Analysis and Multilabel Classification of Quebec Court Decisions in the Domain of Housing Law

Olivier Salaün[1(✉)], Philippe Langlais[1], Andrés Lou[2], Hannes Westermann[3], and Karim Benyekhlef[3]

[1] RALI, Université de Montréal, Montreal, QC, Canada
{salaunol,felipe}@iro.umontreal.ca
[2] CLaC, Concordia University, Montreal, QC, Canada
and_lou@encs.concordia.ca
[3] Cyberjustice Laboratory, Université de Montréal, Montreal, QC, Canada
{hannes.westermann,karim.benyekhlef}@umontreal.ca

Abstract. The Régie du Logement du Québec (RDL) is a tribunal with exclusive jurisdiction in matters regarding rental leases. Within the framework of the ACT (Autonomy Through Cyberjustice Technologies) project, we processed an original collection of court decisions in French and performed a thorough analysis to reveal biases that may influence prediction experiments. We studied a multilabel classification task that consists in predicting the types of verdict in order to illustrate the importance of prior data analysis. Our best model, based on the FlauBERT language model, achieves F1 score micro averages of 93.7% and 84.9% in Landlord v. Tenant and Tenant v. Landlord cases respectively. However, with the support of our in-depth analysis, we emphasize that these results should be kept in perspective and that some metrics may not be suitable for evaluating systems in sensitive domains such as housing law.

Keywords: Natural Language Processing · Court decisions · Legal text · Text mining · Multilabel classification · French text · Housing law

1 Context

Many works related to artificial intelligence and law focus on the creation of tools intended for legal professionals to address, say, legal information retrieval with Natural Language Processing (NLP) [10] or knowledge management [3]. In the context of the ACT project (Autonomy Through Cyberjustice Technologies, https://www.ajcact.org/en), methods are explored in order to facilitate and automate access to justice for laymen unfamiliar with legal procedures. For the purpose of evaluating how far machine learning can fulfill these goals, our work focuses on lawsuits submitted to the Régie du Logement du Québec (RDL), a

E. Métais et al. (Eds.): NLDB 2020, LNCS 12089, pp. 135–143, 2020.
https://doi.org/10.1007/978-3-030-51310-8_13

tribunal specialized in tenant-landlord disputes. To the best of our knowledge, no work investigated this dataset apart from [15] which only studied a tiny fraction of it. One long-term goal of the ACT project is to make a system that allows tenants to gauge their chances of winning a case against their landlord and what outcomes they could expect from it by combining personal situations and relevant laws.

In [2], the authors describe a classification model that can simulate such legal reasoning. We can distinguish a first group of classification works as in [1,11,14] that rely on relatively small datasets (usually at most ten thousand samples) but annotated by legal experts. A second group of more recent works such as [5,12,13] apply text mining and NLP engineering on available metadata, thus relaxing the constraint of scarce human annotation and allowing dramatically larger datasets (at least a hundred thousand instances). Some preprocessing work for extracting labels or categories is shown in [12,14], which emphasizes the importance of performing that step with care in order to design sensible and understandable prediction tasks.

In our work, we deepened that latter point by first conducting a thorough analysis on RDL lawsuits and then presenting one multilabel classification task. Then, we discuss the results obtained and reflect upon how to properly evaluate legal prediction experiments.

2 Dataset Analysis

Understanding the data, especially in a specific domain such as housing law, is paramount to conduct meaningful experiments. The RDL collection consists of 981,112 decisions in French issued from 2001 to early 2018 by 72 judges in 29 tribunals around Quebec. Some of these documents are provided as public data by the SOQUIJ legal documents search engine (http://citoyens.soquij.qc.ca/); however, we obtained access to the entire corpus. Each decision mainly consists of a body of text with three parts that always appear in the following order, as illustrated in Fig. 1:

- fact descriptions and evidence presented by each party (here, a proof of tenant's failure to comply with payment schedule; lines 1 to 3 in Fig. 1);
- a legal reasoning section in which the judge analyses the case in the light of the applicable laws (lines 4 to 6);
- a verdict section with the judge final decisions (e.g. defendant ordered to pay damages to the plaintiff, rejection of the claim; lines 7 to 11).

The decisions also contain metadata (top and bottom of Fig. 1). After cleaning up and removing all documents with missing information and duplicates, we obtained a total of 667,305 texts with an average length of 363 tokens.

BUREAU DE [COURT CITY LOCATION]

No dossier : [FILE NUMBER] No demande : [CLAIM NUMBER]
Date : [SIGNATURE DATE OF THE JUDGMENT]
Régisseure : [JUDGE'S FULL NAME], juge administrative

[PLAINTIFF'S FULL NAME]
Locatrice - Partie demanderesse
c.
[DEFENDANT'S FULL NAME]
Locataire - Partie défenderesse

DÉCISION

[1] La locatrice demande la résiliation du bail et l'expulsion du locataire, le recouvrement du loyer (1 080 $) ainsi que le loyer dû au moment de l'audience, plus l'exécution provisoire de la décision malgré l'appel.
[2] Il s'agit d'un bail du 1er décembre 2016 au 30 juin 2018 au loyer mensuel de 545 $, payable le premier jour de chaque mois.
[3] La preuve démontre que le locataire doit 1 620 $, soit le loyer des mois de février, mars et avril 2017, plus 9 $ représentant les frais de notification prévus au Règlement.
[4] Le locataire est en retard de plus de trois semaines pour le paiement du loyer, la résiliation du bail est donc justifiée par l'application de l'article 1971 C.c.Q.
[5] Le bail n'est toutefois pas résilié si le loyer dû, les intérêts et les frais sont payés avant jugement, conformément aux dispositions de l'article 1883 C.c.Q.
[6] Le préjudice causé à la locatrice justifie l'exécution provisoire de la décision, comme il est prévu à l'article 82.1 de la *Loi sur la Régie du logement*[1].

POUR CES MOTIFS, LE TRIBUNAL :

[7] **RÉSILIE** le bail et **ORDONNE** l'expulsion du locataire et de tous les occupants du logement;
[8] **ORDONNE** l'exécution provisoire, malgré l'appel, de l'ordonnance d'expulsion à compter du 11e jour de sa date;
[9] **CONDAMNE** le locataire à payer à la locatrice la somme de 1 620 $, plus les intérêts au taux légal et l'indemnité additionnelle prévue à l'article 1619 C.c.Q., à compter du 22 mars 2017 sur la somme de 1 080 $, et sur le solde à compter de l'échéance de chaque loyer, plus les frais judiciaires de 83 $;
[10] **RÉSERVE** à la locatrice tous ses recours;
[11] **REJETTE** la demande quant aux autres conclusions.

[JUDGE'S FULL NAME]

Présence(s) : la locatrice
Date de l'audience : [AUDIENCE DATE]

Fig. 1. RDL sample decision from SOQUIJ (available at http://t.soquij.ca/p9TYc)

2.1 Analysis of the Plaintiffs and Defendants

We extracted from the metadata of each decision over a dozen of characteristics using NLP-engineered methods. For instance, we managed to identify the type of each party: legal persons (juridical entities like organizations) and natural (human and physical) persons. The latter encompasses four sub-categories: succession (a liquidator acts on behalf of a deceased person), multiple persons, single female and single male. Overall, 89% of all cases involve landlords suing tenants (Landlord v. Tenant scenarios or LvT) while 11% involve tenants suing landlords (Tenant v. Landlord scenarios or TvL). In the first scenario, plaintiffs are mostly legal persons while defendants are an absolute majority of single males as shown in Table 1. In the TvL setting, plaintiffs and defendants are predominantly single males.

Table 1. Distribution of plaintiff and defendant types (in percent) by case types

Case type		Landlord v. Tenant		Tenant v. Landlord	
Party		Plaintiff	Defendant	Plaintiff	Defendant
Legal person		41.3	0.2	0.3	33.0
Natur. pers.	Single male	36.8	60.1	54.0	40.0
	Single female	11.0	39.5	45.5	14.0
	Multiple (any genders)	10.7	0.0	0.0	12.6
	Succession	0.2	0.2	0.2	0.3
Total		100	100	100	100
Number of decisions		595,808		71,497	

2.2 Analysis of the Verdicts

Extracting consistent outcomes from the judgments (i.e. lines 7 to 11 in Fig. 1) is crucial for the feasibility and interpretability of prediction tasks. Difficulties in making a simple representation encompassing a wide variety of rulings were shown in [14]. One possible solution consists in identifying a "winner" between the plaintiff and the defendant, but this binary approach is not always suitable (e.g. the plaintiff's claims are partly accepted and rejected by the judge). An opposite approach consists in making labels that cover all possible outcomes, implying a high annotation cost partly illustrated in [15], plus the risk of numerous overly specific labels applicable to very few instances as in [5]. We chose an intermediate solution by narrowing all outcomes to three binary labels:

- **penalty:** the defendant receives penalties (e.g. an order for the landlord to pay damages, an eviction from the accommodation for a tenant);
- **agreement:** the judge enforces an agreement between both parties;
- **rejection:** the judge fully or partially rejects the plaintiff's claims.

These three outcomes are not mutually exclusive and can be applied to any case regardless of whether the plaintiff is the landlord or the tenant. We used an approach similar to [8] for determining the labels of each case by relying on key verbs in capital letters that happen to be good proxies of the verdict. In the example of Fig. 1, penalty and rejection labels apply due to the verbs *CONDAMNE* and *REJETTE* on lines 9 and 11. Major trends are shown in Table 2 for each case type: a landlord-plaintiff succeeds in winning over the tenant in 89% of lawsuits while tenant-plaintiffs' demands are totally or partially rejected by the judge in 69% of lawsuits. Such biases must be considered carefully. It might suffice to know whether the plaintiff is a landlord to get a good approximation of the outcome of a lawsuit. So far, all figures found in our analysis reveal that some care is required when developing machine learning applications as such biases and imbalance in the dataset might be the cause of deceptively good results in classification tasks.

Table 2. Distribution of labels (in percent) by case types

Scenario type	Landlord v. Tenant	Tenant v. Landlord
Cases with agreement label	2.0%	5.4%
Cases with penalty label	89.0%	23.6%
Cases with rejection label	38.3%	68.8%
Total number of cases	595,808	71,497

3 Prediction Task and Results

3.1 Models and Features

As seen in the previous section, claims made by landlords are much more successful than those made by tenants. Because of these biases, we decided to make two subtasks for LvT and TvL scenarios. For each of these subtasks, we made a 60:20:20 train-validation-test split for the corresponding datasets. Our baseline is a dummy classifier that returns a label if it occurs in more than half of the training samples. Thus, it will always and only predict the penalty and rejection labels in LvT and TvL respectively. Among the models used, we present the results for logistic regression implemented through a One-versus-Rest approach (OvR, one classifier per label). Three sets of features are used:

- the metadata alone (court location and judge in charge of the audience, presences and types of plaintiff and defendant);
- the metadata plus TF-IDF vectors (2–8-g at character level) fitted on the first line of the decision (line 1 on Fig. 1). These vectors are later replaced by a mean vector of FastText embeddings [4] of all words contained in the first line. These FastText representations are trained beforehand on the first line (window and vector sizes of 5 and 300, 10 training epochs);
- the metadata plus TF-IDF or FastText vectors (same settings as above) fitted this time on all the text before the verdict (lines 1 to 6 on Fig. 1).

The rationale behind using different lengths of the decision is to check whether models can predict the outcome of a case by solely using the factual elements of a case without relying on the legal analysis, as in [16]. As stressed in [13], elements from the legal analysis section may reveal the verdict. In the absence of efficient means to properly isolate the fact descriptions from the legal analysis, we used the first line of each decision as a proxy for factual elements.

For the latter two settings with different input text lengths, we also applied the FlauBERT base cased language model [7], a variant of BERT (Bidirectional Encoder Representations from Transformers [6]) pretrained on French corpora that we finetuned to our input text (metadata were not used). We set the batch size, maximum sequence length and learning rate to 32, 256 and 1e−5 respectively. Training was set to 10 epochs and stopped whenever a lower loss on the evaluation set was not achieved after 10,000 consecutive optimization steps. Our

metrics are accuracy (the predicted labels must exactly match the true ones for a sample to be considered as correctly classified) and micro, macro and weighted F1 score averages. Results are shown in Table 3.

3.2 Discussion of the Results

With the mere use of metadata as features, we managed to beat the baseline model in the LvT scenario for almost all metrics, but got little to no improvement in TvL cases. This may be due to the fact that TvL labels distribution is less imbalanced compared to LvT (see Table 2). When given access to the first line of each document, all models outperform the baseline and the TF-IDF method performs slightly better compared to FastText and FlauBERT across all metrics and scenarios. One possible explanation is that the TF-IDF representation partially preserves characters order while FastText and FlauBERT expressiveness suffer from the shortness of the first line. When the input contains all the text before the verdict, the performances obtained with TF-IDF vectors either stagnate or slightly degrade. The FastText method, on the contrary, significantly improves across all metrics, beating TF-IDF. FlauBERT outperforms FastText with dramatic improvements across all metrics. The fact that FastText and FlauBERT achieve better performance across all metrics with respect to TF-IDF can also be explained by longer text inputs that lead to richer embeddings. On the other hand, the stagnation of TF-IDF may be due to a dramatically larger number of n-grams as the text inputs become longer, leading to longer and sparser TF-IDF vectors that could not be leveraged by our models. FastText and FlauBERT performances also need to be kept in perspective as the paragraphs at the end of the longer text input may reveal information about the verdict as mentioned earlier from [13]. All in all, in both scenarios and with all text before verdict as input, our best model is the FlauBERT one that achieves 93.7% and 85.2% on micro average F1 score and accuracy in the LvT scenario, and 84.9% and 74.6% for the TvL cases.

We must emphasize that regardless of the models and input used, because of the labels imbalance shown in Table 2, one can easily maximize the individual F1 score of the most frequent label in each subtask (individual F1 score exceeds 94% for penalty label in LvT and 81% for rejection label in TvL cases for any model). As a consequence, one can achieve a relatively high micro-average F1 score that is based on recall and precision of all labels altogether: almost all models score above 80% and 70% in LvT and TvL scenarios, even in the first-line setting. The same phenomenon applies to the weighted F1 score average that is also influenced by the most frequent label (provided it can be easily predicted). As the task of a judge consists in applying general legal rules to individual cases with their own particularities, evaluating a classifier for legal outcomes with micro or weighted F1 score averages may convey deceptively good results as these metrics can be influenced by ubiquitous patterns in the data. On the other hand, accuracy and macro F1 score seem to be less sensitive to data imbalance and may be preferred for getting a more rigorous evaluation of predictive systems in sensitive domains such as housing law, though accuracy may also be considered as a metric biased

Table 3. Multilabel classification results for Landlord v. Tenant and Tenant v. Landlord scenarios in percent (for the last two features sets, the highest value of each column is bold)

	Landlord v. Tenant				Tenant v. Landlord			
	F1 micr.	F1 macr.	F1 weig.	Accu.	F1 micr.	F1 macr.	F1 weig.	Accu.
Dummy	77.5	31.4	64.7	56.9	69.8	27.2	57.6	58.3
Metadata only								
OvR Log. reg.	84.3	50.7	82.4	65.4	69.8	35.5	63.4	57.7
Metadata (except FlauBERT) + the first line								
TFIDF representation (2-8 grams at character level)								
OvR Log. reg.	**87.0**	**66.7**	**85.9**	**70.3**	**78.1**	**65.4**	**77.4**	**65.4**
Mean vector of FastText embeddings (vector size 300 and window size 5)								
OvR Log. reg.	85.1	54.6	83.6	66.6	72.7	48.3	69.9	59.3
FlauBERT (batch size 32, max seq length 256, learning rate 1e-5)								
Transformers	83.4	60.9	80.2	64.0	74.0	58.7	73.0	60.1
Metadata (except FlauBERT) + all text before verdict								
TFIDF representation (2-8 grams at character level)								
OvR Log. reg.	86.9	66.8	85.9	70.3	78.1	65.1	77.3	65.2
Mean vector of FastText embeddings (vector size 300 and window size 5)								
OvR Log. reg.	88.7	81.0	88.2	73.7	80.1	77.1	79.7	66.6
FlauBERT (batch size 32, max seq length 256, learning rate 1e-5)								
Transformers	**93.7**	**90.8**	**93.7**	**85.2**	**84.9**	**84.7**	**85.1**	**74.6**

against the majority class when applied to an imbalanced dataset. We would not have been aware of all these fine details without a prior thorough examination of the dataset itself.

4 Conclusion

In this work, we built and analyzed thoroughly an original collection of court decisions in French about landlord-tenant disputes. We were able to extract over a dozen of characteristics for each decision and to detect biases contained in the dataset such as landlords being much more successful plaintiffs with respect to tenants. Such analysis was only feasible thanks to carefully engineered NLP tools combined with background knowledge of the housing law domain. This preliminary step allowed us to suggest one multilabel classification task for predicting legal rulings. Two distinct subtasks were designed for Landlord v. Tenant (LvT) and Tenant v. Landlord (TvL) lawsuits. We could observe that TF-IDF based methods perform relatively well when given the first line of each decision while FastText and FlauBERT approaches excel when all text before verdict is given as input. The latter achieved micro F1 score average and accuracy of 93.7% and 85.2% in LvT cases and 84.9% and 74.6% for TvL cases respectively. Thanks to our prior in-depth study of the strong trends present in the data, we emphasized

the risk of using micro and weighted F1 score averages which can be artificially maximized in the presence of overly frequent labels. This remark is particularly important in the evaluation of legal classification models as judges must apply general legal rules to individual cases with their own particularities.

As future work, we consider pursuing our study with a regression task (predicting the amount of indemnities awarded that the judge orders the losing defendant to pay), improving our input corpora by isolating the text sections related to fact descriptions from those related to legal analysis, and further investigation of CamemBERT [9] for the multilabel classification task.

Acknowledgements. We would like to thank the Social Sciences and Humanities Research Council for funding this research through the Autonomy through Cyberjustice Technologies and Artificial Intelligence Project (ACT).

References

1. Aletras, N., Tsarapatsanis, D., Preoţiuc-Pietro, D., Lampos, V.: Predicting judicial decisions of the European Court of Human Rights: a natural language processing perspective. PeerJ Comput. Sci. **2**, e93 (2016)
2. Ashley, K.D., Brüninghaus, S.: Automatically classifying case texts and predicting outcomes. Artif. Intell. Law **17**(2), 125–165 (2009). https://doi.org/10.1007/s10506-009-9077-9
3. Boella, G., Di Caro, L., Humphreys, L., Robaldo, L., Rossi, P., van der Torre, L.: Eunomos, a legal document and knowledge management system for the web to provide relevant, reliable and up-to-date information on the law. Artif. Intell. Law **24**(3), 245–283 (2016). https://doi.org/10.1007/s10506-016-9184-3
4. Bojanowski, P., Grave, E., Joulin, A., Mikolov, T.: Enriching word vectors with subword information. Trans. Assoc. Comput. Linguist. **5**, 135–146 (2017)
5. Chalkidis, I., Fergadiotis, E., Malakasiotis, P., Androutsopoulos, I.: Large-scale multi-label text classification on EU legislation. In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pp. 6314–6322. Association for Computational Linguistics, Florence, July 2019
6. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: 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), pp. 4171–4186. Association for Computational Linguistics, Minneapolis, June 2019. https://doi.org/10.18653/v1/N19-1423, https://www.aclweb.org/anthology/N19-1423
7. Le, H., et al.: FlauBERT: unsupervised language model pre-training for French. arXiv preprint arXiv:1912.05372 (2019)
8. de Maat, E., Winkels, R.: Automated classification of norms in sources of law. In: Francesconi, E., Montemagni, S., Peters, W., Tiscornia, D. (eds.) Semantic Processing of Legal Texts. LNCS (LNAI), vol. 6036, pp. 170–191. Springer, Heidelberg (2010). https://doi.org/10.1007/978-3-642-12837-0_10
9. Martin, L., et al.: CamemBERT: a tasty French language model. arXiv preprint arXiv:1911.03894 (2019)

10. Maxwell, T., Schafer, B.: Natural language processing and query expansion in legal information retrieval: challenges and a response. Int. Rev. Law Comput. Technol. **24**(1), 63–72 (2010)
11. Nallapati, R., Manning, C.D.: Legal docket-entry classification: where machine learning stumbles. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing, pp. 438–446. Association for Computational Linguistics (2008)
12. Soh, J., Lim, H.K., Chai, I.E.: Legal area classification: a comparative study of text classifiers on Singapore Supreme Court judgments. In: Proceedings of the Natural Legal Language Processing Workshop 2019, pp. 67–77. Association for Computational Linguistics, Minneapolis, June 2019
13. Şulea, O.M., Zampieri, M., Vela, M., van Genabith, J.: 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, pp. 716–722. INCOMA Ltd., Varna, September 2017
14. Vacek, T., Schilder, F.: A sequence approach to case outcome detection. In: Proceedings of the 16th Edition of the International Conference on Artificial Intelligence and Law, pp. 209–215. ACM (2017)
15. Westermann, H., Walker, V.R., Ashley, K.D., Benyekhlef, K.: 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, pp. 133–142 (2019)
16. Zhong, H., Guo, Z., Tu, C., Xiao, C., Liu, Z., Sun, M.: Legal judgment prediction via topological learning. In: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, pp. 3540–3549 (2018)