing hardware [10, 32]. These PLMs, however, perform poorly when applied to domain-specific areas such as medical [33] and legal domain [29].

Domain adaptation uses BERT-base vocabularies and texts to adjust the weights and enhance the performance of the general PLMs on domain-specific data. Many studies have demonstrated that this is the right path to take [29, 33, 34]. To better domain adopt PLMs, we utilized Facebook's pre-trained RoBERTa[4] and Google's pre-trained BERT model and continued training the pre-trained models using approximately 4 million cases from Harvard's case law project. After consultation with our legal partners, it was determined that the legal language of American case law and Canadian case law had significant overlap as they were both derived from British law. Cases before 1960 were evaluated by final-year law students who determined these cases to be linguistically different from present-day legal documents, and thus, these cases were removed. We note that the Harvard case law project involved collecting case law from the United States. Furthermore, we experimented with domain adapting RoBERTa on 10,000 complete insurance cases. A visualization of the domain adaptation process is shown in Figure 2.

3.3. Data and preprocessing

We use case summaries that were generated by human annotators and provided by Westlaw[5] (Figure 1). The Westlaw Quantum Service[6] provides concise summaries of cases, often including the judgment on the notice period granted as part of the severance package. An example of a case summary used in our research is given in Figure 1. The summaries provide a quick overview of the case and a description of the plaintiff, which usually consists of around 150 words compared to the 1000+ words in a full case description to avoid noise and information overload that can negatively influence the results. From the Westlaw website, we had to manually select and download the case summaries for this work. Next, we developed a method to extract selected information from this summary to prepare the data needed to train our models. To focus on predicting an appropriate notice period, we carefully excluded all references to the judgment about the notice period, retaining only the factual details regarding the plaintiff and the specifics of the case. Additionally, to train some of the models, we extracted key information from the structured part of the summaries, such as the year of the judgment, occupation type, age, salary, job title, and length of employment, and enriched the unstructured data with these details. The information was not prepended to the data if it was not found.

Figure 2
Flow diagram of domain adaption



To map the problem of reasonable notice prediction as a classification problem, we preprocessed the data for model development. We labeled each case with a class value of 1 to 25 based on the number of months awarded as the notice period, with the exception of class 25, which represents 25 months or more.

Figure 3 shows an example of the prepared data, which includes the case number, case reference (citation), judgment (months of notice period), and summary. To train most of our DL models, we only use the summary and outcomes as input, with the exception of our domain-specific model, which utilize additional hand-annotated labels. We apply additional data preprocessing using SpaCy [35] as it is a quick and lightweight tool commonly used for tokenization for all models except for BERT-based ones. The data preprocessing pipeline is shown in Figure 4.

In the next section, we explain the model architecture and implementation details.

4. Implementation

The different model implementation details are described below.

4.1. Model implementations

4.1.1. Baseline models

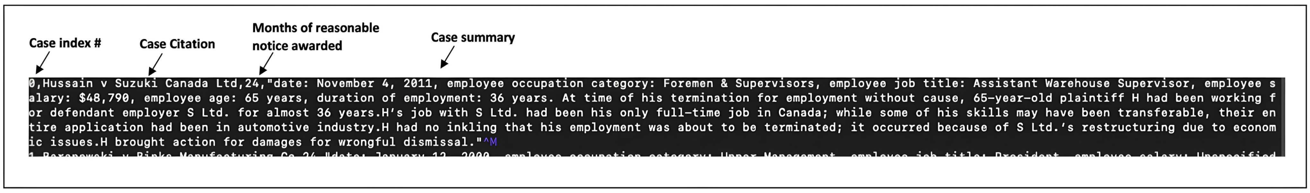
We trained and evaluated numerous statistical models for an accurate comparison. We, thereby, selected the two best-performing statistical ML models, namely, *MLR* and *GBR*. A team of law students extracted data features from the legal case descriptions and

[4] https://github.com/pytorch/fairseq

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

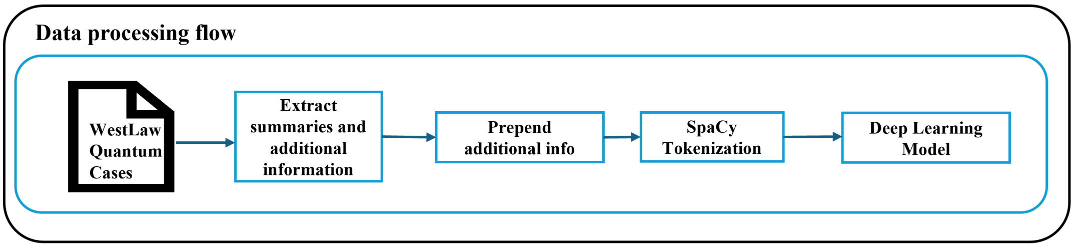
[6] https://www.westlawnextcanada.com/quantums/

Figure 3
Example of preprocessed data for calculating reasonable notice



```
0,Hussain v Suzuki Canada Ltd,24,"date: November 4, 2011, employee occupation category: Foremen & Supervisors, employee job title: Assistant Warehouse Supervisor, employee s
alary: $48,790, employee age: 65 years, duration of employment: 36 years. At time of his termination for employment without cause, 65-year-old plaintiff H had been working f
or defendant employer S Ltd. for almost 36 years.H's job with S Ltd. had been his only full-time job in Canada; while some of his skills may have been transferable, their en
tire application had been in automotive industry.H had no inkling that his employment was about to be terminated; it occurred because of S Ltd.'s restructuring due to econom
ic issues.H brought action for damages for wrongful dismissal."^M
1,Baranowski v Binks Manufacturing Co,24,"date: January 12, 2000, employee occupation category: Upper Management, employee job title: President, employee salary: Unspecified
```

Figure 4
Data flow for predicting reasonable notice



summaries based on the factors defined explicitly in the Bardal case law, such as the age of the employee, character of employment, availability of similar jobs, and length of employment. These data features were entered into a CSV (comma-separated values) file in categorical format or natural language. We extracted the values from the CSV file and used them as input data and labels to train the baseline statistical ML models.

After removing entries with missing variables, our dataset comprised 938 cases, of which 750 were used to train the models, and 188 were used as a holdout set for testing. Although there was some overlap in the cases used to train the baseline models with those used to train the DL models, the cases in this dataset were significantly different from those used in our DL experiments. Hyperparameters for both statistical models were selected through an exhaustive grid search using 10-fold cross-validation.

4.1.2. Approach I: multi-attention model with few-shot

The model we created was heavily inspired by Hu et al. [26], with a key difference. Instead of generating a single attention vector for each attribute (e.g., a person's age), we adopted the sentence embeddings presented by Lin et al. [15] and generated r number of attention vectors for each attribute. The model attended to multiple parts of the text at once for a single attribute and computed the sum of the resulting attention matrices to form the final attention weights.

We took the facts of a case (Westlaw summaries), tokenized and parsed them using SpaCy [35], and embedded the tokens using a pre-trained GloVe [36] text embedding model as shown in the Text Encoder component in Figure 5. The process created 840 billion tokens, each having a dimension of 300. The embeddings were fed into a bidirectional long short-term memory (BiLSTM) [37] model to learn the representation of the text sequence by the next term prediction task. The model produced n hidden states, each of dimension u as shown in Equation (1). The hidden states from both forward and backward directions were then concatenated together at each time step i as shown in Equation (2).

$$H = \ (h_1, h_2, h_3, \ldots h_n) \tag{1}$$

$$h_i = \left[\overleftarrow{h}_t ; \overrightarrow{h}_t \right] \tag{2}$$

where n is the number of words in the sequence and [;] is the concatenation operator, which makes each h_i of dimension $2u$.

The encoded text representation was then sent to two separate modules, Attention and Max Pooling, as shown in Figure 5, to produce a multi-attention and an attribute-free representation. The first attribute-aware module, inspired by Lin et al. [15], took the encoded facts from the BiLSTM encoder into each separate attention mechanism (represented by each row of "n attribute predictors") in the Attention module in Figure 5. Together, r rows of the attention mechanism represent the r dimension attention vector each computed using Equation (3). The attention vector helps predict K attributes. The attention mechanism was implemented using a two-layer multi-layer perceptron (MLP) without bias, where the first hidden dimension size was r, and it produced the attention alphas for the encoded states $r{*}u$. The output generated K different sets of attention alphas, each of size $r{*}n$, one for each attribute. The equations stated below are from Lin et al. [15]. The attention alphas are calculated using the two-layer MLPs, as shown in Equation (3).

$$A = softmax(W_j \tanh(WH^T)) \tag{3}$$

where W is a shared feature matrix across all attributes and W_j is the individual context vector of each attribute. Softmax is used to normalize the attention weights and not to create a probability distribution. Our sentence embedding is defined as:

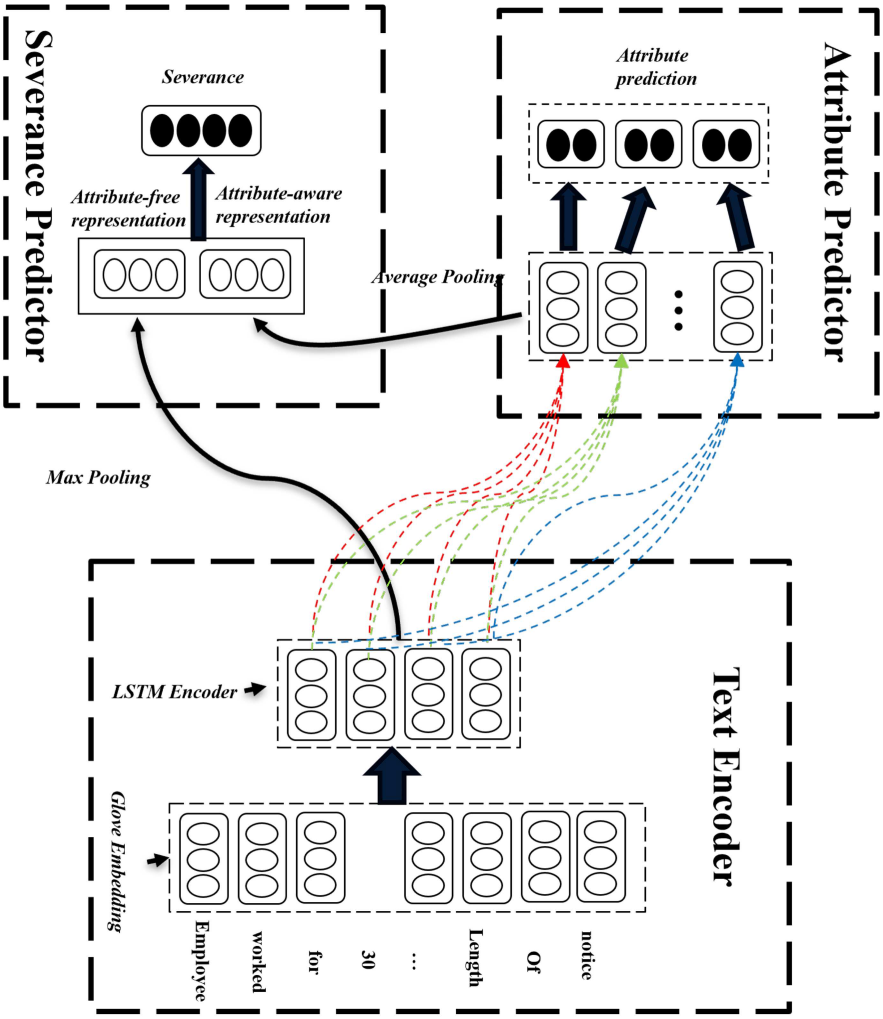
$$\mathrm{S} = AH \tag{4}$$

where $\mathrm{H} \in \mathrm{R}^{n*2u}$ for the resulting final representation $\mathrm{S} \in \mathrm{R}^{r*2u}$.

Consistent with the original approach of Lin et al. [15], we used the Max Pooling module to get our final document representation from H, as shown in Figure 5. The max pooled document representation is combined with the sentence embedding vector from the attention module to predict the severance. The combined representation is passed through a two-layer MLP with a hidden layer of 2000. This was done for each attribute.

$$\mathrm{p_l} = argmax\left(w_{1,l} \tanh(w_{2,l}\left(maxpool\left(s_l\right)\right) + b_l)\right) \tag{5}$$

Figure 5
Overview of few-shot model with few-shot



where $l \in [0, K]$. For the representation to be used in severance prediction:

$$out = [attr_{aware}; attr_{free}] \tag{6}$$

where $attr_{aware}$ is the average pool of the attention weights s_i for $i \in [0, K]$ and $attr_{free}$ is the max pooling of H. The final severance is determined using a linear transformation:

$$severance = argmax(softmax(W_{sev}\, out + b_{sev})) \tag{7}$$

We used a supervised learning approach, and the loss function is defined as:

$$L = L_{severance} + \alpha L_{attributes} \tag{8}$$

where alpha is a hyperparameter, and both losses are defined using cross-entropy.

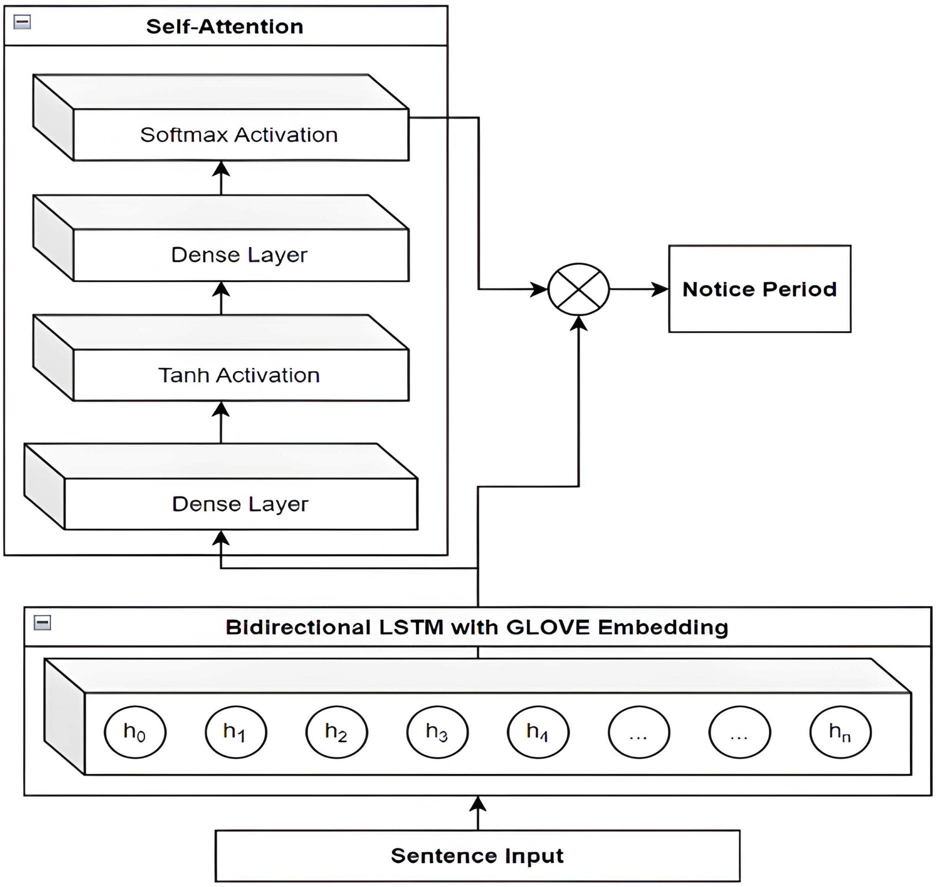
We used pre-trained 300-dimension GloVe vectors [36] that were fine-tuned on legal data with a hidden dimension of 300 and a dropout rate of 0.5. The model employed an attention mechanism with r set to 30. We used an Adam optimizer with a learning rate of 0.001 and reduced the learning rate by a factor of 0.95 when the metric stopped improving. Four labels were used to train the $attr_{aware}$ mechanisms (predicting 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. Alpha had a value of 0.3, and each epoch took approximately 8 min to complete.

4.1.3. Approach II: self-attention model

We believed that a simpler model with enhanced sentence representation could improve the classification of reasonable notice. We wanted to determine how accurately the parts of the text we autonomously learned truly aligned with the Bardal factors affirmed by the Supreme Court of Canada. Theoretically, by only using the severance labels, the self-attention model can be trained to focus on generalizable attributes that are shared across all training cases. The input into this model is the same as the previous attribute-aware model, but the only labels we used for training were the final severance predictions.

This implementation is similar to the multi-attention few-shot models, except we utilized only self-attention and expanded the number of self-attended layers to r. We replicated the model as shown in Figure 6 based on Lin et al. [15], which reported significant improvements in results when compared to techniques using a max pooling approach, the traditional method of creating sentence embeddings.

Figure 6
Overview of self-attention model



A sequence is initially encoded using a multi-attention neural network (usually an LSTM or gated-recurrent unit (GRU)), and the resulting hidden states are concatenated together at each time step. The resulting hidden sequence H is of the size $n*2u$, where n is the number of tokens in the sequence and u is the size of the hidden states as described before. H is then passed onto the self-attention mechanism implemented using a two-layer MLP (without bias) having a hidden layer of size r and producing the attention alphas or weights (A) as the encoded states using Equation (3) as in the bidirectional model. Equations from Lin et al. [15] explain the computation of the attention alphas. The final sentence embeddings are computed using Equation (4). Each sequence generated an attention matrix of $r*2u$ as before, where r is the number of hops of attention. The intuition behind the hops of attention is to force different hops to focus on different areas of the sequence, and to reinforce this, Lin et al. [15] introduced a novel penalty function as shown in Equation (9), where the model is penalized for having similar attention weights in each hop. In the case of no overlap, P equates to 0; otherwise, it returns a positive value. F represents the Frobenius norm.

$$P = ||AA^T - I||_F^2 \tag{9}$$

We utilized a pre-trained and fine-tuned 300-dimension GloVe embedding model [36] as before, along with a two-layer MLP with a hidden dimension of 300, a dropout of 0.5, and an attention r of 30. The Adam optimizer with a batch size of 64 and a learning rate of 0.001 was used, where the learning rate decayed by a factor of 0.95 when the metric for loss stopped improving. Each epoch took approximately 6 min to execute.

4.1.4. *Approach III: Hierarchical Attention Network (HAN)*

To better replicate the human judicial-thinking process, we implemented a HAN, introduced by Yang et al. [14]. In the HAN, we broke down each case summary into its sentences using SpaCy's [35] sentence-bound detection. At the sentence level, we learned the importance of individual tokens tokenized by SpaCy through a multi-attention attention layer. The word-level attention and word embedding were used to create the sentence representation. Since each sentence in our summary can be seen as a separate fact, we learned separate attention weights to represent the document as a weighted sum of factual representations.

Extracting deep semantic and contextual understanding from text data is essential for every NLP task. The words are embedded (e.g., using GloVe, Word2Vec, etc.) and encoded using a bidirectional GRU. A BiLSTM [37] can replace the GRU, as both architectures often produce similar results. The encoded word-level representations are then sent through an attention layer to produce sentence-level attention weights. A weighted average using the attention weights and hidden word representations is used to create sentence vectors. Similar to the word level, the resulting sentence representations of a document are then passed through a different bidirectional GRU [38] to generate hidden representations and a sentence-level attention layer. The final document representation is derived from the weighted sum of the word-level attention weights and the sentence encodings.

We employed pre-trained 200-dimensional GloVe vectors [36] combined with a BiLSTM [37] model, which had 75 hidden dimensions and a dropout rate of 0.5. Additionally, an attention mechanism with 30 attention units and a 50-dimensional attention layer was used. The model was optimized using stochastic gradient descent [39] with a learning rate of 0.06, batch size of 32, and momentum of 0.9. The learning rate was decreased by 0.95 when performance plateaued, and each epoch took approximately 6 min to run.

4.1.5. Approach IV: Legal Adapted BERT

The BERT [5] and BERT-inspired architectures have recently been dominating NLP benchmarks [11]. Although the architecture was not novel as it utilized the Transformer, the combination of transfer learning and domain adaptations made a significant contribution to NLP.

In applying BERT to predict reasonable notice awards, we domain adapted the language model on full reasonable notice case descriptions as well as the Harvard case law dataset instead of only the handwritten summary data by continual pre-training. The resulting Legal Adapted BERT (LA-BERT) was then used to train our downstream classification task of determining the reasonable notice period. To circumvent the 512 token limits of BERT, we used legal case summaries as the input instead of the full case description, where our case summaries had a maximum token length of 414. As BERT has a generalized understanding of the English language, we believe utilizing the contextualized embeddings from BERT could augment our aforementioned models to provide better results.

4.1.6. Approach V: Legal Adapted BERT + self-attend HAN

Rather than using pre-trained GloVe vectors which were derived statistically, we domain adapted our BERT model on full employment case descriptions and the judgments. We then utilized the domain-adapted BERT[7] model to generate embeddings of only the case summaries. Embeddings were further adjusted through back-propagation when we trained the classification model to predict reasonable notice. We used the LA-BERT embeddings in both the self-attention and the HAN models, as they were our best-performing models. Unfortunately, we noticed an approximate 23–15% drop in accuracy to 44% for the BERT+HAN and an accuracy of 51% for the BERT+Self-Attend. We believe we did not have enough data to be able to correctly leverage the large (768-dimension) embedding size. We note that our BERT embeddings were frozen and were not fine-tuned.

We also utilized the pre-trained BERT model from HuggingFace to generate word embeddings, setting the hidden dimension to 300, a dropout rate of 0.5, and an attention parameter r at 30. The training was performed using the Adam optimizer with a learning rate of 0.001 and a batch size of 32. The learning rate was reduced by 0.95 when performance plateaued. Each epoch took around 14 min to complete.

4.1.7. Approach VI: Legal Adapted RoBERTa

For our Legal Adapted RoBERTa (LA-RoBERTa$_{\text{base}}$) implementation, we continually pre-trained the masked language model using five epochs on the full reasonable notice judgment case descriptions and Harvard case law dataset. The classification model was trained on 1,695 case summaries. The last model was used to perform a classification on our 409-case holdout set. The effectiveness of each model is evaluated based on accuracy, as detailed in the next section.

4.2. Experimental setup

All experiments were conducted on an IBM Power8 server featuring 512 GB of RAM, 64 hyper-threaded cores (128 threads), and 4 K80 GPUs with Nvidia. The system ran on Red Hat Enterprise Linux Server 7.6 with a ppc64le architecture.

5. Validation and Results

We combined multiple SOTA models to implement our novel multi-attention DL model for predicting judgments. Attention-based models are well-suited for legal text analysis, as their mechanisms enable them to focus on the most relevant parts of complex documents such as key legal factors. By visualizing attention weights, we gain insight into what the “black-box” DL model is focusing on, providing a clearer understanding of its decision-making process. This makes attention-based models particularly effective and transparent for applications in legal domains where interpretability and context-sensitive analysis are vital. It is difficult to compare the results of our experimentation as there is no comparable work for reasonable notice prediction. Therefore, we implemented baseline models to compare model performances and develop the best-performing model. During the period of this research, black-box large language models such as Legal BERT had not yet been released. Consequently, we did not incorporate it into the comparative analysis.

Our dataset comprises 1,695 cases for training and an additional 409 cases for testing. We did not use a validation set. The prediction was classified as correct if it was within +/−2 of the ground truth labels to account for situational variability.

5.1. Results

Our RoBERTa$_{\text{base}}$ model had the highest accuracy of 69%. HAN had an accuracy of 67%, self-attention had 62% accuracy, few-shot achieved 57%, LA-BERT$_{\text{base}}$ achieved 64%, and LA-RoBERTabase achieved 65%. The results are summarized in Table 1. For explainability, we utilized the attention areas of the text in our model to create heat maps. We compare the attention weights from two separate models, HAN and self-attention, as shown in Figures 7 and 8, respectively. To achieve this, we overlaid attention weights from our models onto the original text to generate a heat

Table 1
Summary of results

Approach	Acc. (+/–02)
Few-shot w/ Multi Attention	51%
Self-attention	62%
HAN	67%
LA-BERT$_{\text{base}}$	64%
LA-BERT$_{\text{base}}$ + Self-Attend	51%
LA-BERT$_{\text{base}}$ + HAN	44%
LA-RoBERTa$_{\text{base}}$	65%
RoBERTa$_{\text{base}}$	69%
MultiRegression (baseline)	55%
GradientBoost Regressor (baseline)	59%

[7] https://github.com/huggingface/transformers

Figure 7
Heat map of a correct reasonable notice prediction using HAN

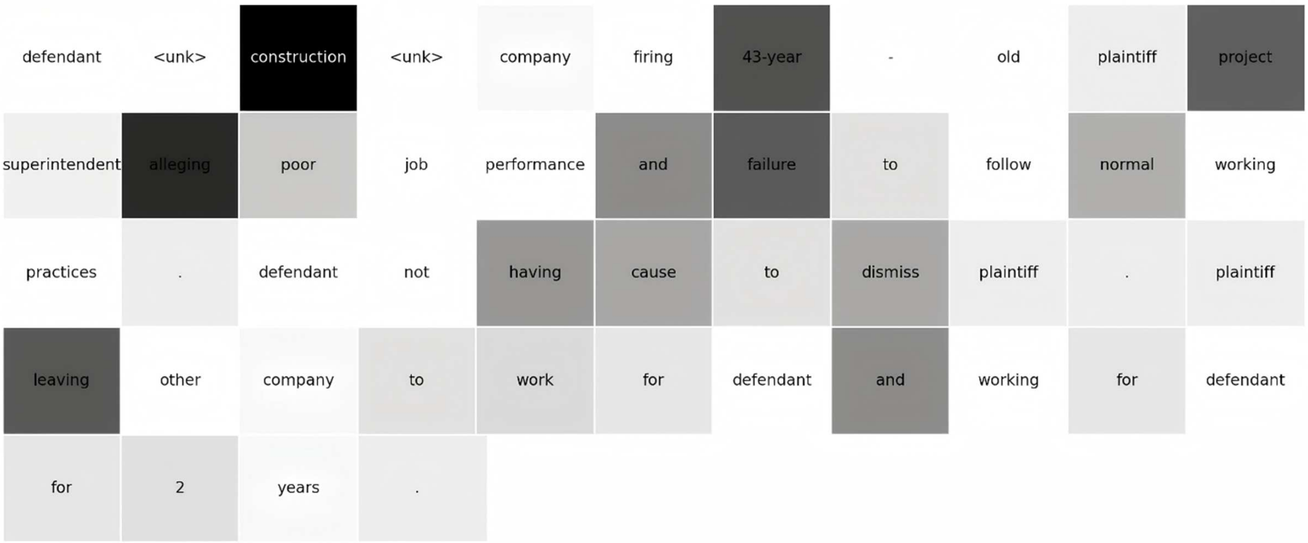
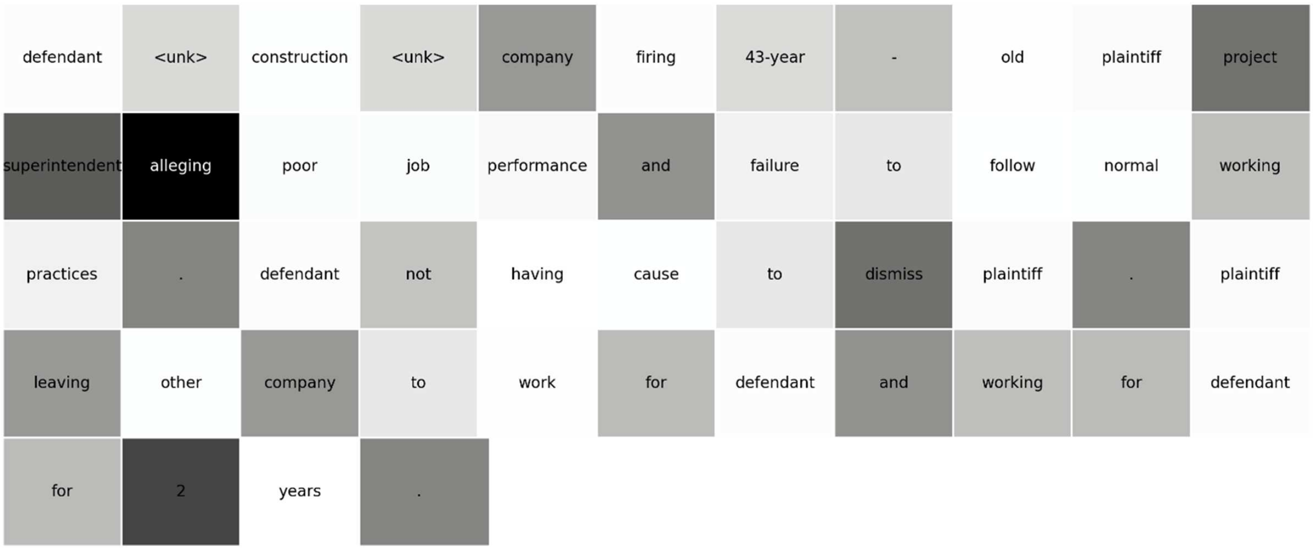


Figure 8
Heat map of a correct reasonable notice prediction using self-attention



map, where darker shaded areas represent higher attention. While the attention weights seen in both figures resulted in an exact prediction as the ground truth, we observe that different emphasis was placed on different areas of the text. While there was a heavy focus on the defendant's field of work in Figure 7, this emphasis was not a part of the prediction in Figure 8. Instead, the self-attention model focused more on the job category, "superintendent." We also note that the majority of high-attention areas of the text do align with the Bardal factors, such as the age of the employee, length of employment, and character of employment.

5.2. Critical discussion

Several insights were gained from the study as summarized below.

Domain adaptation: Interestingly, similar to the observation of Lam et al. [29] and Li et al. [28], the multi-attention RoBERTa model performed worse than the out-of-the-box version. The RoBERTa$_{base}$ model was trained on 160 GB of English language corpora that spanned a variety of sentence lengths and domains [6], giving RoBERTa a generalized understanding of language. We believe that our handwritten summaries from Westlaw's Quantum service are more aligned with everyday language than common legal parlance, giving RoBERTa an edge on the legal classification. An additional factor contributing to the underperformance of LA-RoBERTa compared to the out-of-the-box version could be attributed to the approach of continual pre-training. This approach resulted in maintaining an identical vocabulary for LA-RoBERTa as the original version. Consequently, despite engaging in domain-adapted pre-training, the vocabulary of LA-RoBERTa failed to accurately represent the legal domain [18]. A potential strategy to address this issue is to perform domain adaptation from scratch, similar to the methodology employed in PubMedBERT [18].

Use of summaries versus full case description: Typically, longer descriptive text is preferred over summaries to extract more useful information. Our use of summaries may have introduced some error due to lack of information [40, 41]. Although the full case text contained more information than a case summary, a survey by Cui et al. [42] notes that many LJP systems struggle with the non-standardized nature of legal documents. It complicates the extraction of relevant facts and often results in incomplete or biased inputs to the model. The case summaries also posed an issue of filtering as we needed only the factual statements of the case and had to exclude any reference to the analysis of the presiding judge or the outcome, which we were attempting to predict. Unfortunately, the court cases did not have a standardized structure, making it difficult to extract the factual statements from the case summaries. In addition, some case summaries were truncated as they contained more than 512 tokens, which may have resulted in further information loss.

Model performance: Overall, we observed that attention-based models with pre-training, particularly RoBERTa, proved to be significantly more effective in predicting legal outcomes compared to other methods, achieving the highest accuracy of 69%. Among all models tested, RoBERTa's attention mechanism allowed it to capture contextualized, task-relevant information across the text, contributing directly to its superior performance. In contrast, we found that other attention-based models, such as the "domain adaptation w/ Multi Attention," "LA-BERT$_{\text{base}}$ + Self-Attend," and "LA-BERT$_{\text{base}}$ + HAN," did not outperform traditional statistical methods. "Few-shot w/ Multi Attention" used GloVe embeddings, which provided static, fixed embeddings for each word. While effective in some applications, this approach limits the model's ability to capture the nuanced, contextualized meanings present in legal text, where a word's meaning often depends on the surrounding context. This lack of contextual flexibility may have contributed to its comparatively lower performance. "LA-BERTbase + Self-Attend" and "LA-BERT$_{\text{base}}$ + HAN" struggled to outperform baseline methods, which we attribute to the additional layers in these configurations. While BERT-style models effectively captured hierarchical language features from syntax to semantics through their layered structure. Adding more layers can introduce overfitting or even "forgetting" effects, where some learned representations from the pre-training phase could be lost. This limits the model's ability to generalize effectively on new data.

Other factors: As research on applying deep learning to the field of law is relatively new, we built numerous models that have shown success in other fields. Although the models we implemented performed well in other domains, they did not perform as expected, possibly due to the stark contrast between the language used in law and the other domains. Furthermore, using deep learning to classify legal cases removes the human element involved in judgments. It is possible that emotional components are being considered during the court case, which our models are unable to process.

We strongly believe that there may be too many idiosyncrasies in law and discrepancies between judgments for cases to be perfectly predicted. Throughout our research, our legal team found that judges often decide cases using factors that are beyond the precedent affirmed by the Supreme Court of Canada. For example, in Brien v Niagara Motors Ltd., the plaintiff was awarded a higher severance due to the employer acting in bad faith. Combined with the subjective nature of judges and consistent with the literature, we believe it is not possible to achieve very high accuracy scores [3]. Often, judgments are overturned on appeal, where appellate courts determine that trial courts have failed to give an appropriate consideration to the Bardal factors. For example, in Bohemier v Storwal International Inc., an appellate judge increased the amount of reasonable notice from 8 to 11 months after determining that the trial judge did not put enough weight on the length of service.

Our analysis acknowledges the inherent subjectivity and variability in legal case outcomes as key factors influencing model performance. It is also reflected in our model's performance, as achieving an accuracy beyond 69% proved challenging due to inconsistencies in judgment patterns. We suggest that the tool can be used for decision support as a supplementary for the legal experts to highlight the facts and not for full automation of the judgment system. This context is essential for framing model performance realistically and understanding its limitations in the legal domain.

6. Conclusion

In this paper, we explored the domain of legal case data for predicting judgments. We implemented an end-to-end pipeline to preprocess the legal data using NLP techniques to filter and extract important tokens. Then, we converted these tokens to word embeddings to feed into a predictive model to generate the notice period as the output. We explored leading deep learning models for creating text encoding and classification and implemented a variety of attention-based models. For explainability, we generated heat maps from the attention scores to demonstrate which words are attended to for correct predictions, which interestingly correspond to the established list of Bardal factors. The reported accuracy of 69% alongside other performance metrics provides a solid foundation for pursuing future studies to improve the accuracy further considering that there were no studies on Canadian employment law at that time. For practical applications such as decision support systems, more work would be needed to not only test the efficacy of the AI models but also to add explainability and context analysis through future work.

In the future, we plan to develop an end-to-end data extraction pipeline for the full text focusing on the Bardal factors and the information learned from the analysis of the summaries. The heat maps presented in this paper provide important information that led to the judgments, which can be used to extract further contextual information from the full text. Exploring other models and different learning strategies such as knowledge distillation can also help with exploring the legal text data for judgment prediction.

Acknowledgment

We acknowledge the collaboration of Queen's University Department of Law and the data provided by Westlaw Canada.

Funding Support

This study is funded by the New Frontiers Research Fund and the Natural Sciences and Engineering Research Council of Canada (NSERC) Discovery Grant (CRDPJ 543845-19).

Conflicts of Interest

The authors declare that they have no conflicts of interest to this work.

Data Availability Statement

The data that support the findings of this study are openly available at https://static.case.law/.

Author Contribution Statement

Jason Lam: Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation, Resources, Data curation, Writing – original draft, Writing – review & editing, Visualization. **Yuhao Chen:** Writing – review & editing, Visualization. **Farhana Zulkernine:** Conceptualization, Resources, Data curation, Writing – review & editing, Visualization, Supervision, Project administration, Funding acquisition. **Samuel Dahan:** Validation, Data curation, Writing – review & editing, Supervision, Project administration, Funding acquisition.

References

[1] Lamond, G. (2006). Precedent and analogy in legal reasoning. In E. N. Zalta (Ed.), *The Stanford encyclopedia of philosophy*. Stanford University.

[2] Bardal v. Globe & Mail Ltd., 24 D.L.R. (2d). 140. (1960). Retrieved from: https://aab.legal/files/AMBX545453KPG5T/documents/Bardal-v-Globe-and-Mail.pdf

[3] Dahan, S., Touboul, J., Lam, J., & Sfedj, D. (2020). Predicting employment notice period with machine learning: Promises and limitations. *McGill Law Journal*, *65*(4), 711–753. http://doi.org/10.7202/1076909ar

[4] Hawkes v. Max Aicher (North America) Limited, ONSC 4290. (2021). Retrieved from: https://www.btlegal.ca/assets/pdfs/2021.06.23-Case-Comment-Hawkes-v-Max-Aicher-(North-America)-Limited.pdf

[5] Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 1*, 4171–4186.

[6] Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., ..., & Stoyanov, V. (2019). RoBERTa: A robustly optimized BERT pretraining approach. *arXiv Preprint:1907.11692*. https://doi.org/10.48550/arXiv.1907.11692

[7] Malik, V., Sanjay, R., Nigam, S. K., Ghosh, K., Guha, S. K., Bhattacharya, A., & Modi, A. (2021). ILDC for CJPE: Indian legal documents corpus for court judgment prediction and explanation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, 1*, 4046–4062. https://doi.org/10.18653/v1/2021.acl-long.313

[8] Kauffman, M. E., & Soares, M. N. (2020). AI in legal services: New trends in AI-enabled legal services. *Service Oriented Computing and Applications*, *14*(4), 223–226. https://doi.org/10.1007/s11761-020-00305-x

[9] Bansal, N., Sharma, A., & Singh, R. K. (2019). A review on the application of deep learning in legal domain. In *Artificial Intelligence Applications and Innovations: 15th IFIP WG 12.5 International Conference*, 374–381. https://doi.org/10.1007/978-3-030-19823-7_31

[10] Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., ..., & Rush, A. (2020). Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, 38–45. http://doi.org/10.18653/v1/2020.emnlp-demos.6

[11] Wang, A., Singh, A., Michael, J., Hill, F., Levy, O., & Bowman, S. R. (2018). GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, 353–355. https://doi.org/10.18653/v1/W18-5446

[12] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ..., & Polosukhin, I. (2017). Attention is all you need. In *31st Conference on Neural Information Processing Systems*, 6000–6010.

[13] Bahdanau, D., Cho, K. H., & Bengio, Y. (2015). Neural machine translation by jointly learning to align and translate. In *3rd International Conference on Learning Representations*, 1–15.

[14] Yang, Z., Yang, D., Dyer, C., He, X., Smola, A., & Hovy, E. (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. http://doi.org/10.18653/v1/N16-1174

[15] Lin, Z., Feng, M., dos Santos, C., Yu, M., Xiang, B., Zhou, B., & Bengio, Y. (2017). A structured self-attentive sentence embedding. In *5th International Conference on Learning Representations*, 1–15.

[16] Chalkidis, I., Fergadiotis, M., Malakasiotis, P., Aletras, N., & Androutsopoulos, I. (2020). LEGAL-BERT: The Muppets straight out of law school. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, 2898–2904. https://doi.org/10.18653/v1/2020.findings-emnlp.261

[17] Zheng, L., Guha, N., Anderson, B. R., Henderson, P., & Ho, D. E. (2021). When does pretraining help? Assessing self-supervised learning for law and the CaseHOLD dataset of 53,000+ legal holdings. In *Proceedings of the Eighteenth International Conference on Artificial Intelligence and Law*, 159–168. https://doi.org/10.1145/3462757.3466088

[18] Gu, Y., Tinn, R., Cheng, H., Lucas, M., Usuyama, N., Liu, X., ..., & Poon, H. (2022). Domain-specific language model pretraining for biomedical natural language processing. *ACM Transactions on Computing for Healthcare*, *3*(1), 2. https://doi.org/10.1145/3458754

[19] Yasunaga, M., Leskovec, J., & Liang, P. (2022). LinkBERT: Pretraining language models with document links. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*, *1*, 8003–8016. https://doi.org/10.18653/v1/2022.acl-long.551

[20] Thakur, N., Cui, S., Khanna, K., Knieling, V., Duggal, Y. N., & Shao, M. (2023). Investigation of the gender-specific discourse about online learning during COVID-19 on Twitter using sentiment analysis, subjectivity analysis, and toxicity analysis. *Computers*, *12*(11), 221. https://doi.org/10.3390/computers12110221

[21] Qorib, M., Oladunni, T., Denis, M., Ososanya, E., & Cotae, P. (2023). Covid-19 vaccine hesitancy: Text mining, sentiment analysis and machine learning on COVID-19 vaccination Twitter dataset. *Expert Systems with Applications*, *212*, 118715. https://doi.org/10.1016/j.eswa.2022.118715

[22] Lam, J. T., Liang, D., Dahan, S., & Zulkernine, F. (2020). The gap between deep learning and law: Predicting employment notice. In *Proceedings of the 2020 Natural Legal Language Processing*, 52–56.

[23] Yin, Y., Zulkernine, F., & Dahan, S. (2020). Determining worker type from legal text data using machine learning. In *2020 IEEE International Conference on Dependable,*

Autonomic and Secure Computing, International Conference on Pervasive Intelligence and Computing, International Conference on Cloud and Big Data Computing, International Conference on Cyber Science and Technology Congress, 444–450. https://doi.org/10.1109/DASC-PICom-CBDCom-CyberSciTech49142.2020.00084

[24] Howe, J. S. T., Khang, L. H., & Chai, I. E. (2019). Legal area classification: A comparative study of text classifiers on Singapore Supreme Court judgments. In *Proceedings of the Natural Legal Language Processing Workshop 2019*, 67–77. http://doi.org/10.18653/v1/W19-2208

[25] Luo, B., Feng, Y., Xu, J., Zhang, X., & Zhao, D. (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. http://doi.org/10.18653/v1/D17-1289

[26] Hu, Z., Li, X., Tu, C., Liu, Z., & Sun, M. (2018). Few-shot charge prediction with discriminative legal attributes. In *Proceedings of the 27th International Conference on Computational Linguistics*, 487–498.

[27] Medvedeva, M., Vols, M., & Wieling, M. (2020). Using machine learning to predict decisions of the European Court of Human Rights. *Artificial Intelligence and Law*, *28*(2), 237–266. https://doi.org/10.1007/s10506-019-09255-y

[28] Li, J., Bhambhoria, R., & Zhu, X. (2022). Parameter-efficient legal domain adaptation. In *Proceedings of the Natural Legal Language Processing Workshop 2022*, 119–129. https://doi.org/10.18653/v1/2022.nllp-1.10

[29] Lam, J., Chen, Y., Zulkernine, F., & Dahan, S. (2021). Detection of similar legal cases on personal injury. In *2021 International Conference on Data Mining Workshops*, 639–646. https://doi.org/10.1109/ICDMW53433.2021.00084

[30] Zhang, H., Dou, Z., Zhu, Y., & Wen, J.-R. (2023). Contrastive learning for legal judgment prediction. *ACM Transactions on Information Systems*, *41*(4), 113. https://doi.org/10.1145/3580489

[31] Billon Polak, P., Prusa, J. D., & Khoshgoftaar, T. M. (2024). Low-shot learning and class imbalance: A survey. *Journal of Big Data*, *11*(1), 1. https://doi.org/10.1186/s40537-023-00851-z

[32] Larsen, P., & Proserpio, D. (2023). Information signals in sponsored search: Evidence from Google's BERT. *SSRN*. http://doi.org/10.2139/ssrn.4614402

[33] Lee, J., Yoon, W., Kim, S., Kim, D., Kim, S., So, C. H., & Kang, J. (2020). BioBERT: A pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, *36*(4), 1234–1240. https://doi.org/10.1093/bioinformatics/btz682

[34] Chen, Y., Zhang, Y., Wang, J., & Zhang, X. (2023). YNU-HPCC at SemEval-2023 Task 6: LEGAL-BERT based hierarchical BiLSTM with CRF for rhetorical roles prediction. In *Proceedings of the 17th International Workshop on Semantic Evaluation*, 2075–2081. http://doi.org/10.18653/v1/2023.semeval-1.286

[35] Jugran, S., Kumar, A., Tyagi, B. S., & Anand, V. (2021). Extractive automatic text summarization using SpaCy in Python & NLP. In *2021 International Conference on Advance Computing and Innovative Technologies in Engineering*, 582–585. https://doi.org/10.1109/ICACITE51222.2021.9404712

[36] Pennington, J., Socher, R., & Manning, C. D. (2014). GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*, 1532–1543. http://doi.org/10.3115/v1/D14-1162

[37] Al-Selwi, S. M., Hassan, M. F., Abdulkadir, S. J., Muneer, A., Sumiea, E. H., Alqushaibi, A., & Ragab, M. G. (2024). RNN-LSTM: From applications to modeling techniques and beyond—Systematic review. *Journal of King Saud University-Computer and Information Sciences*, *36*(5), 102068. https://doi.org/10.1016/j.jksuci.2024.102068

[38] Su, Y., & Kuo, C.-C. J. (2022). Recurrent neural networks and their memory behavior: A survey. *APSIPA Transactions on Signal and Information Processing*, *11*(1), e26. http://doi.org/10.1561/116.00000123

[39] Santra, S., Hsieh, J.-W., & Lin, C.-F. (2021). Gradient descent effects on differential neural architecture search: A survey. *IEEE Access*, *9*, 89602–89618. https://doi.org/10.1109/ACCESS.2021.3090918

[40] Zhao, Q., Gao, T., Zhou, S., Li, D., & Wen, Y. (2022). Legal judgment prediction via heterogeneous graphs and knowledge of law articles. *Applied Sciences*, *12*(5), 2531. https://doi.org/10.3390/app12052531

[41] Chalkidis, I., Androutsopoulos, I., & Aletras, N. (2019). Neural legal judgment prediction in English. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 4317–4323. https://doi.org/10.18653/v1/P19-1424

[42] Cui, J., Shen, X., & Wen, S. (2023). A survey on legal judgment prediction: Datasets, metrics, models and challenges. *IEEE Access*, *11*, 102050–102071. http://doi.org/10.1109/ACCESS.2023.3317083

How to Cite: Lam, J., Chen, Y., Zulkernine, F., & Dahan, S. (2025). Legal Text Analytics for Reasonable Notice Period Prediction. *Journal of Computational and Cognitive Engineering*, *4*(3), 319–331. https://doi.org/10.47852/bonviewJCCE52024104