Exploiting Domain-Specific Knowledge for Judgment Prediction is no Panacea

Olivier Salaün[1], **Philippe Langlais**[1], and **Karim Benyekhlef**[2]

[1]RALI, DIRO, University of Montréal
{salaunol,felipe}@iro.umontreal.ca
[2]Cyberjustice Laboratory, Faculty of Law, University of Montréal
karim.benyekhlef@umontreal.ca

Abstract

Legal judgment prediction (LJP) usually consists in a text classification task aimed at predicting the verdict on the basis of the fact description. The literature shows that the use of articles as input features helps improve the classification performance. In this work, we designed a verdict prediction task based on landlord-tenant disputes and we applied BERT-based models to which we fed different article-based features. Although the results obtained are consistent with the literature, the improvements with the articles are mostly obtained with the most frequent labels, suggesting that pre-trained and fine-tuned transformer-based models are not scalable as is for legal reasoning in real life scenarios as they would only excel in accurately predicting the most recurrent verdicts to the detriment of other legal outcomes.

1 Introduction

At the intersection of machine learning and law, legal judgment prediction (LJP) is a task that consists in predicting either the outcome of a lawsuit (Skalak, 1989; Nallapati and Manning, 2008; Katz et al., 2017; Aletras et al., 2016; Liu and Chen, 2017) or some other case attributes such as legal areas (Şulea et al., 2017; Soh et al., 2019) or charges (Xiao et al., 2018).

One specificity of court rulings is that they are based on the application of legal articles to the factual description of the case. That is to say, a judge must determine whether some law articles are relevant to a case, and if applicable, whether the legal principles they embody are violated. Therefore, articles as domain-specific knowledge can be used as leverage for improving LJP performance, as shown by Luo et al. (2017) and Long et al. (2019) for charge prediction and divorce judgment prediction respectively. Xu et al. (2020) also went further by using articles for distinguishing confusing charges in a charge prediction task.

Meanwhile, transformers (Vaswani et al., 2017) and BERT models (Devlin et al., 2019) in particular have been widely used in NLP tasks with the assumption that such models, if first pre-trained on massive corpora and then fine-tuned on the dataset of a given task, could suffice for achieving significant improvements. On one hand, this turned out to be true with the CAIL2018 dataset (charge prediction task) as shown by Wang et al. (2020). On the other hand, Holzenberger et al. (2020) mentioned in a statutory reasoning entailment task that a transformer model does worse than a rule-based model, even after further pre-training on the domain corpus. Furthermore, in an employment notice prediction task, Lam et al. (2020) emphasized that domain adaption of such models could even harm performance. These elements raise the question of how well a pre-trained transformer model can handle a legal NLP task and how well the input from domain-specific knowledge such as legislative text can improve the LJP performance. To the best of our knowledge, in the case of LJP tasks aimed at verdict prediction, no experiment has tested so far the application of pre-trained BERT models on both tribunal decision text and cited law articles text combined altogether, which we intend to do in this work.

We designed a multilabel classification task in which the model must predict the ruling outcomes on the basis of the facts description. One can imagine that such a predictive engine could be used for legal assistance for those who may not afford the services of a legal expert. Unlike Luo et al. (2017), we put the article prediction aside in order to focus solely on the verdict prediction and assess in which conditions input article-based features can improve classification. For our experiments, we use a landlord-tenant disputes corpus used by West-

Proceedings of Recent Advances in Natural Language Processing, pages 1234–1243
Sep 1–3, 2021.
https://doi.org/10.26615/978-954-452-072-4_139

ermann et al. (2019) and Salaün et al. (2020) from which we extracted fine-grained targets labels and article features in order to encompass as much as possible the variety of rulings, thus making the task more representative of real life cases. We present the preprocessing of the dataset along with the creation of article-based features in Section 2. The architectures of the models are shown in Section 3 along with three methods for integrating the information from the articles mentioned in the decisions. Discussion of the results and concluding remarks are provided in Sections 4 and 5 respectively.

2 Preparation of the Dataset

The Administrative Housing Tribunal is a court of Quebec in Canada with an exclusive jurisdiction in landlord-tenant disputes. We got access to an exhaustive corpus of 667,305 decisions in French issued from 2001 to 2018 publicly accessible through SOQUIJ portal[1]. Documents average and median lengths amount to 307 and 235 tokens respectively with a standard deviation of 371.

The landlord asks for termination of the lease and eviction of the tenant, collection of the rent ($1500) as well as the rent due at the time of the hearing, plus provisional execution of the decision despite the appeal.
This is a lease from September 1, 2010 to June 30, 2011 at a monthly rent of $750, payable on the first day of each month.
The evidence shows that the tenant owes $750, i.e. the rent for March 2011, plus $6 representing the costs of service provided for in the Regulation.
The tenant is more than three weeks late in paying the rent, and the cancellation of the lease is therefore justified by the application of Article 1971 C.C.Q.
However, the lease is not resiliated if the rent due, interest and costs are paid before judgment, in accordance with the provisions of article 1883 C.C.Q.
The prejudice caused to the landlord does not justify provisional enforcement of the decision, as provided for in article 82.1 L.R.L.

FOR THESE REASONS, THE COURT :
TERMINATES the lease and ORDERS the eviction of the Tenant and all the occupants of the dwelling;
ORDERS the Tenant to pay the landlord the sum of $750, plus interest at the legal rate and the additional indemnity provided for in article 1619 C.C.Q., as of March 4, 2011, plus legal fees of $72.

Figure 1: Excerpt of a decision translated from French. The text in italics is the verdict while underlined text contains references to articles.

Each decision is split in two by applying heuristics based on the syntax of the documents: the pre-verdict section, used as text input (text before the italics in Figure 1), and the verdict section containing the legal solution chosen by the judge in charge of the case (text in italics in Figure 1). The pre-processing of both sections are described in subsections 2.1 and 2.2 respectively.

2.1 Preprocessing of Input Features

As one of our goal is to assess the conditions in which articles help to improve predictions, we applied heuristics on the pre-verdict text to extract a total of 1,790 unique cited law articles, 33.8% of which were mentioned only once across the entire corpus. Also, not all articles are related to housing law. We address this by keeping only 445 articles from the *Book Five - Obligations* of Civil Code of Quebec (C.C.Q.) which establishes the rules concerning the contractual obligations between landlords and tenants and whose frequency in the corpus has a minimum of two. Three examples are shown on Table 1 for the decision in Figure 1. Article distribution is heavily skewed: the 3 most frequent articles cover 72%, 42% and 27% of all documents respectively while all other articles do not exceed 4%. Mean and median frequency of the articles amount to 2571 and 17 respectively. Section 3 further describes the use of these articles as input.

Articles	Target labels
1619	monetary penalty for defendant
1971	eviction
1883	termination of the lease

Table 1: Civil Code of Quebec articles and verdict labels extracted from the decision shown in Figure 1.

The pre-verdict section contains both fact description and legal analysis. As the latter can give hints about the verdict that the model is expected to predict, we removed from the pre-verdict section any paragraphs containing citations of articles (underlined text in Figure 1) and we capped the maximum input text length at 128 tokens. By doing so, we force the model to make predictions on the sole basis of fact descriptions.

2.2 Making Target Labels from the Verdict Section of the Decisions

We carefully combined NLP-engineered tools (regular expressions and the like) and some housing law expertise in order to pseudo-automatically identify 23 labels that we believe are representative of the rulings and that cover the diversity of the verdicts at a fine grain. These labels are cumulative and three are shown as an example in Table 1 for the decision in Figure 1.

The average and median numbers of labels per decision both amount to 3 with a standard deviation of 1.5. Nearly half of all rulings involves an eviction (48.1%) and a termination of the lease (46.1%), hinting that a significant part of the cases

[1] Société québécoise d'information juridique https://soquij.qc.ca/

involve an unfavourable outcome for tenants sued by landlords. Further investigation confirms a bias favourable for landlords as 80.3% of cases with the top frequent label *monetary_penalty_for_defendant* have a tenant as the (penalized) defendant.

Overall, 0.05% of all instances were not assigned any labels and 18.2% did not contain any articles. For the design of our experiments, all instances with no article or no verdict label were excluded, thus resulting in a final corpus of 544,857 documents with an average of 3.3 labels per instance (standard deviation of 1.4 and median of 4). The average and median numbers of articles per document both amount to 2. Our instances are randomly divided into training, validation and test sets with a 60-20-20 ratio.

3 Models

We aim at designing a multilabel classification task in which a model has to return the labels corresponding to the verdict on the basis of the pre-verdict section of each decision. Our baseline is a One-Versus-Rest Logistic Regression, i.e. each label has its own classifier. The input text is vectorized through character-based TF-IDF spanning bi-grams to 8-grams (character-based features outperformed token-based). Only the top 100k n-grams are retained. We also use CamemBERT (Martin et al., 2019), a variant of BERT (Devlin et al., 2019) that was pre-trained on generic French corpora. We further pre-trained the *camembert-base* default parameters (unsupervised masked language modelling task) from Wolf et al. (2020) on the train set during 20 epochs[2]. We eventually fine-tuned them during the multilabel classification task. The batch size and the maximum number of fine-tuning epochs are set at 50 and 20 respectively. Training is stopped when no further improvement is obtained in terms of exact match on validation set. We use the Adam optimizer (Kingma and Ba, 2015) with a learning rate at 10^{-5} (all other hyperparameters are left at default value). The optimization criterion is the binary cross entropy with logits loss for numerical stability during optimization. Therefore the final output is that of a logit function, with scores ranging from *-inf* to *+inf*. A label is returned whenever its associated output value exceeds 0. We use a vanilla CamemBERT model whose only input is the pre-verdict text and three other variants described in the next subsections.

3.1 One-Hot and Node2Vec Encoding of Articles

For each instance, the mention/absence of each article is one-hot encoded through a 445-dimensional vector, each dimension corresponding to one article. We have one model named BERT-OH (Figure 2 part a) in which the BERT output of a decision text (768-dimensional vector corresponding to the first token from hidden states) and the articles one-hot vector are concatenated and passed through fully connected layers before outputting the verdict labels. Given the heavily skewed distribution of these articles among the documents, these discrete one-hot vectors are sparse and likely not very expressive (in the case of Figure 1 for instance, all dimensions except three would be zeroed because of only three articles extracted as shown in Table 1).

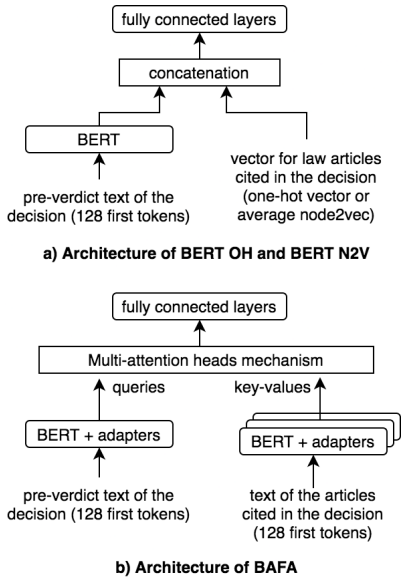

Figure 2: Architecture diagrams of BERT-OH, BERT-N2V both on part a) and BAFA on part b).

As a result, we thought of a continuous and more expressive representation that could embed the articles organization within the law. In the one-hot vector approach, each article is assigned one dimension independent from the other as if all articles were completely unrelated to each other. But when having a closer look at the C.C.Q. *Book Five - Obligations*[3] on Figure 3 that concentrates the rules related to landlord-tenant disputes, the articles are

[2]Exact match achieved by further pre-trained models is around one percent point greater than models with default pre-trained parameters.

[3]Screenshot from CANLII http://canlii.ca/t/533nd last accessed on January 20th 2021

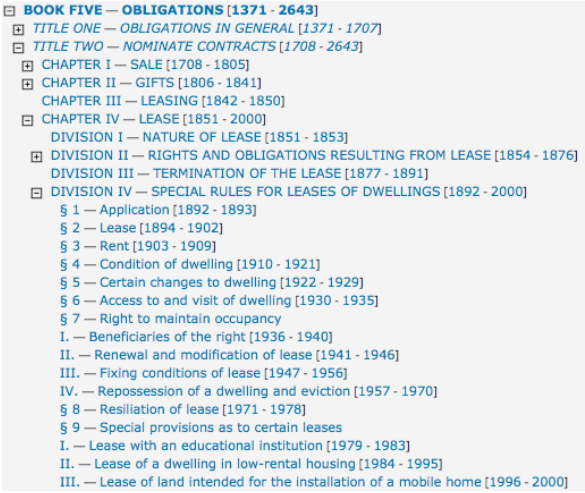

Figure 3: Excerpt of the Book Five of the Civil Code of Quebec with articles related to dwelling rental lease.

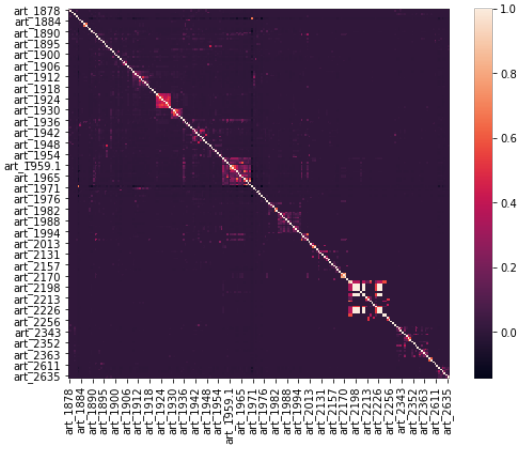

Figure 4: Correlation matrices for 200 articles (sorted by increasing article numbers).

actually organized into titles, chapters, divisions, paragraphs and so on, down to the articles themselves. As each subcategory becomes more and more precise, the articles encompassed in it are related to closer and closer legal concepts. For instance, articles 1957 to 1970 are especially dedicated to repossession of a dwelling and eviction and can be expected to relate to the same legal objects. Therefore, we wanted to make an embedding that could capture the structural closeness between articles, that is, two articles located in the same subsection would have closer representations. Another argument in favour of using embeddings based on the topological relatedness among articles is the fact that articles with close numbers have a tendency to co-occur together in the decisions, as shown on the diagonal of Figure 4 along which distinct articles with close numbers tend to belong to the same subsections and to have higher correlation values.

One method for representing the topological organization of the law is Node2Vec (Grover and Leskovec, 2016). We first built a tree whose root is *Book Five - Obligations* and added all of the subsequent sections as nodes. Articles were added as leaves. Next, the edges were placed by linking each node/section to the subsequent nodes/sections that it encapsulates. For instance, in Figure 3, *CHAPTER IV - LEASE* is linked to *DIVISION IV - SPECIAL RULES FOR LEASES OF DWELLINGS* which is linked to *§ 7 - Right to maintain occupancy*, and so on, until placing the edges between *I. - Beneficiaries of the right* and each of the leaves/articles 1936 to 1940. An edge between two nodes cannot "skip" any node of intermediate level between the two if there is any. Overall, we made a graph with 1,565 nodes (subsections and articles) and 1,564 edges. From this graph, we made an embedding for each article such that articles belonging to the same subsection and related to the same legal concepts, would have embeddings close to each other. Following the Node2Vec technique, we generated 200 random walks from each node (article or section) that gave a total of 313k sequences of 200 nodes by following the edges. Then, we trained a Word2Vec model (Mikolov et al., 2013) on these sequences during 10 epochs and with a window size of 10, so that each node is assigned a Node2Vec embedding (256-sized vector) that captures the proximity of the neighbouring nodes. We eventually only kept the Node2Vec representations of the leaves that correspond to the 445 retained articles. For the sake of clarity, the vectors of two articles randomly drawn would have a higher cosine similarity if the articles belong to the same subsection than if they belong to distinct ones. For the BERT-N2V (node2vec) model, the input contains the pre-verdict text (passed through CamemBERT) plus an average of the Node2Vec embeddings of all articles mentioned in the documents (e.g. average embedding of articles 1619, 1971 and 1883 in Table 1 for the case in Figure 1).

3.2 Applying BERT on the Law Articles Text

The one-hot and the Node2Vec encodings are still shallow representations of the articles as they do not even use the text of Civil Code of Quebec. At most, Node2Vec is only capturing a topological representation of how the 445 articles are organized

in the law. This is why we considered another model called BAFA (BERT model with Adapters applied on Facts and Articles, Figure 2 part b) that is given as input the pre-verdict text of the decision and the text of all articles cited in it. The 445 retained articles have a average length of 34 tokens (median of 32 and standard deviation of 17). The pre-verdict section and the articles are encoded through two distinct CamemBERT models so that one is fine-tuned on the decision text and the other on the law text. The BERT output (i.e. first token of hidden states) of the decision is then passed through a 12-head attention mechanism as a query while the BERT output obtained from each cited articles are concatenated (up to 22 cited articles per decision) and passed as a key-value pair. The output of the attention mechanism is then passed through two fully connected layers. As we use two distinct CamemBERT modules, the batch size is reduced to 4 and we added adapters (Pfeiffer et al., 2020) as an attempt to speed up computation.

To the best of our knowledge, there is no other work that uses the fine-tuning of pre-trained BERT models on the text of cited articles for verdict prediction of court decisions.When it comes to LJP tasks formalized as text classification, many of most recent works usually aimed at charge prediction or law articles prediction on the sole basis of the facts description (Xiao et al., 2018; Zhong et al., 2018). Şulea et al. (2017) made a ruling prediction task comparable to ours but without the text of the articles. When it comes to experiments that actually use the text of law articles, Hu et al. (2018) and Xu et al. (2020) used it for improving prediction of confusing charges only while Luo et al. (2017) and Long et al. (2019) used a combination of recurrent neural nets with attention mechanisms for encoding it into their models for charge prediction and divorce verdict respectively. Still, none of these works involve transformer architecture. Concerning the experiments that use BERT, Chalkidis et al. (2019) and Wang et al. (2020) used pre-trained models for prediction of violation of human rights article and of charges respectively on the basis of the facts only.

4 Results and Discussion

The classification results are shown in Table 2. For each label, we compute the F1 score (harmonic mean between precision and recall) obtained by each model and add the label distribution across the test documents. For each model, the last two rows of Table 2 present two overall scores based on metrics that we believe are constraining enough and appropriate for the evaluation of systems that could one day be deployed in real life scenarios. F1 macro average is unweighted average of all labels F1 scores, and thus penalizes models that delivers poor F1 scores for a large number of labels. It measures the ability of the model to predict a large variety of rulings. We also compute exact match which corresponds to the ratio of instances for which a model is able to return the exact set of labels assigned to them. Therefore, an instance is considered as misclassified whenever its prediction has a label in excess or one missing.

4.1 Gains Obtained across Verdict Labels with Article-Based Features

One goal of our experiment is to assess how articles can improve the prediction of cases in which they are cited. Figure 5 shows a heatmap detailing the correlation among articles and verdict labels. In the top left corner, *monetary_penalty_for_defendant*, *eviction*, *termination_lease* and *provisional_enforcement* are strongly correlated with articles 1619, 1971 and 1883 which respectively define: the computation of an additional indemnity that can be added to damages ; a rule that allows the termination of the lease if rent is over three weeks late in payment ; a rule so that the tenant may avoid lease termination by paying the due rent plus interest before the judgment. Although these articles make a consistent legal ground with the aforementioned verdict labels, inputting them into the models through any representation (be it one-hot/node2vec/BERT encoding) added very little improvement for the F1 scores of these labels (by at most 2.8 percent point on average relative to CamemBERT alone), very likely because of their already high frequency in the corpus. Furthermore, on the heatmap on Figure 5, *tenant_ordered_pay_rent* is strongly correlated with article 1973 that defines the conditions allowing the judge to grant lease termination (unless the payment of the rent is over three weeks late, the judge may choose to either terminate the lease immediately or either order the tenant to pay the rent) and the article-based features help in dramatically improving the corresponding F1 score by 18-22 percent points compared to a sole CamemBERT setting. We also observe that *landlord_repossesses_rental_unit* and

Verdict label	Baseline	CamemBERT	CamemBERT +one-hot	CamemBERT + node2vec	BAFA	Support
monetary_penalty_for_defendant	98.2	98.3	98.4	98.5	**98.6**	92.7
eviction	94.8	94.0	96.7	96.5	**97.1**	57.7
termination_lease	95.2	94.2	96.8	96.6	**97.2**	55.6
applicant_request_denied	71.6	77.6	78.0	**78.5**	77.9	37.0
provisional_enforcement	87.4	88.8	89.3	89.3	**89.7**	25.5
applicant_is_reserved_recourses	76.3	**80.9**	80.7	80.8	79.8	12.4
lease_already_terminated	85.8	86.8	**88.2**	88.0	88.1	4.3
tenant_ordered_pay_rent	65.8	69.0	90.7	88.0	**91.3**	1.7
one_party_ordered_some_action	51.3	60.9	66.4	65.4	**67.4**	1.6
landlord_reposesses_rental_unit	85.1	86.2	89.7	88.5	**92.6**	1.4
monetary_penalty_for_applicant	69.3	72.5	83.6	78.3	**86.0**	1.1
agreement_between_parties	74.3	75.7	76.7	**76.8**	76.0	1.0
tribunal_sets_new_rent	59.6	69.6	70.3	**71.1**	66.9	0.6
defendant_ordered_some_action	39.0	**57.3**	51.1	56.0	51.9	0.4
penalty_misc	10.5	33.0	**36.6**	28.7	22.6	0.2
tribunal_cancels_past_ruling	61.5	**76.2**	75.8	72.3	71.9	0.2
discountinuance_claim	22.0	**50.6**	46.6	49.3	33.7	0.1
tribunal_declines_jurisdiction	14.3	1.5	31.2	39.8	**50.5**	0.1
schedule_new_audience	49.7	**52.9**	50.6	50.0	49.4	0.1
tribunal_upholds_past_ruling	19.2	38.9	42.5	**42.9**	16.3	0.1
applicant_forbidden_seek_recourse	0.0	0.0	0.0	**17.4**	0.0	0.1
applicant_ordered_some_action	0.0	0.0	0.0	0.0	0.0	<0.1
trib_asserts_jurisdiction	0.0	0.0	0.0	0.0	0.0	<0.1
F1 across all labels (macro-average)	53.5	59.4	62.6	**63.2**	61.1	
Exact match	58.6	63.9	**67.0**	66.4	66.7	

Table 2: F1 scores for each label and model (percentage, the highest score of each label is in bold). The last two rows show macro-averaged F1 and exact match across all labels for each of the four settings. The last column on the right shows the distribution of the labels in the test set.

monetary_penalty_for_applicant are strongly associated with several articles, especially 1963 and 1967 which establish respectively the conditions by which the judge can authorize a landlord-applicant to repossess their dwelling from which a tenant refuses to depart and the indemnities that the landlord must pay to the tenant for moving expenses when repossession is granted. Article-based models also improved the F1 scores of these two labels, though not as important as for *tenant_ordered_to_pay_rent*, with average gains of 4.1 and 10.1 points respectively. All in all, the inclusion of article-based features has a negligible impact when the labels already have a high support in the documents, but the improvement is more significant for labels that are rarer (*landlord_repossesses_rental_unit*, *monetary_penalty_for_applicant* and *tenant_ordered_to_pay_rent* have supports below 5%) and that have a high correlation value with the articles cited in them (for the three aforementioned labels and their articles, the average correlation is around 0.75). A counter-example to that principle would be *agreement_between_parties* and *tribunal_sets_new_rent* whose correlation values with cited articles are not that important (below 0.5) and for which no significant improvement on F1 scores is observed with article-based features.

4.2 Comparing Performance among Models

In general, BERT-based models do better than the baseline in terms of macro-averaged F1 score and exact match. Furthermore, the scores show that article-based features help in outperforming a sole CamemBERT model with higher exact match scores by up to 3.1 percent points and a higher F1 score macro-average by up to 3.8 points. The best F1 macro average score is achieved by the model with node2vec (63.2%) while best exact match score is obtained by CamemBERT with one-hot vectors (67.0%). Still, such results must be nuanced: the performance gains are mostly obtained with either high frequency labels (five most frequent plus *lease_already_terminated*) or labels that are strongly correlated with certain articles, which can explain the marginal improvements achieved in the coarse scores at the bottom of Table 2. Furthermore, the use of article-based features

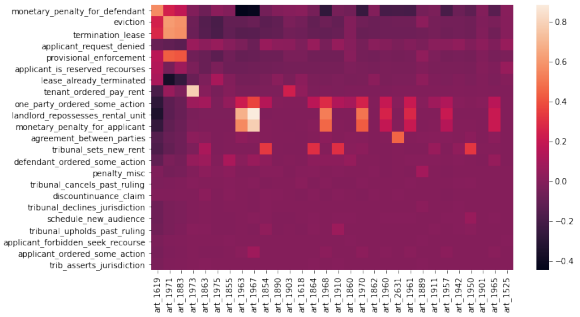

Figure 5: Heatmap of the correlation matrix of the verdict labels and the 30 most frequent articles. The verdict labels and the articles are sorted by decreasing frequency on their respective axes.

seem to sometimes harm the performance for low-frequency labels relative to vanilla BERT (*defendant_ordered_some_action*, *penalty_misc*, *tribunal_cancels_past_ruling*, *discontinuance_claim*, *schedule_new_audience*, *tribunal_upholds_past_ruling*), which suggests that such features add noise rather than help the model in accurately predicting these verdicts.

Although BAFA used the text of both decisions and articles, both encoded through two distinct BERT models, the overall performance is disappointing compared to the variants with one-hot and node2vec features: despite achieving the best F1 scores for some top frequent labels, the overall coarse scores remain below those of BERT-OH and BERT-N2V. Plus, the training of BAFA also is significantly longer[4].

For illustrative purposes, if we were to compare the performance achieved by our models in this task with other similar works, we could cite:

- Aletras et al. (2016) and Chalkidis et al. (2019) who achieved respectively an accuracy of 79% and a F1 macro-average score of 80.2% in binary classifications for violation of human rights ;
- Şulea et al. (2017) who achieved an accuracy of 92.8% for an 8-mutually-exclusive classes classification task for ruling prediction (our task has 23 cumulative labels) ;
- Luo et al. (2017) got a macro-average F1 score of 95.4% in a charge prediction task that use articles text as input ; in another charge prediction task, Zhong et al. (2018) achieved 49.1% and 70.9% for that score on two other datasets (our task is about verdict prediction).

Two main points can be made from these results. First of all, shallow articles embeddings (one-hot and node2vec) do better than BERT-encoding of law text at allowing a marginal improvement over some low-support labels (though not all), with BERT-N2V reaching the highest F1 macro-average score at 63.2%. A tentative explanation is that directly inputting the text of the cited articles adds noisy information that confuses rather than helps the model in the task while a "fuzzier" representation of the articles gives a broader information about articles (Node2Vec embeds the topological position of articles in Civil Code of Quebec) without forcing the model to combine the legal terminology of the law and the text input of the decisions. The second point is that although article-based models outperform vanilla CamemBERT, this is mainly due to the marginal improvement over some of the top frequent labels and to the improvement over some verdict that are strongly correlated with certain articles. This suggests that such models only excel in predicting the most recurrent and stereotypical landlord-tenant disputes (eviction of a non-paying tenant; moving indemnities for a tenant whose rental unit is repossessed by the landlord).

All in all, these models would be unusable for providing legal assistance for a large variety of cases as they would only excel in predicting accurately the most frequent rulings to the detriment of other types of cases. Also, as these models deal with housing law domain that is related to sensitive social issues (Gallié et al. (2016) emphasized that tenants are less confident in dealing with judicial proceedings), we tried to extract some significant patterns from the self-attention weights in the CamemBERT architecture that could help in understanding what causes the model to return some prediction, but found nothing prone to interpretation, which is consistent with a statement from Jain and Wallace (2019).

4.3 Discussion about the Experiment Setting

The results obtained seem consistent with observations from Holzenberger et al. (2020) who stated that even a further pre-trained BERT model struggles with a legal entailment task, thus suggesting that the fine-tuning of pre-trained BERT models on statutes and law articles text is not sufficient

[4]Over 1 hour per epoch for a one-BERT model, over 14 hours per epoch for two BERTs combined.

for solving tasks in very specific domains such as tax law or housing law. Regardless of the method used for inputting articles into the models, all of the approaches combining description of the facts and articles just excel at predicting the most frequent verdicts, which suggests that they would be unusable as is at a higher scale as they would not be able to provide satisfactory legal assistance for cases different from the most recurrent ones. Bender et al. (2021) emphasized the risks involved in using large pre-trained models that tend to encode and amplify biases already present in the training data. To paraphrase the title of their paper, their remarks are consistent with our observations as we end up with "legal parrots" which would not be able to accurately address the variety of real world landlord-tenant disputes.

The fact that article-based features allow for significant improvement under the condition that the labels are strongly correlated with articles also raises questions about the setting of LJP experiments: in the charge prediction task made by Xiao et al. (2018), charges and laws with frequency below 30 were removed from the CAIL2018 dataset and each charge label is strongly associated with one specific article. In contrast, in our corpus, each label is not always strongly correlated with some law article, as shown in Figure 5. Some works using the CAIL2018 dataset such as that from Wang et al. (2020) made further changes in the dataset by removing target labels with frequency below 100. Unlike them, in our work, we were much more permissive during the creation of our corpus as we retained articles with a frequency of at least 2 and made labels to exhaustively cover as many verdicts as possible (i.e. 1102 unique combinations of labels), even though some labels could have been merged together (e.g. *landlord_repossesses_rental_unit* and *monetary_penalty_for_applicant* tend to co-occur together) or discarded/weighted down. For instance, *schedule_new_audience* and *applicant_forbidden_seek_recourse* have a low frequency and are rather technical legal details that would be more relevant for a legal expert than for a layman seeking general advice. If we computed F1 average score weighted by each label's support, BERT models would have an average performance of 91.6%, but that coarse metric is mostly pulled upwards by scores achieved for most frequent labels. We must also emphasize that in our dataset there is no 1-to-1 correspondence between labels and articles as in CAIL2018 in which articles not relevant to specific charges were removed beforehand. This illustrates the difficulty in automating legal reasoning over cases and unfiltered law articles in a realistic context.

5 Concluding Remarks

We designed a LJP task as text multilabel classification for verdict prediction based on a collection of landlord-tenant disputes in French for which we used a further pre-trained CamemBERT model and applied different types of features derived from the articles cited in the decisions (one-hot, Node2Vec, BERT encoding of the text of articles). By doing so, we noticed that leveraging articles as input features (regardless of the representation used) made either marginal improvements for F1 scores of most frequent labels, either significant improvements for labels that are strongly correlated with certain articles. The use of article-based one-hot features achieves best exact match score (67.0%) while node2vec features achieve best F1 macro average score (63.2%). The model that encodes the text of the articles with BERT does not outperform the two previous methods.

As future work, we plan on comparing how models perform under both "realistic" setting (several rare target labels with no or few connections with the law available, as we did in this work) and "laboratory" setting (where low frequency targets and laws are aggressively filtered out). We also plan to assess whether the patterns observed in our work (performance improves when articles are strongly correlated with labels) also exist in other LJP datasets beyond housing law and Canadian cases. Furthermore, we plan to study further the attention weights and the mechanisms underlying the significant prediction improvement observed for certain labels when the input contains the text of articles that are highly correlated with the corresponding verdicts.

Acknowledgements

We would like to thank the Cyberjustice Laboratory at the Université de Montréal, the LexUM Chair on Legal Information and the Autonomy through Cyberjustice Technologies project for supporting this research.

References

N Aletras, D Tsarapatsanis, D Preoţiuc-Pietro, and V Lampos. 2016. Predicting judicial decisions of the European Court of Human Rights: a natural language processing perspective. peerj comput sci 2: e93.

Emily M Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the dangers of stochastic parrots: Can language models be too big. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency; Association for Computing Machinery: New York, NY, USA*.

Ilias Chalkidis, Ion Androutsopoulos, and Nikolaos Aletras. 2019. Neural legal judgment prediction in english. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4317–4323.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 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, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Martin Gallié, Julie Brunet, and Richard-Alexandre Laniel. 2016. Les expulsions pour arriérés de loyer au québec: un contentieux de masse. *McGill Law Journal/Revue de droit de McGill*, 61(3):611–666.

Aditya Grover and Jure Leskovec. 2016. node2vec: Scalable feature learning for networks. In *Proceedings of the 22nd ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 855–864.

Nils Holzenberger, Andrew Blair-Stanek, and Benjamin Van Durme. 2020. A dataset for statutory reasoning in tax law entailment and question answering. In *Proceedings of the 2020 Natural Legal Language Processing (NLLP) Workshop, 24 August 2020, San Diego, US*.

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*, pages 487–498.

Sarthak Jain and Byron C Wallace. 2019. Attention is not explanation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 3543–3556.

Daniel Martin Katz, Michael J Bommarito II, and Josh Blackman. 2017. A general approach for predicting the behavior of the Supreme Court of the United States. *PloS one*, 12(4):e0174698.

Diederik P Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *ICLR (Poster)*.

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*.

Zhenyu Liu and Huanhuan Chen. 2017. A predictive performance comparison of machine learning models for judicial cases. *2017 IEEE Symposium Series on Computational Intelligence (SSCI)*, pages 1–6.

Shangbang Long, Cunchao Tu, Zhiyuan Liu, and Maosong Sun. 2019. Automatic judgment prediction via legal reading comprehension. In *China National Conference on Chinese Computational Linguistics*, pages 558–572. Springer.

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*, pages 2727–2736.

Louis Martin, Benjamin Muller, Pedro Javier Ortiz Suárez, Yoann Dupont, Laurent Romary, Éric Villemonte de la Clergerie, Djamé Seddah, and Benoît Sagot. 2019. CamemBERT: a tasty French language model. *arXiv preprint arXiv:1911.03894*.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Ramesh Nallapati and Christopher D Manning. 2008. Legal docket-entry classification: Where machine learning stumbles. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 438–446. Association for Computational Linguistics.

Jonas Pfeiffer, Andreas Rücklé, Clifton Poth, Aishwarya Kamath, Ivan Vulić, Sebastian Ruder, Kyunghyun Cho, and Iryna Gurevych. 2020. Adapterhub: A framework for adapting transformers. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 46–54.

Olivier Salaün, Philippe Langlais, Andrés Lou, Hannes Westermann, and Karim Benyekhlef. 2020. Analysis and Multilabel Classification of Quebec Court Decisions in the Domain of Housing Law. In *International Conference on Applications of Natural Language to Information Systems*, pages 135–143. Springer.

David B Skalak. 1989. Taking advantage of models for legal classification. In *Proceedings of the 2nd international conference on Artificial intelligence and law*, pages 234–241. ACM.

Jerrold Soh, How Khang Lim, and Ian Ernst Chai. 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*, pages 67–77, Minneapolis, Minnesota. Association for Computational Linguistics.

Octavia-Maria Şulea, Marcos Zampieri, Mihaela Vela, and Josef van Genabith. 2017. Predicting the Law Area and Decisions of French Supreme Court Cases. In *Proceedings of the International Conference Recent Advances in Natural Language Processing, RANLP 2017*, pages 716–722, Varna, Bulgaria. INCOMA Ltd.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.

Yongjun Wang, Jing Gao, and Junjie Chen. 2020. Deep learning algorithm for judicial judgment prediction based on bert. In *2020 5th International Conference on Computing, Communication and Security (ICCCS)*, pages 1–6. IEEE.

Hannes Westermann, Vern R Walker, Kevin D Ashley, and Karim Benyekhlef. 2019. Using factors to predict and analyze landlord-tenant decisions to increase access to justice. In *Proceedings of the Seventeenth International Conference on Artificial Intelligence and Law*, pages 133–142.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Chaojun Xiao, Haoxi Zhong, Zhipeng Guo, Cunchao Tu, Zhiyuan Liu, Maosong Sun, Yansong Feng, Xianpei Han, Zhen Hu, Heng Wang, et al. 2018. Cail2018: A large-scale legal dataset for judgment prediction. *arXiv preprint arXiv:1807.02478*.

Nuo Xu, Pinghui Wang, Long Chen, Li Pan, Xiaoyan Wang, and Junzhou Zhao. 2020. Distinguish confusing law articles for legal judgment prediction. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3086–3095.

Haoxi Zhong, Zhipeng Guo, Cunchao Tu, Chaojun Xiao, Zhiyuan Liu, and Maosong Sun. 2018. Legal judgment prediction via topological learning. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3540–3549.