Université de Montréal

Traitement automatique du langage naturel pour les textes juridiques : Prédiction de verdict et exploitation de connaissances du domaine

par

Olivier Salaün

Département d'informatique et de recherche opérationnelle
Faculté des arts et des sciences

Thèse présentée en vue de l'obtention du grade de
Philosophiæ Doctor (Ph.D.)
en Informatique

22 décembre 2023

Université de Montréal

Faculté des arts et des sciences

Cette thèse intitulée

Traitement automatique du langage naturel pour les textes juridiques : Prédiction de verdict et exploitation de connaissances du domaine

présentée par

Olivier Salaün

a été évaluée par un jury composé des personnes suivantes :

Guy Lapalme
(président-rapporteur)

Philippe Langlais
(directeur de recherche)

Jian-Yun Nie
(membre du jury)

Luc Lamontagne
(examinateur externe)

Marion Vacheret
(représentant du doyen de la FESP)

Résumé

À l'intersection du traitement automatique du langage naturel et du droit, la prédiction de verdict (*legal judgment prediction*) est une tâche permettant de représenter la question de la justice prédictive, c'est-à-dire tester les capacités d'un système automatique à prédire le verdict décidé par un juge dans une décision de justice. La thèse présente de bout en bout la mise en place d'une telle tâche formalisée sous la forme d'une classification multilabel, ainsi que différentes stratégies pour tenter d'améliorer les performances des classifieurs. Le tout se base sur un corpus de décisions provenant du Tribunal administratif du logement du Québec (litiges entre propriétaires et locataires).

Tout d'abord, un prétraitement préliminaire et une analyse approfondie du corpus permettent d'en tirer les aspects métier les plus saillants. Cette étape primordiale permet de s'assurer que la tâche de prédiction de verdict a du sens, et de mettre en relief des biais devant être pris en considération pour les tâches ultérieures. En effet, à l'issue d'un premier banc d'essai comparant différents modèles sur cette tâche, ces derniers tendent à exacerber des biais préexistant dans le corpus (p. ex. ils donnent encore moins gain de cause aux locataires par rapport à un juge humain). Fort de ce constat, la suite des expériences vise à améliorer les performances de classification et à atténuer ces biais, en se focalisant sur CamemBERT.

Pour ce faire, des connaissances du domaine cible (droit du logement) sont exploitées. Une première approche consiste à employer des articles de loi comme données d'entrée qui font l'objet de différentes représentations, mais c'est cependant loin d'être la panacée. Une autre approche employant la modélisation thématique s'intéresse aux thèmes pouvant être extraits à partir du texte décrivant les faits litigieux. Une évaluation automatique et manuelle des thèmes obtenus démontre leur informativité vis-à-vis des motifs amenant des justiciables à se rendre au tribunal.

Avec ce constat, la dernière partie de notre travail revisite une nouvelle fois la tâche de prédiction de verdict en s'appuyant à la fois sur des systèmes de recherche d'information (RI), et des thèmes associés aux décisions. Les modèles conçus ici ont la particularité de s'appuyer sur une jurisprudence (décisions passées pertinentes) récoltée selon différents critères de recherche (p. ex. similarité au niveau du texte et/ou des thèmes). Les modèles utilisant des critères de RI basés sur des sacs-de-mots (Lucene) et des thèmes obtiennent des gains significatifs en termes de scores F1 Macro. Cependant, le problème d'amplification des biais persiste encore bien qu'atténué.

De manière globale, l'exploitation de connaissances du domaine permet d'améliorer les performances des prédicteurs de verdict, mais la persistance de biais dans les résultats décourage le déploiement de tels modèles à grande échelle dans le monde réel. D'un autre côté, les résultats de la modélisation thématique laissent entrevoir de meilleurs débouchés pour ce qui relève de l'accessibilité et de la lisibilité des documents juridiques par des utilisateurs humains.

Mots-clés : apprentissage automatique, traitement automatique du langage naturel, droit, prédiction de verdict, classification multilabel, modélisation thématique, recherche d'information

Abstract

At the intersection of natural language processing and law, legal judgment prediction is a task that can represent the problem of predictive justice, or in other words, the capacity of an automated system to predict the verdict decided by a judge in a court ruling. The thesis presents from end to end the implementation of such a task formalized as a multilabel classification, along with different strategies attempting to improve classifiers' performance. The whole work is based on a corpus of decisions from the Administrative housing tribunal of Québec (disputes between landlords and tenants).

First of all, a preliminary preprocessing and an in-depth analysis of the corpus highlight its most prominent domain aspects. This crucial step ensures that the verdict prediction task is sound, and also emphasizes biases that must be taken into consideration for future tasks. Indeed, a first testbed comparing different models on this task reveals that they tend to exacerbate biases pre-existing within the corpus (i.e. their verdicts are even less favourable to tenants compared with a human judge). In light of this, the next experiments aim at improving classification performance and at mitigating these biases, by focusing on CamemBERT.

In order to do so, knowledge from the target domain (housing law) are exploited. A first approach consists in employing articles of law as input features which are used under different representations, but such method is far from being a panacea. Another approach relying on topic modeling focuses on topics that can be extracted from the text describing the disputed facts. An automatic and manual evaluation of topics obtained shows evidence of their informativeness about reasons leading litigants to go to court.

On this basis, the last part of our work revisits the verdict prediction task by relying on both information retrieval (IR) system, and topics assigned to decisions. The models designed here have the particularity to rely on jurisprudence (relevant past cases) retrieved

with different search criteria (e.g. similarity at the text or topics level). Models using IR criteria based on bags-of-words (Lucene) and topics obtain significant gains in terms of Macro F1 scores. However, the aforementioned amplified biases issue, though mitigated, still remains.

Overall, the exploitation of domain-related knowledge can improve the performance of verdict predictors, but the persistence of biases in the predictions hinders the deployment of such models on a large scale in the real world. On the other hand, results obtained from topic modeling suggest better prospects for anything that can improve the accessibility and readability of legal documents by human users.

Keywords: machine learning, natural language processing, law, legal judgment prediction, multilabel classification, topic modeling, information retrieval

Table des matières

Liste des tableaux

Liste des figures

Liste des sigles et des abréviations

[A] c. [B]	La nomenclature juridique pour nommer une décision est de la forme "[nom du demandeur] c. [nom du défendeur]" (exemple : "Tremblay c. Roger"). "c." est l'abréviation de "contre" et est équivalent à "v." pour "versus" utilisé en anglais (exemple : "Roe v. Wade").
BCE	De l'anglais *Binary Cross Entroy*, entropie croisée binaire.
BERT	De l'anglais *Bidirectional Encoder Representations from Transformers*, modèle de langage mis au point par [**Devlin et al., 2019**].
c.-à-d.	Abréviation de "c'est-à-dire".
C.c.Q.	Code civil du Québec.
CQ1s et CQ2s	Dans le cadre de la tâche de modélisation thématique, désignent les thèmes **c**andidats **q**ualifiés pertinents par au moins **1** ou par les **2** annotateurs non-experts du droit du logement.

CTs	De l'anglais ***c**andidate **t**opic**s***, traduisible par "thèmes candidats". Il s'agit des thèmes générés par les modèles thématiques.
DSV	**D**écomposition de matrice en **v**aleurs **s**ingulières.
e.g.	Du latin "exempli gratia" pour signifier "par exemple".
EM	De l'anglais ***E**xact **m**atch* pour l'exactitude, soit le ratio d'instances pour lesquelles le modèle a pu prédire l'ensemble exact de labels attendus. Cette métrique est aussi parfois appelée *accuracy*, mais l'expression *exact match* est préférable ici pour éviter les ambiguïtés.
ex.	Abréviation de "exemple".
i.e.	Du latin "id est" pour signifier "c'est-à-dire".
LDA	De l'anglais, *Latent Dirichlet Allocation* ou allocation de Dirichlet latente par [**Blei et al., 2003**].
L.R.L.	Loi sur la Régie du logement (depuis 2020, la Régie est désormais appelée le Tribunal administratif du logement).

LSI	De l'anglais, *lantent semantic indexing* ou analyse sémantique latente par [**Dumais et al., 1988**]. Aussi connu sous le nom de *Latent Semantic Analysis*.
MLM	De l'anglais *Masked Language Modeling* ou modélisation de langage masqué, tâche non supervisée dans laquelle un modèle doit retrouver les jetons préalablement masqués dans une séquence.
p. ex.	Abréviation de "par exemple".
RALI	Laboratoire de **R**echerche **a**ppliquée en **l**inguistique **i**nformatique à l'Université de Montréal.
RTs	De l'anglais ***r**eference **t**opic**s***, traduisible par "thèmes de référence". Il s'agit des 44 facteurs identifiés manuellement par [**Westermann et al., 2019**].
TAL	**T**raitement **a**utomatique du **l**angage naturel.
T.a.l.	**T**ribunal **a**dministratif du **l**ogement du Québec.
TF-IDF	De l'anglais *term frequency-inverse document frequency*.

token — Traduit en français par “jeton”. Selon le contexte, peut désigner les jetons qui constituent le vocabulaire des *tokenizers* utilisés par les modèles de type BERT [**Devlin et al., 2019**] (p. ex. le jeton [CLS]), ou encore les termes qui ont fait l’objet d’un prétraitement préalable avec d’éventuelles fusions pour les expressions multimots (p. ex. “eau_chaude” correspond à un jeton unique avec un trait de soulignement reliant deux termes).

Nomenclature des nombres : la virgule est utilisée pour séparer les milliers et le point pour distinguer les décimales (p. ex. “1,234,567.89”).

Remerciements

La route a été longue avant d'entrer dans le cursus doctoral et surtout pour en venir à bout. Plusieurs rencontres ont été cruciales tout au long de ce cheminement. Il est temps ici de rendre justice à celles et à ceux sans qui je n'aurais pu mener à bien la vaste entreprise qu'a été la thèse de doctorat à l'Université de Montréal (UdeM).

Je tiens en premier lieu à exprimer ma plus profonde gratitude à Philippe Langlais, professeur titulaire à l'UdeM au DIRO (Département d'informatique et de recherche opérationnelle), qui m'a donné l'opportunité de travailler au RALI (Recherche appliquée en linguistique informatique) sous sa supervision et d'éprouver ma valeur. Rien n'aurait été possible sans sa précieuse expérience, ses sages conseils et son soutien de tous les instants. Il est un scientifique engagé et dévoué avec qui mes échanges ont été aussi riches que passionnants. C'est grâce à lui et à son entremise que je dois d'avoir pu travailler sur un sujet de niche aussi pointu que celui de cette thèse. Tout au long de ce parcours, il m'a permis d'aiguiser une certaine sensibilité à "l'amour des données", ainsi que de développer une rigueur et une exigence scientifique qui m'ont guidé au cours de mon travail. C'est une très grande chance que d'avoir pu travailler auprès de lui, et dont je suis très honoré. J'espère que le travail effectué a été à la hauteur de ses attentes.

Je souhaite remercier Guy Lapalme, professeur associé et émérite à l'UdeM, 聶建雲 (Jian-Yun Nie), professeur titulaire à l'UdeM, et Luc Lamontagne, professeur titulaire à l'Université Laval, d'avoir accepté d'être les membres de mon jury, et d'avoir pris la peine et le temps d'examiner la présente thèse.

Je tiens à remercier les organismes suivants d'avoir permis le financement du doctorat : le laboratoire de Cyberjustice à la Faculté de droit, qui a aussi permis l'accès au corpus étudié ; le Département d'informatique et de recherche opérationnelle ; le Vice-Décanat ; la Faculté des arts et des sciences, ainsi que ses généreux donateurs ; la société Lexum Inc.

Je sais gré à Fabrizio Gotti, développeur spécialisé en TAL au RALI et conseiller scientifique chez IVADO (Institut de valorisation des données), d'avoir été disponible tout au long de ces années, et de m'avoir aidé à de nombreuses reprises au cours de mon travail grâce à son savoir et ses compétences aiguisées par sa riche expérience. Son avis a été précieux à différents instants dans ma recherche que ce soit pour un dépannage technique, une précision concernant un protocole expérimental, une relecture de papier, ou simplement me remonter le moral dans certaines situations ardues. Il est aussi un collègue fabuleux avec qui travailler et converser sont un véritable plaisir.

Je tiens à remercier les différents amis, (anciens) membres et collègues du RALI dont la compagnie a pu rendre ces années intellectuellement stimulantes, ponctuées notamment par différents jeux de société dont Exploding Kittens, Hanabi, Ricochet Robot ou le tarot (liste non exhaustive) qui ont animé notre groupe : David Alfonso Hermelo (*muchísimas gracias por tu generosidad y grandeza de alma*), Shivendra Bhardwaj, Khalil Bibi, Ilan Elbaz, Juan Felipe Duran, Arnaud Ferré, Xavier Frenette, Abbas Ghaddar, Laurent Jakubina, Guillaume Le Berre (merci pour les parties de tarot, les joueurs sont rares), William Léchelle, Vincent Letard, Nathan Migeon, Uros Petricevic, Frédéric Piedboeuf, Sriram Sanjeev Pratti, Louis Van Beurden, 高天健 (Tianjian Gao), 吳傑晨 (Jason Wu ; 我對你、姍姍和Emily也有美好的回憶), 聶一凡 (Yifan Nie), 杜攀 (Pan Du), 盧鵬 (Peng Lu), 呂志斌 (Lü Zhibin, dit Louis), 楊澤 (Yang Ze ;非常感謝妳給我的電鍋), 吳思凡 (Sifan Wu), 이현진 (Hyeonjin Lee, dite Daisy), 邵琪偉 (Qiwei Shao) et 婧涵 (Jing Han).

Je souhaite remercier aussi les membres d'autres laboratoires et les gens issus d'autres milieux avec qui mes échanges ont été fort enrichissants : Larissa Avononmadegbe, Rim

Ben Salem, Nicolas Garneau, Younès Kamel, Sylvain Longhais, Andrés Lou, Myriem Hnini, Capucine Marteau, Timothée Samou, Adam Samson, Florence So, Max Sobroza, 譚晉哲 (Tan Jinzhe), et Aurore Troussel.

Je tiens aussi à exprimer ma reconnaissance envers Raouf Bencheraiet et toute l'équipe du support technique du DIRO pour la maintenance des machines du département. Merci aussi à Céline Bégin, technicienne en gestion des dossiers étudiants au DIRO.

Il me faut aussi remercier Marc-André Morissette, Benjamin Cérat et Noreddine Ben Jillali, respectivement Directeur de la technologie, programmeur logiciel et programmeur analyste chez Lexum, qui m'ont accordé leur confiance pour un projet de génération de mots-clés. L'expérience de travail avec eux qui en a découlé m'a permis d'aborder ma recherche sous de nouveaux angles et de constater que je pouvais être utile aux autres.

Il me paraît aussi important de remercier ici les docteurs que j'avais rencontrés à Toulouse il y a plusieurs années et qui m'ont donné l'inspiration pour poursuivre un doctorat en informatique : Anaïs Cadilhac, Philippe Muller, Camille Pradel, Patrick Séguéla, Damien Sileo, et Lavoisier Wapet.

Au cours du doctorat, l'aube des années deux mille vingt s'est levée dans l'ombre des vagues d'une calamité sans précédent qui a fait sombrer de nombreuses vies. Il me faut remercier différentes personnes qui m'ont permis de rester un tant soit peu sain d'esprit et de tenir la barre contre vents et marées : 陳志群 (Chih-chun Chen), ami précieux et partenaire de langue d'une patience hors du commun, qui a accompli l'exploit de me faire voyager à Taïwan depuis Montréal tout au long d'innombrables heures de conversation (感激不盡) ; Hugo Duguay dont les nombreuses leçons de piano m'ont permis de passer le temps autrement que devant un moniteur et d'élargir mes passions en alternant entre les touches du clavier et celles du piano ; 方小姐 (Mlle. Fang) qui a été d'une très

grande patience et bienveillance à la 滿地可圖書閱覽室[1] ; 張莘培 (Hsin-pei Chang), une polyglotte hors pair qui a dû quitter Montréal en urgence au tout début de la pandémie, mais avec qui j'ai eu une correspondance nourrie malgré les douze heures de décalage horaire.

Enfin, il me faut remercier ma mère, mon père, ma sœur et Kucing pour leur soutien moral et matériel au cours de ce très long cheminement, notamment durant les moments les plus ardus qui l'ont précédé ou ponctué. Je suis très chanceux et heureux de les avoir comme proches, et je serai toujours reconnaissant à leur égard pour m'avoir donné autant de liberté. Cette thèse leur est dédiée.

[1]Située au dernier étage du 112 rue de la Gauchetière Ouest à Montréal, cette salle de lecture n'est hélas pas référencée sur Google Maps et ne possède pas de site internet. C'est pourtant une perle rare qui mérite davantage de visiteurs.

Introduction

Le traitement automatique du langage naturel (TAL) se présente généralement comme une discipline à la croisée de l'informatique et de la linguistique, visant à automatiser la compréhension et la génération du langage humain, que ce soit via la parole ou l'écrit. L'objet de la présente thèse se focalise sur un type bien particulier de langage qui préexiste aux sociétés humaines : le langage juridique (*legal language*). À mesure que les sociétés se complexifient, les liens sociaux (ex. : famille, contrats de mariage), marchands (ex. : commerce, contrats de travail) et de pouvoir (c.-à-d. la vie politique) qui s'y manifestent font l'objet de contrats, de lois, et de jugements qui régulent le comportement des individus et les relations entre différents groupes. Ces écrits et ces documents qui organisent les relations humaines s'appuient sur un langage technique et précis émanant entre autres selon [**Maley, 2014**] de l'exercice législatif des parlementaires et de la résolution de litiges par les juges dans les tribunaux. Ce langage juridique, dont les usages se ramifient en divers types de discours oraux et écrits [**Kurzon, 1997**], se distingue du langage courant en ceci qu'il constitue un support sur lequel s'inscrivent les choix organisationnels et les systèmes de régulation d'une société. C'est ainsi que [**Gibbons, 1999**] ajoute que "le droit est langage" (*"Law is language"*), et qu'au-delà d'être un langage juridique et une "institution sociale", il constitue une "institution profondément linguistique" (*"profoundly linguistic institution*).

Un tel langage juridique se spécialise de plus en plus au fur et à mesure que les concepts légaux qu'il vise deviennent de plus en plus précis[2]. De plus, ce langage fait aussi preuve d'une grande variété et diversité selon les différents contextes professionnels et culturels dans lesquels il est employé [**Goźdź-Roszkowski, 2012**]. Il faut aussi souligner que ce langage est devenu particulièrement foisonnant et accessible dans les dernières décennies avec

[2]Par exemple, un "meurtrier" du langage courant subira différents traitements selon qu'il lui est reproché un "homicide volontaire" ou "involontaire" dans le langage juridique.

l'essor d'Internet. Cette accessibilité instantanée a ainsi donné lieu, d'une part, à la mise à disposition massive de documents juridiques (p. ex. : lois votées par le parlement, décisions émises par les tribunaux), d'autre part, à la multiplication des recours par les justiciables (p. ex. : meilleure connaissance de leurs propres droits que l'on peut revendiquer, facilité à lancer une mise en demeure via un modèle ou un formulaire en ligne).

Dans le même temps, bien que les écrits en langage juridique soient devenus en apparence plus facilement accessibles, la technicité de ce langage et son décalage par rapport au langage courant font qu'il peut demeurer peu compréhensible pour les citoyens sans expertise en droit [**Charrow and Charrow, 1979**], voire même sans un minimum de compétence informatique [**Paley et al., 2021**]. Par ailleurs, la massification, la dématérialisation et la densification des documents juridiques a fait que la technologie, a fortiori ici le TAL, est devenue quasiment indispensable dans le travail des professionnels du droit (avocats, juristes, juges, greffiers). Ces derniers font ainsi appel à des moteurs de recherche et systèmes de web sémantique pour trouver des jurisprudences et des précédents pertinents vis-à-vis du cas qu'ils cherchent à élucider. Une fois le ou les documents pertinents trouvés, ceux-ci peuvent s'avérer particulièrement longs à lire et à comprendre, d'où le recours à d'autres outils pour par exemple identifier les citations nichées au sein du texte. D'autres cas d'usage professionnels concernent aussi la rédaction de contrats, qui débouche sur la production de longs documents à l'aide d'outils informatiques, et la traduction de textes légaux, qui requiert de retranscrire fidèlement les concepts visés d'une langue à l'autre. En raison de la densité et de la sensibilité de toutes ces opérations, la prolifération des documents juridiques s'est ainsi accompagnée d'un essor d'entreprises identifiables sous le vocable *LegalTech* [**Dale, 2019**] qui fournissent une panoplie d'outils allant de la veille documentaire à l'aide à la rédaction, en passant par l'aide à la décision. Sur le plan scientifique, [**Nazarenko and Wyner, 2017**] soulignent cependant que le TAL juridique présente encore des défis particulièrement difficiles :

- le langage juridique varie grandement d'un pays à un autre, et même d'un sous-domaine à un autre à l'intérieur d'une même langue ou d'un même pays, ce qui complique la généralisation d'un outil mis au point pour un sous-domaine et une langue en particulier ;

- le travail d'un expert juridique comprend tout un ensemble de tâches cognitives bien précises posant des défis techniques du plus bas au plus haut niveau d'abstraction : entre autres, la normalisation et le stockage optimal des documents consultables (*document engineering*) ; l'analyse syntactique et sémantique du texte au sein d'un document (p. ex. résolution des ambiguïtés et des anaphores) ; la gestion des arguments mis en jeu au sein du raisonnement juridique avec la visée de fournir une aide à la prise de décision. Chacune de ces étapes constitue une tâche expérimentale à elle seule. À l'heure où ces lignes sont écrites, nous n'avons pas connaissance d'un modèle unique pouvant prendre en charge toutes ces tâches cognitives de façon unifiée. S'il existe différents *benchmarks* formalisant un tel défi tel celui de [**Rasiah et al., 2023**], il s'agit généralement d'un empilement de différentes tâches à partir du même corpus où des modèles distincts sont déployés séparément à chaque étape du travail juridique.

Dans le cas qui nous concerne, la présente thèse se focalise sur une tâche maîtresse dans le TAL juridique : la prédiction de verdict (*legal judgement prediction*). Cette tâche a ceci de particulier que les modèles sont amenés à simuler le raisonnement juridique que ferait un juge humain avant d'aboutir au verdict. Le but ici n'est pas de chercher à automatiser le travail des juges, encore moins de concevoir des juges-robots, mais plus précisément d'analyser jusqu'à quel point les modèles de langage modernes sont en mesure de produire ce type de travail. Contrairement aux corpora de TAL qui concernent généralement des documents employant un langage courant (p. ex. nouvelles, réseaux sociaux), le langage utilisé en droit cherche à exprimer un raisonnement logique et rationnel, que certains chercheurs (évoqués ultérieurement) comparent à un langage mathématique (p. ex. "Si tel critère et tel critère sont satisfaits, alors telle conséquence a lieu.") À partir d'un corpus de décisions du Tribunal administratif du logement du Québec, le but du présent travail a été de formaliser la tâche de justice prédictive d'un point de vue apprentissage machine, et d'y tester et évaluer différents modèles, le tout en gardant en considération les implications métiers d'un point de vue juridique. D'autres tâches telles que la modélisation thématique ont aussi été mises en place afin de montrer d'autres approches possibles vis-à-vis d'un corpus juridique. La présente thèse se compose des chapitres suivants :

(1) Une revue de littérature des différents travaux et corpora mobilisés dans le cadre des recherches à l'intersection de l'informatique et du droit, avec une focalisation sur les enjeux du TAL juridique ;

(2) Une analyse approfondie du corpus de décisions du Tribunal administratif du logement du Québec (anciennement connu sous le nom de Régie du logement). Ce chapitre décrit aussi les différentes étapes de nettoyage et préparation des instances ;

(3) La formalisation et la mise en place d'une tâche de classification pour la prédiction de verdict, le tout avec un banc d'essai comprenant des modèles non neuronaux et neuronaux ;

(4) L'exploitation d'articles de lois, connaissances spécifiques au domaine juridique, comme soutien à la prédiction de verdict. Il s'agit de déterminer les conditions dans lesquelles ces connaissances légales peuvent améliorer les performances ;

(5) L'usage de la modélisation thématique comme moyen d'extraire les sujets saillants des différents litiges. Le but est de montrer comment les modèles thématiques permettent d'effectuer une extraction des motifs de litiges récurrents à l'échelle de tout le corpus comparé à une approche manuelle ;

(6) Une tentative de rehausser la performance des prédicteurs de verdict au moyen de systèmes basés sur la modélisation thématique et la recherche d'information.

Publications

Le contenu de cette thèse repose en partie sur différents travaux publiés au cours du doctorat, présentés ci-dessous en ordre chronologique :

(1) **Salaün, O.**, Langlais, P., Lou, A., Westermann, H., and Benyekhlef, K. (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. [mentionné dans le Chapitre 3] ;

(2) Lou, A., **Salaün, O.**, Westermann, H., and Kosseim, L. (2021). Extracting facts from case rulings through paragraph segmentation of judicial decisions. In *International Conference on Applications of Natural Language to Information Systems*, pages 187–198. Springer. [mentionné dans le Chapitre 2] ;

(3) **Salaün, O.**, Langlais, P., and Benyekhlef, K. (2021). Labels distribution matters in performance achieved in legal judgment prediction tasks. In *Proceedings of the Eighteenth International Conference on Artificial Intelligence and Law*, ICAIL '21, page 268–269, São Paulo, SP, Brazil. Association for Computing Machinery. [mentionné dans le Chapitre 4] ;

(4) **Salaün, O.**, Langlais, P., and Benyekhlef, K. (2021). Exploiting domain-specific knowledge for judgment prediction is no panacea. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021)*, pages 1234–1243. [mentionné dans le Chapitre 4] ;

(5) **Salaün, O.**, Gotti, F., Langlais, P., and Benyekhlef, K. (2022a). Why do tenants sue their landlords? answers from a topic model. In *Legal Knowledge and Information Systems*, pages 113–122. IOS Press. (**Best Student Paper Award**). [mentionné dans le Chapitre 5 et la Conclusion] ;

(6) **Salaün, O.**, Troussel, A., Longhais, S., Westermann, H., Langlais, P., and Benyekhlef, K. (2022b). Conditional abstractive summarization of court decisions for laymen and insights from human evaluation. In *Legal Knowledge and Information Systems*, pages 123–132. IOS Press. [mentionné dans la Conclusion] ;

(7) Cérat, B., **Salaün, O.**, Ben Jillali, N., Morissette, M.A., Pocovnicu, I., Elliott, E., Harvey, F. (2023). LexKey: A Keyword Generator for Legal Documents. In *Proceedings of the 6th Workshop on Automated Semantic Analysis of Information in Legal Text (ASAIL 2023) co-located with the 19th International Conference on Artificial Intelligence and Law (ICAIL 2023)*, pages 111-121. [mentionné dans la Conclusion] ;

(8) **Salaün, O.**, Piedboeuf, F., Le Berre, G., Hermelo, D. A., & Langlais, P. (2024). EUROPA: A Legal Multilingual Keyphrase Generation Dataset. *arXiv preprint arXiv:2403.00252*. (soumis à ACL RR) [mentionné dans la Conclusion].

Chapitre 1

Informatique et droit : Revue de littérature

Au sein de l'institution judiciaire, les magistrats, juges, juristes, avocats et autres experts du droit doivent gérer et traiter une masse importante de documents juridiques (p. ex. les décisions de tribunal) et législatifs (p. ex. les textes de loi), de telle sorte que l'informatique fournit un secours précieux pour les assister dans leur travail. Ainsi, les systèmes de recherche d'information constituent en premier lieu un outil fort utile pour chercher une jurisprudence ou un texte de loi pertinent vis-à-vis d'un litige en cours de traitement. Cela est vrai d'une part dans le droit anglo-saxon (*common law* qui prédomine dans les pays anglophones et du Commonwealth) où chaque décision de justice repose sur des précédents ou des jugements précédents. Cela est vrai d'autre part dans le droit romano-civiliste ou continental (*civil law* qui prédomine notamment en Europe continentale dont la France et l'Allemagne, en Amérique latine, en Chine) dans lequel une décision de tribunal repose sur l'interprétation des lois existantes au moment où un juge doit trancher. D'autres applications pratiques de l'informatique concernent par ailleurs la rédaction et le traitement de contrats, des documents particulièrement longs et complexes, où les juristes peuvent être assistés par des systèmes pour rédiger des documents à partir de modèles ou récupérer des informations clés au sein de documents existants.

Avec l'amélioration des puissances de calcul et la numérisation massive des ressources légales, certains cabinets d'avocats et entreprises juridiques, notamment aux États-Unis, ont également commencé à concevoir d'autres outils qui visent à sonder les chances de succès de leur client pour un litige donné (c.-à-d. quels sont les gains ou pertes potentielles à aller en cour ? Le cas échéant, quels juges ou quels facteurs externes ont le plus d'effet sur la

décision finale du tribunal ?). Du côté des magistrats, toujours aux États-Unis, il existe même des systèmes pour estimer le risque de récidive d'un prévenu, dont COMPAS analysé par [**Stevenson, 2018**].

Ces activités ont peu à peu permis l'émergence de ce qui est présentement appelé la justice prédictive, pouvant être définie comme un domaine visant à prédire ou du moins à estimer les probabilités qu'un juge tranche une décision dans un sens ou dans un autre. Les professionnels du droit préfèrent employer le terme "jurimétrie" proposé il y a plus d'un demi-siècle par [**Loevinger, 1948, Loevinger, 1963**]. En son temps, celui-ci réclamait et désignait par ce terme une approche scientifique du droit et une modernisation de la réflexion juridique par l'apport de méthodes quantitatives. Il opposait ainsi la jurimétrie, comme approche scientifique du droit, à la jurisprudence, en tant que philosophie du droit. Plus récemment, la jurimétrie désigne l'ensemble des applications des méthodes statistiques et d'apprentissage automatique aux documents juridiques. Elle est surtout le fait de jeunes entreprises informatiques regroupées sous les vocables *legal AI* et *legal tech* qui fournissent des services pour optimiser le travail des juristes, comme le décrit [**Barthe, 2017**]. L'effervescence de ces diverses applications a non seulement ouvert un marché pour les firmes juridiques et informatiques, mais a aussi posé des questions éthiques et morales quant au bien-fondé de tels systèmes prédictifs : faut-il s'attendre à ce que les acteurs judiciaires, par un instinct grégaire, finissent par conformer leurs comportements aux prédictions émises par ces systèmes ? Est-ce que ces derniers peuvent amplifier ou réduire les biais préexistants dans les jugements passés, ce qui peut avoir des conséquences majeures pour certaines minorités sociales ?

Au-delà de ces considérations sociétales, la justice prédictive, plus communément appelée en anglais *legal judgment prediction* dans la communauté du TAL, a aussi ouvert un champ de recherche à la croisée du droit et de l'informatique en posant des problèmes scientifiques complexes : alors que les jugements reposent en grande partie sur des raisonnements logiques exprimés en langage naturel, jusqu'à quel point un système informatique peut-il reproduire le même genre de réflexion juridique effectué par un juge ? Dans quelle mesure est-ce que les techniques d'intelligence artificielle et d'apprentissage machine peuvent effectuer de telles tâches ? La revue de littérature qui suit donne un premier aperçu de ce qui a être réalisé en ce sens en prenant pour point de départ les données utilisées.

1.1. Les corpora juridiques pour le traitement automatique du langage naturel (TAL)

Contrairement à la plupart des jeux de données et *benchmarks* disponibles en TAL qui sont accessibles sans grandes restrictions à la communauté académique, les corpora juridiques, notamment issus de tribunaux, sont plus ou moins faciles d'accès. Cela peut s'expliquer par différents facteurs:

- un choix de la part des juges ou des institutions de ne pas permettre une large diffusion des jugements qu'ils produisent, parfois dans le but de protéger les magistrats [1] ;
- l'absence de plate-forme et/ou de pratique de diffusion pour le grand public ;
- quand elles existent, les plates-formes sont protégées par des systèmes contre la collecte de données (*web scraping*) ;
- l'existence de lois pour la préservation des informations personnelles contenues dans les décisions, ce qui limite les possibilités de diffusion (d'autres pays comme l'Allemagne résolvent ce problème en anonymisant les informations à la source).

De tels documents sont généralement collectés par des éditeurs spécialisés qui vendent par la suite des abonnements payants à destination des experts juridiques et des facultés de droit.

Pour ce qui est des tâches de TAL qui ont pu intégrer des corpora juridiques dans leurs protocoles expérimentaux, elles consistent pour la plupart en des tâches de classification ayant pour cibles le verdict décidé par les magistrats (sous la forme de classes) ou des métadonnées qui étaient déjà immédiatement disponibles avec les documents. Il est important de souligner ici des caractéristiques et des différences majeures de ces corpora spécialisés vis-à-vis de ceux plus communément utilisés en TAL :

- Les textes juridiques se présentent la plupart du temps soit : sous la forme de textes de loi ou de comptes-rendus issus d'organes législatifs (p. ex. le corpus Europarl du Parlement européen par [**Koehn, 2005**]) qui sont souvent largement diffusés sur

[1] En France, par exemple, une loi de réforme de la justice votée en 2019 dispose : "Les données d'identité des magistrats et des membres du greffe ne peuvent faire l'objet d'une réutilisation ayant pour objet ou pour effet d'évaluer, d'analyser, de comparer ou de prédire leurs pratiques professionnelles réelles ou supposées." Les contrevenants s'exposent à cinq années de détention.

internet ; soit sous la forme de décisions judiciaires (décisions de tribunal écrites par un juge) dont l'accès peut être plus difficile pour le grand public et le milieu de la recherche pour les raisons évoquées précédemment. Bien que [**Katz et al., 2023**] se félicitent d'une plus large diffusion de *benchmarks* en TAL légal (p. ex. LexGLUE par [**Chalkidis et al., 2021b**], Multi-LexSum par [**Shen et al., 2022**]) permettant la reproductibilité des expériences au cours des dernières années, il n'en demeure pas moins que la facilité d'accès aux documents reste surtout vraie pour ceux émis par les tribunaux au sommet de la hiérarchie de l'appareil judiciaire (p. ex. la Cour suprême). Elle reste cependant plus ardue pour ceux provenant de cours de première instance, comme le corpus du Tribunal administratif du logement ;

- Ces documents sont généralement beaucoup plus longs que ceux utilisés en TAL. Alors que chaque instance d'un corpus générique consiste à minima en un gazouillis[2] ou au plus en quelques phrases, une décision juridique ou un document législatif est plus susceptible de s'étaler sur plusieurs paragraphes voire pages ;
- Le domaine juridique nécessite une certaine expertise pour mettre au point des tâches de TAL qui aient du sens. Dans le cadre de tâches de justice prédictive ou *legal judgment prediction*, la mise au point de cibles/classes/labels pour caractériser le verdict que le modèle doit prédire n'est pas une tâche triviale comme l'a montré [**Vacek and Schilder, 2017**]. Cela s'explique par plusieurs facteurs : la très grande variété des verdicts décidés par les juges ; la rareté des experts qui sont en mesure d'annoter ces données (l'emploi d'annotateurs de Mechanical Turk d'Amazon n'est pas envisageable ici) ; le coût d'annotation (il faut rémunérer les juristes ou étudiants faisant les annotations). L'annotation de textes longs et très spécialisés est telle que les jeux de données annotés auront une taille plus modeste par rapport à ceux touchant à des domaines plus génériques ou accessibles au grand public.

C'est ainsi qu'il est possible d'identifier une première famille de corpora avec des annotations manuelles comprenant les *court dockets* (registres de tribunaux américains) [**Nallapati and Manning, 2008, Vacek and Schilder, 2017**] (respectivement 5.6k instances pour une classification binaire et près de 10k instances pour une

[2]Une publication sur X, réseau social anciennement connu sous le nom Twitter.

tâche combinant classification de texte et étiquetage séquentiel), des décisions portugaises [**Gonçalves and Quaresma, 2005**] (8k documents pour de la classification multiclasse), les décisions de la Cour Suprême des États-Unis [**Segal, 1984**] (123 instances pour de la classification binaire) et la Cour européenne des droits de l'homme [**Aletras et al., 2016, Liu and Chen, 2017**] (un total de 584 instances pour des classifications binaires). En raison du coût d'annotation, ces corpora sont de taille modeste de l'ordre de quelques centaines d'instances, à l'exception de [**Katz et al., 2017**] qui jouissent de 240k instances pour de la classification multiclasse.

Par la suite, une deuxième famille de corpora émerge dans laquelle la réutilisation de labels ou de métadonnées déjà disponibles permettent de s'émanciper du besoin d'annotateurs et d'obtenir ainsi des volumes de données bien plus importants. C'est ainsi que plusieurs *benchmarks* s'inscrivent dans la tâche de prédiction du domaine juridique d'un document (p. ex. droit de la famille, droit commercial) avec des documents de la Cour de cassation française [**Şulea et al., 2017**] (127k instances), des documents de l'Union européenne [**Chalkidis et al., 2019b**] (57k instances) et des décisions singapouriennes [**Soh et al., 2019**] (126k instances). Cette croissance dans le volume des données est surtout notable avec les corpora chinois de [**Long et al., 2019**] (186k décisions liées à des divorces, classification binaire) et CAIL2018 (*Chinese AI and Law challenge dataset*) par [**Xiao et al., 2018**] (2.7 millions de jugements criminels, deux tâches de classification multilabel et une tâche de régression). De tels ordres de grandeur rendent désormais possible l'entraînement de modèles neuronaux qui nécessitent de larges volumes de données. Dans le même temps, la massification des corpora fait aussi passer au second plan les aspects singuliers propres aux documents juridiques, notamment tout ce qui touche au raisonnement amenant le magistrat à trancher un litige : quels sont les faits et les preuves rapportées par les parties ? Quels sont les articles pertinents pour résoudre le litige ? En quoi les règles contenues dans les articles de loi sont-elles satisfaites ou contredites par les éléments apportés par les différentes parties ?

1.2. Les documents juridiques ne sont pas des documents comme les autres en TAL : le défi de rendre compte du raisonnement juridique sous-jacent

Bien avant l'émergence des modèles à la transformeur (*transformer*) de [**Vaswani et al., 2017**] et de la standardisation de la tâche de classification de texte comme façon de modéliser la prédiction de jugement, les travaux à la croisée de l'intelligence artificielle et du droit privilégiaient les approches par raisonnement à partir de cas (*case-based reasoning*, les documents/cas sont représentés sous la forme de facteurs) ou raisonnement à partir de règles (*rule-based reasoning*), à l'image des travaux de [**Skalak, 1989**] et [**Ashley and Brüninghaus, 2009**]. L'objectif était alors de reproduire et d'imiter le plus possible la démarche que ferait un juriste humain et ainsi de mieux rendre compte du chemin emprunté par un modèle pour délivrer sa prédiction comme décrit par [**Bench-Capon, 2021**]. Une telle approche s'accompagne cependant d'un coût d'annotation manuel important puisqu'il faut traduire chaque instance du langage naturel (juridique) vers un langage formel et convertir toutes les possibilités de raisonnement juridique en un enchaînement de règles logiques. Malgré ce coût, des travaux menés durant la période voyant le succès grandissant des modèles neuronaux et/ou BERT [**Devlin et al., 2019**] ont continué à utiliser de telles approches par règles afin de mieux rendre compte et conserver des traces des mécanismes juridiques sous-jacents :

- ainsi, [**Dragoni et al., 2016**] ont utilisé un parseur grammatical provenant de [**Manning et al., 2014**] et la base de connaissances WordNet conçue par [**Miller, 1995**] pour extraire des règles logiques (sous la forme d'obligations, permissions et interdictions) à partir d'une section du Code de protection des consommateurs de télécommunications en Australie. La qualité des règles extraites n'a cependant pas pu être vérifiée faute de références/*gold standard* disponibles ;
- dans une tâche avec environ 300 questions dans le domaine du droit fiscal américain, [**Holzenberger et al., 2020**] montrent qu'un modèle BERT [**Devlin et al., 2019**], même après pré-entraînement (*masked language modeling*) reste en deçà d'un système `Prolog` qui résout correctement toutes les questions ;

- le même système `Prolog` a aussi été employé par [**Collenette et al., 2020**] dans un raisonneur dont la tâche est de déterminer si l'article de loi européen donnant droit à un procès équitable a été violé. Cette approche avait surtout l'avantage de fournir automatiquement une explication du cheminement logique aboutissant au verdict.

En dehors de ces différents exemples d'approches par règle ou cas, il faut aussi citer certains *benchmarks* dont la construction tient compte de la nature juridique particulière des textes utilisés :

- la compétition annuelle COLIEE par [**Kano et al., 2018**] et [**Rabelo et al., 2020**] propose ainsi des tâches de recherche d'information et d'inférence (*entailment*) tirées d'examens du barreau japonais et de décisions canadiennes ;
- un *benchmark* analogue est aussi proposé par [**Zhong et al., 2020**], cette fois basé sur des questions d'examen du barreau chinois ;
- l'équivalent d'un SQuAD [**Rajpurkar et al., 2016**] chinois juridique (tâche de *machine reading comprehension*) a été conçu par [**Duan et al., 2019**] avec l'intervention d'experts du domaine.

Malgré l'existence de ces *benchmarks* qui restent plus spécifiques au domaine du droit, le format de la classification de texte reste la plupart du temps privilégié pour mettre en place une expérience de prédiction de verdict. Cependant, il faut souligner que ce format est davantage employé pour des litiges avec des circonstances et des verdicts relativement faciles à délimiter et à discerner (p. ex. le refus ou l'accord d'un permis, d'une pension ou d'une indemnisation). Cela débouche sur des classes binaires faciles à comprendre et formaliser. Un tel format ne peut cependant pas convenir à tous les domaines juridiques et à toutes les décisions au fur et à mesure qu'augmentent la quantité et la diversité des informations en entrée (p. ex. : confrontation d'allégations contradictoires entre deux voire davantage de parties, les parties comprennent plusieurs personnes, présence/absence de preuves, témoignages, citation d'articles ou d'un jugement antérieur) et des verdicts possibles en sortie, comme l'ont montré [**Vacek and Schilder, 2017**].

Ces différents éléments illustrent ainsi une certaine dichotomie entre deux approches opposées dès qu'une tâche de TAL implique d'utiliser des textes juridiques :

- d'une part, une approche voulant conserver des traces de raisonnement juridique à des fins explicatives. [**Bench-Capon, 2021**] avance même que les approches par règle gardent un avantage de par leur transparence et leur explicabilité contrairement aux approches d'apprentissage automatique dont le fonctionnement sous-jacent reste plus difficile à interpréter. Les approches par règle s'accompagnent cependant d'un coût d'annotation élevé, car il faut manuellement retranscrire le sens des textes en langage naturel en un langage formel ;
- d'autre part, une approche plus répandue aujourd'hui qui minimise le coût d'annotation en se contentant d'extraire des métadonnées ou des cibles faciles à identifier à partir des corpora juridiques. Cette approche emploie les mêmes protocoles expérimentaux que ceux utilisés pour les corpora génériques de TAL, mais tend aussi à sursimplifier la réalité du travail du juge (p. ex. recours à une classification binaire).

Le prochain chapitre présente un corpus en français de décisions provenant du Tribunal administratif du logement du Québec et quelles informations peuvent en être extraites en vue de mettre en place des tâches d'apprentissage machine. Il sera vu notamment en quoi le contenu et le traitement de ces données sont originaux par rapport aux corpora légaux existants.

Chapitre 2

Corpus : analyse des décisions du Tribunal administratif du logement du Québec

2.1. Contexte

Le Tribunal administratif du logement (anciennement dénommé Régie du logement jusqu'au 31 août 2020) est un organisme judiciaire au Québec qui traite exclusivement les litiges survenant entre locateurs[1] et locataires avec l'application du droit du logement, notamment des articles de lois tirés du Code civil du Québec et de la Loi de la Régie du Logement. Différentes raisons peuvent amener des justiciables à demander gain de cause devant cette cour: les locateurs réclameront par exemple des compensations en raison d'un locataire qui ne paye pas son loyer tandis qu'un locataire peut poursuivre un locateur qui laisserait le logement loué dans un état insalubre.

Certaines de ces décisions de justice sont accessibles publiquement via le portail SOQUIJ[2]. Dans le cadre du projet de recherche ACT/AJC[3], le laboratoire de Cyberjustice[4] de la faculté de droit de l'Université de Montréal a obtenu l'accès à une grande partie de ces données. Le

[1]Terme juridique utilisé au Québec pour désigner les propriétaires qui mettent en location des biens immobiliers. Il est aussi important de noter que "locatrice" est le féminin de "locateur" tandis que "locataire" (la personne qui loue un logement) est un terme qui peut être aussi bien féminin que masculin.

[2]`http://citoyens.soquij.qc.ca/`

[3]ACT : *Autonomy Through Cyberjustice Technologies* ; AJC : Autonomisation des acteurs Judiciaires par la Cyberjustice et l'intelligence artificielle ; `https://www.ajcact.org/`

[4]`https://www.cyberjustice.ca/`

laboratoire RALI (Recherche appliquée en linguistique informatique) s'inscrit dans ce projet en ayant pour but de traiter les enjeux de TAL (Traitement automatique des langues) autour des textes légaux. Bien que nous n'avons pas la permission de partager nos données avec la communauté, nous nous efforcerons néanmoins d'illustrer nos travaux avec des exemples de notre corpus chaque fois que cela est pertinent.

2.2. Description et aperçu préliminaire des motifs récurrents du corpus à partir des métadonnées

L'ensemble du corpus comprend 981,112 documents en français émanant de 72 juges et 29 tribunaux à travers le Québec de 2001 à 2018. Ils étaient à l'origine au format `doc` ou `docx` que nous avons convertis en `HTML` pour faciliter l'extraction de métadonnées autour du corps de texte principal. Un exemple illustratif est donné à la Figure 2.1 où ces métadonnées sont, par exemple, le lieu, la date de l'audience ou des informations personnelles concernant le juge et les parties impliquées. Cette décision est un exemple fréquent d'un locateur qui attaque son locataire pour loyer impayé. D'autres jugements concernent des cas où le locataire poursuit son propriétaire en justice en raison par exemple d'un logement mal entretenu. Pour ce qui est du corps de texte d'une décision, les paragraphes suivent généralement la structure suivante :

- d'abord une **description des faits** et des éléments de preuves de chaque partie (ex. : un locateur prouve que le locataire est en retard dans le paiement de son loyer, lignes 1 à 4 sur la Figure 2.1) ;
- ensuite une **analyse juridique** dans laquelle le juge étudie si certains articles de loi sont applicables ou non (lignes 5 à 9) ;
- enfin, le **verdict** du juge (ici, le locataire doit payer des dommages et intérêts au locateur, lignes 10 à 12).

Il faut noter que la séparation entre les deux premières parties est rarement explicite dans l'ensemble des documents ; ils peuvent même être entremêlés, comme ont pu le constater [**Lou et al., 2021**] qui cherchaient à distinguer les phrases relevant des faits et celles relevant de l'analyse juridique. Seul le verdict peut être facilement séparé grâce à une formule récurrente qui le précède ("Pour ces motifs/raisons, le tribunal :").

Bureau de [VILLE]
Régisseure : M[e] [NOM DU JUGE]

N° de la demande : [IDENTIFIANT]

[NOM ET ADRESSE DU DEMANDEUR]
Locateur - Partie demanderesse

c.

[NOM ET ADRESSE DU DÉFENDEUR]
Locataire - Partie défenderesse

Logement concerné : [ADRESSE DU BIEN LOUÉ]
Date de l'audience : [DATE DE L'AUDIENCE]
Présence(s) : le locateur
le locataire

D é c i s i o n

[1] Le locateur demande la résiliation du bail et l'expulsion du locataire, le recouvrement du loyer 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 à durée indéterminée au loyer mensuel de 585 $, payable le premier jour de chaque mois.
[3] La preuve démontre que le locataire doit 2 765 $, soit des arrérages de loyer depuis octobre 2007, plus 6 $ représentant les frais de signification prévus au Règlement.
[4] Le locataire croit que le montant est moindre, ce qu'il ne peut prouver. Il invoque qu'il n'a pas de bail écrit ni de reçu. Il invoque différents problèmes relatifs au logement suivant lesquels il s'est vu opposer une fin de non-recevoir par ce tribunal, à l'occasion d'une décision précédemment rendue, suivant le témoignage qu'il fait à l'audience. Il prétend avoir réintégré son logement avec l'aide des policiers (expulsion par le locateur), ce qu'il ne prouve pas autrement que par son témoignage.
[5] Le tribunal précise qu'il n'est pas requis que le bail soit écrit pour justifier une demande en recouvrement de loyer, celui-ci étant exigible du fait de l'occupation des lieux (bail par tolérance).
[6] Il appartient au locataire de prouver qu'il a payé son loyer ou qu'il a une défense d'inexécution à faire valoir, ce qu'il n'a pas fait.
[7] 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.
[8] 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.
[9] Le préjudice causé au locateur justifie l'exécution provisoire de la décision, comme il est prévu à l'article 82.1 L.R.L.

POUR CES MOTIFS, LE TRIBUNAL :

[10] **RÉSILIE** le bail et **ORDONNE** l'expulsion du locataire et de tous les occupants du logement;
[11] **ORDONNE** l'exécution provisoire, malgré l'appel, de l'ordonnance d'expulsion à compter du 11[e] jour de sa date;
[12] **CONDAMNE** le locataire à payer au locateur la somme de 2 765 $, plus les intérêts au taux légal et l'indemnité additionnelle prévue à l'article 1619 C.c.Q., à compter du 20 mai 2008 sur la somme de 2 180 $, et sur le solde à compter de l'échéance de chaque loyer, plus les frais judiciaires de 70 $.

Le [DATE DE SIGNATURE]

Fig. 2.1. Exemple de décision du Tribunal administratif du logement (les informations personnelles ont été masquées).

Durant le prétraitement, de très nombreux doublons ont été retirés du corpus, ainsi que les instances pour lesquelles certaines métadonnées n'ont pas pu être extraites. De tels doublons s'expliquent par le fait que les décisions ne faisaient pas l'objet d'un archivage structuré, d'où

la présence de quelques centaines de milliers de documents en double éparpillés dans différents dossiers. L'ensemble aboutit à un total de 667,305 documents.

Avec des connaissances du domaine et des outils de TAL, nous avons pu extraire des connaissances à partir des métadonnées, notamment la distribution entre les différents types de parties[5]. Ces dernières peuvent être séparées entre deux grands groupes : les personnes morales (entités juridiques immatérielles comme les entreprises, les organisations ou les associations) et les personnes physiques. Ces dernières se divisent en d'autres sous-catégories pour les hommes seuls, les femmes seules (cela signifie que la partie est constituée par une seule personne et n'a aucun rapport avec le statut marital), les groupes de personnes et les défunts représentés par un liquidateur[6]. La distribution est présentée plus en détail dans le Tableau 2.1. Elle révèle que lorsque le locateur est le demandeur (89% des décisions), les plaignants sont généralement des personnes morales suivies par des hommes qui attaquent en majorité des hommes devant le tribunal. Dans les cas où le locataire est le demandeur (11% de tout le corpus), les plaignants sont essentiellement des hommes et des femmes qui attaquent en majorité des hommes suivis par des personnes morales (p. ex. entreprises, organismes).

Des informations disponibles ont aussi pu être extraites, entre autres choses, la présence ou l'absence des défendeur(s) ou plaignant(s) le jour de l'audience et la présence d'avocat ou mandataire (c.-à-d. un représentant) pour chaque partie. Ainsi, pour les litiges de types Locateur c. Locataire[7], le Tableau 2.2 montre que le locateur-demandeur est présent à la quasi-totalité des audiences alors que le locataire-défendeur est présent dans seulement plus d'un tiers des cas. Pour ce qui est des litiges de type Locataire c. Locateur, le Tableau 2.3 montre que les locataires-demandeurs sont présents dans deux tiers des audiences alors que ce taux atteint 91% pour les propriétaires. Un tel absentéisme parmi les locataires s'explique selon [**Gallié et al., 2016**] par le fait que les décisions sont généralement favorables aux

[5]Les "parties" dans un procès désignent le demandeur et le défendeur. Une partie peut être composée d'une ou plusieurs personnes.

[6]Lorsqu'un locataire vivant seul décède, ses héritiers ou la personne responsable de ses avoirs (liquidateur) doivent procéder eux-mêmes à la résiliation du bail et au paiement des loyers non versés par le défunt.

[7]La nomenclature juridique pour nommer une décision est de la forme "Demandeur c. Défendeur", "c." étant l'abréviation de "contre".

Type de litige		Locateur c. Locataire		Locataire c. Locateur	
Partie		Demandeur	Défendeur	Demandeur	Défendeur
Personne morale		41.3	0.2	0.3	33.0
Pers. physique	Homme seul	36.8	60.1	54.0	40.0
	Femme seule	11.0	39.5	45.5	14.0
	Groupe	10.7	0.0	0.0	12.6
	Succession	0.2	0.2	0.2	0.3
Total		100	100	100	100
Nombre de décisions		595,808		71,497	

Tableau 2.1. Distribution des types de personnes par partie et par type de litige (Locateur c. Locataire et Locataire c. Locateur)

		Locateur		Total
		Présent	Absent	
Locataire	Présent	33.6	2.9	36.5
	Absent	63.4	0.1	63.5
Total		97.0	3.0	

Tableau 2.2. Taux de présence des locateurs et locataire dans les audiences de type Locateur c. Locataire (595,808 litiges, taux en pourcentages).

propriétaires, ce qui nourrit un sentiment d'impuissance des locataires qui ne prennent pas la peine de se présenter au tribunal.

Le manque de familiarité des locataires avec les procédures judiciaires s'illustre aussi par leur plus faible recours à un avocat par rapport aux locateurs, comme le montre le Tableau 2.4. Ces premières remarques à partir des métadonnées présagent l'existence de biais importants qui sont explorés en profondeur dans la section qui suit.

		Locateur		Total
		Présent	Absent	
Locataire	Présent	59.9	8.6	68.5
	Absent	31.2	0.3	31.5
Total		91.1	8.9	

Tableau 2.3. Taux de présence des locateurs et locataire dans les audiences de type Locataire c. Locateur (71,497 litiges, taux en pourcentages).

	Locateur c. Locataire	Locataire c. Locateur	Tous les litiges
Locateur représenté par un avocat	4.2%	14.6%	4.8%
Locataire représenté par un avocat	1.9%	10.9%	2.4%

Tableau 2.4. Taux de locateurs et locataires représentés par un avocat selon le type de litige.

2.3. De la difficile et délicate préparation des données pour des tâches d'apprentissage machine

Le corps de texte central (p. ex. lignes 1 à 12 sur la Figure 2.1) se prête plus difficilement à l'extraction de connaissances, contrairement aux métadonnées, mais nous avons pu identifier néanmoins à l'aide d'expressions régulières :

- les articles de lois cités par le juge. Nous nous focalisons ici notamment sur ceux tirés du Code civil du Québec [**Assemblée nationale, 2018**] et de la Loi de la Régie du Logement [**Assemblée nationale, 2016**];
- les types de verdicts prononcés par le juge, notamment à travers l'utilisation d'expressions régulières appliquées au verdict. Ainsi, dans la Figure 2.1, la seule présence des termes "CONDAMNE" et "RÉSILIE" sont de bons relais pour indiquer que le locataire a été condamné à payer des indemnités et que son bail a été résilié ;

- le montant total des indemnités à payer en dollars canadiens par le défendeur qui est condamné.

Le fait de pouvoir extraire des informations concernant l'issue des jugements, que ce soit en termes de types de verdicts (exemples : condamnation du défendeur, résiliation du bail, rétractation d'un jugement antérieur, rejet des demandes du plaignant) ou de montants à payer pour un prévenu ouvre la voie à des tâches de prédiction d'articles ou de verdict, par exemple.

2.3.1. Extraction des articles de lois et structuration des verdicts sous forme de labels

Afin de mettre au point un corpus utilisable pour nos expériences, nous identifions avec des expressions régulières les articles de loi cités qui apparaissent au moins 100 fois dans tout le corpus, soit un total de 154 articles distincts, dont 134 tirés du Code civil du Québec et 20 de la Loi sur la Régie du logement. Si un article de loi est mentionné à plusieurs reprises dans une décision, il est comptabilisé une seule fois pour ce document.

Pour ce qui est des verdicts, nous nous sommes fixé l'exigence de distinguer le plus possible les différentes nuances dans la variété des sentences prononcées par les juges. Cette étape est complexe, car contrairement à d'autres tâches de prédiction de verdict le plus souvent formalisées sous la forme de classification binaire qui sursimplifie la réalité (p. ex. acceptation ou refus d'un recours), un certain soin a été apporté ici pour que les étiquettes de verdict obtenues traduisent le plus fidèlement possible la réalité, ce qui nécessite une connaissance métier assez fine. C'est ainsi qu'à l'inverse de certaines cours où le juge remet une décision pouvant être interprétée de façon binaire (p. ex. est-ce que le prévenu a violé un article de la Convention européenne des droits de l'Homme dans la tâche de [**Chalkidis et al., 2019a**]), notre jeu de données comprend des situations plus complexes à l'issue desquelles aucune des parties n'est ni forcément "gagnante" ni "perdante" (p. ex. les parties choisissent de faire un accord à l'amiable devant le juge ; la cour estime que l'audience doit être reportée). Il faut aussi ajouter que les labels sont ici cumulables[8]. À défaut d'avoir une armée d'annotateurs

[8]Pour clarifier, les tâches de classification multiclasse impliquent que chaque instance appartient à une seule catégorie tandis que les tâches multilabels impliquent que chaque instance peut cumuler plusieurs étiquettes à la fois.

humains avec une expertise juridique pour labelliser toutes les décisions, c'est par divers essais-erreurs et l'expérience accumulée que la complexité des verdicts a pu être annotée de plus en plus finement. Dans le détail, c'est la lecture manuelle de nombreuses décisions, et les quelques collaborations en parallèle avec le laboratoire de Cyberjustice pour des travaux hors du champ de la prédiction de verdict, qui ont permis d'avoir une meilleure appréciation des données.

Initialement, un partitionnement préliminaire non supervisé (*clustering*) a été effectué au niveau des lignes des verdicts (c.-à-d. des segments de texte séparés par des sauts de lignes) représentés par des vecteurs TF-IDF afin d'en déceler les traits les plus saillants (p. ex. verbes en lettres capitales, formulations récurrentes dans les verdicts). Par la suite, un partitionnement[9] sur la base de représentations Sentence-BERT [**Reimers and Gurevych, 2020a**] des lignes de chaque verdict[10] a été effectué dans une tentative d'avoir directement des *clusters* associables à des étiquettes. Les résultats ont été relativement décevants dans la mesure où les *clusters* n'étaient pas particulièrement homogènes (p. ex. des tournures de phrases très proches mais aux sens différents étaient regroupées ensemble) et restaient relativement nombreux. C'est ainsi que des segments de verdict avec des significations identiques se retrouvaient dans des groupes séparés.

Ces partitionnements ont tout de même été utiles pour repérer des tournures de phrases récurrentes, et ainsi mettre en place des règles et expressions régulières permettant de mieux distinguer différentes nuances. Par exemple, un label spécifiant au premier abord une sanction financière quelconque (une information grossière et peu détaillée) a été remplacé par deux labels distinguant la direction du paiement (du locataire au propriétaire ou du propriétaire au locataire). Cela a permis par exemple de révéler que lorsqu'un propriétaire demande et obtient la reprise d'un logement (pour s'y loger ou y loger un membre de sa famille), il se retrouve aussi condamné par le juge à payer une compensation aux occupants expulsés. Une

[9]La technique de *clustering* utilisée dans la librarie sentence_tranformers est nommée *fast community detection*. Elle consiste à regrouper des instances selon que la similarité cosinus entre leurs *embeddings* dépasse un seuil donné. Une communauté est constituée dès que le nombre d'instances la composant dépasse un nombre minimal.

[10]Nous avions utilisé le modèle multilingue `https://huggingface.co/sentence-transformers/paraphrase-multilingual-MiniLM-L12-v2`.

telle pseudoannotation apporte ainsi un degré plus riche de nuances qui sort de la dichotomie gagnant/perdant.

Les étiquettes de verdict dont la fréquence est au moins égale à 100 dans l'ensemble du corpus sont conservées, ce qui donne un total de 28 labels. Il est utile de rappeler ici qu'une décision peut cumuler plusieurs articles et labels de verdict. Seules les décisions contenant au moins un article cité et au moins un label de verdict sont conservés[11], ce qui donne un jeu de données final de 564,900 décisions, avec une longueur moyenne de 394 jetons (tokénization effectuée avec l'analyseur français[12] de `spaCy` [**Honnibal et al., 2020**]) et un écart-type de 448 jetons. Les 28 labels de verdict retenus sont les suivants (par ordre décroissant) :

(1) tenant_pays_landlord (88.2% du corpus) : le locataire doit payer des dommages-intérêts / une compensation au locateur ;

(2) eviction (56.5%) : le/la locataire et/ou les occupants du logement sont expulsés ;

(3) termination_lease (53.7%) : le contrat de location est résilié par le tribunal[13] ;

(4) applicant_request_denied (38.2%) : les requêtes du demandeur sont totalement ou partiellement rejetés. Il se peut par exemple qu'un propriétaire obtienne la condamnation d'un locataire à payer des dommages-intérêts pour loyer impayé sans qu'une expulsion soit prononcée ;

(5) provisional_enforcement (35.9%) : le verdict fait l'objet d'une exécution provisoire, c'est-à-dire qu'il est appliqué sans attendre un éventuel appel du défendeur ;

(6) reserve_recourse (24.8%) : le tribunal réserve ses recours au demandeur ; celui-ci peut lancer une nouvelle demande au juge pour le même litige le cas échéant ;

(7) lease_already_terminated (12.1%) : le demandeur avait demandé la résiliation du bail par le juge, mais ce dernier constate que le contrat de location est déjà résilié de fait ;

(8) applicant_gets_no_reimbursement_for_court_fees (3.4%) : les frais judiciaires payés par le demandeur à l'origine de l'audience restent à sa charge ;

[11] Nous posons l'hypothèse qu'une décision citant au moins un article a un fondement légal mieux construit et argumenté que si elle n'en mentionnait aucun.

[12] `https://spacy.io/models/fr#fr_core_news_md`

[13] Pour une lectrice ou un lecteur ne vivant pas au Québec, il faut savoir qu'il est plus difficile au Québec de résilier un contrat de bail unilatéralement à moins d'attendre la date anniversaire du contrat généralement fixé au 1er juillet, de transférer le bail à un nouveau locataire, ou bien de recourir à un juge.

(9) landlord_pays_tenant (4.0%) : le locateur doit payer des dommages-intérêts / une compensation au locataire ;

(10) tenant_ordered_pay_rent (4.2%) : le juge ordonne au locataire de payer le loyer à temps ;

(11) agreement (0.9%) : le juge valide et donne une valeur juridique à un accord entre les deux parties (p. ex. un locataire s'engage à payer le loyer à temps pour éviter une expulsion) ;

(12) cancel_ruling (1.4%) : le tribunal rétracte une décision passée ;

(13) landlord_repossesses_rental_unit (1.5%) : le tribunal autorise le locateur à reprendre possession de son logement ;

(14) tribunal_sets_new_rent (1.0%) : le tribunal fixe un nouveau montant de loyer (p. ex. diminution) ;

(15) conditional_sentence (1.3%) : le verdict est assorti d'une peine qui n'entre en vigueur que si une condition (n')est(pas) remplie (p. ex. une expulsion est prononcée si le locataire ne paye pas tout le loyer dû avant une date limite) ;

(16) uphold_ruling (0.9%) : le tribunal maintient une décision qui a été produite précédemment ;

(17) new_audience (0.9%) : le tribunal ordonne une nouvelle audience ultérieure entre les parties ;

(18) no_more_recourse (0.3%) : le juge déclare le demandeur forclos d'introduire une nouvelle demande, c'est-à-dire que le demandeur ne peut plus effectuer de nouveau recours pour le même litige (cette mesure permet d'éviter qu'une personne multiplie abusivement des procédures pour gagner du temps) ;

(19) order_landlord_repairs (0.4%) : le juge ordonne au locateur de procéder à des réparations ou travaux dans le logement ;

(20) tenant_can_deduct_from_rent (0.4%) : le juge donne au locataire la possibilité de déduire du loyer la compensation que doit lui verser le locateur ;

(21) withdrawal_demand (0.2%) : le juge prend acte du désistement par le demandeur d'une partie de sa demande (p. ex. un locateur décide de ne plus demander l'expulsion du locataire au cours de l'audience) ;

(22) tenant_must_provide_access (0.2%) : le juge ordonne au locataire de permettre au locateur l'accès au logement (p. ex. pour permettre des visites, l'intervention d'ouvriers pour des travaux) ;

(23) verdict_related_to_peaceful_enjoyment (0.1%) : le verdict a trait à la jouissance paisible du logement ou de l'immeuble dans lequel le logement est situé (p. ex. bruit, moisissure) ;

(24) outside_jurisdiction (0.1%) : le juge déclare que le tribunal n'est pas compétent pour se saisir du litige (p. ex. le litige doit être résolu devant un autre type de cour) ;

(25) verdict_related_to_pets (0.1%) : le verdict concerne des animaux domestiques ;

(26) assignment_of_lease (0.1%) : le juge tranche un litige concernant une cession de bail (le fait qu'un locataire transmette son bail à un autre locataire, c'est un procédé que le propriétaire ne peut empêcher que dans des conditions bien précises) ;

(27) verdict_related_to_land_use_change ($< 0.1\%$) : le juge tranche un litige impliquant un changement d'affectation du bien loué (p. ex. transformation en zone commerciale) ;

(28) inside_jurisdiction($< 0.1\%$) : le juge déclare que le tribunal est bien compétent pour se saisir du litige.

Parmi les décisions du corpus final, 94% sont émises suite à une demande par un locateur et 6% suite à une demande par un locataire. La réduction du ratio des demandes par les locataires (11% à 6%) est due au filtrage de litiges pour lesquels le locataire-demandeur ne s'est pas présenté à l'audience. En raison de ces absences, les décisions produites sont très pauvres (aucun article de loi cité, demande automatiquement rejetée) et sont ainsi exclues du jeu de données final.

Une vérification manuelle a été opérée sur 100 instances prises au hasard (une moitié dans laquelle le demandeur est le propriétaire, une autre moitié dans laquelle le demandeur est le locataire) afin d'évaluer la qualité de cette pseudoannotation, c'est-à-dire vérifier si les étiquettes d'articles et de verdict ont été correctement attribuées (pour rappel, une instance peut cumuler plusieurs étiquettes). Sur cet échantillon, l'exactitude dans l'attribution des articles et des labels de verdict est de respectivement 0.96 et 0.98. Certains articles n'ont pas pu être capturés par les expressions régulières en raison de leur format (p. ex. citation d'un alinéa à l'intérieur d'un article au lieu de citation de l'article en tant que tel). La

conversion des décisions du format DOC(X) vers HTML a aussi entraîné la disparition de certains espaces spéciaux, ce qui empêche une bonne capture (p. ex. “1854al.1C.c.Q.”). C'est sans compter aussi les fautes de frappe. Pour ce qui est des labels, l'annotation se heurte à des verdicts évoquant des dédommagements en nature ou des clarifications de règles qui sont trop particulières à un cas précis pour permettre la capture par regex. Ces cas de figure sont aussi trop rares pour faire des labels qui y sont dédiés : “DÉCLARE que le locataire a l'usage d'un espace de stationnement intérieur dont le coût est inclus dans son loyer; DÉCLARE que locateur assume le déneigement de la voie d'accès à cet espace de stationnement;”, “ORDONNE au locataire d'enlever l'ensemble des rebuts de toute sorte”. Enfin, pour l'ensemble du corpus, le nombre moyen d'articles identifiés s'élève à 2.55 (médiane à 2) et la moyenne de labels de verdict par instance s'élève à 3.31 (médiane à 4).

2.3.2. Segmentation intradocument

Comme indiqué dans la section 2.2 et montré dans la Figure 2.1, le verdict de chaque décision est facile à isoler, car systématiquement précédé par une formulation prévisible de la forme : “POUR CES MOTIFS/RAISONS, LE TRIBUNAL”. La segmentation la plus délicate concerne la partie du document avant verdict, en particulier la séparation entre les faits et l'analyse juridique. Idéalement, seuls les paragraphes des faits doivent faire partie des données d'entrée dans une tâche de prédiction de verdict. Ce filtrage est essentiel pour éviter que le modèle ait accès à des indices révélant directement ou indirectement le verdict même. Ainsi, plus un paragraphe est proche du début du verdict, plus il est susceptible de faire fuiter des informations présageant le résultat final. Ce risque de fuite n'est pas propre à notre corpus et avait déjà été soulevé par [**Şulea et al., 2017**] au moment de mettre en place une tâche semblable avec la Cour de cassation française. Dans une tâche de segmentation portant sur un sous-ensemble de notre corpus (plus de 5k instances avec une délimitation faits-analyse explicite, ce qui est très exceptionnel), [**Lou et al., 2021**] avaient utilisé un classifieur binaire définissant si un paragraphe (délimité par des sauts de lignes) relevait des faits ou de l'analyse juridique. Une fois tous les paragraphes classés, le but était alors de trouver une “coupe” au milieu d'eux de façon à isoler les X premiers paragraphes contenant la plus forte densité de faits. Ce travail a révélé que faits et analyses n'étaient pas toujours strictement séparés, ni ne se suivaient séquentiellement, voire même qu'ils pouvaient

être entremêlés. Une approche par segmentation automatique pour l'ensemble du corpus paraissait ainsi peu adéquate, et c'est pourquoi une approche par règles a été employée à la place. À l'issue d'un examen approfondi des données et différents tâtonnements, plusieurs heuristiques ont été développées pour filtrer au mieux le texte des faits :

- retenir tous les paragraphes précédant un sous-titre annonçant l'analyse juridique (p. ex. "Analyse du droit applicable", "La discussion", "Le droit") ;
- retirer les paragraphes qui contiennent des citations de législations ;
- retenir tous les paragraphes précédant celui contenant une phrase dans laquelle le sujet, isolé par un parseur grammatical `spaCy`, désigne un magistrat ou le tribunal[14]. Cette dernière technique s'est avérée relativement efficace pour détecter l'analyse juridique qui contient souvent des tournures de phrase telles que "Le tribunal estime que" ou "la cour est d'avis que". Pour l'exemple de la Figure 2.1, cela permet de retenir les quatre premiers paragraphes pour les données d'entrée.

Une vérification manuelle a été opérée sur 100 instances (chaque moitié avec respectivement le propriétaire et le locataire en demandeur) pour évaluer la qualité de cette segmentation. La segmentation a été jugée correcte dans 92% des cas. Les erreurs décelées concernent des segments laissant présager la direction du verdict (p. ex. "Considérant que la demande du locateur est bien fondée"), ou encore des coupes qui ont lieu trop tôt dans le document. Le segment suivant est un exemple de faux positif qui prive le modèle d'autres informations factuelles qui apparaissent à la suite : "Le locataire voisin a aussi reçu un avis d'éviction qui a également fait l'objet d'une opposition sur laquelle le tribunal se prononce simultanément". De façon globale, il est important de resouligner la difficulté à séparer les faits et l'analyse juridique qui sont soit peu délimités, soit entremêlés, comme l'avaient remarqué [**Lou et al., 2021**]. Pour l'ensemble des documents, la section des faits a une longueur moyenne de 210 jetons avec un écart-type de 207[15].

2.3.3. *Split* temporel des documents

Il est d'usage dans une tâche d'apprentissage machine de constituer trois jeux d'entraînement, de validation et de test à partir d'instances préalablement randomisées, de sorte que la

[14]Nous utilisons l'expression régulière suivante: `"auteure?s?|collègues?|juge(?:s|\s)|régie|tribuna(?:l|ux)"`.

[15]La tokénization a encore une fois été réalisée avec `spaCy`.

distribution des données dans chaque jeu soit à peu près identique. [**Søgaard et al., 2021**] ont cependant montré qu'un découpage aléatoire pouvait amener à une surestimation de la performance des modèles. Un tel découpage a surtout le défaut de ne pas rendre compte de la capacité du modèle à généraliser à des instances futures dont la distribution diffère de celles des instances passées utilisées durant l'entraînement, comme souligné par [**Lazaridou et al., 2021**]. Afin d'avoir un cadre aussi réaliste que possible pour une telle tâche de prédiction et en suivant le protocole de [**Chalkidis et al., 2021a, Medvedeva et al., 2021**], le choix a été fait de constituer des ensembles d'entraînement, validation et test qui se suivent dans l'ordre chronologique :

- l'ensemble d'entraînement comprend les instances datant de 2001 à 2011 (371,989 instances soit 66% du corpus) ;
- l'ensemble de validation comprend les instances datant de 2012 à 2014 (101,757 instances soit 18% du corpus) ;
- l'ensemble de test comprend les instances datant de 2015 à 2018 (91,154 instances soit 16% du corpus).

Ce découpage a aussi été choisi de façon à s'assurer à ce que chaque période temporelle contient suffisamment d'instances pour toutes les étiquettes, en particulier celles minoritaires. La distribution temporelle des labels majoritaires reste relativement stable comme le montre la Figure 2.2, mis à part la chute en 2015 qui est due à des départs à la retraite et à des absences des magistrats pour cause de maladie. Le même graphique est présenté sur la Figure 2.3 avec des valeurs relatives où l'on voit clairement que le locataire doit payer des dommages-intérêts dans plus de 80% des cas, quelle que soit l'année.

Si l'on se concentre sur la distribution des 18 labels les moins fréquents, présentée en valeur absolue et relative sur les Figures 2.4 et 2.5 respectivement, on constate davantage de volatilité à travers le temps.

2.4. Des biais du corpus

2.4.1. Des labels déséquilibrés

Il a été vu précédemment un déséquilibre important entre les ratios de litiges de la forme Locateur c. Locataire (94%) et Locataire c. Locateur (6%). Les 28 étiquettes utilisées pour

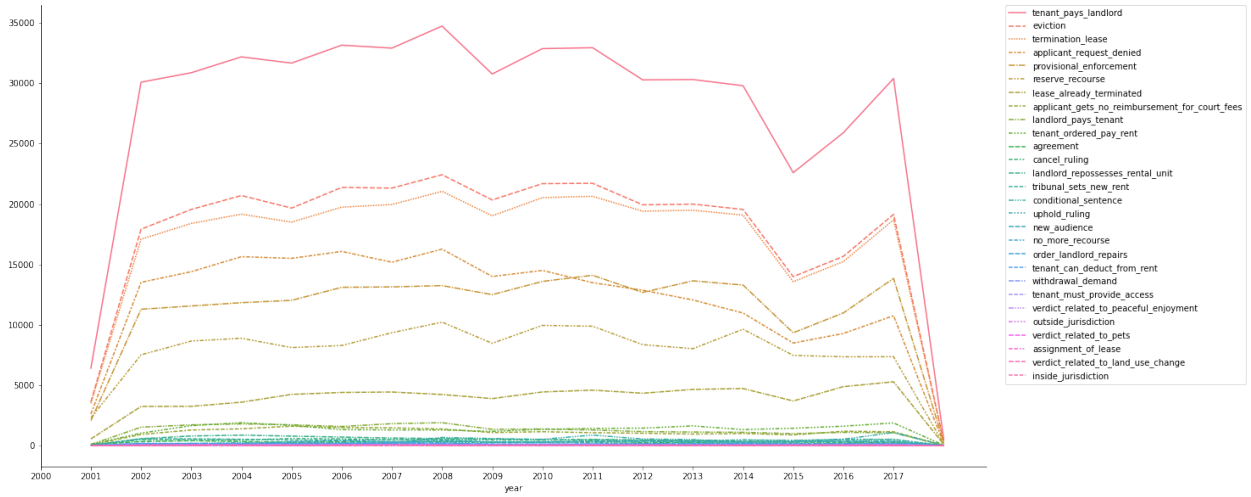

Fig. 2.2. Distribution des 28 labels de verdict par année (valeur absolue). Comme le corpus comprend les quelques premières décisions de 2018, les courbes convergent vers 0 à la fin de l'axe des abscisses.

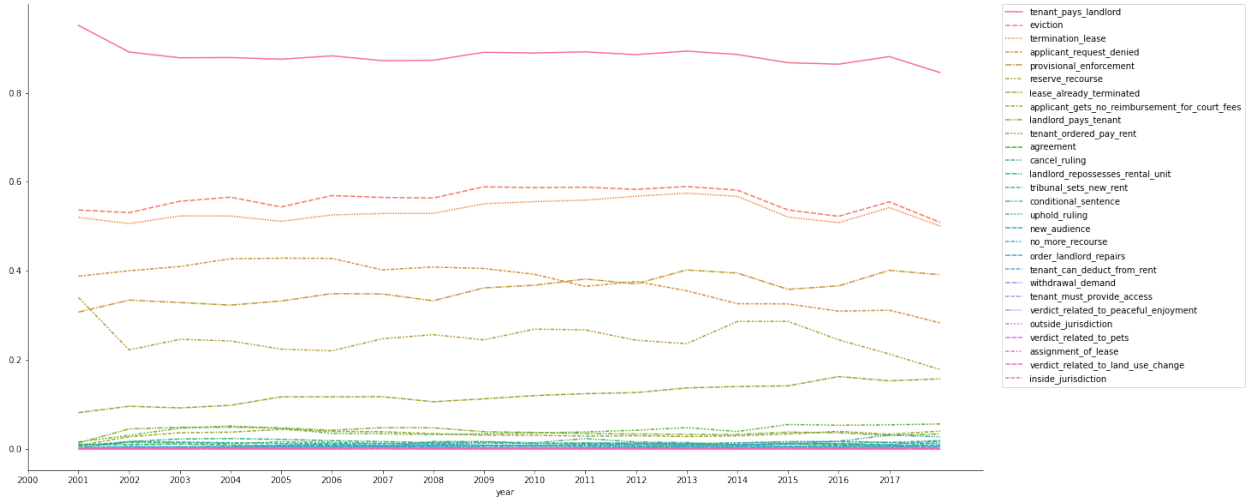

Fig. 2.3. Distribution des 18 labels de verdict les moins fréquents par année (valeur absolue).

caractériser les différents verdicts associés révèlent d'autres biais encore plus importants. Ainsi, les trois labels de verdict les plus fréquents parmi les litiges où le demandeur est un propriétaire sont : tenant_pays_landlord (94.2%), eviction (60.2%) et termination_lease (57.2%). En d'autres termes, les locateurs-demandeurs ont gain de cause dans plus de 90% des cas, notamment pour des demandes d'expulsion. Pour ce qui est des litiges où le demandeur est un locataire, les trois labels les plus fréquents sont : applicant_request_denied (60.6%), landlord_pays_tenant (37.6%) et applicant_gets_no_reimbursement_for_court_fees

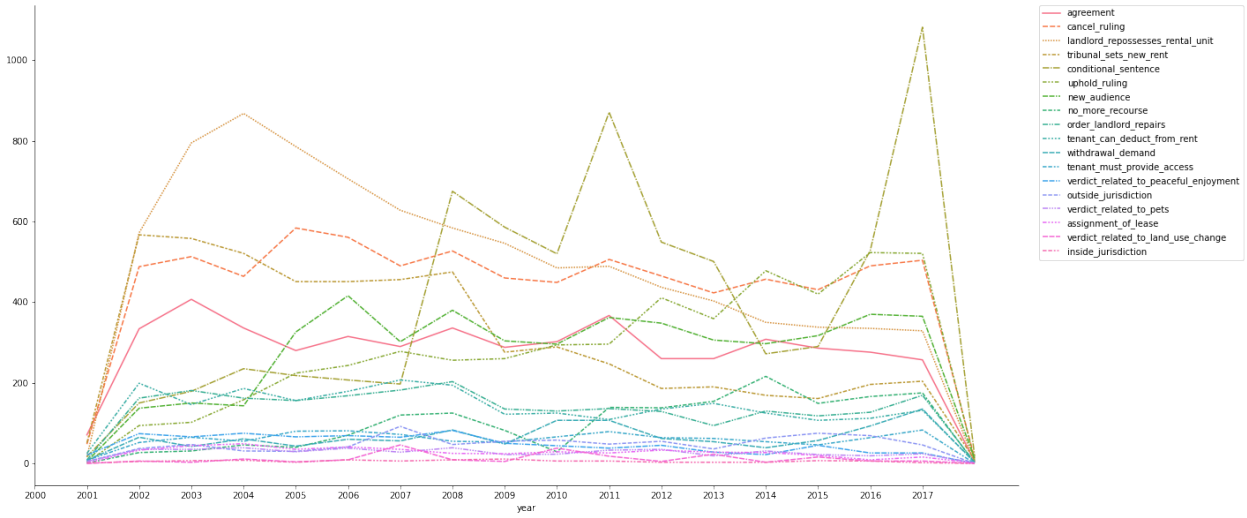

Fig. 2.4. Distribution des 18 labels de verdict les moins fréquents par année (valeur absolue).

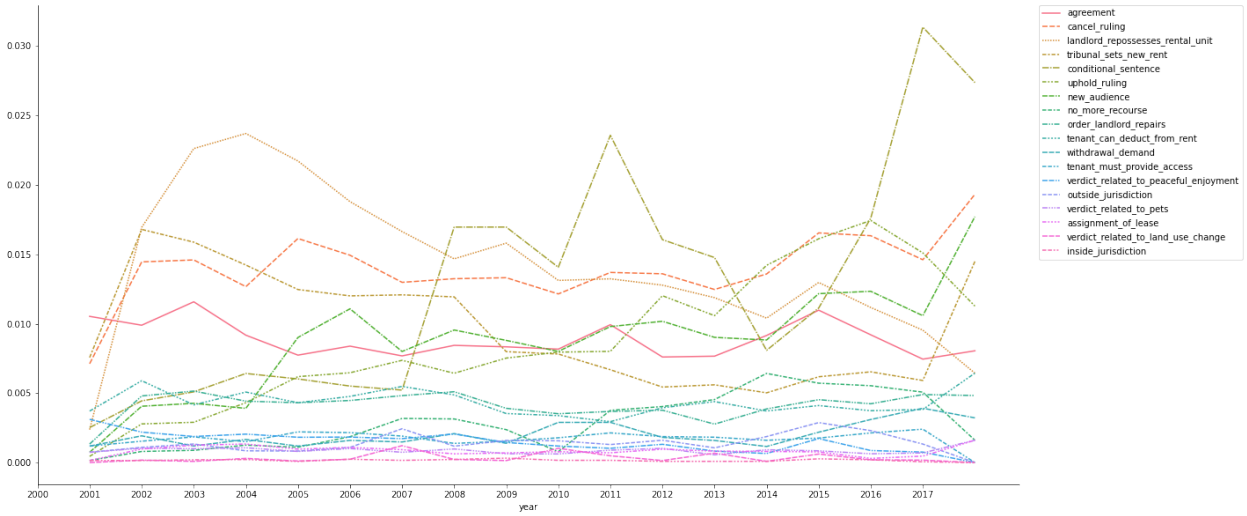

Fig. 2.5. Distribution des 18 labels de verdict les moins fréquents par année (valeur relative).

(18.4%). Autrement dit, un locataire voit ses demandes généralement rejetées par le tribunal, avec parfois des frais judiciaires qui restent à sa charge. Il obtiendrait gain de cause dans seulement un peu plus d'un tiers des cas (versement de dommages-intérêts par le locateur).

Ce déséquilibre se mesure aussi dans le montant des dommages-intérêts à verser. Le défendeur paye en moyenne une somme plus importante s'il est locataire (2,421$) que s'il est locateur (1,595$). Ces biais et cette différence de succès entre les demandes émises par les propriétaires et celles émises par les locataires sont des informations qui doivent être prises en compte au moment d'analyser des résultats de prédiction de verdict.

2.4.2. Des corrélations article-verdict et entre les labels de verdict

Chaque instance peut se voir attribuer plusieurs articles et labels de verdict. Une matrice de corrélation entre les labels de verdict, illustrée par la carte de chaleur sur la Figure 2.6, montre des groupements de labels qui apparaissent ensemble dans les décisions du juge.

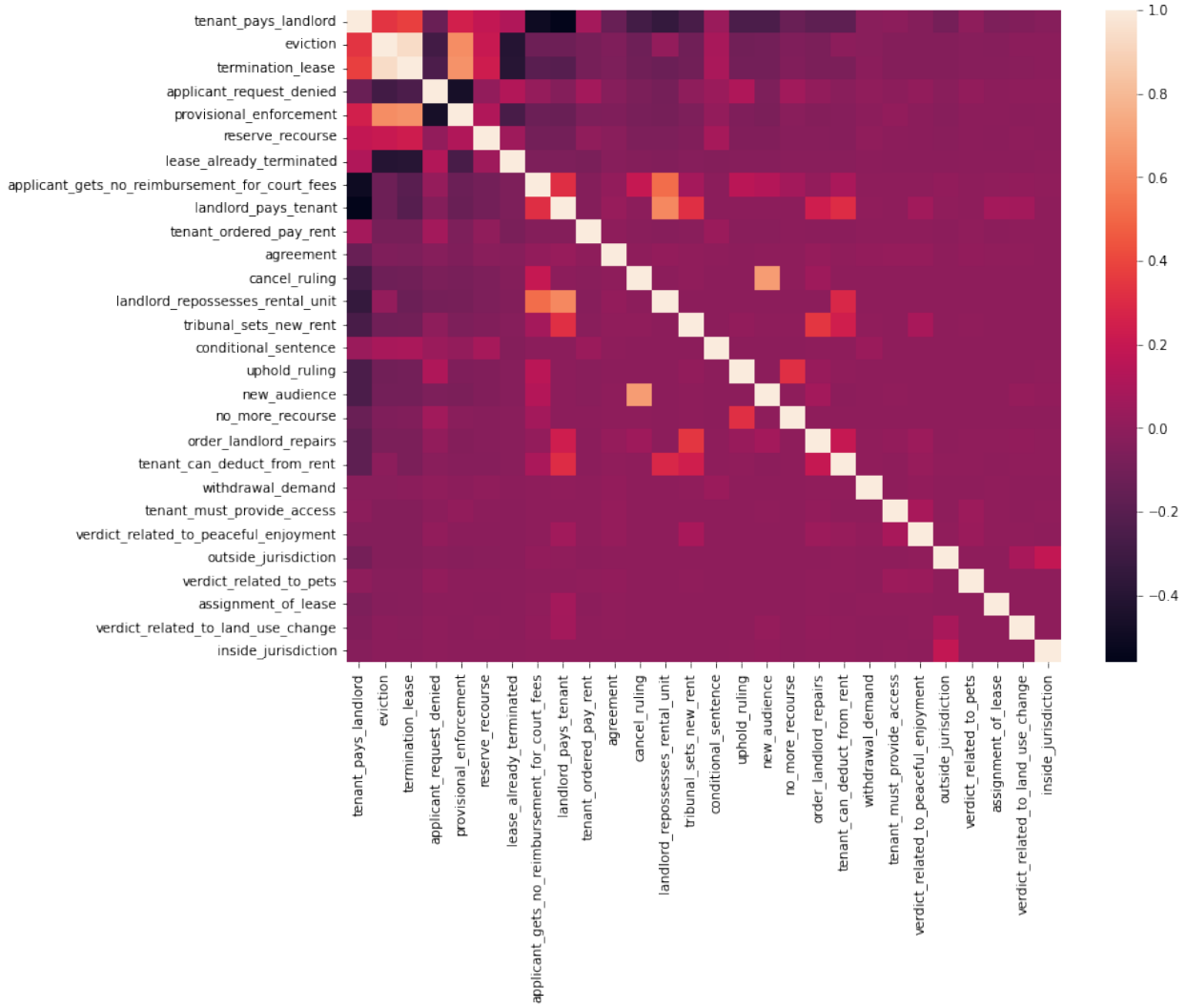

Fig. 2.6. Carte de chaleur représentant la matrice de corrélation entre les labels de verdict, triés sur chaque axe par ordre de fréquence décroissante.

Ainsi, les étiquettes tenant_pays_landlord, eviction, termination_lease et provisional_enforcement dans le coin supérieur gauche, sont une combinaison très courante de verdict dans laquelle le non-paiement de loyer est sanctionné par une sanction financière, une éviction du logement et une résiliation de bail, le tout étant effectif malgré un éventuel

appel du défendeur (la décision du juge est mise à exécution même si un recours contre elle a été fait en appel). Un autre groupement peut également être identifié avec les étiquettes landlord_repossesses_rental_unit, applicant_gets_no_reimbursement_for_court_fees et landlord_pays_tenant : lorsqu'un propriétaire obtient l'accord du tribunal pour reprendre son logement malgré la présence d'un locataire, il doit payer les frais judiciaires et une indemnité à l'occupant. Un dernier groupement peut enfin être mentionné avec le duo new_audience et cancel_ruling : une décision antérieure est annulée et une nouvelle audience est organisée pour rejuger un litige.

Le même exercice peut être effectué pour ce qui est des corrélations entre les étiquettes de verdict et les articles. La carte de chaleur illustrant la matrice de corrélation entre les deux à la Figure 2.7 montre ainsi que certaines étiquettes sont fortement associées à un ou plusieurs articles.

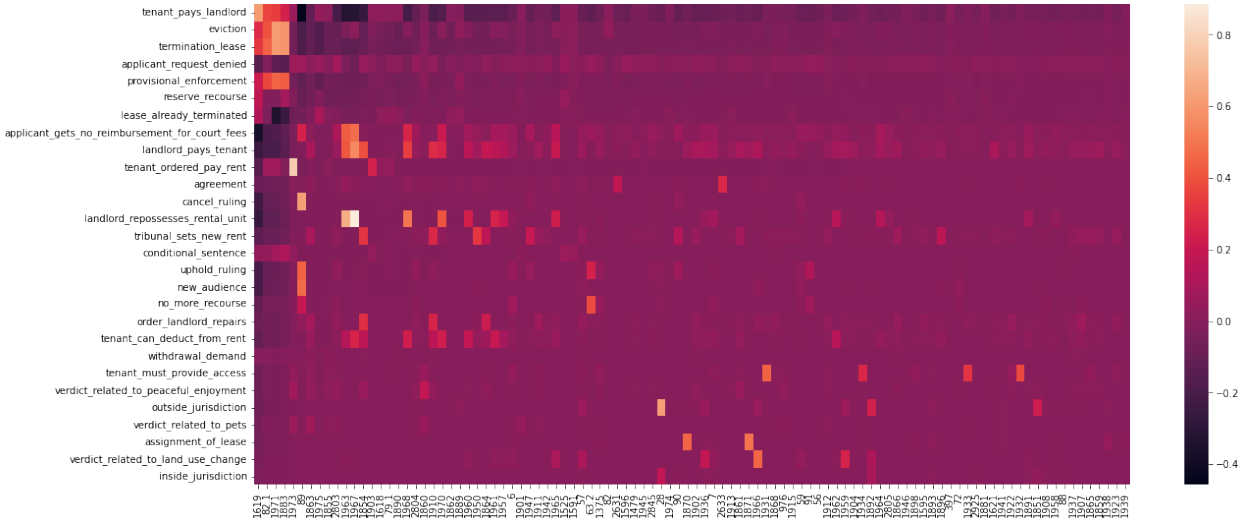

Fig. 2.7. Carte de chaleur représentant la matrice de corrélation entre les labels de verdict et les articles. Les éléments figurant sur les axes sont triés par ordre de fréquence décroissante.

Pour donner quelques exemples de combinaisons verdict-articles pertinents, nous pouvons relever quelques exemples en reprenant les groupements de labels précédents :

- les labels tenant_pays_landlord, eviction, termination_lease, provisional_enforcement avec les articles 1619, 82.1, 1971 et 1883. Ces articles définissent respectivement le

calcul des dommages-intérêts, les conditions de l'exécution provisoire[16] et enfin, les conditions permettant la résiliation du bail[17] ;

- les labels landlord_repossesses_rental_unit, applicant_gets_no_reimbursement_for_court_fees, landlord_pays_tenant avec les articles 1963 et 1967. Le premier dispose que le locateur peut, en cas de refus de quitter le logement du locataire, reprendre son bien avec l'accord du tribunal. Le second précise qu'en cas d'autorisation de reprise, le tribunal peut fixer l'indemnité que le locataire peut recevoir du locateur pour les frais de déménagement ;
- les labels new_audience, cancel_ruling avec l'article 89. Ce dernier est invoqué par un défendeur qui, n'ayant pas pu se rendre à une audience qui s'est déroulée en son absence, réclame une annulation du verdict et la tenue d'une nouvelle audience.

Il est utile de souligner que, contrairement à certaines corpora comme les litiges criminels chinois par [**Xiao et al., 2018, Xu et al., 2020**], il n'existe pas vraiment de correspondance un pour un entre un article et un label de verdict. Si cela avait été le cas, nous aurions pu nous attendre à une sorte de diagonale avec des corrélations élevées dans la matrice représentée à la Figure 2.7. Il existe ainsi certains cas de figure comme les labels agreement (accord entre les parties auquel le juge donne un caractère légal) ou conditional_sentence (des sanctions sont assorties d'un sursis) qui ne font pas l'objet d'une définition précise dans la loi et dont les modalités sont laissées à l'appréciation du magistrat.

2.5. Conclusion : Les aspects originaux à retenir de ce corpus

À l'issue des analyses précédentes, cette section récapitule les particularités des documents en notre possession.

Un langage technique: il comprend des documents en français (et non en anglais qui est la langue majoritaire en TAL, comme le déplore [**Søgaard, 2022**]) focalisés sur un domaine très spécialisé qui diffère grandement du langage courant/générique plus fréquent

[16]Autrement dit, est-ce que la décision est définie comme effective même si le défendeur fait appel.

[17]Au-delà de trois semaines de retard dans le paiement du loyer, le locataire peut voir son bail résilié, à moins qu'il ne parvienne à tout payer avant le jugement.

dans les corpora de TAL. Du fait de cette spécialisation, le vocabulaire et le style utilisés sont relativement homogènes. Ainsi, les juges emploient d'une part un jargon technique qui peut nécessiter une certaine compréhension du domaine et qui peut être peu compréhensible pour des citoyens ordinaires (non-experts ou *laymen/laypeople* en anglais), ce qui ouvre la question de comment rendre l'information juridique accessible. D'autre part, la présence d'expressions redondantes et plus ou moins longues (p. ex. enchaînement de plusieurs propositions dans la même phrase tout au long d'un même paragraphe) tend à renforcer l'homogénéité syntaxique des documents. Il faut aussi ajouter que le fait que les textes soient en français fait que nos expériences ne pourront pas bénéficier de certaines ressources ou outils mis au point pour les documents légaux en anglais, comme par exemple un outil de *semantic role labeling* par [**Ali et al., 2021**] pour des décisions indiennes en anglais.

Un domaine légal de niche: les corpora de TAL les plus utilisés en justice prédictive proviennent en grande partie de cours suprêmes ou de tribunaux spécialisés dans des litiges portant sur des enjeux majeurs confiés à des magistrats expérimentés. Par exemple, les jeux de données de [**Aletras et al., 2016**] et [**Chalkidis et al., 2019a**] s'attardent sur des décisions de la Cour européenne des droits de l'homme qui tranche des litiges entre des particuliers et des États concernant le respect des droits de l'homme (p. ex. est-ce que les conditions de détention d'un particulier dans un État donné sont conformes à la Convention européenne des droits de l'homme ?). Un autre exemple d'enjeu majeur est illustré par le corpus de [**Xiao et al., 2018**] qui concentre des affaires criminelles collectées par la Cour Suprême du Peuple en Chine. Pour ce qui est du corpus du Tribunal administratif du logement, les litiges opposant locateurs et locataires peuvent être décrits comme étant des "conflits de basse intensité" et un "contentieux de masse" par [**Benyekhlef and Zhu, 2018**]. Cela implique que ces litiges :

- posent des questions judiciaires d'une complexité relativement faible. Il s'agit davantage de juger des faits relativement simples au regard d'articles de loi préétablis (p. ex. est-ce que le propriétaire a bien rempli ses obligations vis-à-vis de son locataire), que de traiter de questions purement juridiques plus complexes (p. ex. est-ce que la décision rendue par telle juge est bien conforme à une procédure prédéfinie par telle loi). Ce corpus est ainsi une bonne occasion pour voir jusqu'à quel point un modèle d'apprentissage machine peut résoudre des questions juridiques assez simples ;

- impliquent des sommes en jeu (p. ex. pénalités, dommages-intérêts, loyer(s) non payé(s)) assez modestes comparés à d'autres types de contentieux. Certains litiges relatifs aux petites créances, au droit du logement ou de la consommation par exemple, pourraient ainsi être traités en dehors des tribunaux avec des arrangements à l'amiable, voire par l'intermédiaire de plateformes de "réglement en ligne des litiges" (en anglais, *online dispute resolution* ou ODR) décrits par [**Benyekhlef and Zhu, 2020**]. Ces résolutions de conflits sans juge ont l'avantage d'être moins coûteuses, plus accessibles et flexibles pour les parties prenantes. C'est pourquoi le corpus du Tribunal administratif du logement peut constituer un banc d'essai pertinent pour des techniques d'apprentissage machine qui pourraient un jour être amenées à être intégrées à des systèmes d'information, d'assistance ou de médiation juridique pour les justiciables. En effet, ces derniers peuvent se retrouver confrontés à des obstacles tels que le manque d'accès à un avocat (barrière économique) ou un système judiciaire sous-dimensionné pour un très large volume de contentieux pour le même type de problème ;
- ont déjà été traités à de nombreuses reprises, ce qui débouche sur un cadre judiciaire aux contours délimités et prévisibles. Cette remarque est appuyée par les redondances observées entre différentes décisions traitant d'affaires similaires entre elles[18]. La présence de ces redondances est importante d'un point de vue apprentissage machine, car elle implique aussi un déséquilibre et des biais dans les données qui auront des impacts sur les tâches en aval.

Maintenant que le corpus a fait l'objet d'une annotation pseudoautomatique avec un apport métier important, il sera vu dans le chapitre qui suit comment de telles données peuvent être utilisées dans le cadre d'une tâche d'apprentissage machine de justice prédictive.

[18]Au vu de plusieurs instances examinées manuellement, il faut aussi souligner la présence de quasi-copiés-collés, ce qui suggère que les tribunaux utilisent fort probablement des *templates* ou modèles à remplir pour les cas les plus simples à traiter. Cela est surtout vrai pour les cas de retards de loyer qui finissent avec le même verdict (expulsion, résiliation, dommages-intérêts à payer pour le locataire, exécution provisoire) et dont seuls les variables monétaires (montants de loyer) et temporelles (dates) varient.

Chapitre 3

Prédiction de verdict par l'intermédiaire d'une classification multilabel

3.1. Introduction

Une première tâche très commune en TAL appliqué au domaine juridique consiste à faire de la prédiction de verdict dans laquelle les modèles, sur la base de faits décrits en langage naturel, doivent donner les classes/labels correspondant au verdict. Ce type de tâche est communément appelé *legal judgment prediction*, appellation qui peut quelque peu induire en erreur. Comme l'ont souligné [**Medvedeva et al., 2023**], le terme "prédire" peut signifier "donner la catégorie à laquelle appartient une instance", ce qui est le sens le plus communément admis en apprentissage machine, mais il peut aussi avoir le sens de "faire une prédiction concernant un événement à venir". Or, dans le cadre d'une tâche de justice prédictive, faire de telles prédictions nécessite d'utiliser en entrée des documents disponibles avant même que l'audience n'ait lieu. C'est ainsi que [**Medvedeva et al., 2023**] font la distinction entre deux tâches :

- Classification de verdict (*Outcome-based judgement categorisation*) : le modèle doit donner la ou les étiquettes correspondant au verdict sur la base du texte de la décision du juge préalablement tronqué. En l'espèce, les données d'entrée ne conservent que les éléments factuels et excluent les indices pouvant révéler le verdict.
- Prédiction de verdict (*Outcome forecasting*) : le modèle a pour seules données d'entrée les documents disponibles avant même que la décision finale du juge ne soit rendue publique. L'existence de tels documents préalables à l'audience est possible (p. ex.

[**Medvedeva et al., 2021**] exploite des documents rendus publics avant procès à la Cour européenne des droits de l'homme), mais reste très rare.

En ce qui concerne notre tâche, elle s'inscrit dans le groupe majoritaire de la classification de verdict. Celle-ci est le plus souvent formalisée sous forme de classification binaire, comme l'ont fait [**Nallapati and Manning, 2008**], [**Aletras et al., 2016**], [**Liu and Chen, 2017**], [**Long et al., 2019**] ou [**Chalkidis et al., 2019a**]. Notre tâche a cependant ceci de particulier que le verdict de chaque litige est encodé via un ensemble de labels cumulables.[1] La préparation des données a déjà été décrite dans la section 2.3. L'enjeu ici consiste à savoir si un modèle, sur la base du texte des faits présentés par les justiciables, est capable de retourner le verdict formalisé sous la forme de label. Ce chapitre reprend et étend le travail effectué dans notre contribution [**Salaün et al., 2020**]. Il peut aussi être considéré comme un élargissement du travail de [**Westermann et al., 2019**] qui se contentaient de plus d'une centaine d'instances pour prédire si le juge allait octroyer ou non une réduction de loyer (classification binaire).

3.2. Modèles utilisés

3.2.1. Modèles non neuronaux

Plusieurs méthodes non-neuronales sont utilisées pour notre tâche ;

- **MostFreq** (*most frequent*) : une méthode naïve qui retourne systématiquement un label s'il est majoritaire dans l'ensemble d'entraînement. Ce modèle ne tient donc pas compte des données d'entrée ;
- **OvRLogReg** (*One-versus-rest logistic regression*) : pour chaque label possible, un prédicteur basé sur une régression logistique indique sa présence ou son absence pour chaque instance donnée. Comme les labels sont cumulables, chaque régresseur[2] prédit ainsi la présence ou l'absence de son label indépendamment des autres régresseurs. Autrement dit, on empile plusieurs classifieurs binaires pour donner une prédiction multilabel.

[1] Il sera vu dans la Sous-Section 3.4.3 que certaines étiquettes peuvent être incompatibles entre elles.

[2] Le terme "régresseur" désigne ici un modèle basé sur une régression logistique retournant une sortie binaire 0 ou 1.

- **OvRSVM** (*One-versus-rest Support Vector Machine*) : ce modèle est analogue à OvRLogReg à la différence près que les régressions logistiques sont remplacées par des machines à vecteur de support.

Pour les deux derniers modèles, le texte des faits donné en entrée fait l'objet d'une représentation TF-IDF (*term frequency-inverse document frequency*) [**Salton, 1983**]. Après avoir essayé différentes configurations de n-grams et de prétraitement, le choix s'est arrêté sur des n-grams au niveau des mots (séparés par des *whitespaces*) avec $n \in [1, 3]$ et un vocabulaire maximal de 100,000 jetons. Les *stopwords* français sont préalablement retirés sur la base d'une liste de NLTK [**Bird et al., 2009**]. Aussi, les textes en entrée sont mis en minuscules.

À ces modèles est ajouté un autre basé sur la recherche d'informations et mis en place grâce à Lucene[3] implémenté via ElasticSearch[4]. Ce modèle cherche, pour chaque instance d_{unseen} de l'ensemble de validation/test, une instance d'entraînement d_{train} la plus ressemblante sur la base du texte des faits qui fait l'objet d'une représentation sac-de-mots[5]. Les scores de correspondance entre les textes s'effectuent grâce à la fonction more_like_this. Trois variantes sont employées :

- **IRtop1** (*information retrieval based on top 1*) : les labels du document d_{train} avec le meilleur score sont retournés comme verdict pour le document d_{unseen} ;
- **IRtop10** (*information retrieval based on top 10*) : les labels qui apparaissent dans une majorité des 10 documents d_{train}, avec les meilleurs scores de similarité vis-à-vis de d_{unseen}, sont retournés comme verdict pour le document d_{unseen}. Cette majorité est calculé uniquement sur la base des top k documents obtenus et ne tient pas compte du score obtenu par chaque instance d_{train} retenue dans le top k ;
- **IRtop100** (*information retrieval based on top 100*) : comme le précédent, mais avec les 100 documents les plus similaires à d_{unseen}.

L'intérêt de cette approche par recherche d'information est qu'elle est une façon de formaliser l'approche juridique de la *common law* : un juge anglo-saxon va trancher un litige en fonction

[3] `https://lucene.apache.org/`

[4] `https://www.elastic.co/`

[5] La représentation sac-de-mots utilisée par les systèmes basés sur la recherche d'information est propre et interne à Lucene/ElasticSearch ; elle n'a aucun lien avec la représentation n-grams décrite pour les modèles non neuronaux dans le paragraphe précédent.

de la façon dont d'autres litiges similaires ont été tranchés auparavant. Dans le cas du droit du logement, l'approche juridique est mixte : elle est en théorie romano-civiliste (*civil law*) car les magistrats analysent les litiges à travers des lois déjà prédéfinies (c.-à-d. le Code civil du Québec) ; elle est en pratique imprégnée de la *common law* de par la culture anglo-saxonne au Canada (il faut aussi ajouter que les juges évitent de dévier de la jurisprudence existante pour éviter que leurs décisions ne soient remises en cause ultérieurement).

3.2.2. Modèles neuronaux

Depuis l'apparition du modèle transformeur de [**Vaswani et al., 2017**] et le succès de BERT par [**Devlin et al., 2019**] sur plusieurs tâches de *NLU* (*natural language understanding*) parmi lesquelles des classifications, plusieurs modèles ont été rendus publics avec des pré-entraînements non supervisés (généralement avec la tâche de *Masked Language Modeling*[6] utilisé par [**Devlin et al., 2019**] pour BERT) parfois basés sur une langue ou un domaine en particulier. Pour les besoins de notre tâche, nous retenons trois modèles disponibles pour le français :

- **FlauBERT** par [**Le et al., 2019**] : son corpus d'entraînement se compose principalement de textes encyclopédiques et de nouvelles (actualités) provenant de la tâche WMT 2019 [**Li et al., 2019**], de la collection OPUS [**Tiedemann, 2012**] et de Wikimedia ;
- **CamemBERT** par [**Martin et al., 2020**] : ce modèle a été entraîné sur la partie française du corpus OSCAR [**Suárez et al., 2019**] ;
- **JuriBERT** par [**Douka et al., 2021**] : le corpus d'entraînement est basé sur des données tirées de Légifrance[7] et de décisions de la Cour de cassation française.

Les deux premiers modèles ont la particularité d'être entraînés sur des corpora dits "génériques", c'est-à-dire non spécifiques à un domaine (p. ex. comprenant des articles encyclopédiques ou des nouvelles[8]). Le dernier, à l'inverse, a été pré-entraîné sur un corpus

[6]En français, "modélisation masquée du langage", tâche où 15% des jetons dans une séquence sont masqués et doivent être retrouvés par le modèle.

[7]`https://www.legifrance.gouv.fr`

[8]Pour un lecteur francophone vivant hors Québec, "nouvelles" désignent ici les actualités ou *news* en anglais.

spécifique au droit français, et peut être considéré comme l'équivalent francophone de Legal-BERT [**Chalkidis et al., 2020**] qui lui a été entraîné sur des textes légaux anglophones en provenance majoritairement des États-Unis, de l'Union européenne et du Royaume-Uni. Il sera ainsi intéressant de voir si les capacités d'un modèle de langage adapté pour le droit français peuvent se généraliser à une tâche dans la même langue, mais pour une juridiction québécoise. Par ailleurs, il est utile de préciser que ces trois modèles francophones ne font pas l'objet de pré-entraînement non supervisé supplémentaire (*unsupervised further pretraining*) avec la tâche de *masked language modeling*[9] ; les paramètres par défaut des différents *checkpoints* sont utilisés tels quels et directement affinés (*fine-tuned*) sur la tâche. Cela permet ainsi de voir si un corpus de pré-entraînement par rapport à d'autres se distingue lors de la tâche de classification.

Pour chacun de ces modèles, la taille du *batch* est de 20, le taux d'apprentissage à $1e^{-5}$, et le nombre maximum d'époques à 20. L'optimisation s'effectue avec Adam [**Kingma and Ba, 2015**] et la fonction de perte est l'entropie croisée binaire (*binary cross-entropy*) qui est adaptée pour la classification multilabel. Comme chaque instance peut cumuler plusieurs étiquettes, on peut considérer qu'il existe aussi autant de classifications binaires que de labels dans chacune des instances à prédire. C'est ainsi que pour obtenir la probabilité de chaque étiquette, une fonction sigmoïde[10] est appliquée à la sortie du modèle. Chaque label est prédit dès que sa probabilité dépasse un seuil t de 0.5.

Les modèles transformeur utilisés ici ont des tailles de séquences d'entrée maximales de 512 jetons. Bien que les documents juridiques sont en général très longs, notre extraction du seul texte des faits, décrite dans la sous-section 2.3.2, permet d'avoir une longueur moyenne de séquence d'entrée inférieure à 512 quel que soit le *tokenizer* choisi. Le Tableau 3.1 montre par ailleurs que le pourcentage d'instances dépassant le seuil de 512 jetons se situe entre 3.89 et 5.17% selon le vocabulaire de jetons utilisé. Ainsi, malgré la limite de longueur de *context input* pour chacun des différents modèles transformeur, la perte d'information reste limitée.

[9]Le *further pretraining* non supervisé sera utilisé dans le chapitre suivant, une fois avoir sélectionné un modèle à la BERT à l'issue du présent chapitre. Il est utile de rappeler que le pré-entraînement non supervisé est particulièrement long et coûteux à effectuer sur une carte graphique.

[10]Dans un souci de clarté, il faut préciser qu'une fonction softmax n'est pas utilisable pour notre tâche car elle implique qu'il existe une seule catégorie pour chaque instance.

Modèle lié au *tokenizer*	FlauBERT	CamemBERT	JuriBERT
Nombre de jetons dans le vocabulaire	68,729	32,005	32,000
Longueur moyenne	217.20	242.71	258.70
Écart-type	216.10	236.50	258.99
Longueur médiane	185	214	223
Pourcentage d'instances dépassant 512 jetons	3.98	4.65	5.17

Tableau 3.1. Statistiques concernant les séquences de texte données en entrée du modèle selon le *tokenizer* utilisé. Les longueurs et l'écart-type sont donnés en nombres de jetons.

À la fin de chaque époque, la performance du modèle est mesurée sur l'ensemble de validation avec l'**exactitude** (abrégée **EM** pour *exact match*) EM_{valid}, c'est-à-dire le ratio d'instances de validation pour lesquelles le modèle est capable de retourner l'ensemble exact des étiquettes correspondant au verdict. Cette métrique, qui est utilisée durant l'optimisation du modèle et durant l'évaluation sur l'ensemble de test, a été préférée par rapport au F1-micro et au F1-macro qui sont plus communément utilisés en classification multilabel. La raison est double : d'une part, l'exactitude est plus difficile à maximiser par rapport aux moyennes F1 (cela sera démontrée dans la sous-section dédiée aux résultats) ; d'autre part, elle permet de rendre compte de la capacité du modèle à donner le verdict exact pour chaque instance, ce qui est crucial d'un point de vue métier. Cette démarche vise aussi à éviter les conclusions triomphalistes tirées de scores considérés comme élevés en apprentissage machine, donnant alors la fausse impression que la tâche est résolue dans le domaine cible visé [**Bench-Capon, 2021**]. Les autres métriques sont discutés plus en détail dans la sous-section qui suit.

La patience est fixée à 5 époques ; ainsi, si EM_{valid} ne s'améliore pas après 5 époques consécutives, l'entraînement s'arrête, et le modèle dont l'état des paramètres a permis la meilleure performance EM^*_{valid} est utilisé pour effectuer des prédictions sur l'ensemble de test. Ce procédé est répété à 5 reprises pour chacun des modèles transformeur avec des

chiffres d'amorce (*seed* ou *random state*) distincts. Les scores sont ensuite moyennés et reportés avec leurs écarts-types dans le Tableau 3.3.

D'autres métriques en validation telles que Micro F1, Macro F1 (expliquées plus bas), et la perte (*loss*) ont été essayées à la place de EM_{valid} pour guider l'entraînement du modèle (procédé du *early stopping*, c.-à-d. interrompre l'entraînement selon la performance dans une métrique cible). Sans grande surprise, le choix d'une métrique cible m en validation entraîne un meilleur score dans cette métrique m sur l'ensemble de test. Cependant, l'amélioration apportée par m pour une autre métrique m' reste en deçà de celle apportée par un entraînement guidé par m'.

3.3. Métriques

Dans le cadre d'une classification multilabel, chaque instance peut cumuler plusieurs étiquettes, ce qui implique quelques différences dans les métriques utilisées par rapport à la classification multiclasse (chaque instance a une étiquette unique). Trois métriques sont utilisées ici : l'**exactitude**, **Micro F1** et **Macro F1**.

L'exactitude (*exact match*) indique le ratio d'instances pour lesquelles le modèle est capable de retourner l'ensemble exact de labels correspondant au verdict. Pour le score Micro F1, on compte pour l'ensemble de tous les labels le nombre de vrais positifs (TP), faux positifs (FP), vrais négatifs (TN), faux négatifs (FN) pour calculer la précision (P) et le rappel (R) globaux :

$$P = \frac{TP}{TP + FP} \tag{3.3.1}$$

$$R = \frac{TP}{TP + FN} \tag{3.3.2}$$

Le Micro F1 correspond ensuite à la moyenne harmonique entre ces deux valeurs :

$$F1 = 2 \times \frac{P \times R}{P + R} \tag{3.3.3}$$

Pour ce qui est du Macro F1, on calcule d'abord le score F1 spécifique à chaque label selon l'équation 3.3.3 avant d'effectuer une moyenne non pondérée entre tous les scores obtenus. La nuance entre le Micro F1 et le Macro F1 réside dans le fait que le premier évalue la capacité du modèle à prédire correctement le plus de labels possible sans distinction, tandis que le deuxième évalue la capacité du modèle à délivrer une bonne performance à travers les différents labels, chaque label ayant ainsi la même importance. En d'autres termes, un

Modèle	Micro F1	Macro F1	Exactitude
MostFreq	61.78 (0.00)	8.59 (0.00)	5.62 (0.00)
OvRLogReg	83.98 (0.00)	46.22 (0.00)	40.59 (0.00)
OvRSVM	82.28 (0.00)	32.65 (0.00)	37.04 (0.00)
IRtop1	61.92 (0.00)	31.86 (0.00)	14.57 (0.00)
IRtop10	71.22 (0.00)	34.08 (0.00)	25.40 (0.00)
IRtop100	73.80 (0.00)	31.97 (0.00)	27.20 (0.00)
FlauBERT	**88.69 (0.26)**	**60.83 (1.04)**	49.75 (0.86)
CamemBERT	88.68 (0.15)	54.43 (2.66)	**50.67 (0.23)**
JuriBERT	88.21 (0.15)	56.46 (2.62)	48.91 (0.79)

Tableau 3.2. Performance des modèles dans la tâche de classification multilabel sur l'ensemble de test (scores sur 100, écarts-types entre parenthèses).

modèle capable de prédire correctement un label très répandu dans le corpus aura un Micro F1 plus élevé grâce à ce label. Il sera cependant sanctionné par le Macro F1 s'il n'est pas capable de prédire correctement l'ensemble des labels disponibles. L'exactitude quant à elle est beaucoup plus stricte en ceci que pour une instance donnée, une classification sera considérée comme mauvaise si la prédiction contient une étiquette en plus ou en moins par rapport à la référence. L'exactitude tient ainsi compte de la capacité du modèle à délivrer de bonnes prédictions vis-à-vis de chaque instance, ce qui est crucial d'un point de vue métier pour le domaine juridique.

3.4. Résultats et discussion

3.4.1. Comparaison entre les différentes approches non neuronales, par recherche d'information et par BERT

Les résultats des modèles non neuronaux sont présentés dans le Tableau 3.3. Le premier modèle "naïf" MostFreq montre une différence de plus de 50 points entre le score Micro F1 d'une part, et le Macro F1 et l'exactitude d'autre part. Ce décalage illustre l'importance

Modèle	Micro F1	Macro F1	Exactitude
MostFreq	61.78 (0.00)	8.59 (0.00)	5.62 (0.00)
OvRLogReg	83.98 (0.00)	46.22 (0.00)	40.59 (0.00)
OvRSVM	82.28 (0.00)	32.65 (0.00)	37.04 (0.00)
IRtop1	61.92 (0.00)	31.86 (0.00)	14.57 (0.00)
IRtop10	71.22 (0.00)	34.08 (0.00)	25.40 (0.00)
IRtop100	73.80 (0.00)	31.97 (0.00)	27.20 (0.00)
FlauBERT	**88.69 (0.26)**	**60.83 (1.04)**	49.75 (0.86)
CamemBERT	88.68 (0.15)	54.43 (2.66)	**50.67 (0.23)**
JuriBERT	88.21 (0.15)	56.46 (2.62)	48.91 (0.79)

Tableau 3.3. Performance des modèles dans la tâche de classification multilabel sur l'ensemble de test (scores sur 100, écarts-types entre parenthèses).

Modèle	Micro F1	Macro F1	Exactitude
MostFreq	61.8	8.6	5.6
OvRLogReg	84.0	46.2	40.6
OvRSVM	82.3	32.7	37.0
IRtop1	61.9	31.9	14.6
IRtop10	71.2	34.1	25.4
IRtop100	73.8	32.0	27.2
FlauBERT	**88.7**	**60.8**	49.8
CamemBERT	88.7	54.4	**50.7**
JuriBERT	88.2	56.5	48.9

de prendre en compte les biais des données ; comme indiqué dans la Section 2.3.1, seuls trois labels recouvrent plus de la moitié du corpus et sont retournés systématiquement par MostFreq pour n'importe quelle instance. Cela montre que dans une classification multilabel

utilisant un corpus avec une distribution très piquée[11] dans ses étiquettes, le Micro F1 est une métrique relativement aisée à maximiser lorsque peu de labels sont majoritaires, tandis que le Macro F1 et l'exactitude permettent d'avoir une perspective plus fine sur la véritable capacité des modèles.

Pour ce qui est des autres modèles non neuronaux, il faut noter que les classifieurs OvR-LogReg et OvRSVM (qui, pour rappel, sont des regroupements de classifieurs binaires qui sont chacun spécifique à un label), font mieux que IRtop1 , IRtop10 et IRtop100 pour toutes métriques confondues. Cette sous-performance des systèmes basés sur la recherche d'information peut laisser supposer que chercher des précédents juridiques sur la seule base du texte des faits n'est pas suffisant pour donner des prédictions correctes. Cela peut aussi suggérer que la représentation sac-de-mots de Lucene désavantage les modèles IRtop* par rapport aux autres approches. Le fait d'utiliser l'*analyzer*[12] pour le français ou celui par défaut ne débouche pas sur des différences particulières en termes de performances[13]. Les meilleures performances sont accomplies par les modèles transformeur parmi lesquels FlauBERT et CamemBERT. Le premier obtient la meilleure performance en termes de Micro F1 et Macro F1 tandis que le deuxième obtient la meilleure exactitude. JuriBERT, bien qu'entraîné sur des données juridiques issues des institutions françaises, a des performances en deçà des deux autres modèles francophones. Cela suggère que JuriBERT peut être trop spécifique au droit français pour pouvoir être employé pour une juridiction d'un autre pays, fût-elle francophone. Il se peut aussi que FlauBERT et CamemBERT aient une capacité supérieure due à des corpora de pré-entraînement plus variés et plus volumineux que celui de JuriBERT. En dehors de ces comparaisons, il faut aussi souligner que le niveau des scores en exactitude laisse présager le degré de difficulté de la tâche : les meilleurs modèles parviennent à retrouver les labels de référence pour à peine la moitié des instances.

[11] En anglais, *highly skewed distribution*.

[12] Les *analyzers* sont des modules dans ElasticSearch opérant un prétraitement sur les textes (p. ex. liste de mots-outils, tokénization, *stemming*), ce qui va avoir une influence sur la façon dont les documents seront indexés.

[13] Les modèles IRtop* dont les résultats sont rapportés dans cette section utilisent l'*analyzer* pour le français.

IRtop100	OvRLogReg	CamemBERT	Fréquence relative	Combinaison unique d'étiquettes
53.57	71.55	**84.18**	22.9	(a) tenant_pays_landlord ; eviction ; termination_lease ; provisional_enforcement
48.29	**62.58**	50.26	9.0	(b) tenant_pays_landlord
19.07	26.54	**41.61**	8.0	(c) tenant_pays_landlord ; eviction ; termination_lease ; provisional_enforcement ; reserve_recourse
5.47	30.71	**32.71**	7.4	(d) tenant_pays_landlord ; lease_already_terminated
21.27	**28.96**	25.89	5.6	(e) tenant_pays_landlord ; eviction ; termination_lease
28.88	50.43	**65.2**	5.5	(f) tenant_pays_landlord ; applicant_request_denied
6.81	17.19	**64.79**	3.4	(g) tenant_pays_landlord ; eviction ; termination_lease ; applicant_request_denied ; reserve_recourse
16.91	50.16	**71.59**	3.0	(h) tenant_pays_landlord ; applicant_request_denied ; lease_already_terminated
33.73	24.77	**76.56**	2.8	(i) tenant_pays_landlord ; eviction ; termination_lease ; applicant_request_denied
1.86	**5.82**	3.62	2.2	(j) tenant_pays_landlord ; reserve_recourse ; lease_already_terminated
4.18	21.57	**41.31**	2.1	(k) tenant_pays_landlord ; applicant_request_denied ; reserve_recourse ; lease_already_terminated
17.42	**34.54**	11.14	1.9	(l) tenant_pays_landlord ; eviction ; termination_lease ; reserve_recourse

Tableau 3.4. Performance en exactitude de chaque modèle (sur 100) pour chaque combinaison unique de labels pour les litiges de type **Locateur c. Locataire**. Les combinaisons uniques de labels sont présentées en ordre décroissant de fréquence relative (le ratio qu'occupe chaque combinaison dans l'ensemble de test).

Prenons à présent un modèle de chacun des trois groupes et voyons en quoi leurs performances diffèrent : OvRLogReg pour les modèles non-neuronaux, IRtop100 pour l'approche par recherche d'information et CamemBERT pour les modèles BERT. Il a été vu dans le Tableau 3.3 que le modèle BERT faisait de meilleures prédictions que OvRLogReg qui lui-même était supérieur à IRtop100. Pour tous les modèles confondus, les litiges de type "Locataire

IRtop100	OvRLogReg	CamemBERT	Fréquence relative	Combinaison unique d'étiquettes
57.90	58.64	**68.93**	1.6	(a) applicant_request_denied
28.14	**29.71**	27.09	0.8	(b) applicant_request_denied ; applicant_gets_no_reimbursement_for_court_fees
0.14	16.95	**44.21**	0.8	(c) applicant_request_denied ; uphold_ruling
13.62	**35.05**	**35.05**	0.7	(d) landlord_pays_tenant
46.90	**53.59**	53.27	0.7	(e) applicant_request_denied ; landlord_pays_tenant
9.67	32.32	**66.02**	0.4	(f) cancel_ruling ; new_audience
0.00	6.31	**31.08**	0.2	(g) applicant_gets_no_reimbursement_for_court_fees ; cancel_ruling ; new_audience
0.92	1.84	**15.67**	0.2	(h) applicant_request_denied ; applicant_gets_no_reimbursement_for_court_fees ; uphold_ruling
0.00	1.89	**4.4**	0.2	(i) applicant_request_denied ; uphold_ruling ; no_more_recourse
1.34	8.05	**11.41**	0.2	(j) applicant_request_denied ; landlord_pays_tenant ; tribunal_sets_new_rent
25.17	**36.05**	34.69	0.2	(k) cancel_ruling
0.0	**1.69**	0	0.1	(l) applicant_request_denied ; no_more_recourse

Tableau 3.5. Performance en exactitude de chaque modèle (sur 100) pour chaque combinaison unique de labels pour les litiges de type **Locataire c. Locateur**. Les combinaisons uniques de labels sont présentées en ordre décroissant de fréquence relative (ratio qu'occupe chaque combinaison dans l'ensemble de test).

c. Locateur" restent plus difficiles à prédire que les litiges "Locateur c. Locataire".[14] Afin

[14]Une longue réflexion a eu lieu concernant le choix entre créer un modèle unique et créer deux modèles distincts pour les deux types de litiges. La deuxième option aurait non seulement alourdi et complexifié l'analyse des résultats (avec deux jeux distincts d'étiquettes cibles), mais elle aurait aussi posé la question de créer un modèle distinct pour chaque cas de figure posant problème à un modèle unique. Pour ce qui est de l'approche avec un modèle unique, nous nous sommes posé la question de donner en entrée un signal au modèle pour indiquer le type de litige à traiter, ce qui aurait pu fournir une aide, voire une anti-sèche n'ayant pas lieu d'être. Nous avons finalement préféré laisser un modèle unique se débrouiller pour définir lui-même le type de litige et décider du verdict qui s'en suit.

de pouvoir procéder à une analyse encore plus fine des performances des trois modèles, nous avons considéré que chaque instance avait une classe ou catégorie unique formée par l'ensemble des labels attribués. Avec 28 labels cumulables, il existe en théorie 2^{28} combinaisons possibles, soit plus de 268 millions de possibilités. En pratique, 728 combinaisons uniques ont été décelées. Les 12 premières combinaisons pour les litiges initiés par les propriétaires et les locataires sont respectivement montrées dans les Tableaux 3.4 et 3.5. Pour les décisions de type "Locateur c. Locataire", le verdict majoritaire correspond à la combinaison (a) du Tableau 3.4 avec 4 labels, tandis que les combinaisons suivantes sont la plupart du temps des variantes de (a) avec un label en plus ou en moins (ils ont tous en commun d'inclure une condamnation du locataire à faire un paiement pour le propriétaire). C'est ainsi qu'une combinaison distincte de (a) cumulant 4 ou 5 labels va susciter de grands écarts de performance entre les modèles de par le nombre d'étiquettes à prédire en même temps. C'est ce qui est observable avec les combinaisons (c), (g), (i), (k). Par ailleurs, la présence d'un label apparaissant de façon sporadique tel que reserve_recourse dans (c) ou (g) complique aussi la tâche de prédiction de verdict.

Pour les décisions de type "Locataire c. Locateur" dont les combinaisons de labels sont présentées dans le Tableau 3.5, le label applicant_request_denied est présent dans 8 des 12 premières combinaisons, ce qui implique un refus total (combinaison (a)) ou partiel des demandes du locataire (la combinaison (j) implique qu'une partie des demandes du locataire a été refusée et que le propriétaire a été quand même condamné à lui payer des indemnités). Les combinaisons sont aussi beaucoup plus variées que celles du Tableau 3.4 qui se concentrent surtout autour des évictions et des résiliations de bail. Les décisions initiées par des locataires-demandeurs concernent ainsi des demandes d'annulation d'une décision antérieure (combinaisons (c), (f), (g), (h), (i), (k)), de pénalités contre le propriétaire ((d), (e)) ou de fixation de loyer ((j)).

À la suite d'un examen manuel de quelques instances, nous proposons des hypothèses concernant les disparités de performance entre les modèles :

- IRtop100 se base sur un "vote majoritaire" des 100 premières décisions dont le texte des faits ressemble le plus à celui de l'instance en test. En d'autres termes, pour que IRtop100 parvienne à donner le verdict exact, il faudrait qu'au moins 51 décisions tirées de l'ensemble d'entraînement possèdent les labels correspondant au verdict

attendu. Une telle condition peut être délicate à remplir lorsque la combinaison attendue contient un grand nombre d'étiquettes ou lorsque les étiquettes attendues sont déjà en très faible nombre dans l'ensemble du corpus (pour rappel, un label a une fréquence minimale de 100 pour la totalité de notre corpus). Il faut aussi noter que le calcul de similarité entre le document en test et les documents en entraînement s'opère au niveau du seul texte des faits, ce qui semble insuffisant ici pour trouver de bon "matchs" ;

- OvRLogReg a un avantage sur IRtop100 en ceci qu'il est un empilement de classifieurs binaires complètement indépendants les uns des autres. Cela signifie aussi qu'un label sera retourné ou non, sans tenir compte des prédictions effectuées pour les composantes possibles du verdict. C'est ainsi qu'on a vu des prédictions de OvRLogReg avec aucune étiquette retournée ou avec des étiquettes manquantes (p. ex. une décision avec des pénalités financières contre le propriétaire dans laquelle OvRLogReg omet de préciser que ce dernier doit aussi effectuer des travaux).

La suite de l'analyse se focalise sur l'approche transformeur.

3.4.2. Analyse quantitative : des disparités selon le demandeur à l'origine du litige et la fréquence des cibles

Il est utile de souligner que les scores figurant sur le Tableau 3.3 sont des moyennes pour l'ensemble des litiges, dissimulant ainsi d'importantes disparités qui figurent dans le Tableau 3.6. Les performances restent en effet bien meilleures pour les litiges dans lesquels le propriétaire attaque en justice le locataire que lorsque le locataire attaque le propriétaire (pour rappel, ce dernier cas de figure correspond à 6% du corpus). Ces disparités font écho aux biais, précédemment décrits dans la Section 2.4, qui ont pour conséquence que les labels de verdicts habituellement associés aux cas minoritaires (en l'occurrence les litiges initiés par les locataires) font l'objet de prédictions moins bonnes. Ainsi, avec les modèles à la transformeur, le score F1 de l'étiquette tenant_pays_landlord (attribuée à 79,427 instances de test parmi les cas majoritaires "Locateur c. Locataire) est en moyenne de 98.90% tandis que celui pour landlord_pays_tenant (attribuée à 2,098 instances de test parmi les cas minoritaires "Locataire c. Locateur") s'élève à une moyenne de 76.23%, soit une différence de 22.67 points. Cette différence de performance s'observe aussi avec la corrélation entre les scores

	Locateur c. Locataire			Locataire c. Locateur		
Modèle	Micro F1	Macro F1	Exactitude	Micro F1	Macro F1	Exactitude
FlauBERT	**89.65**	**56.19**	50.87	68.05	**33.40**	36.63
CamemBERT	89.64	51.44	**51.83**	**68.17**	28.75	**36.96**
JuriBERT	89.25	53.29	50.17	65.89	29.01	34.07

Tableau 3.6. Performance des modèles dans la tâche de classification multilabel sur l'ensemble de test selon le demandeur et le défendeur (scores sur 100). Chaque modèle a été entraîné à cinq reprises avec différents chiffres d'amorce.

F1 de chaque label et leurs fréquences relatives qui monte à une moyenne de 0.62 pour les modèles à la BERT. En d'autres termes, plus un label est fréquent, plus son score F1 est susceptible d'être élevé.

3.4.3. Analyse qualitative : des incomplétudes, contradictions et exagérations dans les verdicts prédits

Au-delà du fait que les modèles tendent à donner de meilleures prédictions pour les cas de figure les plus fréquents, il faut aussi souligner que les verdicts prédits manquent de nuance et d'exactitude par rapport aux cibles attendues. L'analyse et les exemples ci-dessous sont tirés de prédictions par un modèle CamemBERT (ce modèle a la meilleure exactitude, ce qui est crucial dans le domaine juridique).

Il faut tout d'abord constater que certains verdicts peuvent être avantageux ou désavantageux pour chaque partie selon les personnes visées. Dans plusieurs décisions visant un locataire mauvais payeur, comme l'exemple (a) dans le Tableau 3.7, le juge décide seulement d'ordonner au locataire de payer le loyer à temps (tenant_ordered_pay_rent) avec des pénalités financières tandis que CamemBERT va prédire une expulsion (eviction) et une résiliation du bail (termination_lease). Le modèle tend à surgénéraliser cette prédiction pour l'ensemble des litiges dans lesquels un locateur attaque un locataire en justice.

Ce cas de figure où le locataire se retrouve désavantagé par le modèle de langue s'observe aussi lorsque le locataire est le demandeur avec l'exemple (b). Dans ce cas-ci, le juge a estimé que le propriétaire devait verser un dédommagement (landlord_pays_tenant) et effectuer des

	Cibles attendues (juge)	Labels prédits (modèle)
(a)	tenant_pays_landlord tenant_ordered_pay_rent	tenant_pays_landlord eviction termination_lease applicant_request_denied
(b)	landlord_pays_tenant tribunal_sets_new_rent order_landlord_repairs	applicant_request_denied
(c)	tenant_pays_landlord eviction termination_lease	tenant_pays_landlord eviction termination_lease applicant_request_denied
(d)	applicant_request_denied landlord_pays_tenant tribunal_sets_new_rent	applicant_request_denied landlord_pays_tenant

Tableau 3.7. Comparaison entre labels cibles et labels prédits pour différentes instances de l'ensemble de test avec CamemBERT.

travaux (order_landlord_repairs). De plus, le loyer a été ajusté à la baisse en faveur du locataire en guise de dédommagement (tribunal_sets_new_rent). De son côté, CamemBERT prédit uniquement un rejet des demandes du locataire (applicant_request_denied).

Les exemples (a) et (b) correspondent à des situations dans lesquelles le modèle donne la prédiction majoritaire pour les cas de figure "Locateur contre Locataire" (c.-à-d. l'expulsion du locataire avec annulation du contrat de bail de façon systématique) et "Locataire contre Locateur" (c.-à-d. que le locataire voit ses demandes refusées de façon systématique). En dehors de ces deux situations, il a aussi été observé que le modèle peut manquer de finesse dans ces prédictions dans la mesure où il retourne un label en trop ou en moins par rapport aux cibles attendues. L'exemple (c) est une illustration de ce manque de finesse où CamemBERT prédit un refus partiel de la demande du locateur (applicant_request_denied) de façon erronée. De la même façon, l'exemple (d) donne une instance pour laquelle le modèle omet de prédire une fixation de loyer par le juge (tribunal_sets_new_rent).

(tenant_pays_landlord ; landlord_pays_tenant)
(tenant_pays_landlord ; landlord_repossesses_rental_unit)
(tenant_pays_landlord ; order_landlord_repairs)
(tenant_pays_landlord ; verdict_related_to_land_use_change)
(termination_lease ; verdict_related_to_land_use_change)
(tenant_ordered_pay_rent ; order_landlord_repairs)
(tenant_ordered_pay_rent ; assignment_of_lease)
(tenant_ordered_pay_rent ; verdict_related_to_land_use_change)

Tableau 3.8. Paires de labels impossibles. Exemple pour la première ligne : un juge ne prononce pas de pénalités à la fois à l'encontre du propriétaire et du locataire dans le même verdict. La sixième donne un autre verdict impossible avec un ordre au locataire de payer le loyer à temps et un ordre au propriétaire d'entreprendre des réparations.

Ces divers exemples montrent à quel point il est difficile pour un modèle, fût-il *transformer-based*, de trouver l'ensemble exact de labels pour chacune des instances. Jusqu'ici, les manquements de ces modèles sont surtout mesurés avec des métriques utilisées de façon générique pour les tâches de classification multilabel, à savoir l'exactitude (*exact_match*) et le score (Micro/Macro) F1. Nous avons voulu ajouter deux autres métriques qui puissent mieux rendre compte des performances des modèles pour le cadre juridique.

La première métrique est un **score de contradiction** : il donne le ratio d'instances en test pour lesquelles le modèle donne des labels qui ne devraient pas coexister. Dans un premier temps, nous avons procédé à un examen manuel des paires de labels cibles qui ne coexistent jamais au niveau de toutes les instances de notre corpus. Ce sont finalement huit paires de labels impossibles (ne pouvant coexister dans le même verdict) qui ont été retenues et qui sont présentées dans le Tableau 3.8.

La deuxième métrique est un **score d'exagération** : il mesure le ratio d'instances en test pour lesquelles le modèle ne parvient à trouver les labels cibles en se contentant de retourner une combinaison de labels majoritaires. Dans le cas d'un jugement "Locateur contre Locataire", un verdict exagéré correspond à une prédiction reprenant la combinaison (a) de labels du Tableau 3.4 alors que le verdict attendu est différent. Par exemple : le

juge a donné l'ordre au locataire de payer le loyer à temps sans prononcer d'expulsion tandis que le modèle prédit les quatre labels signifiant des pénalités financières contre le locataire, l'expulsion de ce dernier, l'annulation du contrat de bail ainsi que l'exécution provisoire de la décision. Dans le cas d'un jugement "Locataire contre Locateur", une prédiction exagérée correspond à la combinaison (a) du Tableau 3.5 qui correspond à la prédiction (b) du Tableau 3.7 : le locataire-demandeur se voit opposer un simple refus par le modèle alors que le verdict du juge est différent.

Les mesures de contradictions et d'exagération sont données dans le Tableau 3.9. Le score (ratio en pourcentage) de contradiction ne permet pas ici de départager clairement les modèles contrairement au score d'exagération. En effet, les prédictions de verdict contenant des combinaisons impossibles de labels sont rares au point qu'une colonne avec le nombre absolu de verdicts contradictoires[15] a été ajoutée. De façon générale, pour toutes les instances de test, il faut remarquer que FlauBERT, bien qu'ayant les meilleurs scores en termes de Micro et Macro F1, obtient le pire score d'exagération à 8.16%. De son côté, CamemBERT, qui avait la meilleure exactitude, obtient le plus bas score en exagération (7.72%) et le nombre absolu le plus faible de contradictions (0.8%). Pour ce qui est de cette dernière métrique, elle contrebalance le bilan des métriques traditionnellement utilisées en classification multilabel (Micro F1, Macro F1, exactitude) en donnant une estimation plus fine des performances dans un cadre juridique : les modèles tendent à donner des verdicts majoritaires dans plus de 7% des instances de test alors qu'ils ne correspondent pas au verdict attendu. Le même Tableau 3.9 présente les mêmes métriques par type de litige. Que ce soit pour les litiges de type "Locateur contre Locataire" ou "Locataire contre Locateur", CamemBERT se démarque encore une fois par les scores d'exagération les plus bas (respectivement 7.17% et 14.27%). Il en va de même pour ce qui est des métriques pénalisant les verdicts contradictoires.

De façon général, avec l'ensemble des instances de test et sur la seule base des métriques classiques de classification multilabel (Micro/Macro F1, exactitude), FlauBERT peut donner l'illusion de délivrer la meilleure performance dans cette tâche avec les meilleurs scores F1. L'ajout des deux métriques de contradiction et d'exagération donne cependant une analyse plus nuancée et davantage en phase avec le domaine juridique : FlauBERT est ainsi le

[15]Une instance contenant une ou plusieurs combinaisons de labels impossibles est comptabilisée comme une seule contradiction.

Modèle	Micro F1 ↑	Macro F1 ↑	Exacti. ↑	Contrad. (nombre absolu) ↓	Contrad. (%) ↓	Exag. ↓
Toutes les décisions						
FlauBERT	**88.69**	**60.83**	49.75	1.8	0.1 >	8.16
	(0.26)	**(1.04)**	(0.86)	(0.75)	(0.00)	(1.00)
CamemBERT	88.68	54.43	**50.67**	**0.8**	0.1 >	**7.72**
	(0.15)	(2.66)	**(0.23)**	**(1.17)**	(0.00)	**(0.79)**
JuriBERT	88.21	56.46	48.91	3.0	0.1 >	7.99
	(0.15)	(2.62)	(0.79)	(3.69)	(0.00)	(0.74)
Locateur contre Locataire						
FlauBERT	**89.65**	**56.19**	50.87	1.2	0.1 >	7.53
	(0.25)	**(0.98)**	(0.93)	(0.40)	(0.00)	(0.90)
CamemBERT	89.64	51.44	**51.83**	**0.6**	0.1 >	**7.17**
	(0.17)	(1.83)	**(0.28)**	**(0.80)**	(0.00)	**(0.75)**
JuriBERT	89.25	53.29	50.17	2.6	0.1 >	7.33
	(0.13)	(1.81)	(0.77)	(3.77)	(0.00)	(0.61)
Locataire contre Locateur						
FlauBERT	68.05	**33.40**	36.63	0.6	0.01	15.63
	(0.44)	**(1.13)**	(1.14)	(0.80)	(0.01)	(3.49)
CamemBERT	**68.17**	28.75	**36.96**	**0.2**	**0.1 >**	**14.27**
	(0.42)	(2.41)	**(0.7)**	**(0.40)**	**(0.01)**	**(3.88)**
JuriBERT	65.89	29.01	34.07	0.4	0.01	15.68
	(0.62)	(3.20)	(1.58)	(0.49)	(0.01)	(4.09)

Tableau 3.9. Performance des modèles transformeur dans la tâche de classification multi-label sur l'ensemble de test, avec sous-scores selon que les décisions sont de type "Locateur contre Locataire" ou "Locataire contre Locateur". Sauf mention contraire, les scores sont sur 100. Les scores correspondent à des moyennes sur cinq *runs* avec les écarts-types entre parenthèses. Dans chaque encadré, le meilleur score de chaque colonne est en gras.

modèle qui donne le plus de "verdicts exagérés" (c.-à-d. que la prédiction correspond à une combinaison surreprésentée de labels alors que les labels cibles attendus sont autres). Pour ce qui est de CamemBERT, il faut souligner qu'il obtient les meilleurs scores en matière d'exactitude, de contradictions et d'exagération. Cela suggère qu'une exactitude élevée va de pair avec un score d'exagération en baisse. L'exactitude pourrait ainsi être considérée comme une métrique "classique"[16] qui serait la plus à même de refléter les performances d'un modèle dans un contexte de justice prédictive.

3.5. Tentatives d'amélioration de la performance lors de l'affinage

Dans la sous-section 3.2.2 décrivant les détails de l'affinage pour les modèles à la BERT, un seuil $t = 0.5$ a été défini pour retourner un label donné : dès que la probabilité d'une étiquette dépasse t, l'étiquette est incluse dans la prédiction. Nous avons repris CamemBERT et avons fixé le seuil t à différentes valeurs pour voir ses potentiels effets sur la performance : 0.25, 0.75 et 0.8.

Au vu du fort déséquilibre entre les proportions de décisions de type "Locateur c. Locataire" et "Locataire c. Locateur", nous avons aussi effectué un autre affinage où les deux types de litiges sont à proportions égales dans l'ensemble d'entraînement d'une part, et dans les ensembles d'entraînement et de validation d'autre part, à la manière de [**Aletras et al., 2016**]. Les nombres d'instances dans les ensembles d'entraînement et de validation passent ainsi de 371,989 et 101,757 à respectivement 45,228 et 13,006 (c'est-à-dire le nombre de litiges initiés par les locataires multiplié par 2). Les instances correspondant à des litiges initiées par des locateurs sont tirées au hasard.

Enfin, il est important de soulever deux difficultés de la classification multilabel où chaque instance peut cumuler plusieurs étiquettes :

- d'une part, le nombre de labels possibles et le déséquilibre entre eux peuvent être tels que le modèle soit amené à prédire des étiquettes obéissant à une distribution à longue traîne (*long-tail distribution*). En d'autres termes : d'une part, un sous-ensemble des labels forme un petit nombre de combinaisons d'étiquettes attribuées

[16] C'est-à-dire communément utilisée en classification multilabel dans n'importe quel domaine.

à une majorité d'instances ; d'autre part, les autres labels constituent des combinaisons plus nombreuses et plus diversifiées, mais représentées chacune par bien moins d'instances. Cela a été illustré pour notre corpus à 28 labels avec les Tableaux 3.4 et 3.5 ;

- d'autre part, les labels majoritaires peuvent apparaître dans le verdict de certaines instances avec d'autres labels plus rares. Il y a ainsi un biais qui peut amener des modèles à "sur-prédire" certaines étiquettes majoritaires et à omettre celles plus rares, comme cela a été vu dans la section précédente.

Ces deux obstacles, nommément le déséquilibre entre étiquettes (*labels imbalance*) et la concomitance entre labels (*labels co-occurrence*), peuvent mettre en difficulté la fonction de perte (*loss function*) communément utilisée en classification multilabel, soit l'entropie croisée binaire (*binary cross-entropy*). Elle est définie dans l'Équation 3.5.1 ci-dessous :

$$L_{bce} = -[(y_i^k) \times \log(p_i^k) + (1 - y_i^k) \times \log(1 - p_i^k)]$$

$$L_{bce} = \begin{cases} -\log(p_i^k) & \text{si } y_i^k = 1 \\ -\log(1 - p_i^k) & \text{sinon} \end{cases} \tag{3.5.1}$$

avec :

- $k = 0, 1, \ldots, n-1$ l'indice de l'instance parmi n instances ;
- $i = 0, 1, \ldots, 27$ l'indice du label parmi 28 étiquettes cumulables ;
- $y_i^k = 0, 1$ indiquant si le label i est ou non attribué à l'instance k ;
- p_i^k la probabilité que l'instance k ait le label i.

Cette fonction de perte a l'inconvénient de donner la même pénalité aux instances mal classées sans tenir compte de la différence de difficulté entre elles. Ce type de problème a fait objet d'un travail en vision ordinateur par [**Lin et al., 2017**] qui ont créé une *focal loss* dédiée à la détection d'objets dans une image. Dans le cadre d'une classification binaire avec des classes déséquilibrées, cette fonction de perte donne une pénalité moins importante aux instances faciles à classer, grâce à un système de pondérations présenté dans l'Équation 3.5.2 ci-dessous :

$$L_{focal} = \begin{cases} -(1-p_i^k)^\gamma \log(p_i^k) & \text{si } y_i^k = 1 \\ -(p_i^k)^\gamma \log(1-p_i^k), & \text{sinon} \end{cases} \tag{3.5.2}$$

où $(1-p_i)^\gamma$ est un facteur de modulation avec l'hyperparamètre $\gamma \geq 0$ qui détermine à quel point la perte L_{focal} est réduite pour les instances faciles à classer. Ainsi pour une instance facile à classer, plus p_i^k est grand, plus le facteur $(1-p_i^k)^\gamma$ réduit la perte L_{focal}.

Cette fonction introduite par [**Lin et al., 2017**] a ensuite été reprise et adaptée par [**Huang et al., 2021**] pour la tâche de classification multilabel de texte avec une perte de distribution équilibrée (*distribution-balanced loss*) présentée par l'Équation 3.5.3 ci-dessous :

$$L_{DB} = \begin{cases} -\hat{r}_{DB}(1-q_i^k)^\gamma \log(q_i^k) & \text{si } y_i^k = 1 \\ -\hat{r}_{DB}\frac{1}{\lambda}(q_i^k)^\gamma \log(1-q_i^k), & \text{sinon} \end{cases} \tag{3.5.3}$$

avec :

- $\hat{r}_{DB}$, un facteur de rééquilibrage des pondérations (*rebalanced weighting*) qui donne à chaque instance un poids déterminant son influence sur la perte L_{DB}. En bref, une instance avec des labels rares aura plus d'influence sur L_{DB} qu'une instance avec des labels majoritaires.
- $\frac{1}{\lambda}$ et q_i^k sont des éléments permettant une régularisation de la perte vis-à-vis des exemples négatifs (*negative-tolerant regularization*). En somme, ce procédé fait que les instances négatives et positives pour un même label i auront des pénalités différentes. Il faut aussi souligner que q_i^k est une variante de p_i^k en ceci qu'elle est calculée différemment selon que $y_i^k = 1$ ou 0. De plus amples détails sont disponibles dans l'article de [**Huang et al., 2021**].

La fonction de perte appelée *DBloss* est reprise dans notre expérience avec les mêmes configurations que celles utilisées dans l'article de [**Huang et al., 2021**] qui était basée sur des classifications multilabel de documents de nouvelles et de médecine. Une autre fonction de perte semblable adaptée pour les jeux de données aux labels déséquilibrés a aussi été conçue par [**Fallah et al., 2023**], mais nous n'avons pas pu en retrouver le code malgré nos sollicitations aux auteurs.

Les résultats de chaque modification sont présentés dans le Tableau 3.10 qui donne les scores moyens de 5 exécutions (*runs*) pour chaque configuration (chacune a son propre chiffre d'amorce ou *random seed*). De façon générale, un seuil à 0.25 permet de retourner davantage

Fine-tuning modification	t=0.25	t=0.5	t=0.75	t=0.8	Balanced train set	Balanced train and val. sets	DB loss
Micro F1 ↑	87.79***	**88.68**	88.26*	87.94**	87.21***	87.07**	88.09***
Macro F1 ↑	**58.44***	54.43	50.04	47.14*	49.45*	49.60*	57.69
Exact Match ↑	45.05***	50.67	**51.16**	50.74	47.50**	47.56*	49.59*
Exageration ↓	**4.70*****	7.72	9.09*	9.39**	9.12*	9.06*	8.32

Tableau 3.10. Performance de CamemBERT suite à différentes modifications dans le processus d'affinage (sur 100). Le modèle avec le seuil $t = 0.5$ correspond au CamemBERT *vanilla* utilisée précédemment. Les différences significatives par rapport à $t = 0.5$ sont dénotées par *, **, *** pour des valeurs p inférieures à 0.05, 0.01, 0.001 respectivement.

de labels minoritaires, ce qui aura un effet positif sur le Macro F1 (moyenne non pondérée de tous les scores F1), mais influence négativement le Micro F1 et l'exactitude. La configuration avec $t = 0.25$ permet aussi d'obtenir de façon significative le meilleur et plus bas score d'exagération. À l'inverse, en relevant le seuil t au-dessus de 0.5, aucune amélioration significative n'est observée. Pour ce qui est de rééquilibrer les types d'instances au sein des ensembles d'entraînement et de validation, seules des dégradations significatives sont observées. Cela peut signifier que le rééquilibrage des jeux de données provoque une telle réduction en volume et en diversité dans les *sets* subséquents que le modèle peine à généraliser à d'autres instances. Il faut aussi souligner que nous avons pris ici le parti de formaliser la tâche de classification de verdict sous forme de classification multilabel, et non de classification binaire qui est le format majoritaire utilisé par exemple par [**Aletras et al., 2016**]. De ce fait, construire un corpus balancé est plus délicat dans notre cas où chaque instance peut se voir attribuer plusieurs étiquettes au lieu d'une catégorie unique. Enfin, pour ce qui est de la fonction de perte *DB loss*, aucune amélioration n'est observée au niveau des métriques, excepté pour le Macro F1 mais dont le score moyen n'est pas significativement meilleur. Cette amélioration pour la seule métrique Macro F1 semble concordante avec les travaux de [**Huang et al., 2021**] qui avaient utilisé des corpora de classification multilabel (tirés de Reuters et PubMed). Pour ces deux jeux de données, une forte amélioration du Macro F1 a aussi été observée, notamment pour les labels dans la traîne de la distribution, ce qui est

aussi le cas dans notre corpus. Pour ce qui est de leur bond de performance en termes de F1 Micro, rien de semblable ne prévaut dans notre cas. Il est utile de mentionner que les auteurs se contentent de rapporter les performances en termes de Micro et Macro F1, sans donner de scores en termes d'exactitude (*exact match*). Enfin, il se peut aussi que les améliorations de *DB loss* soient plus tangibles lorsque le nombre de labels cibles est très grand. Alors que notre corpus comprend 28 labels de verdict, [**Huang et al., 2021**] ont employé deux corpora comprenant respectivement 90 et 18211 étiquettes cibles.

3.6. Conclusion

Cette première expérience a été l'occasion d'avoir un aperçu de la difficulté de la tâche et de comparer différents modèles, dont des modèles transformeur conçus pour le français. Contrairement à de nombreuses tâches de prédiction de verdict basées sur des étiquettes binaires (p. ex. acceptation/rejet de la demande), les verdicts sont caractérisés ici par des labels cumulables, ce qui oblige les modèles à faire des prédictions retranscrivant plus fidèlement la complexité du verdict d'un juge. En l'état, aucun des modèles testés n'est satisfaisant d'un point de juridique (p. ex. le désavantage des demandes des locataires par rapport à celles des propriétaires est accentué par les prédicteurs), et aucun ne pourrait être déployé à large échelle dans la société pour une quelconque assistance dans la résolution de litiges. Pour la suite des expériences, nous reprenons le modèle CamemBERT qui a obtenu la meilleure exactitude, et projetons de voir comment ses performances seraient améliorées en la présence de connaissances spécifiques au droit.

Chapitre 4

TAL légal et articles de loi : de l'utilisation de connaissances spécifiques au domaine

4.1. Introduction

Le travail d'un juge est particulier en ceci que le verdict est le fruit d'une réflexion où différents faits présentés par les parties sont appréciés et évalués au regard d'articles de loi existants (c.-à-d. est-ce que certains agissements ou certaines situations sont conformes ou contraires à des législations préétablies). Cela est particulièrement vrai dans les pays de droit romano-civiliste (*civil law*) dans lesquels les lois (p. ex. le Code civil) définissent les permissions, les obligations et les interdictions des citoyens. Ces mêmes lois constituent une grille de lecture permettant au magistrat de choisir l'issue et les éventuelles sanctions appropriées pour résoudre un litige. Les articles de loi constituent donc une forme de connaissance spécifique au domaine juridique qui peut être exploitée dans une tâche de justice prédictive.

C'est ainsi que [**Luo et al., 2017**] et [**Long et al., 2019**] conçoivent des modèles prenant en entrée des articles pour une tâche de prédiction des chefs d'accusation[1] et une tâche de prédiction de divorces, respectivement. Dans une autre tâche de prédiction de chef d'accusation, [**Xu et al., 2020**] va plus loin en montrant comment les articles permettent de démêler des cibles susceptibles d'être confondues (p. ex. corruption de fonctionnaires et corruption de non-fonctionnaires).

[1]En anglais, *charge prediction*. Il s'agit non pas de déterminer le verdict, mais de trouver la bonne qualification légale pour le comportement supposé litigieux présenté au juge.

En parallèle de ces travaux, les modèles à la BERT inspirés par [**Devlin et al., 2019**] et l'architecture transformeur (*transformer*) de [**Vaswani et al., 2017**] sont devenus de plus en plus répandus dans différentes tâches de TAL et doivent leur succès au mode opératoire suivant : un pré-entraînement non supervisé sur des corpora massifs (généralement avec une tâche de *masked language modeling*) suivi par un affinage (*fine-tuning*) supervisé suffit à obtenir des gains significatifs de performances pour une tâche cible. Cela s'est avéré vrai pour différentes tâches liées au TAL juridique employant des modèles à la BERT comme l'ont montré [**Wang et al., 2020**], [**Chalkidis et al., 2020**], [**Douka et al., 2021**], [**Chalkidis et al., 2021a**] ou encore [**Garneau et al., 2021**]. Il existe cependant quelques cas dans lesquels un tel procédé n'a pas forcément permis d'améliorer les performances. Dans une tâche de prédiction de préavis (*employment notice prediction*), [**Lam et al., 2020**] ont ainsi observé que l'adaptation de RoBERTa par [**Liu et al., 2019**] au domaine juridique nuisait à la performance. De même, dans une tâche liée au droit fiscal américain (*statutory reasoning in tax law entailment*), [**Holzenberger et al., 2020**] ont remarqué qu'un modèle à la BERT donnait une performance moindre par rapport à un modèle par règles, même après pré-entraînement additionnel (*further pretraining*) sur un corpus du domaine.

Ces différents éléments soulèvent différentes questions :

- jusqu'à quel point un modèle à la transformeur adapté au domaine (*further pretraining*) puis affiné (*fine-tuned*) peut-il résoudre une tâche de TAL juridique, en l'occurrence ici la prédiction de verdict (classification multilabel) ?
- dans quelle mesure des données d'entrée tirées de connaissances du domaine (c.-à-d. les articles de lois mentionnés tantôt) peuvent-elles améliorer la performance pour une telle tâche ?

Notre contribution consiste ici en l'évaluation de différentes architectures qui, en plus d'avoir le texte des faits en entrée, utilisent des informations tirées des articles via différentes représentations. Il est important de souligner ici que le texte des faits est dépourvu de références aux articles. Pour autant que nous sachions, du moins au moment où nous avions réalisé notre contribution [**Salaün et al., 2021**], nous avons été les premiers à utiliser des modèles BERT pré-entraînés pour encoder à la fois le texte des faits et celui des articles. Il est important de souligner que contrairement à [**Luo et al., 2017**] dont le modèle intègre

la prédiction des articles de lois pertinents avant de prédire les chefs d'accusation, nous mettons ici de côté la tâche de prédiction d'articles. En d'autres termes, les articles pertinents à chaque litige sont déjà connus du modèle dès le départ, comme si un **oracle** les avait fournis. Le but dans ce protocole est en effet de déterminer quelle représentation et quelle intégration des articles dans les données d'entrée permettent la meilleure performance dans la prédiction de verdict, le tout en supposant que les articles pertinents sont connus dès le départ. Contrairement à d'autres travaux comme ceux de [**Luo et al., 2017, Zhong et al., 2018**] qui utilisent la prédiction d'articles comme une tâche intermédiaire dont les résultats sont réutilisés pour une autre tâche finale (c.-à-d. la prédiction de charges accusatoires), notre but ici est d'examiner les circonstances dans lesquelles la présence d'information issue des articles permet d'améliorer la classification. La majeure partie de ce chapitre consiste donc à sonder quelle serait la **borne supérieure théorique** (*performance upper bound*) que pourrait atteindre un prédicteur de verdict dont les données d'entrée comprennent le texte des faits et les articles de lois pertinents. Vers la fin de ce chapitre, il sera vu quelle serait la **performance réelle** (*realistic performance*) d'un tel modèle lorsque les articles sont prédits au lieu d'être fournis par un oracle.

4.2. La préparation des données d'entrée

La préparation du texte des décisions et des labels cibles a déjà été expliquée en détail dans la Section 2.3. Le point le plus important à en retenir est que le texte précédant le verdict a été prétraité de façon à ce que les paragraphes des faits soient filtrés avant d'être donnés en entrée au modèle. Plus précisément, les textes donnés en entrée ont été séparés des paragraphes contenant des citations ou des références à des lois/jurisprudence, ou bien des réflexions concernant l'application du droit à la situation litigieuse. Une première motivation de ce procédé est que nous voulons forcer le modèle à trouver le verdict sur la seule base des faits et à opérer lui-même l'analyse juridique, sans accéder à celle déjà effectuée par le juge. Ce mode opératoire permet non seulement de minimiser le risque de fuite d'information, soulevé par [**Şulea et al., 2017**] et qui est susceptible de révéler le verdict, il permet également de retirer des données d'entrée l'information liée aux articles. Nous serons ainsi en mesure de comparer la performance des modèles avec et sans information des articles. Bien que le contenu de ce chapitre soit analogue à [**Salaün et al., 2021,**

Salaün et al., 2021], des différences subsistent au niveau de la préparation des données déjà décrite dans la Sous-Section 2.3 (ex. : *split* temporel ici plutôt que *random split* comme précédemment ; davantage de labels ; meilleure segmentation intradocument ; 154 articles au lieu de 445), ce qui aboutit à des résultats différents par rapport à notre article initial, mais souligne surtout la difficulté à mettre en place cette tâche pour la rendre aussi proche que possible de la réalité.

Dans une première partie du protocole dans laquelle les articles sont donnés en entrée du modèle comme introduits par un oracle, le but est d'observer s'il existe une représentation des articles parmi d'autres qui est plus à même d'améliorer les performances dans la tâche de prédiction de verdict. Il sera vu plus tard vers la fin du chapitre quelle performance serait obtenue avec la représentation d'articles considérée la meilleure lorsque les articles sont donnés par un prédicteur dédié plutôt que par un oracle.

Les articles, au nombre de 154, sont représentés de trois façons différentes. La première est l'encodage **one-hot** (abrégé OH) à travers lequel chaque instance est accompagnée d'un vecteur de taille 154, chaque dimension correspondant à un article. Cette représentation a cependant le désavantage d'être creuse (*sparse vector*, rappelons que le nombre moyen d'articles par instance s'élève à 2.55) et d'impliquer que chaque article-dimension est strictement indépendante des autres. Ces vecteurs portent donc une information relativement pauvre. C'est pourquoi nous nous intéressons à une autre représentation qui tient compte de la hiérarchisation des articles dans les textes de loi et que nous appelons **node2vec** (abrégé N2V). Cette représentation est expliquée dans la sous-section qui suit. La troisième approche consiste à donner directement le **texte des articles** cités au modèle lui-même.

4.2.1. Représentation des articles en Node2Vec

Les articles proviennent de deux sources qui sont le Code civil du Québec (C.c.Q.) [**Assemblée nationale, 2018**] et la Loi sur la régie du logement (L.R.L.) [**Assemblée nationale, 2016**]. Une partie du C.c.Q. comprenant les articles les plus couramment cités dans notre corpus est présentée dans la Figure 4.1. Ces dispositions relatives au droit locatif sont organisées dans une structure hiérarchique organisée en titres,

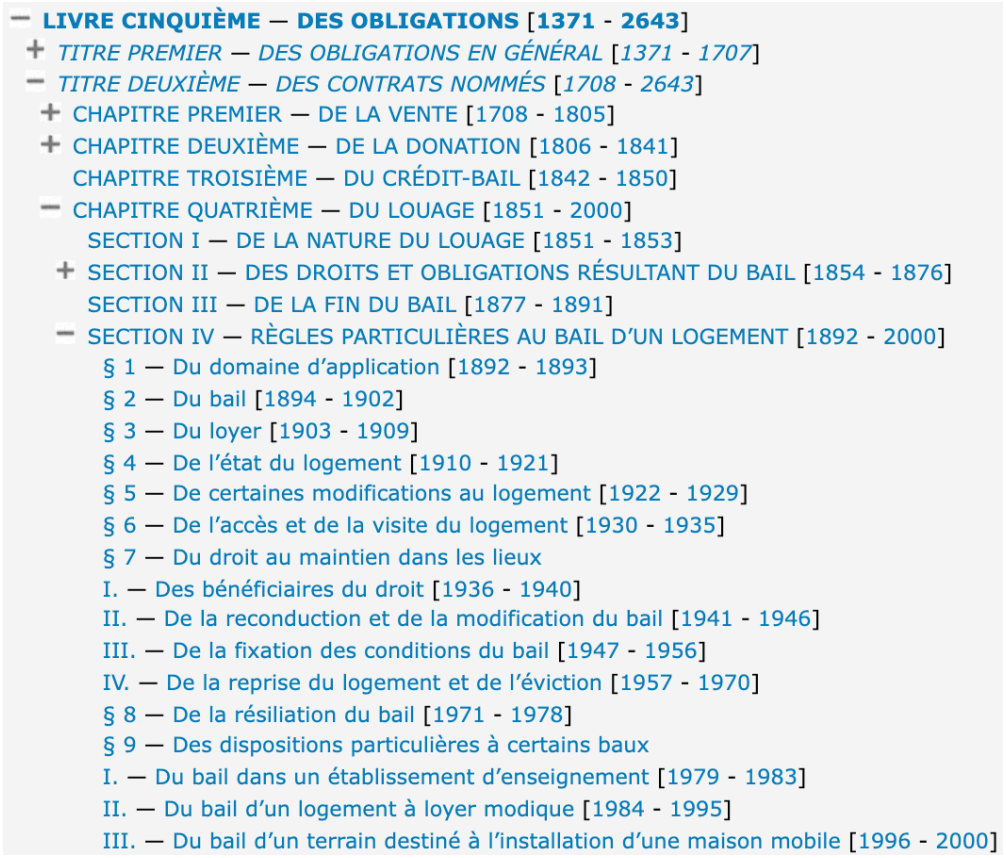

Fig. 4.1. Hiérarchie des articles du Code civil du Québec au niveau du livre cinquième.

chapitres, divisions, paragraphes et ainsi de suite, jusqu'à atteindre les articles eux-mêmes[2]. Le C.c.Q. peut ainsi être comparé à un arbre avec des sous-graphes concentriques dont les feuilles sont les articles. Au fur et à mesure que l'on parcourt en profondeur ce graphe, on atteint des sous-catégories dont les articles couvrent des concepts légaux de plus en plus précis. Par exemple, sur la Figure 4.1, les articles 1910 à 1921 traitent de l'état du logement et on peut donc s'attendre à ce qu'ils traitent de concepts légaux similaires. C'est ainsi qu'est apparue l'idée de créer une représentation des articles qui tienne compte de la proximité entre eux au sein du C.c.Q. de façon à ce que deux articles se situant dans la même sous-section aient des représentations proches. Un autre argument plaidant en faveur d'une telle représentation est le fait que les articles avec des numéros proches tendent à être cités ensemble dans les décisions. Cela est illustré par la carte de chaleur de la Figure 4.2 avec des "régions chaudes" assez rapprochées le long de la diagonale.

[2]Le détail complet de l'organisation des articles entre eux est disponible sur `https://canlii.ca/t/6b4rq`

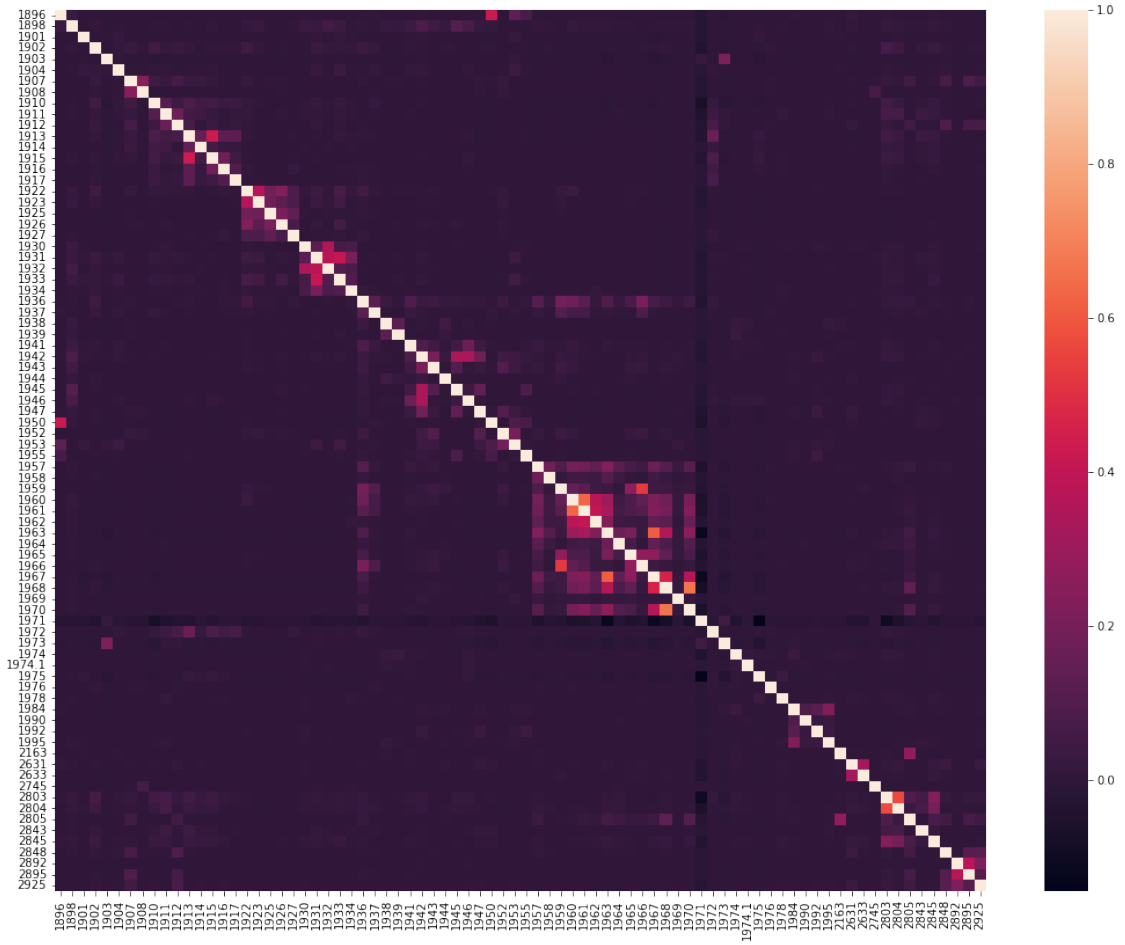

Fig. 4.2. Carte de chaleur représentant la matrice de corrélation entre 90 articles tirés du Code civil du Québec cités dans notre corpus. Les articles sont triés par ordre croissant sur les axes, suivant l'ordre de numérotation.

Afin de rendre compte de cette structuration des articles au sein du C.c.Q., nous optons pour une représentation de type Node2Vec [**Grover and Leskovec, 2016**]. Initialement, deux graphes sont construits à partir de la structure hiérarchique des lois au sein du C.c.Q. et du L.R.L. , sur le même modèle que l'arbre qui transparaît à la Figure 4.1. Les articles font office de feuilles tandis que les livres, sections, paragraphes et autres sections constituent des nœuds. Ces nœuds et ces feuilles sont ensuite reliés entre eux par des arêtes de façon à ce que chaque nœud/section soit relié à la section qui lui est immédiatement supérieure. Ainsi, sur la Figure 4.1, le Chapitre quatrième est lié à la Section IV, elle-même liée au Paragraphe 6, lui-même relié aux articles 1930 à 1935 par des arêtes. Une arête entre deux nœuds ne peut pas contourner un nœud de niveau hiérarchique intermédiaire entre eux. Les deux graphes

obtenus à partir du C.c.Q. et du L.R.L. contiennent ainsi respectivement 4021/204 nœuds et 4030/209 arêtes. En suivant la technique de [**Grover and Leskovec, 2016**], pour un graphe donné, on génère à partir de chaque nœud 200 marches aléatoires (*random walks*) de 1000 nœuds de longueur, ce qui donne des séquences de nœuds à partir desquelles est entraîné un modèle Word2Vec [**Mikolov et al., 2013**] pendant 20 époques, avec une fenêtre de taille 10, et des vecteurs de dimension 300. De cette façon, chaque nœud du graphe, et donc chacun des 154 articles, a une représentation qui reflète la proximité vis-à-vis des nœuds voisins. En d'autres termes, la similarité cosinus entre les vecteurs Node2Vec de deux articles pris au hasard est plus grande lorsque les articles appartiennent à la même sous-section. Pour une décision donnée, les articles cités sont représentés par la moyenne de leurs vecteurs Node2Vec.

4.3. Modèles utilisés

Pour cette tâche de classification multilabel, le modèle CamemBERT par [**Martin et al., 2020**] est repris du chapitre précédent (il avait obtenu la meilleure exactitude, ce qui est crucial pour le domaine juridique, ainsi que les meilleurs scores d'exagération) et fait ici l'objet d'un pré-entraînement additionnel (*further pretraining*) avec la tâche de *MLM* pendant 2 millions de pas (*steps* ou *updates*), soit presque 120 époques. Ce pré-entraînement additionnel est motivé par [**Zheng et al., 2021**] qui suggère qu'un tel procédé peut être bénéfique pour des tâches de TAL juridique selon le degré de spécificité et de spécialisation du domaine légal visé. À cet effet, le texte entier des décisions (soit les faits, l'analyse juridique, plus le verdict, décrits dans la Section 2.2) de l'ensemble d'entraînement est utilisé pour le pré-entraînement additionnel du modèle. Ce corpus de pré-entraînement comprend aussi l'intégralité du Code civil du Québec (C.c.Q.) [**Assemblée nationale, 2018**] et de la Loi sur la régie du logement (L.R.L.) [**Assemblée nationale, 2016**]. L'évaluation est effectuée sur le texte entier des décisions de l'ensemble de validation. Cette adaptation au domaine juridique s'effectue avec les librairies de HuggingFace [**Wolf et al., 2020**] sur une carte graphique NVIDIA GeForce RTX 4090 pendant 6 jours et 4 heures avec un *batch size* de taille 20 et un taux d'apprentissage à $1e-4$. Au terme de cette opération, la perte (*loss*) en validation passe de 0.3622 à 0.2813.

Le modèle pré-entraîné obtenu est nommé **FPTCamemBERT** (FPT pour *further pretrained*). En tant que tel, il ne peut accueillir que le texte des faits comme données d'entrée,

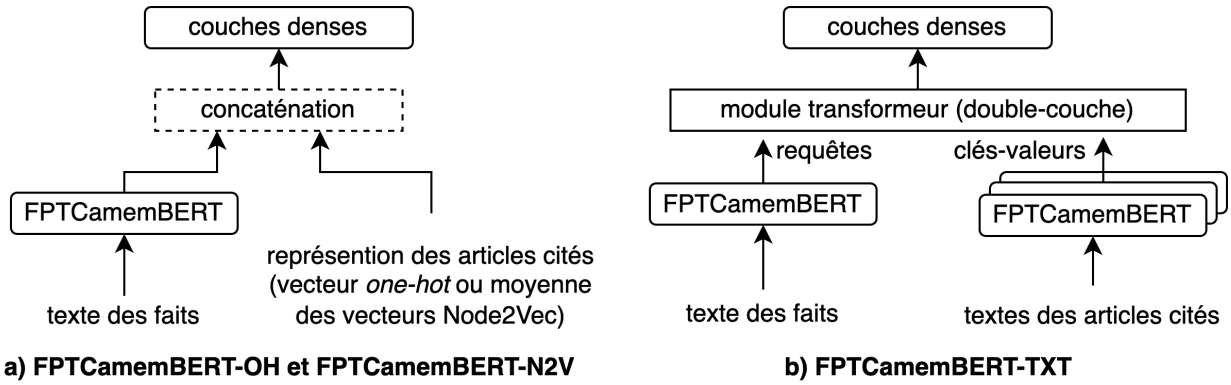

Fig. 4.3. Diagrammes des architectures de FPTCamemBERT-OH/N2V/TXT.

mais il est incorporé dans d'autres architectures, montrées dans la Figure 4.3, pouvant aussi accueillir les différentes représentations d'articles décrites dans la section précédente. Pour le modèle **FPTCamemBERT-OH**, la sortie du premier jeton de BERT ([CLS]) est concaténée au vecteur one-hot des articles. Le tout est ensuite envoyé dans deux couches linéaires denses (*fully connected layers*) afin de donner les labels prédits. Dans le cas du deuxième modèle **FPTCamemBERT-N2V**, la moyenne des vecteurs node2vec des articles cités remplace le vecteur **one-hot**. Pour le troisième modèle **FPTCamemBERT-TXT**, le texte des faits et celui des articles sont encodés dans des FPTCamemBERT distincts et leurs représentations [CLS] sont envoyées dans un module transformeur (mécanisme d'attention par [**Vaswani et al., 2017**]) à double couche en tant que requête (*query*) et clé-valeur (*key-value*) respectivement. Plus précisément, un premier encodeur reçoit la représentation des faits en requêtes et la représentation des articles comme clés-valeurs. La sortie de ce premier encodeur est ensuite envoyée dans un deuxième encodeur en tant que requêtes et clés-valeurs. Ces deux encodeurs sont chacun dotés de 12 têtes d'attention.

Pour l'affinage (*fine-tuning*), les mêmes hyperparamètres et métriques que dans le Chapitre 3 sont réutilisés. Seule la taille de *batch* de FPTCamemBERT-TXT est réduite de moitié à 10 en raison de la présence de deux encodeurs BERT qui prennent davantage de mémoire sur la carte graphique. Chaque modèle est exécuté cinq fois avec des chiffres d'amorces distincts afin de mesurer la performance moyenne.

		CamemBERT	FPTCamemBERT
Tout type de litiges	Micro/Macro F1	88.68 / 54.43	**88.71 / 59.06***
	Exactitude	**50.67**	50.13
Locateur c. Locataire	Micro/Macro F1	**89.64** / 51.44	89.63 / **55.2***
	Exactitude	**51.83**	51.21
Locataire c. Locateur	Micro/Macro F1	68.17 / 28.75	**69.1 / 33.25***
	Exactitude	36.96	**37.32**

Tableau 4.1. Performance des modèles dans la tâche de classification multilabel sur la base de 5 *runs* par modèle. Comparaison entre CamemBERT (paramètres par défaut) et FPTCamemBERT (scores sur 100, l'astérisque indique une amélioration significative de FPTCamemBERT par rapport à CamemBERT avec valeur $p < 0.05$).

4.4. Résultats et discussions

Pour un souci de clarté, l'analyse des résultats obtenus est effectuée via différentes questions de recherche.

4.4.1. En quoi le pré-entraînement additionnel permet-il d'améliorer les performances ?

Les scores de CamemBERT (paramètres d'origine de [**Martin et al., 2020**]) et FPTCamemBERT sont présentés dans le Tableau 4.1. En prenant en compte l'ensemble des litiges (peu importe qui est le demandeur), le pré-entraînement permet un gain de performance pour le Micro F1 et le Macro F1. Le gain pour cette dernière métrique est d'ailleurs significatif. L'exactitude tend à se dégrader légèrement. Si on observe cette fois les scores en tenant compte de qui sont les demandeur et défendeur, il faut souligner que les performances de FPTCamemBERT stagnent voire se dégradent pour les litiges "Locateur c. Locataire", mis à part pour le Macro F1. À l'inverse, le pré-entraînement a des effets bénéfiques pour toutes les métriques pour les litiges initiés par un locataire contre son propriétaire. Ces observations complètent celles effectuées par [**Zheng et al., 2021**]. Ces derniers, sur la base de trois tâches (deux de classification binaire et une de question à choix multiple), affirmaient

		FPTCamemBERT	FPTCamemBERT-OH	FPTCamemBERT-N2V	FPTCamemBERT-TXT
Tous les litiges	Micro/Macro F1	88.71 / 59.06	88.60 / 57.38	88.74 / 58.63	**89.14** / 63.54**
	Exactitude	50.13	49.50	50.30	**50.48**
Locateur c. Locataire	Micro/Macro F1	89.63 / 55.20	89.49 / 53.47	89.63 / 53.83	**89.96* / 58.90****
	Exactitude	51.21	50.49	**51.25**	51.23
Locataire c. Locateur	Micro/Macro F1	69.10 / 33.25	69.31 / 31.53	69.57 / 32.28	**71.76*** / 39.41****
	Exactitude	37.32	37.87	39.08*	**41.62*****

Tableau 4.2. Performance des modèles selon le type de litige (meilleurs scores en gras). Les différences significatives par rapport à FPTCamemBERT sont dénotées par *, **, *** pour des valeurs p inférieures à 0.05 0.01, 0.001 respectivement.

que le pré-entraînement permettaient des gains importants pour des domaines légaux très spécialisés. Les résultats de notre tâche de classification multilabel apportent une réponse plus nuancée : le pré-entraînement semble ici surtout bénéfique pour améliorer les performances pour des instances assez rares (les litiges de type Locataire c. Locateur recouvrent 6% de tout le corpus). Un examen approfondi des scores F1 par label révèle aussi que le pré-entraînement est surtout bénéfique pour les labels minoritaires (qui couvre moins de 5% des instances) avec des gains moyens de 4.6 points.

4.4.2. En quoi l'injection de connaissances spécifiques au domaine (articles de loi) permet-elle d'améliorer la classification ?

Notre analyse se focalise cette fois-ci sur FPTCamemBERT, FPTCamemBERT-OH, FPTCamemBERT-N2V et FPTCamemBERT-TXT dont les résultats sont présentés dans le Tableau 4.2. En termes de performance brute (en prenant en compte tous les litiges), l'injection d'articles sous forme de *one-hot* ou Node2Vec dans les données d'entrée débouche généralement sur une stagnation voire détérioration pour les scores Micro F1, Macro F1 et l'exactitude. Cela suggère que les représentations *one-hot* et Node2Vec n'apportent pas d'information utile à la classification, voire qu'ils apportent du bruit. Les gains de performance sont beaucoup plus importants pour FPTCamemBERT-TXT qui obtient les meilleurs scores

pour les trois métriques analysées, avec des différences significatives par rapport à FPTCamemBERT.

En analysant la performance des classifieurs selon le type de litige, il faut souligner que l'ajout du texte des articles dans les *input features* permet des gains plus importants dans les cas où un locataire poursuit son propriétaire en justice par rapport à ceux où le propriétaire est le demandeur. Pour les situations de type "Locataire c. Locateur" du Tableau 4.2, les gains de FPTCamemBERT-TXT sont en effet de plus grande amplitude avec une plus grande significativité statistique (valeurs p inférieures à 1% ou 0.1%).

Jusqu'ici, l'analyse des résultats s'est porté sur des moyennes concernant l'ensemble des litiges, ce qui porte un regard encore superficiel sur l'apport des articles dans les données d'entrée. La suite de l'examen se poursuit avec le Tableau 4.3 qui permet de décortiquer les performances des classifieurs à un niveau plus fin encore, avec le détail du score F1 individuel à chaque label. Cet examen se focalise en particulier sur les gains de performance obtenus par FPTCamemBERT-TXT en raison de leur significativité statistique par rapport à FPTCamemBERT qui fait office de point de comparaison (*baseline*).

Il a été vu précédemment dans la Section 2.4.2, l'existence de corrélations : d'une part, parmi certains labels de verdict (Figure 2.6), et d'autre part, entre des labels de verdict et des articles (Figure 2.7). Cela a ainsi permis d'identifier au préalable certains groupements de labels et d'articles qui circonscrivent certaines situations juridiques récurrentes pouvant nous servir de grille de lecture.

4.4.2.1. Litiges de type "Locateur c. Locataire".

Ainsi, dans le cas des litiges initiés par un propriétaire, un cas emblématique est l'expulsion du locataire (eviction) avec résiliation du bail (termination_lease), condamnation à verser des dommages-intérêts (tenant_pays_landlord) et exécution de la peine malgré un éventuel appel (provisional_enforcement). Ces étiquettes sont fortement corrélées aux articles 1619, 1971, 1883 C.c.Q. et 82.1 L.R.L. L'inclusion du texte de ces articles dans les données d'entrée a permis une augmentation des scores F1 de ces labels, mais avec une différence statistiquement significative pour seulement deux d'entre eux. Cela peut être dû au fait que, comme ces verdicts sont majoritaires dans le corpus (plus de la moitié des litiges de l'ensemble de

Label de verdict cible	FPTCamemBERT	FPTCamemBERT-OH	FPTCamemBERT-N2V	FPTCamemBERT-TXT	Support
tenant_pays_landlord	98.96	99.01	99.09	**99.26****	87.1
eviction	97.01	96.75	96.89	**97.29**	53.9
termination_lease	97.16	97.02	96.99	**97.24**	52.5
applicant_request_denied	68.95	68.92	68.61	**70.42****	31.5
provisional_enforcement	94.57	94.59	94.60	**95.00*****	37.8
reserve_recourse	55.93	56.04	**56.58**	54.33	24.5
lease_already_terminated	80.09	78.68	78.86	**82.33****	15.3
legal_costs_paid_by_applicant	56.20	54.08	56.18	**57.60**	3.6
landlord_pays_tenant	82.47	82.88	83.46	**88.45*****	3.5
tenant_ordered_pay_rent	81.47	86.12**	83.20*	**90.30*****	5.4
agreement	76.90	**77.35**	76.99	76.96	0.9
cancel_ruling	76.40	75.80	75.73	**77.04**	1.6
landlord_repossesses_rental_unit	90.51	91.47*	90.43	**95.51*****	1.1
tribunal_sets_new_rent	53.18	52.52	52.93	**55.54**	0.6
conditional_sentence	**22.85**	22.66	21.55	22.17	2.1
uphold_ruling	**62.90**	60.59	59.76	62.60	1.6
new_audience	**59.49**	58.07	57.24	59.39	1.2
no_more_recourse	34.47	32.65	30.81	**41.77**	0.5
order_landlord_repairs	38.11	34.09**	37.33	**41.75***	0.5
tenant_can_deduct_from_rent	39.19	40.65	**43.66**	39.29	0.4
withdrawal_demand	29.32	28.20	26.27	**32.49**	0.3
tenant_must_provide_access	65.13	62.47	63.42	**67.36**	0.2
verdict_related_to_peaceful_enjoyment	10.91	5.35	6.63	**23.07**	0.1
outside_jurisdiction	42.15	36.77	45.29	**73.74*****	0.2
verdict_related_to_pets	40.13	28.07	32.25	**60.62**	0.1
assignment_of_lease	**64.28**	63.99	61.18	59.33	0.1
verdict_related_to_land_use_change	34.96	21.79	45.84	**58.39**	0.1
inside_jurisdiction	0.00	0.00	0.00	0.00	0.1
Micro F1	88.71 (0.15)	88.60 (0.39)	88.74 (0.17)	**89.14** (0.19)**	
Macro F1	59.06 (1.88)	57.38 (1.59)	58.63 (2.06)	**63.54** (0.40)**	
Exactitude	50.13 (0.99)	49.50 (0.99)	50.30 (0.51)	**50.48 (0.46)**	

Tableau 4.3. Score F1 par label pour chaque modèle (meilleurs scores en gras, écarts-types entre parenthèses). Les différences significatives par rapport à FPTCamemBERT sont dénotées par *, **, *** pour des valeurs p inférieures à 0.05 0.01, 0.001 respectivement. La fréquence des labels dans l'ensemble de test est indiquée dans la dernière colonne (sur 100).

test débouchent sur des évictions et des résiliations de bail), la marge de progression reste alors plus restreinte.

En parallèle de ces observations, il est important de souligner que les modèles à la BERT (sans articles en entrée) avaient une tendance fâcheuse à systématiquement condamner le locataire à l'expulsion et résiliation de bail, tandis que FPTCamemBERT-TXT semble capable d'apporter une plus grande nuance grâce à un score F1 en hausse de plus de 8 points pour l'étiquette tenant_ordered_pay_rent. Cette dernière décrit la situation où un locataire est simplement rappelé à l'ordre de payer son loyer à temps. Cette amélioration semble possible grâce à la forte corrélation entre tenant_ordered_pay_rent et l'article 1973 C.c.Q. qui permet d'éviter l'expulsion et la résiliation en cas de retard de paiement de loyer inférieur à 3 semaines.

Enfin, il existe une autre situation récurrente à soulever : celle du propriétaire demandant au juge de pouvoir récupérer son logement pour son usage personnel, ce que le juge lui accorde généralement (landlord_repossesses_rental_unit) en le condamnant à verser des frais de déménagement au locataire existant (landlord_pays_tenant). Ce cas de reprise du logement en cours de location est fortement corrélé aux articles 1963 et 1967 C.c.Q. définissant les conditions d'une reprise. Dans le modèle FPTCamemBERT-TXT, l'apport du texte de ces articles permet une hausse significative (valeur p inférieure à 0.1%) des F1 scores de ces deux labels.

4.4.2.2. Litiges de type "Locataire c. Locateur".

Dans l'ensemble des instances de test dans lesquelles un locataire poursuit son propriétaire, environ 69% des cas aboutissent à un rejet total ou partiel des requêtes du demandeur, ce qui est dénoté par le label applicant_request_denied. Cette étiquette trouve sa plus forte corrélation avec l'article 2803 C.c.Q. qui dispose : "Celui qui veut faire valoir un droit doit prouver les faits qui soutiennent sa prétention." L'injection de ce segment de texte dans les données d'entrée de FPTCamemBERT-TXT semble expliquer l'amélioration du score F1 de applicant_request_denied avec une valeur p inférieure à 1%.

Un autre cas de figure majeur est celui où le propriétaire se voit ordonner de procéder à des réparations dans le logement (p. ex. un locataire qui veut que son locateur

rénove la plomberie, ce que ce dernier tarde à faire). Ce verdict est dénoté par le label order_landlord_repairs qui est fortement corrélé aux articles 1854, 1864 et 1910 qui définissent les obligations de maintenir le "bon état de réparation" et "d'habitabilité" du logement. Là encore, l'incorporation des articles dans FPTCamemBERT-TXT permet une amélioration significative du score F1.

Jusqu'ici, il serait tentant d'affirmer que la performance d'un classifieur vis-à-vis de certains labels peut être grandement améliorée par l'ajout de connaissances du domaine, sous réserve que lesdits labels soient fortement corrélés à des articles de loi pertinents. Il existe cependant certaines situations où cette affirmation ne se concrétise pas. Par exemple, le fréquent duo de labels cancel_ruling et new_audience caractérise le cas de figure où un locataire parvient à faire annuler une décision passée et convoquer une nouvelle audience (p. ex. le locataire ne pouvait être présent à la première audience pour raisons de santé et veut faire annuler le jugement qui avait été émis en son absence). Ces deux étiquettes sont fortement corrélées à l'article 89 L.R.L. qui encadre les rétractations de décisions. Pour autant, l'inclusion de cet article dans FPTCamemBERT-TXT ne permet pas de gains significatifs pour ces deux labels. La même observation peut être faite pour l'étiquette uphold_ruling, lui aussi fortement corrélé à l'article 89 L.R.L.. Un constat similaire peut encore être effectué avec le label assignment_of_lease (passation du bail du locataire à une autre personne) corrélé avec l'article 1871 C.c.Q. qui encadre les sous-locations et cession de bail (p. ex. le propriétaire ne peut s'opposer à ces transferts à moins d'un motif sérieux) : le F1 score de cette étiquette se dégrade dans tous les modèles ayant accès aux articles.

4.4.3. Quelle serait la performance des modèles si les articles étaient prédits au préalable ?

Comme mentionné précédemment à la fin de la Sous-Section 4.1, les articles pertinents à chaque décision sont déjà connus d'avance par le modèle, comme s'ils avaient été fournis par un oracle. Cela signifie que les scores de prédiction de verdict présentés jusqu'à présent sont des bornes supérieures (*upper bounds*) théoriques que les modèles pourraient atteindre avec l'hypothèse que tous les articles pertinents soient fournis par un oracle ou un prédicteur

d'articles parfait. Dans cette section, un modèle CamemBERT analogue au FPTCamemBERT décrit à la Section 4.3[3] est cette fois entraîné pour prédire les articles associés à chaque litige[4]. Les hyperparamètres utilisés sont les mêmes que ceux décrits à la Sous-Section 4.3. Deux variantes ont été utilisées au niveau de la fonction de perte du prédicteur d'article : l'entropie binaire croisée (*binary cross-entropy* ou *BCE*), et la *DBloss* (*distribution-balanced loss*) qui serait à priori davantage adaptée pour les tâches de classification multilabel selon [**Huang et al., 2021**].

En moyenne, comme illustrée dans le Tableau 4.4, la performance en termes d'exactitude et de Micro F1 reste équivalente, quelle que soit la fonction de perte utilisée. La fonction de perte *DBloss* apporte un gain substantiel uniquement pour le Macro F1, ce qui suggère de meilleures prédictions pour des articles rares situés dans la longue traîne de la distribution.

Fonction de perte	BCE	DBloss
Micro F1	**91.09**	91.03
Macro F1	31.31	**38.92**
Exactitude	76.25	**76.39**

Tableau 4.4. Performance d'un modèle CamemBERT pré-entraîné sur le corpus du droit du logement et affiné pour la prédiction d'articles selon deux fonctions de perte : l'entropie croisée binaire (BCE) et *DBloss* par [**Huang et al., 2021**]. Les scores sont des moyennes à partir de 5 *runs* et les meilleurs résultats sont en gras.

Une fois obtenus les articles prédits, ceux-ci sont réutilisés comme données d'entrée dans le modèle FPTCamemBERT-TXT qui avait eu les meilleures performances. Ce prédicteur de verdict est ensuite entraîné avec strictement les mêmes hyperparamètres que dans la Sous-Section 4.3. Seule l'entropie binaire croisée est utilisée pour ce modèle, étant donné que la fonction de perte *DBloss* ne donnait pas de gain significatif pour la prédiction de verdict comme cela avait été vu dans la Sous-Section 3.5.

[3]Il s'agit d'un *further pretrained* CamemBERT, pré-entraîné de façon non supervisée sur le corpus du droit du logement, puis ici affiné de façon supervisée cette fois pour la prédiction d'articles et non de verdict.

[4]Il existe 154 articles possibles. Une décision pouvant cumuler plusieurs articles, la tâche est formalisée comme une classification multilabel.

Dans le Tableau 4.5, les modèles avec les articles prédits ont des performances inférieures à celle du modèle sans (FPTCamemBERT), en termes de Micro F1 et d'exactitude. Elles le surpassent cependant pour le Macro F1 et le score d'exagération, bien que ces améliorations ne soient pas statistiquement significatives. Il reste ainsi une certaine marge de progression avant de se rapprocher du niveau de performance de CamemBERT-TXT. Le principal levier d'action semble être ici d'améliorer le Macro F1 de la prédiction d'articles en amont de la prédiction de verdict : ainsi le meilleur prédicteur d'article du Tableau 4.4 reste encore incapable de prédire plus de la moitié des articles les moins fréquents dans la traîne de la distribution, comme en témoigne le Macro F1 plafonnant à 38.92 obtenu par un CamemBERT optimisé avec *DBloss* .

Métrique	Sans articles	Articles prédits avec BCE	Articles prédits avec DBloss	Articles données par un oracle
Micro F1 ↑	<u>88.71</u>	88.15**	88.43*	**89.14****
Macro F1 ↑	59.06	60.94	<u>61.23</u>	**63.54****
Exactitude ↑	<u>50.13</u>	49.29	49.02	**50.48**
Exagération ↓	8.31	<u>7.78</u>	**7.43**	7.98

Tableau 4.5. Performance moyenne sur 5 *runs* de FPTCamemBERT pour la tâche de prédiction de verdict, selon la présence d'articles en données d'entrée fournis ou bien par un prédicteur d'articles (troisième et quatrième colonnes), ou bien par un oracle (dernière colonne). Les deuxième et dernière colonnes correspondent respectivement aux modèles FPTCamemBERT et FPTCamemBERT-TXT du Tableau 4.3. Les meilleurs et deuxièmes meilleurs scores sont en gras et soulignés. Les différences significatives par rapport à FPT-CamemBERT (Sans articles) sont dénotées par *, ** pour des valeurs p inférieures à 0.05, 0.01 respectivement.

4.4.4. Discussion et conclusion

Le chapitre précédent avait mis en lumière la différence de performance d'un classifieur à la BERT selon qu'une instance comprenait un litige de type "Locateur c. Locataire" ou "Locataire c. Locateur", avec une dégradation des performances pour ce dernier. Dans le

présent chapitre, il a été vu que le pré-entraînement étendu (*further pretraining*) d'un modèle CamemBERT avec la tâche de *MLM* permet de réduire en partie ces disparités grâce à une adaptation au domaine.

Par la suite s'est posée la question de savoir si l'ajout d'informations spécifiques au domaine dans les données d'entrée, soit les articles de loi, permet d'améliorer les performances des classifieurs. Il est important de rappeler ici une nouvelle fois que les informations tirées des articles de loi n'étaient pas disponibles dans le texte des faits en entrée.

À l'issue de ces expériences, il semble que l'accès du modèle au texte des articles permet d'atteindre le maximum de performance. En examinant les performances par label plus finement, certains gains de performance significatifs ont notamment été observés pour des labels de verdict qui avaient de fortes corrélations avec certains articles de lois, mais ce phénomène n'est pas systématique. Ainsi, pour des combinaisons verdict-articles telles que (cancel_ruling, new_audience ; article 89 L.R.L.) ou (agreement ; article 2633 C.c.Q.), les scores peuvent stagner ou se dégrader, et les éventuels gains de performance ne sont pas significatifs. La même remarque peut être faite pour le label verdict_related_to_peaceful_enjoyment et l'article article 1860 C.c.Q. qui définit pourtant l'obligation de "ne pas troubler la jouissance normale" des habitants : malgré un score F1 en hausse de 12 points, le gain n'est pas significatif.

En d'autres termes, l'exploitation de connaissances spécifiques au domaine juridique n'est pas la panacée pour relever les performances d'un prédicteur de verdict. Aussi, bien qu'un tel procédé permet d'atténuer les biais et les déséquilibres en améliorant les performances pour les litiges de type "Locataire c. Locateur" (Tableau 4.2), nous sommes encore loin d'avoir un système automatique pouvant donner de façon fiable l'issue d'un jugement ; les scores en exactitude restent stagnants malgré les différentes améliorations apportées à l'architecture des modèles. C'est sans compter la difficulté à prédire les articles pertinents pour chaque instance. En d'autres termes, en l'état actuel des connaissances, il est difficile d'espérer de l'apprentissage machine de pouvoir rendre justice comme le ferait un magistrat humain.

Jusqu'ici, les protocoles expérimentaux, comme la majorité des travaux en justice prédictive, se focalisent sur une formalisation *input-model-ouput* sans trop se soucier des informations mêmes qui sont traitées et qui transitent à travers les systèmes. Le prochain chapitre

se concentre sur une tâche de modélisation thématique qui vise à rendre visibles les éléments discutés par les magistrats dans leurs décisions et par les législateurs dans leurs textes de loi.

Chapitre 5

Modélisation thématique : un moyen de dresser une carte de la pratique juridique

5.1. Introduction

La modélisation thématique (*topic modeling*) désigne l'ensemble des techniques consistant à extraire, à partir d'un certain volume de documents, des thèmes ou *topics*[1] caractérisés par des groupes (*clusters*) de termes saillants. La constitution de ces *clusters* dépend de la distribution des termes au sein des documents. La modélisation thématique est notamment appliquée par des spécialistes de diverses disciplines pour avoir un aperçu rapide des sujets traités dans un corpus trop volumineux pour être examiné manuellement. Il est possible de dénombrer plusieurs applications en sciences sociales comme l'analyse de réseaux sociaux [**Tan et al., 2013**], d'articles d'actualités [**Wu et al., 2012**] ou encore en science politique [**Bertalan and Ruiz, 2019**]. D'autre part, d'autres usages existent aussi dans le domaine scientifique ou technique, comme par exemple pour le traitement d'articles de recherche [**Griffiths and Steyvers, 2004**], de données médicales [**Song et al., 2019**] ou encore le génie logiciel (*software engineering*) [**Thomas, 2011**]. Une telle flexibilité et adaptabilité à autant de domaines hétérogènes ont fait que la modélisation thématique est aussi considérée comme une méthode de "lecture à distance" (*distant reading*) par [**Moretti, 2013**] ou encore d'analyse de texte assistée par ordinateur (*computer-assisted text analysis*) par [**Krippendorff, 2018**], grâce à sa capacité à extraire des thèmes à partir de textes.

[1] Dans ce chapitre et celui qui suit, les termes "thème" et "*topic*" sont utilisés comme des synonymes pour éviter de trop répéter l'un ou l'autre.

5.1.1. La modélisation thématique pour le droit

Un tel succès n'a pas échappé aux spécialistes du droit, dont [**Dyevre, 2021**], qui voient dans la modélisation thématique une méthode non supervisée de partitionnement/catérigorisation souple de documents juridiques, sans la nécessité de devoir conceptualiser des classes au préalable. C'est ainsi que le *topic modeling* a été utilisé pour de nombreux corpora : des textes législatifs britanniques [**O'Neill et al., 2017**], des actes juridiques lettons [**Vīksna et al., 2020**], des décisions de cours australiennes [**Carter et al., 2016**], de la Cour suprême néerlandaise [**Remmits, 2017**], du Tribunal suprême fédéral [**Luz De Araujo and De Campos, 2020**] et de la Cour de justice de l'État du Ceará au Brésil [**Aguiar et al., 2022**], ainsi que la Cour suprême des États-Unis [**Silveira et al., 2021**]. Pour l'ensemble de ces travaux, il est important de préciser que les thèmes modélisés à partir des documents étaient surtout destinés à être lus et analysés par des experts du domaine juridique, à l'exception de [**Remmits, 2017**] qui a effectué une évaluation manuelle par des experts et des non-experts.

Or, [**Branting et al., 2020**] avait montré l'existence d'une dichotomie (*gap*) entre d'une part, la terminologie technique et spécialisée employée par les juristes, et d'autre part, le langage générique ou courant employé par les citoyens non-initiés à la matière juridique. En d'autres termes, les juristes et les profanes[2] décrivent la même réalité factuelle, mais à travers des langages différents. Cette dichotomie linguistique (*language gap*) existe d'ailleurs dans le domaine du droit du logement québécois. Nous faisons ici le postulat que la modélisation thématique peut aider à combler le fossé entre les langages juridiques et profanes, notamment en créant ce que [**Carter et al., 2016**] appelle une "taxonomie pratique" (*taxonomy of practice*). En d'autres termes, plutôt que de concevoir un tableau du paysage juridique construit à priori par des juristes à travers des concepts abstraits, les auteurs estiment que la modélisation thématique permet de dresser une carte plus fidèle des pratiques juridiques à partir des éléments concrets discutés dans les documents. C'est ainsi qu'ils ont pu détecter par exemple que les décisions émises dans le cadre de l'*administrative law* australienne avaient surtout trait à des problématiques d'immigration.

[2]Le terme "profanes" est utilisé ici comme équivalent de *laypeople* ou *laymen*, pour désigner les citoyens ordinaires non initiés au droit. Il n'est point de connotation religieuse ou négative ici.

5.1.2. Un moyen de dresser un pont linguistique entre le langage juridique et le langage profane

Comme vu dans le Chapitre 4, les articles de lois constituent une connaissance spécifique au domaine juridique, mais qui ont le plus souvent été utilisés soit comme cibles dans le cadre d'une tâche de classification [**Xiao et al., 2018, Chalkidis et al., 2019a**], soit comme moyens (*input features*) d'améliorer la performance dans une tâche de classification pré-existante avec ses propres cibles [**Luo et al., 2017, Long et al., 2019, Xu et al., 2020, Salaün et al., 2021**].

Rares sont les travaux qui s'attardent sur une analyse fine des articles eux-mêmes, si ce n'est à travers une approche syntaxique ou par règle. Par exemple, [**Dragoni et al., 2016**] avaient cherché à extraire des règles à partir du texte du code de protection australien du consommateur en télécommunication, notamment avec un parser syntaxique (Stanford Parser). Leur projet était de construire des règles formelles à partir de texte libre (*unstructured free-form text*) qui peuvent être utilisées dans des raisonnements de logique déontique[3]. Un autre exemple est le travail de [**Holzenberger et al., 2020**] qui ont retranscrit des règles du droit fiscal américain en `Prolog`. La même approche avait été utilisée par [**Collenette et al., 2020**] pour représenter les différents facteurs permettant de définir si l'article 6 de Convention européenne des droits de l'homme (le droit à un procès équitable) avait été violé ou non. Le but du présent chapitre n'est pas de retranscrire les lois du droit du logement en règles formelles, mais de donner un éclairage sur l'usage même qui est fait de ces lois, à l'image de la *taxonomy of practice* de [**Carter et al., 2016**].

Pour revenir à notre corpus, la plupart des textes de loi cités, tirés du Code civil du Québec (C.c.Q.) et de la Loi sur la Régie du logement (L.R.L.), définissent un cadre relativement prévisible et assez bien délimité pour codifier et dénouer les relations entre locateurs et locataires (p. ex. les motifs et les éléments nécessaires et suffisants pour une résiliation de bail). Il demeure cependant certaines zones d'ombres concernant les modalités d'application concrètes concernant certains articles de loi. C'est ainsi que [**Westermann et al., 2019**] ont cherché à identifier les facteurs (ils utilisent le terme *factors*) susceptibles de permettre à un locataire d'obtenir une réduction de loyer (p. ex. punaises de lits, manque d'accès à l'eau

[3]Dans le détail, il s'agit d'écrire des règles logiques en langage formel pour retranscrire trois types de normes qui sont l'obligation, la permission et l'interdiction.

1854	*Le locateur est tenu de délivrer au locataire le bien loué en* ***bon état de réparation*** *de toute espèce et de lui en procurer la* ***jouissance paisible*** *pendant toute la durée du bail.* *Il est aussi tenu de garantir au locataire que le bien peut servir à l'usage pour lequel il est loué, et de l'entretenir à cette fin pendant toute la durée du bail.*
1864	*Le locateur est tenu, au cours du bail, de faire* ***toutes les réparations nécessaires*** *au bien loué, à l'exception des menues réparations d'entretien; celles-ci sont à la charge du locataire, à moins qu'elles ne résultent de la vétusté du bien ou d'une force majeure.*
1910	*Le locateur est tenu de délivrer un logement en* ***bon état d'habitabilité****; il est aussi tenu de le maintenir ainsi pendant toute la durée du bail.* *La stipulation par laquelle le locataire reconnaît que le logement est en bon état d'habitabilité est sans effet.*

Tableau 5.1. Texte des articles 1854, 1864 et 1910 du Code civil du Québec (`https://canlii.ca/t/6b4rq`). Certaines mentions ont été mises en gras par nos soins.

ou à l'électricité). Dans ces litiges-là, les juges analysent les faits au regard des articles de loi 1854, 1864 et 1910 C.c.Q., présentés dans le Tableau 5.1, qui définissent certaines obligations contractuelles du propriétaire telles que le *bon état de réparation*, la *jouissance paisible* ou le *bon état d'habitabilité*.

Ces notions restent cependant vagues et abstraites. Comme la loi ne dresse pas de liste exhaustive des situations concrètes qui sont recoupées ou non par elles, les modalités d'application sont laissées à l'appréciation et à la discrétion des juges. Cela a amené [**Westermann et al., 2019**] à annoter manuellement 149 décisions et à en extraire 44 facteurs présentés dans le Tableau 5.2. Ce procédé présente cependant quelques limites : l'extraction manuelle est coûteuse (les documents sont plus longs que dans les corpora courants de TAL, un certain niveau d'expertise est requis) et ne couvre qu'une faible portion des documents citant les articles 1854, 1864 et 1910 C.c.Q. Dans notre contribution, l'objectif a été de reproduire ce travail effectué manuellement à plus grande échelle et plus efficacement

grâce à la modélisation thématique. Pour la suite de ce chapitre, les 44 facteurs identifiés manuellement et présentés dans le Tableau 5.2 sont dorénavant nommés **thèmes de références** (*reference topics*).

5.1.3. De la difficulté d'évaluer les modèles thématiques

Comme vu au tout début de ce chapitre, les modèles thématiques sont des outils relativement accessibles dont le succès est appuyé par ses nombreuses applications dans divers domaines. Il faut néanmoins souligner que ceux qui travaillent sur ces modèles et leurs outils d'évaluation font généralement du *topic modeling* sur la base de corpora dits génériques tels que les nouvelles ou Wikipédia [**Aletras and Stevenson, 2013, O'callaghan et al., 2015**]. Cela a suscité une concentration des efforts de recherche sur la mise au point d'outils et de métriques qui se veulent universelles (*one-fits-all*) et généralisables pour l'ensemble des corpora susceptibles de faire l'objet de modélisation thématique. Cependant, bien que certains modèles sont devenus des approches incontournables à travers diverses disciplines (d'après nos lectures, LDA semble ainsi être la plus connue), l'évaluation des thèmes obtenus reste encore une question ouverte, et ce d'autant plus que les domaines visés sont très hétérogènes. Cette remarque est aussi valable pour les différents sous-domaines du droit.

Ainsi, pour certains travaux, les documents peuvent être accompagnés de *gold labels* (p. ex. des catégories de documents), ce qui permet la mise en place d'une évaluation extrinsèque via une classification de texte. Pour donner un exemple concret, [**Luz De Araujo and De Campos, 2020**] et [**Aguiar et al., 2022**] ont utilisé les *topics* obtenus à partir des documents en tant qu'*input features*, afin de voir s'ils permettaient d'améliorer les performances de classification en prenant les catégories pour cibles. Plus la classification est bonne, plus les thèmes obtenus sont jugés de bonne qualité. Mais un tel protocole d'évaluation présuppose que les décisions soient catégorisées manuellement au préalable, ce qui n'est pas le cas de notre corpus[4]. En dehors de la classification, il est aussi possible, à la manière de [**Silveira et al., 2021**], d'évaluer manuellement les thèmes d'un

[4]Dans notre cas précis, il faudrait que chaque décision soit assignée à une catégorie thématique qui soit représentative et sémantiquement concordante avec les *topics* qu'un bon modèle thématique pourrait générer (p. ex. dégâts matériels, agressions verbales, manque d'entretien). Les labels de verdicts utilisés précédemment (par ordre décroissant : landlord_pays_tenant, applicant_request_denied, tribunal_sets_new_rent,

Electricity Access	Mice
Warm Water Access	Other Infestation
Water Access	Noise
Heating	Smell
Cooling	Constant Repairs
Intruder Protection	Flooding
Rain Leakage	Water Leakage
Isolation	Neighborhood Issues
Kitchen Devices	Danger
Bathroom Utilities	Snow Removal
Other Accessories	Exterior Issues
Risk In Emergency	Repairs Not Conducted
Furniture Missing	Structural Integrity
Access To Premises (or Blockage Of Rooms)	Wall Repair
Blockage Of Entry	Floor Repair
Parking Access	Landlord is Unresponsive
Dirty	Landlord Intervention
Moisture	Harassment
Mold	Violence
Bedbugs	Pets
Cockroaches	Commercial Use
Rats	Relocalisation

Tableau 5.2. Noms des 44 facteurs extraits manuellement sur la base de 149 décisions. Les intitulés sont originellement en anglais et tirés d'un formulaire en ligne qui contient aussi le descriptif de chaque facteur. Ces textes sont traduits par nos soins en français pour nos expériences.

order_landlord_repairs, tenant_can_deduct_from_rent) sont moins adaptés pour l'évaluation des *topics* obtenus. Néanmoins, il sera vu dans le chapitre suivant comment les *topics* pourraient aider dans la prédiction de verdict.

document en les comparant avec l'intitulé du *gold label* ou de la *ground truth category* qui lui avait été attribuée.

Lorsque de telles étiquettes sont indisponibles, les méthodes d'évaluation manuelle les plus souvent employées sont l'échelle de Likert (*Likert scale*, score ordinal sur 3 ou 5 points par exemple) [**Aletras and Stevenson, 2013, Remmits, 2017**] et le test d'intrusion[5] (*word intrusion test*) [**Chang et al., 2009, Lau et al., 2014**]. Une autre possibilité, illustrée par les auteurs comme [**Carter et al., 2016**] qui ont un bagage en sciences humaines, consiste à faire uniquement une description et des commentaires extensifs sur les thèmes obtenus plutôt que d'utiliser des métriques.

Étant donné la difficulté de l'obtention d'un corpus juridique annoté ou d'une évaluation manuelle [**Dyevre, 2021**], il est fort tentant de recourir à une métrique automatique qui ferait office de relais (*proxy*) du jugement humain. Une telle métrique est par exemple le score de cohérence (*topic coherence score*, il en existe plusieurs variantes) [**Aletras and Stevenson, 2013, O'callaghan et al., 2015**] qui s'est fortement répandu, et qui se base sur les co-occurrences de termes au sein d'un corpus externe de référence (*external reference corpus*). L'usage d'une telle métrique considérée comme "standard" n'est pas sans risque : [**Hoyle et al., 2021**] soulevaient ainsi le manque d'homogénéité (*consistency*) dans ses modalités d'application dans plusieurs travaux avec les problèmes de reproductibilité associés. Ils soulèvent par ailleurs le fait qu'une telle métrique n'est pas nécessairement alignée avec les nuances du jugement humain (p. ex. l'écart de performance entre deux modèles peut être faible pour des annotateurs humains, mais artificiellement amplifié par la *topic coherence score*). C'est donc ainsi qu'ils réfutent l'existence de corpora génériques et donc l'existence d'une évaluation universelle adaptée pour tous les cas de figure. De la même façon, ils appellent aussi à repenser à la conception de métriques qui soient en phase avec les besoins du monde réel, un conseil que nous avons suivi ici.

5.2. Description et prétraitement des données

Le jeu de donnés de [**Westermann et al., 2019**] contient 149 litiges de 2017. Nous prenons en considération les litiges qui, comme dans le travail précédent, citent les articles

[5]Un mot-intrus est inséré dans les termes descriptifs d'un thème. Celui-ci sera jugé cohérent et de bonne qualité si l'évaluateur parvient à identifier l'intrus.

1854, 1864 et 1910 C.c.Q., mais sur une échelle temporelle plus large de 2001 à 2018, ce qui donne en tout 1,381 documents. Tout le contenu d'une décision n'est pas pertinent ; [**Remmits, 2017**] avait ainsi suggéré que certaines sections pouvaient être plus pertinentes que d'autres pour obtenir des thèmes de bonne qualité. Dans notre cas, seul le texte de la description des faits nous intéresse. Il est par ailleurs important de souligner que le texte des faits peut rassembler plusieurs éléments litigieux, et c'est pourquoi le texte filtré est segmenté au niveau des paragraphes. Par ailleurs, l'exclusion des parties de la décision touchant à des concepts purement juridiques est utile à notre tâche dans la mesure où ce sont des phrases quasi copiées-collées d'un document à l'autre, peu distinctives, et donc susceptibles de générer du bruit au niveau des thèmes.

Le jeu de données comprend ainsi un total de 34,685 paragraphes qui font l'objet d'un prétraitement (*preprocessing*) particulièrement soigneux : tokénization avec un *tokenizer* `spaCy`[6] de [**Honnibal et al., 2020**] ; retrait des dates, des valeurs monétaires, des symboles, des *stopwords* en français, des *stopwords* spécifiques au droit du logement[7] ; toujours avec `spaCy`, lemmatisation des jetons (*tokens*), puis conservation des jetons avec des caractères alphabétiques comprenant certaines étiquettes morpho-syntaxiques[8] (*part-of-speech tags* ou *POS tags*) ; mise en minuscules des paragraphes (*lowercasing*) ; fusion des bigrammes avec des traits de soulignement (p. ex. "eau_chaude" avec un *underscore* au milieu). Pour finir, le jeu de données ne conserve que les paragraphes avec un minimum de 5 jetons, ce qui donne un total de 26,815 instances. Un exemple de document avant et après prétraitement est présenté dans le Tableau 5.3.

L'aboutissement à ce long *pipeline* a été le résultat de différents essais-erreurs qui soulignent à quel point le prétraitement des textes est crucial afin de minimiser la quantité de bruit textuel qui pourrait s'immiscer dans les *topics* obtenus en bout de chaîne. Il est utile

[6]Disponible via `https://spacy.io/` et `https://github.com/explosion/spaCy`. Le modèle utilisé pour le français est disponible à cette adresse : `https://github.com/explosion/spacy-models/releases/tag/fr_core_news_lg-3.3.0`.

[7]Notre liste inclut : "locataire", "logement", "locateur", "loyer", "demande", "locataires", "mois", "bail", "locatrice", "tribunal', "parties", "audience", "dommages-intérêts", "intérêts", "monsieur", "madame", "audience". Ces termes étant quasi-omniprésents dans des documents relativement homogènes car traitant du même domaine juridique, il a semblé préférable de les considérer comme mots-outils.

[8]Dans le détail, seuls ont été conservés les jetons avec les étiquettes suivantes : ADJ, ADV, NOUN, PROPN, VERB, X.

La demande

Le locataire demande que le tribunal rende une ordonnance obligeant la locatrice à fournir un chauffage adéquat, à nettoyer les conduits de chauffage de deux logements, l'exécution provisoire de la décision malgré l'appel et les frais.

Par amendement, il réclame une diminution de loyer de 50$ par mois de [DATE] à [DATE] et demande d'ajouter comme défendeur le nouveau propriétaire [ANONYME]. Néanmoins à l'audience, il se désiste de sa réclamation contre [ANONYME].

Les faits

Au moment de la demande, les parties étaient liées par un bail du [DATE] au [DATE] à un loyer mensuel de 593$.

La preuve, non contredite en raison de l'absence de la locatrice à l'audience, démontre que le locataire ne pouvait utiliser la fournaise sans être présent au logement, car il devait la repartir chaque fois qu'elle s'éteignait. La locatrice a changé la fournaise, mais le logement est plus long à chauffer et coûte plus cher.

Discussion

L'article 1960 du Code civil du Québec permet que:

«1910. [TEXTE DE L'ARTICLE ANALOGUE À CELUI DU TABLEAU 5.1] »

Il appert de la preuve que cette condition n'a pas été respectée et que le locataire n'a pas eu la pleine jouissance des lieux, en raison du manque de chauffage. De plus, les conduits n'ont jamais été nettoyés.

Le locataire aura droit à une diminution de loyer, mais le tribunal ne peut faire droit aux ordonnances réclamées, attendu la vente de la maison et le désistement contre le nouveau locateur.

POUR CES MOTIFS, LE TRIBUNAL :

ACCUEILLE en partie la demande du locataire;

DIMINUE le loyer de 50$ par mois du mois de [DATE] au mois [DATE] et CONDAMNE la locatrice à payer au locataire 250$ et les frais de 70$.

REJETTE la demande quant aux autres conclusions.

---

rendre ordonnance obliger fournir chauffage adéquat nettoyer conduit chauffage logement exécution provisoire décision appel frais

amendement réclame diminution ajouter défendeur propriétaire [ANONYME] désiste réclamation [ANONYME]

preuve non contredire raison absence démontre utiliser fournaise présent devoir repartir fois éteindre changer fournaise long chauffer coûte cher

Tableau 5.3. Exemple de document utilisé pour la modélisation thématique. En haut, le texte des faits retenu pour la tâche est en noir, le texte exclu est en gris. En-dessous, le texte des faits après prétraitement est présenté en gras. Cet exemple donné à titre illustratif a été anonymisé manuellement par nos soins.

d'insister ici sur le fait que la seule utilisation d'un quelconque modèle thématique, fût-il le meilleur, s'avèrera peu efficace et infructueuse tant que le prétraitement des documents en entrée n'aura pas fait l'objet d'une attention particulière. Cela est d'autant plus vrai lorsque les textes utilisés sont relativement homogènes ou rattachés au même domaine de niche, comme c'est le cas ici avec le droit du logement et ses *stopwords* retirés par nos soins.

5.3. Modèles

Deux modèles non neuronaux et un modèle neuronal ont été utilisés ici :

- l'analyse sémantique latente ou *Latent Semantic Indexing* (abrégé **LSI**, aussi connu sous le nom de *Latent Semantic Analysis*) par [**Dumais et al., 1988**] ;
- l'allocation de Dirichlet latente ou *Latent Dirichlet Allocation* (abrégé **LDA**) par [**Blei et al., 2003**] ;
- **BERTopic** par [**Grootendorst, 2022**].

Au moment d'opérer l'entraînement pour chacun de ces modèles, le nombre de thèmes (un hyperparamètre défini arbitrairement à l'avance) est placé successivement à 50, 100, et 200.

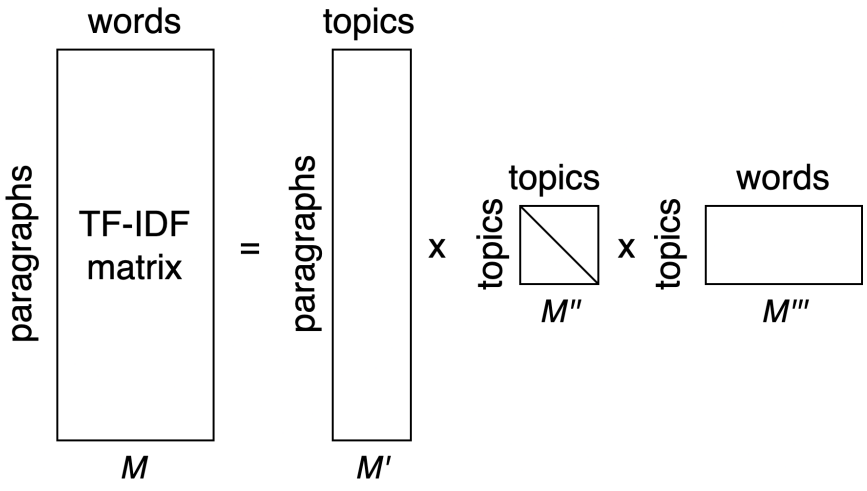

Fig. 5.1. Principe de fonctionnement de LSI avec la décomposition en valeurs singulières de la matrice M de dimension $paragraphs \times words$. La matrice M''' de taille $topics \times words$ est celle qui donne la distribution des jetons de chaque thème.

5.3.1. Analyse sémantique latente ou *Latent Semantic Indexing* (LSI)

Le fonctionnement de LSI repose sur la décomposition en valeurs singulières (en anglais *singular value decomposition*, montrée sur la Figure 5.1 et abrégée DVS) d'une matrice creuse M aux dimensions $P \times W$ ($paragraphs \times words$) dans laquelle chaque valeur $v_{p,w}$ représente le poids TF-IDF (nous devrions dire ici TF-IPF, car c'est le paragraphe qui fait office de document et non le document entier) du terme w pour le paragraphe p. Dit autrement, $v_{p,w}$ augmente avec la fréquence de w dans p, mais décroît lorsque w est très répandu parmi les paragraphes du jeu de données. La DVS de la matrice M produit trois matrices M', M'' et M''' qui ont pour dimensions respectives : $paragraphs \times topics$ ($P \times T$), $T \times T$ et $T \times W$. La dernière matrice M''' est celle qui nous intéresse le plus, car elle fournit la distribution des jetons pour chaque thème. La librairie Gensim [**Řehůřek and Sojka, 2010**] est utilisée pour entraîner les modèles LSI. Le nombre de *power iteration steps* (hyperparamètre) est fixé à 100 afin d'améliorer la finesse de l'approximation effectuée par la DVS sur de grandes matrices creuses.

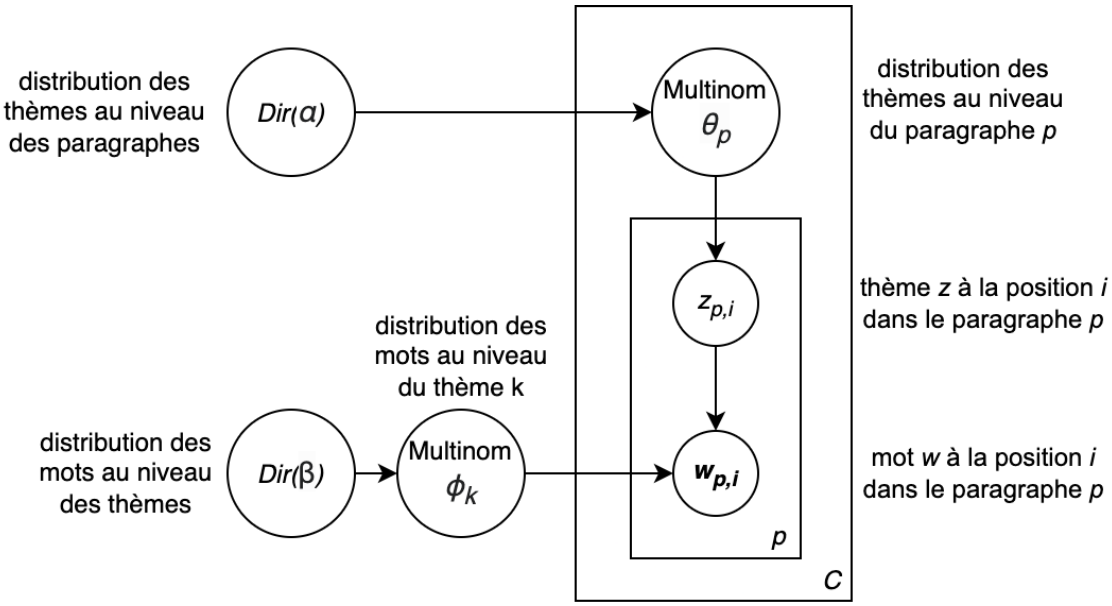

Fig. 5.2. Représentation graphique du processus génératif du modèle LDA pour le mot $w_{p,i}$ à la position i du pseudoparagraphe p dans le pseudocorpus C.

5.3.2. Allocation de Dirichlet latente ou *Latent Dirichlet Allocation* (LDA)

LDA, l'une des approches les plus utilisées en modélisation thématique, est un modèle génératif stochastique qui vise à recréer le corpus original à travers la génération d'un pseudocorpus[9]. Sur la base d'un nombre prédéfini de thèmes, le modèle génère une collection de pseudoparagraphes dans lesquels les distributions des mots et des thèmes sont des approximations aussi proches que possible de celles du jeu de données original. Plus précisément, il est considéré d'une part, que les paragraphes définissent une distribution des thèmes, tandis que d'autre part, les thèmes définissent une distribution des mots. $Dir(\alpha)$ est la distribution Dirichlet des thèmes au niveau des paragraphes tandis que $Dir(\beta)$ correspond à la distribution Dirichlet des mots au niveau des thèmes. α et β correspondent à un paramètre de la loi de Dirichlet pour le degré de concentration : plus ce paramètre est bas, plus la distribution sera déséquilibrée. Dans le cas de $Dir(\alpha)$, cela signifie que la distribution des thèmes dans un pseudoparagraphe sera plus piquée. Idem avec $Dir(\beta)$ qui fera que la distribution des mots sera davantage concentrée autour de quelques termes pour un *topic* en particulier.

En suivant le diagramme de la Figure 5.2, $Dir(\alpha)$ constitue le *prior* pour la distribution multinomiale des thèmes θ_p. Quant à $Dir(\beta)$, il est le *prior* pour la distribution multinomiale des mots ϕ_k pour le thème k. Par la suite, à la position i du pseudoparagraphe p du pseudocorpus C, le mot $w_{p,i}$ est défini, d'une part par ϕ_k, d'autre part par θ_p qui définit le thème $z_{p,i}$ à i dans p.

Après une initialisation aléatoire préalable de ces distributions, suivie par plusieurs passes sur le corpus, le modèle LDA est normalement capable de générer un pseudocorpus ressemblant le plus possible au jeu de données original. Dans le même temps, LDA a aussi pu approximer les différentes lois de distributions mentionnées précédemment, et est donc en mesure d'identifier les termes les plus saillants pour chacun des thèmes k. La librairie Gensim est à nouveau utilisée ici avec un nombre de passes fixé à 100.

[9] "pseudo" est un préfixe utilisé ici pour désigner les éléments générés par le modèle LDA dans sa tentative de reconstituer le corpus d'origine.

LSI : porte fenêtre chambre lieu dernier réparer propriétaire jour problème plancher
LDA : partie propriétaire infiltration estime habitation concerne chauffer finir liste recevoir_avis
BERTopic : cuisine fenêtre intimité ouvrir hinges hennas adjust shelves plateforme transiter

Tableau 5.4. Les dix premiers termes d'un thème pris au hasard pour chaque modèle (nombre de thèmes prédéfini à 100). Seuls les 5 premiers termes de chaque thème (en noir) sont retenus pour l'évaluation. Dans un souci de clarté : les termes d'un thème sont séparés par des espaces ; un terme peut correspondre à un n-gram dont les composantes sont reliées par un trait de soulignement (*underscore*).

5.3.3. BERTopic

Le fonctionnement de BERTopic repose sur des représentations basées sur le contexte émanant des textes. Ces représentations sont tirées d'un modèle transformeur [**Vaswani et al., 2017, Devlin et al., 2019**], ce qui permet au modèle d'accéder à l'information sémantique des paragraphes. Ceux-ci sont d'abord encodés par des représentations `sentence-BERT` [**Reimers and Gurevych, 2019**] qui sont adaptées pour la détection de paraphrases et le partitionnement de documents. Comme notre corpus est en français, nous avons utilisé un modèle multilingue[10] (plus de 50 langues) [**Reimers and Gurevych, 2020b**] afin de créer des représentations de nos paragraphes. Par la suite, ces représentations voient leurs dimensions réduites avec UMAP [**McInnes et al., 2018**] avant de faire l'objet d'un partitionnement (*clustering*) souple et hiérarchique avec HDBSCAN [**McInnes et al., 2017**]. Chaque partition (*cluster*) obtenue contient des paragraphes supposés être liés au même thème. Chaque thème est enfin représenté par les termes les plus saillants de son *cluster* grâce à une approche TF-IDF.

5.3.4. Comment évaluer les modèles et leurs sorties ?

Suite à un examen des thèmes obtenus avec les différents modèles et dont quelques-uns sont présentés dans le Tableau 5.4, le choix a été fait de ne retenir que les 5 premiers termes de

[10]Le modèle pré-entraîné sentence-transformer est disponible à cette adresse : `https://huggingface.co/sentence-transformers/paraphrase-multilingual-MiniLM-L12-v2`

chaque thème-candidat pour l'évaluation. Ceci est motivé par le fait que les termes restants sont moins représentatifs du thème et constituent du bruit textuel.

Afin d'évaluer les différents thèmes obtenus à partir des trois modèles, deux types d'évaluation sont utilisés : une **évaluation automatique quantitative** (Section 5.4) suivie par une **évaluation manuelle qualitative** (Section 5.5). La première évaluation vise à mesurer la qualité des thèmes candidats en utilisant des scores automatiques. Ceux-ci comparent les thèmes candidats à :

- des **thèmes de référence** (équivalent de *gold standard*) qui avaient été extraits manuellement d'une part. Une telle approche permet de savoir jusqu'à quel point les modèles thématiques peuvent reproduire le travail humain d'extraction de *topics* ;
- des **corpora externes de référence** d'autre part. Cela permet de calculer des scores de *topic coherence* qui permettent de mesurer jusqu'à quel point la distribution des termes dans les *topics* candidats est cohérente (*consistent*) avec la distribution du vocabulaire dans d'autres corpora. Les *topics* sont ainsi pénalisés par cette métrique s'ils contiennent des termes trop hétéroclites pour désigner un sujet précis.

Une fois cette première étape franchie, les deux modèles thématiques avec les meilleurs scores sont soumis à une évaluation manuelle[11] se décomposant comme suit :

- deux annotateurs **non experts** examinent les *topics* candidats en spécifiant pour chacun d'entre eux s'ils peuvent identifier une thématique liée au logement (ils n'ont pas accès aux thèmes de référence obtenus manuellement) ;
- un troisième annotateur **expert du domaine** reprend les thèmes candidats jugés pertinents par les annotateurs précédents. Son travail consiste alors à déterminer si les thèmes retenus se recoupent avec les thèmes de référence mentionnés plus haut ou s'ils révèlent au contraire des sujets inédits.

5.4. Évaluation automatique quantitative

Comme mentionné dans la sous-section 5.1.3, l'évaluation de thèmes n'est pas une tâche triviale. Deux approches automatiques quantitatives sont proposées dans notre cas. La première est basée sur la comparaison entre les thèmes candidats (créés par les modèles) et

[11] L'évaluation manuelle étant coûteuse en main-d'œuvre et en temps, il a été choisi ici de ne soumettre que quelques modèles thématiques au regard d'évaluateurs humains.

des thèmes de référence (extraits par des humains). Elle pose la question de savoir si un *topic model* est capable d'identifier les mêmes thèmes observés et extraits manuellement par des humains. La deuxième consiste à mesurer la cohérence thématique (*topic coherence*) mesurant à quel point les termes constitutifs d'un thème candidat suivent la distribution d'un corpus dit de référence (p. ex. Wikipédia ou n'importe quel corpus massif comportant des textes rédigés manuellement). Une telle mesure permet ainsi de pénaliser les thèmes candidats comportant des termes trop hétéroclites pour désigner un sujet précis.

5.4.1. Comparaison avec des thèmes de référence spécifiques au domaine

Une première approche consiste à comparer les **thèmes candidats** (abrégé CTs pour ***C****andidate* ***T****opics*, dans l'encadré jaune en bas à gauche de la Figure 5.3) générés par les modèles avec les **thèmes de référence** (abrégé TRs pour ***R****eference* ***T****opics*, dans l'encadré bleu en haut à gauche de la Figure 5.3) correspondant aux 44 facteurs identifiés manuellement par [**Westermann et al., 2019**]. Cette approche pose la question de savoir si un modèle thématique est capable de retrouver les RTs. Une comparaison par paires (*pairwise comparison*) entre CTs et RTs est ici envisagée avec une méthode automatique présentée dans la Figure 5.3. Pour chaque paire ($\mathrm{RT_i}$, $\mathrm{CT_j}$), un score de correspondance s_{ij} (*matching score*) est calculé, ce qui donne une matrice de dimensions $|CT| \times |RT|$. De cette matrice, nous retenons les $|CT|$ scores les plus élevés (les 5 cases roses dans la Figure 5.3) et les paires correspondantes ($\mathrm{RT_i}$, $\mathrm{CT_j}$). À partir de ces paires, nous comptons le nombre d de RTs **d**istincts ayant été *matchés* (d vaut 3 dans la Figure 5.3). Le rappel et la précision sont obtenus en divisant d par $|CT|$ et $|RT|$ (5 et 4 dans notre exemple) respectivement. Bien que ces métriques ne soient pas parfaites[12], elles sont un compromis nécessaire ; en effet, les RTs identifiés manuellement ne couvrent qu'une très faible portion des litiges disponibles.

Le calcul du score de similarité s_{ij} est délicat. En effet, les RTs sont des noms ou groupes nominaux suivis par des descriptions en une phrase tandis que les CTs sont des séquences comprenant les 5 termes les plus représentatifs de chaque thème. Deux modes de calcul de similarité sont employés ici :

[12]Le nombre de thèmes de référence RTs reste strictement inférieur au nombre de thèmes candidats CTs. La précision est donc *par design* plafonnée à $|RT|/|CT|$ qui est inférieur à 1.

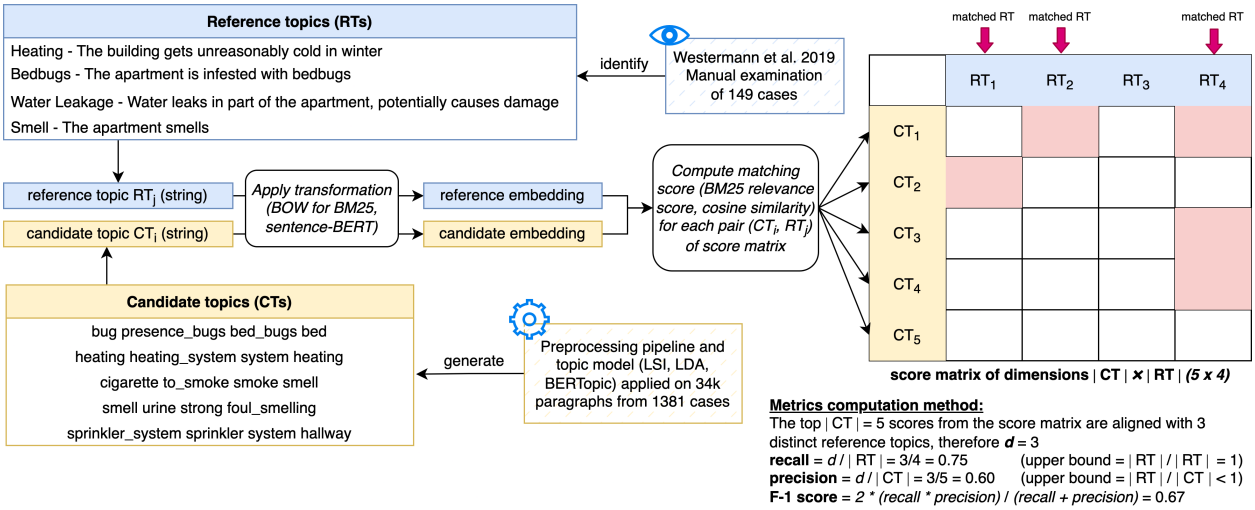

Fig. 5.3. Un exemple jouet (*toy example*) avec 5 CTs comparés avec 4 RTs. Ce diagramme est tiré de [**Salaün et al., 2022a**] qui est publié en anglais, mais tous les thèmes sont en langue française dans nos expériences.

(1) Pour chaque pair (RT_i, CT_j), chaque chaîne de caractères (*string*) est encodée avec un *sentence-transformer embedder* multilingue[13]. La similarité cosinus entre les deux *embeddings* fait ensuite office de score de similarité s_{ij}.

(2) Nous utilisons un système de recherche d'information Okapi BM25 [**Trotman et al., 2014**] dans lequel RTs et CTs sont respectivement des requêtes (*queries*) et des documents. Le score de correspondance s_{ij} correspond au score de proximité que BM25 attribue à CT_j vis-à-vis de RT_i sur la base de sacs-de-mots.

5.4.2. Score de cohérence : évaluation par rapport à un corpus externe de référence

La qualité d'un thème est mesurée dans une grande partie de la littérature avec des métriques de cohérence thématique (*topic coherence*). Elle est généralement mesurée avec la valeur NPMI (*normalized pointwise mutual information* ou information mutuelle spécifique normalisée par [**Bouma, 2009**]). D'après [**Lau et al., 2014**], cette métrique est positivement corrélée à une évaluation humaine basée sur un test d'intrusion, ce qui explique en partie le succès de cette mesure dans d'autres travaux. Selon l'équation 5.4.1, un thème

[13] `https://huggingface.co/sentence-transformers/paraphrase-multilingual-mpnet-base-v2`. Ce modèle est différent du modèle multilingue utilisé dans notre modèle thématique BERTopic.

candidat CT_t aura un score c_NPMI élevé si les N premiers termes qui le constituent ont des probabilités jointes par paires élevées (*high pairwise joint probabilities*).

$$c\_NPMI(CT_t) = \sum_{j=2}^{N} \sum_{i=1}^{j-1} \frac{log \frac{P(w_i,w_j)}{P(w_i)P(w_j)}}{-logP(w_i,w_j)} \tag{5.4.1}$$

Les probabilités jointes $P(w_i,w_j)$ sont calculées au préalable sur la base d'un large corpus dit "de référence", généralement Wikipédia [**Röder et al., 2015**], avec une fenêtre de cooccurrence de 10 mots. Deux corpora de référence sont utilisés ici avec le même prétraitement décrit dans la Section 5.2 :

- le texte de **Wiki**pédia en français datant du 1er septembre 2022 (2.4 millions d'articles) qui fait office de corpus générique (décrit par "**Wiki**" dans le Tableau 5.5) ;
- le corpus de décisions du **T**ribunal **a**dministratif du **l**ogement (531k décisions) qui fait office de corpus spécifique au domaine (décrit par "**T.a.l.**" dans le Tableau 5.5).

Une fois les probabilités jointes calculées, les scores c_NPMI qui en résultent varient entre -1 et 1: -1 implique la complète absence de cooccurrence d'une paire de termes (w_i,w_j) descriptifs d'un thème candidat CT à l'intérieur du corpus de référence ; et 1 signifie au contraire une complète cooccurrence (un terme vient toujours avec l'autre au sein du corpus de référence). L'idée principale à garder de cette métrique est la suivante : plus le score c_NPMI est élevé, plus les termes constitutifs des thèmes candidats obéissent à la distribution des termes composant le corpus externe de référence. Ainsi, un thème avec les termes "ascenseur balançoire fascicule" (termes peu susceptibles d'apparaître ensemble dans le même document d'un corpus de référence) sera davantage pénalisé qu'avec "libellule fourmi coléoptère" qui désignent des insectes (ils pourraient figurer ensemble dans une même rubrique).

5.4.3. Résultats

Comme le montre le Tableau 5.5, LDA et BERTopic surpassent LSI pour l'ensemble des métriques. Pour un modèle donné, augmenter le nombre prédéfini de thèmes générés améliore le rappel pour les métriques BM25 et SBERT, ce qui est attendu. De façon générale, pour un nombre fixe de thèmes et en prenant en compte les scores en précision et F1, LDA fait mieux que BERTopic pour les scores de proximité BM25, tandis que BERTopic obtient les plus hautes performances en termes de similarité cosinus SBERT. Cela peut s'expliquer par

Modèle et n^{bre}		Score proximité BM25			Similar. cosinus SBERT			c_NPMI	
de thèmes		R	P	F1	R	P	F1	Wiki	T.a.l.
LSI	50	0.273	0.240	0.255	0.409	0.360	0.383	0.0090	0.0504
	100	0.409	0.180	0.250	0.591	0.260	0.361	-0.0008	0.0512
	200	0.545	0.120	0.197	0.659	0.145	0.238	-0.0058	0.0361
LDA	50	0.568	**0.500**	**0.532**	0.523	0.460	0.489	0.0215	0.0780
	100	0.614	0.270	0.375	0.705	0.310	0.431	0.0035	0.0607
	200	**0.636**	0.140	0.230	0.818	0.180	0.295	-0.0181	0.0412
BERTopic	50	0.545	0.480	0.511	0.614	**0.540**	**0.574**	**0.1462**	**0.3087**
	100	0.591	0.260	0.361	0.818	0.360	0.500	0.1266	0.2368
	200	0.614	0.135	0.221	**0.909**	0.200	0.328	0.0830	0.1852

Tableau 5.5. Scores par modèle et par nombre de thèmes. Ils se décomposent en scores de similarité entre CTs et RTs (score de proximité BM25 et similarité cosinus basée sur des représentations *sentence-transformer*, avec **r**appel, **p**récision, et score **F1**), et en scores de cohérence c_NPMI. La plus grande et la deuxième plus grande valeurs sont respectivement en gras et soulignée pour chaque métrique.

le fait que LDA et BERTopic sont des modèles basés respectivement sur des sacs-de-mots et des prolongements lexicaux.

Pour ce qui est des scores c_NPMI, quel que soit le corpus de référence utilisé, LSI et LDA ont des scores proches de 0, avec le dernier modèle dépassant légèrement le premier tandis que BERTopic obtient les scores les plus élevés. De tels résultats suggèrent que les termes descriptifs des thèmes générés par LSI et LDA ont des co-occurrences qui surviennent par chance (les distributions des termes sont indépendantes entre elles), tandis que les termes rassemblés au sein des thèmes de BERTopic ont des cooccurrences qui surviennent au-delà du hasard [**Bouma, 2009**]. Étant donné que les scores de rappel de LDA et BERTopic pour les métriques BM25 et SBERT augmentent avec le nombre de thèmes générés, et étant donné que nous cherchons à vérifier jusqu'à quel point les modèles thématiques peuvent aider à identifier des cas d'usages en droit du logement, nous avons décidé d'évaluer manuellement les 200 thèmes générés par chacun de ces modèles.

5.5. Évaluation manuelle qualitative

5.5.1. Évaluation intrinsèque de la pertinence des thèmes candidats

Afin d'évaluer la qualité et l'intelligibilité des thèmes candidats CTs pour les citoyens ordinaires, il a été demandé à deux personnes non expertes en droit d'évaluer les 5 premiers termes décrivant chacun des 400 thèmes obtenus avec LDA et BERTopic (chaque modèle en génère 200). Ces thèmes sont montrés aléatoirement aux évaluateurs qui n'ont pas d'information concernant les modèles qui les ont générés. À partir des termes d'un thème, les évaluateurs devaient dire s'ils étaient en mesure d'identifier un problème ou une situation susceptible de concerner un locataire. Dans l'affirmative, ils devaient décrire succinctement le sujet identifié (p. ex. "moisissure", "désaccord sur le loyer"). Le travail d'évaluation est présenté dans le Tableau 5.6. Les CTs pertinents sont présentés dans les Tableaux 5.7 et 5.8. Les **C**andidats **Q**ualifiés pertinents par au moins **1** annotateur et par les **2** annotateurs sont abrégés respectivement **CQ1**s et **CQ2**s[14].

De façon générale, le score kappa de Cohen pour l'accord inter-annotateur[15] monte à 0.562 pour tous les thèmes, 0.386 pour les 200 thèmes de LDA et 0.649 pour les 200 thèmes de BERTopic. La différence majeure entre les évaluateurs réside dans le fait que l'un d'entre eux a considéré les problèmes matériels comme étant les seuls thèmes pertinents, mettant de côté les problèmes interpersonnels qui sont pourtant bien des objets de litige en droit du logement (ex. : harcèlement avec *harcèlement intimidation harceler comportement victime* ; violence avec *craindre police poste bébé appeler_police*, intrusion avec *donner_accès charger ingénieur verrouiller_porte détailler*). Si nous ignorons les 14 thèmes candidats liés à des problèmes interpersonnels, les scores kappa mentionnés précédemment remontent respectivement à 0.619, 0.440 et 0.706. La différence de kappa score persistant entre LDA et BERTopic peut s'expliquer par le fait que les thèmes de LDA (un exemple est donné dans le Tableau 5.4) basés sur des sacs-de-mots sont plus ardus à interpréter car leurs termes ne permettent pas d'identifier facilement une problématique contrairement à BERTopic. Le faible score kappa

[14]Dans un souci de clarté : toutes les CQ2s sont aussi des CQ1s, mais tous les CQ1s ne sont pas des CQ2s.

[15]Selon [**McHugh, 2012**], Cohen suggère que les scores kappa doivent être interprétés comme des accords interannotateurs passables, modérés et substantiels pour des valeurs comprises dans les intervalles respectifs suivants : $0.21 - 0.40$, $0.41 - 0.60$, et $0.61 - 0.80$.

Thème candidat	Commentaires des évaluateurs	Qualité
• *dater sérieux fissure plâtre produire_photo*	• N/A • dommage dans le plâtre	CQ1
• *fils tapis produit acheter recouvrir*	• N/A • N/A	
• article_code légal_indemnité intérêt_taux civil_québec additionnel_prévoir	• N/A • N/A	
• pluie pleuvoir précipitation neige couler	• infiltration d'eau • infiltration de précipitation	CQ2
• odeur_cigarette fumer fumée cigarette fumeur	• odeur de cigarette • odeur de cigarette	CQ2

Tableau 5.6. Exemples de thèmes candidats examinés par deux évaluateurs non experts du domaine. Par défaut, ils proviennent de BERTopic (LDA si italique). Si un évaluateur peut déceler un sujet pertinent à la lecture des termes d'un *topic*, il doit le décrire brièvement ("N/A" signifie qu'aucune description n'a été fournie). Un thème qualifié pertinent par les deux évaluateurs devient un CQ2 (CQ1 si par un seul évaluateur).

de LDA est cohérent (*consistent*) avec son faible score c_NPMI du Tableau 5.5 et avec le fait que la proportion de thèmes candidats jugés pertinents par les deux annotateurs monte respectivement à 10.5% et 33.0% pour LDA et BERTopic.

5.5.2. Analyse qualitative des thèmes candidats considérés comme pertinents par les évaluateurs

Pour rappel, les **C**andidats **Q**ualifiés pertinents par au moins **1** annotateur et par les **2** annotateurs sont abrégés respectivement **CQ1**s et **CQ2**s. Après l'identification de thèmes candidats pertinents par deux non-experts en droit, un troisième annotateur, cette fois expert du domaine, a manuellement créé des paires entre ces candidats et les 44 thèmes de référence (RTs). Le nombre de RTs distincts ayant pu être combinés aux CQ1s et CQ2 monte respectivement à 28 et 22 sur 44. Les 5 premiers RTs combinés avec le plus de CQ1s et CQ2s

Thème de référence	Nbre. de CQ2s	Nbre. de CQ1s	Exemple de CQ2s
Fuite d'eau	9	10	• eau eau_chaud dégât_eau pression_eau survenir • infiltration_eau infiltration survenir toit toiture • *infiltration_eau garage survenir compléter user* • *laisser dégât_eau croire place matériau*
Bruit	8	10	• bruit marteau_piqueur bruit_excessif scie bruit_provenir • musique musique_fort excessif_voisinage party jouer • bruit enfant enfant_courir déranger jouer • sujet_insonorisation plainte_invraisemblable unité_vérifier admettre_toutefois plainte
Punaises de lit	6	6	• exterminateur traitement extermination punaise procéder_traitement • insomnie stress sommeil piqûre_punaise phobie • *punaise matelas infestation livrer plancher_cuisine*
Chauffage	5	8	• chauffage température thermostat degré fournaise • froid température hiver octobre chauffage • démantèlement_calorifère fermer_début désuet_problème calorifère système_chauffage
Problèmes extérieurs	5	9	• balcon balcon_arrière antenne pourrir rampe • escalier escalier_mener main_courant marche_escalier solidement • accès_terrasse accès arrière donne_accès verrouiller • niveau_ascenseur rafraichissement_mise système_gicleur corridor_besoin manquer_entretien

Tableau 5.7. Sélection de thèmes **C**andidats **Q**ualifiés pertinents par les deux évaluateurs (**CQ2**s) ou au moins un (**CQ1**s) et qui correspondent à des thèmes de référence (RTs). Lecture : pour le thème de référence "fuite d'eau", 10 thèmes candidats ont été jugés pertinents par au moins un annotateur (dont 9 par les deux annotateurs). Les thèmes proviennent de BERTopic par défaut, et de LDA si en italique. La présence de certains termes à première vue peu liés au thème de référence (p. ex. "survenir" pour "Fuite d'eau") s'explique par des expressions récurrentes dans les paragraphes en entrée (p. ex. "une infiltration est survenue").

Thème décelé	Nbre. de CQ2s	Nbre. de CQ1s	Exemples de CQ2s
Plomberie	5	7	• *affecter plomberie chaud manque finition* • plombier plomberie drain clapet batur • cuisine réparer_robinet lavabo_salle eau repas • robinet bruit eau filet_eau régler_définitivement
Qualité de l'air	4	4	• asthme symptôme souffrir docteur nez • ventilation système_ventilation air bouche_évacuation conduit • allergie allergique mélabo teste problème_respiratoire
Accès à internet	3	3	• téléphone numéro_téléphone service_internet téléphoner ligne_téléphonique • câble_origine optique_actualiser passer_fibre bell_vidéotron trou_fait • vidéotron câble panneau technicien câblodistribution
Éclairage	1	1	• lumière éclairage briser_hauteur brûlé_poteau éclairage_déficier
Accessibilité	1	1	• personne_handicapé refaire_juin intercom_rampe hall_entrée automne

Tableau 5.8. Exemples de thèmes candidats qualifiés pertinents par des évaluateurs non-experts qui ne correspondent à aucun thème de référence préexistant. Les CQ2s proviennent de BERTopic sauf s'ils sont en italique (génération par LDA).

sont présentés dans le Tableau 5.7. Au moment de grouper des RTs avec des CQ2s, il a été observé que les RTs pouvaient être abstraits et assez vagues tandis que les CQ2s donnaient des nuances plus fines et précises grâce à des termes saillants. Par exemple, pour le thème de référence concernant le bruit dans le Tableau 5.7, la modélisation thématique permet d'identifier différentes sources sonores telles que les engins de construction ("marteau_piqueur"), de la "musique_fort[e]", des "enfant[s qui] cour[ent]" ou des problèmes d'"insonorisation". Ce gain d'information est encore plus remarquable pour le thème de référence des "Problèmes extérieurs" en nommant des éléments précis : "balcon_arrière", "escalier", "main_courant[e]", "accès_[à_la_]terasse", "niveau_[d']ascenseur".

Modèle thématique	Thèmes générés	Annotations par 2 non-experts		Annotations par 1 expert du domaine			
		Candidats qualifiés		Correspondances avec des RTs		Thèmes inédits	
		CQ1s	CQ2s	CQ1s	CQ2s	CQ1s	CQ2s
LDA	200	64	21	42	17	22	4
BERTopic	200	102	67	64	47	38	20

Tableau 5.9. Répartition des thèmes candidats par modèle selon les annotations attribués par les annotateurs. CQ1s et CQ2s désignent les candidats qualifiés pertinents par au moins 1 et par les 2 annotateurs non-experts respectivement.

L'annotateur a aussi détecté des thèmes candidats pertinents qui n'ont aucune correspondance avec les thèmes de référence (colonne "Thèmes inédits" du Tableau 5.9). Les modèles thématiques permettent ainsi de déceler de nouveaux facteurs litigieux, présentés dans le Tableau 5.8, qui n'étaient pas inclus dans les thèmes identifiés manuellement par [**Westermann et al., 2019**]. Ainsi, plusieurs thèmes candidats qualifiés décrivent divers problèmes de plomberie sans qu'ils n'impliquent de fuites d'eau. D'autres thèmes encore, bien que rares, pointent des sujets sensibles tels que la qualité de l'air ou l'accessibilité pour les personnes handicapées. À l'inverse, les thèmes de référence trouvés manuellement qui n'ont pas été retrouvés par les modèles sont au nombre de 15 sur un total de 44 *reference topics* (soit 34%). En examinant ces thèmes de référence manquants, il s'avère que ces derniers pouvaient être très généraux (p. ex. "Autres accessoires manquants ou non fonctionnels") ou très abstraits (p. ex. "Intégrité structurelle - Le logement présente un manque d'intégrité structurelle") au point qu'un thème candidat aurait du mal à cerner des termes qui y soient spécifiquement dédiés .

5.6. Discussion

De manière globale, les thèmes candidats pertinents sont plus facilement obtenus avec BERTopic qu'avec LDA. Une explication possible réside dans le fait que contrairement aux travaux décrits dans la Section 5.1.1 qui traitent des documents provenant de sous-domaines

juridiques différents [**Luz De Araujo and De Campos, 2020, Aguiar et al., 2022, Silveira et al., 2021**], notre corpus de paragraphes est bien plus homogène car traitant exclusivement de droit du logement, rendant ainsi la tâche de modélisation thématique plus ardue. En conséquence, malgré le prétraitement approfondi effectué en amont à la Section 5.2, l'approche sac-de-mots de LDA donne moins de thèmes pertinents comparée à BERTopic qui a accès à l'information sémantique des termes. Il a aussi été observé qu'en augmentant le nombre de thèmes candidats, LDA était plus susceptible de produire des thèmes bruités et incompréhensibles tels que "berat blood applicances best pilule" (sic) rapporté par les évaluateurs. Une explication possible est que définir un nombre élevé de thèmes peut amener LDA à en produire à partir de bruit textuel. Ce problème n'a pas été détecté avec BERTopic qui était en mesure de déceler des sujets rares et inédits tels que ceux en bas du Tableau 5.8.

Mis à part cette différence de performance entre les deux modèles, il faut soulever que le ratio de thèmes pertinents (colonne "Candidats qualifiés" dans le Tableau 5.9) ne couvre qu'au mieux la moitié des *topics* générés. Ainsi, le ratio de CQ2s s'élève à 10.5% et 33.5% pour LDA et BERTopic respectivement. Concernant les CQ1s, les valeurs s'élèvent à 32.0% et 51.0%. Une explication repose dans le fait que, malgré le prétraitement important effectué en amont (voir la Section 5.2), les paragraphes des faits utilisés en entrée contiennent encore du jargon juridique ou des formulations légales qui décrivent davantage des procédures formelles et routinières (ex. : "les locataires demandent une diminution de loyer") que des situations issues de la vie réelle (ex. : "l'eau s'est mise à couler dans leur logement en raison d'un bris de la toiture"). D'autre part, le manque de familiarité et de connaissances vis-à-vis du domaine du droit du logement a aussi pu empêcher les évaluateurs non-experts de détecter des thèmes pertinents (p. ex. les violences domestiques relèvent aussi du droit du logement, mais cela n'est pas nécessairement évident pour un non-expert). Malgré ces obstacles, il faut souligner que les modèles thématiques ont permis de déceler des thèmes qui n'ont pas pu être découverts manuellement en un temps raisonnable (voir Tableau 5.8).

5.7. Conclusion

Au terme de ce chapitre, il a été vu que les modèles thématiques permettent de dresser efficacement une carte des cas d'usages, une *taxonomy of practice* décrite par

[**Carter et al., 2016**], à partir de décisions de justice concernant le droit du logement du Québec. En l'occurrence ici, les thèmes ont permis d'identifier différentes situations concrètes dans lesquelles des locataires attaquaient des propriétaires au tribunal. Dans un cadre pratique, cela pourrait permettre à des professionnels de réunir des décisions relatives à un même cas pratique (p. ex. fuite d'eau, chauffage, qualité de l'air, éclairage) ou de déceler des problématiques assez rares en droit du logement (p. ex. l'accessibilité pour les personnes handicapés ou à mobilité réduite).

Pour ce qui est de notre tâche, un examen manuel des thèmes obtenus a montré que BERTopic tendait à générer des *topics* de meilleure qualité par rapport à LDA. Cela s'explique en grande partie par l'encodage à la `sentence-BERT` [**Reimers and Gurevych, 2019**] qui jouit d'une information sémantique beaucoup plus riche que les représentations sac-de-mots. Pour celles et ceux qui voudraient mener des expériences analogues aux nôtres, en particulier lorsque le corpus visé est particulièrement homogène ou spécialisé, nous recommandons ainsi fortement l'utilisation de méthodes basées sur des plongements lexicaux (*embeddings*) plutôt que sur des sacs-de-mots (approche majoritaire avec LDA qui est l'un des modèles les plus employés).

Dans le prochain chapitre, nous revisitons la tâche de prédiction de verdict avec les nouvelles informations extraites par la modélisation thématique.

Chapitre 6

L'aide à la prédiction de verdict via la modélisation thématique et la recherche d'information

6.1. Introduction

Il a été vu lors du précédent chapitre que les modèles thématiques permettent de dégager des thèmes saillants à travers l'ensemble des documents disponibles sans devoir recourir à une inspection manuelle longue et coûteuse. L'objectif du présent chapitre est de généraliser cette approche pour l'ensemble du corpus[1], et d'examiner dans quelle mesure les *topics* obtenus permettent d'améliorer les performances dans la tâche de prédiction de verdict. Le modèle faisant office de *baseline* ici est le modèle **FPTCamemBERT** qui a fait l'objet d'un *further pretraining*, comme décrit dans la Sous-Section 4.3 du Chapitre 4. Ce modèle utilise le seul texte des faits pour prédire les labels du verdict, et il sera vu ici dans quelle mesure sa performance pourrait être améliorée.

Par exemple, une première approche consiste à employer les thèmes extraits du texte des faits comme des *input features* directement donnés au modèle. Une autre approche basée sur la **recherche d'information** consiste à aider le modèle dans ses prédictions en lui présentant d'autres documents relatant des situations analogues à celle de l'instance que le modèle doit

[1]Le chapitre précédent s'attardait uniquement sur les litiges dans lesquels les locataires poursuivaient leurs propriétaires, mais non la situation inverse.

traiter (p. ex. trouver d'autres litiges analogues à une instance où le propriétaire a tardé à réparer la plomberie).

Trouver des similarités entre des documents légaux ou chercher les k premiers documents pertinents pour un document donné (p. ex. chercher une jurisprudence qui traite du litige en cours ; trouver des textes législatifs traitant des mêmes concepts qu'une loi particulière) est un travail qu'effectuent déjà les spécialistes juridiques et les magistrats au quotidien, comme l'ont souligné [**Alschner, 2019**] et [**Charmet et al., 2022**]. Une telle tâche a évidemment fait l'objet de différents travaux qui en ont sondé les possibilités d'automatisation, à l'instar du *benchmark* chinois CAIL2019-SCM de [**Xiao et al., 2019**]. Ainsi, [**Kumar et al., 2013**] ont montré que la similarité entre décisions de justice pouvait s'appuyer sur les citations et les similitudes entre paragraphes afin de relier des décisions entre elles. D'autre part, [**Moodley et al., 2019**] ont comparé l'efficacité des représentations syntactiques et sémantiques pour calculer la pertinence entre documents de la Cour de justice de l'Union européenne en se passant de graphe de citations. Cette combinaison de représentations syntactiques et sémantiques est très utilisée dans le TAL légal. C'est le cas de [**Thenmozhi et al., 2017, Ali et al., 2021**] dans le cadre d'une tâche de recherche d'information juridique. C'est aussi le cas de [**Wagh and Anand, 2020**] qui ont agrégé des concepts juridiques en un graphe pour trouver des similarités interdocuments. Par ailleurs, [**Bhattacharya et al., 2020**] ont aussi présenté les bienfaits d'utiliser à la fois des réseaux de citations et la ressemblance textuelle pour mesurer la similarité entre des documents légaux. Il est utile de préciser ici que le but du présent chapitre n'est pas en soi de déceler des similarités entre les différentes décisions du corpus, mais davantage de sonder dans quelle mesure la performance d'un prédicteur de verdict pourrait s'appuyer sur une jurisprudence pertinente vis-à-vis d'un litige à résoudre.

Trouver cette jurisprudence pertinente peut s'effectuer via des techniques de modélisation thématique ou de recherche d'information qui seront déployées dans le présent chapitre. Des travaux comme ceux de [**Luz De Araujo and De Campos, 2020**] et [**Aguiar et al., 2022**] ont utilisé des thèmes comme données d'entrée dans des tâches de classification de documents légaux. Cependant, au meilleur de nos connaissances pour ce qui est du domaine juridique, il n'existe pas de tâche de classification qui s'appuie à la fois sur la modélisation thématique et sur la recherche d'information pour déceler des

similarités interdocuments. Il existe bien le travail de [**Gong et al., 2018**] qui a employé le *topic modeling* pour trouver des similitudes entre documents, mais seulement pour des textes scientifiques et non juridiques. Pour ce qui est de la recherche d'information, [**Chalkidis and Kementchedjhieva, 2023**] avaient proposé un classifieur multilabel augmenté avec un *document retriever* pour des tâches de classification de documents juridiques et biomédicaux, avec des gains substantiels en termes de Macro F1 pour les étiquettes les moins fréquentes. Ce système de recherche d'information était pourtant uniquement basé sur une représentation sémantique des documents, sans recourir à de quelconques *topics*. Pour ce qui est du présent chapitre, différentes représentations et informations tirées des documents seront utilisées pour déceler et fournir la jurisprudence la plus pertinente vis-à-vis de chaque instance :

- une représentation sac-de-mots du texte des faits (encodage propre à Lucene[2]) ;
- une représentation sémantique du texte des faits (encodage avec un FPTCamemBERT affiné) ;
- les *topics* obtenus via le texte des faits (avec la modélisation thématique décrite dans le chapitre précédent).

Les sections suivantes décrivent plus en détail la préparation des données d'entrée.

6.2. Préparation des données

Pour l'heure, chaque instance ou litige de notre corpus comprend le texte des faits comme données d'entrée et des étiquettes cibles décrivant le verdict à prédire. Il sera décrit ici les différents outils déployés pour enrichir les données (c.-à-d. l'identification de *topics* avec la modélisation thématique qui pourraient faire office de *input features*) et pour déceler des similitudes entre les différentes instances (système de recherche d'information). In fine, l'objectif est de bâtir un répertoire associant à chaque instance un identifiant unique (chaîne de caractères), le texte des faits représenté de façon syntaxique (sacs-de-mots) ou sémantique (vecteur tiré d'un transformeur), ainsi que les thèmes y sont associés (des booléens ou des vecteurs aux valeurs continues).

[2] `https://lucene.apache.org/`

6.2.1. Génération de thèmes pour chacune des décisions

Une première partie du travail consiste à générer des thèmes à partir du texte des faits de chaque décision. Le choix de se restreindre au texte des faits s'explique, comme dit dans les précédents chapitres, par la volonté de ne pas exposer le modèle au raisonnement juridique formulé par le juge, que ce soit directement via les paragraphes dédiés ou indirectement via des thèmes tirés de ces paragraphes-là. Le texte des faits passe par le même prétraitement que dans la Sous-Section 5.2 du Chapitre 5, c'est-à-dire : un retrait des *stopwords*, le retrait des paragraphes de moins de 5 mots, le retrait des chiffres, le filtrage des *tokens* ayant certaines étiquettes grammaticales (*part-of-speech tags*), la mise en minuscule, et enfin, la lemmatisation. Un tel procédé permet de réduire la redondance morphologique des différents segments[3] de texte, ce qui permet d'éviter d'avoir un trop grand nombre de thèmes chacun associé à un très faible volume d'instances [4].

Dans le détail, le corpus comprend un total de 564,900 instances (cumulation des ensembles d'entraînement, de validation et de test). Le texte des faits de ces litiges est décomposé en 3,665,663 paragraphes. Après prétraitement et retrait des duplicatas, leur nombre s'élève à 65,674 segments[5]. Une telle réduction s'explique en grande partie par le fait que beaucoup de segments sont des phrases ou des formulations très redondantes de par le caractère routinier du droit du logement (p. ex. le cas "classique" du locataire mauvais payeur que le propriétaire souhaite faire expulser). La fusion de ces segments redondants permet non seulement de réduire le temps de calcul de la modélisation thématique, mais permet aussi d'assurer une certaine diversité dans les *topics* qui seront obtenus à la sortie.

Pour ce qui est de la création des différents thèmes, le mode opératoire est identique à celui du Chapitre 5. Les segments tirés des paragraphes du texte des faits sont donnés en entrée du modèle. Il faut néanmoins souligner certains points ou modifications ci-dessous :

---

[3]Chaque segment ou paragraphe correspond à une chaîne de caractères délimités par des sauts de ligne.

[4]L'extrême opposé consiste à avoir un très petit nombre de thèmes associés à la quasi-totalité des instances, ce qui en ferait des *topics* ne permettant pas de caractériser les instances les unes par rapport aux autres.

[5]À titre de comparaison, le chapitre précédent se concentrait uniquement sur les litiges "Locataire c. Locateur" citant des articles très précis, d'où un total de 1381 instances décomposées en 34,685 paragraphes qui se réduisent à 26,815 segments de texte donnés en entrée du modèle thématique.

- seul le modèle thématique BERTopic par [**Grootendorst, 2022**] est employé ici en raison de la qualité des thèmes qu'il avait pu dégager précédemment ;
- chaque *topic* est caractérisé par 10 termes pouvant être des unigrammes, des bigrammes et/ou des trigrammes ;
- le nombre de thèmes est ici illimité (le chapitre précédent avait une limite à 50, 100 ou 200), d'où l'importance du prétraitement en amont pour réduire la redondance morphologique des segments et ainsi avoir un nombre relativement restreint de *topics* significatifs à la sortie ;
- un thème doit être présent dans au moins 100 segments de textes, ce qui permet de réduire le risque d'avoir de nombreux *topics* couvrant très peu de segments.

Après entraînement, le modèle thématique obtenu contient un total de 738 *topics*. Il est ensuite déployé sur chaque instance du corpus de façon à attribuer, le cas échéant, un thème pour chacun des paragraphes constitutifs du texte des faits. Ainsi, chaque instance se voit attribuer des *topics* correspondant à ceux obtenus par les paragraphes qui la constituent. En moyenne, une instance comprend 4.4 thèmes (médiane à 4).

6.2.2. Mise en place d'indices avec Lucene (ElasticSearch)

Des indices (ou répertoires) sont construits grâce à Lucene implémenté via ElasticSearch[6]. Le but de ce système de recherche d'information est de trouver pour chaque instance-requête les k premiers documents-candidats qui lui ressemblent le plus. Les documents-requêtes proviennent de l'ensemble d'entraînement (litiges de 2001 à 2011), de validation (2012 à 2014), et de test (2015 à 2018). Les documents-candidats sont des instances provenant uniquement de l'ensemble d'entraînement. Cette répartition permet de formaliser la recherche d'information juridique focalisée sur la jurisprudence. En d'autres termes, pour résoudre un litige d'une année donnée, un magistrat ou avocat va chercher des cas semblables qui ont déjà été tranchés au cours des années précédentes. Lorsque l'instance-requête provient elle aussi de l'ensemble d'entraînement, nous nous assurons que la même instance n'apparaît pas parmi les documents-candidats identifiés comme les plus similaires par le système. Au cours de l'indexation de tous les documents, différentes informations accompagnent chacun d'entre eux :

[6] `https://www.elastic.co/`

- un identifiant unique (chaîne de caractères) ;
- le texte des faits encodé sous forme de sacs-de-mots internes à ElasticSearch ;
- les *topics* identifiés parmi les paragraphes constitutifs du texte des faits (chaque thème correspond à une valeur booléenne qui est vraie si elle est assignée à une instance et fausse dans le cas contraire).

Une fois les indices créés, nous cherchons pour chacune des instances-requêtes les 4 premiers documents-candidats les plus pertinents. Ce nombre peut paraître de prime abord très bas par rapport à la littérature en recherche d'information où le nombre de résultats s'élève généralement à plusieurs dizaines voire centaines. Il reflète cependant ici le fait que les experts légaux nécessitent d'étudier des précédents qui sont en faible nombre, mais très significatifs vis-à-vis du litige à résoudre. [**Bench-Capon, 2021**] déclare ainsi : "*legal decisions are primarily concerned with a single case and a handful of relevant precedents*". Par ailleurs, ne prendre que les 4 premiers résultats pour chaque instance est un choix qui est analogue à celui de [**Chalkidis and Kementchedjhieva, 2023**] dans une tâche de classification multilabel avec des classifieurs augmentés par des systèmes de recherche d'information. Les auteurs avaient trouvé que ce nombre était le plus pertinent à travers différentes tâches de classification employant des corpora légaux et biomédicaux.

Plusieurs ensembles de critères sont utilisés pour recueillir les candidats les plus similaires ou pertinents à chaque instance-requête :

- la **similarité au niveau du texte des faits** : les scores de correspondance sont calculés avec la fonction more_like_this sur la base d'une représentation sac-de-mots interne à Lucene ;
- la **similarité au niveau des *topics*** : les thèmes attribués à l'instance-requête constituent alors des valeurs booléennes qui sont intégrées dans une requête booléenne Lucene de type "*should*". Le système Lucene doit donc trouver des documents-candidats dont les *topics* correspondent le plus possible à ceux du document de départ, même s'ils n'y correspondent pas exactement[7] ;

[7]Dans une requête booléenne de type "*must*", Lucene ne retiendrait que les documents-candidats ayant exactement les thèmes demandés. Une telle approche est donc beaucoup plus restrictive et plus susceptible d'avoir beaucoup moins de résultats par rapport à une requête de type "*should*".

	Texte	Thèmes	Labels
Instance-requête avec les labels de référence	Par des recours introduits en [DATE], cinq locataires du même immeuble demandent une diminution de loyer et des dommages suite à des travaux exécutés par le locateur à l'extérieur de l'immeuble. [...]	✓peinture ✓étage ✓preuve ✓bruit ✓travaux ✓nettoyage ✓balcon ✓rénovation	✓applicant_request_denied ✓landlord_pays_tenant ✓tribunal_sets_new_rent
Document-candidat (top 1)	Le [DATE], une demande de diminution de loyer de 150$ par mois rétroactive et exécution en nature d'une obligation des locateurs, était présentée à la Régie par la locataire. [...]	✓bruit ✓travaux ✓nettoyage	✓landlord_pays_tenant ✓tribunal_sets_new_rent order_landlord_repairs verdict_related_to_peaceful_enjoyment
Document-candidat (top 2)	Par un recours introduit le [DATE] et amendé le [DATE], le [DATE] et le [DATE], les locataires demandent l'exécution en nature des obligations des locateurs, la diminution de leur loyer de 50% par mois, des dommages matériels de 475$ [...]	✓peinture ✓étage ✓preuve ✓bruit ✓balcon ✓rénovation	✓applicant_request_denied ✓landlord_pays_tenant
Prédiction du modèle FPTCamemBERT			✓applicant_request_denied
Prédiction du modèle FPTCamemBERT_IR_text_topics			✓applicant_request_denied ✓landlord_pays_tenant ✓tribunal_sets_new_rent

Tableau 6.1. Exemple d'une instance-requête avec des locataires se plaignant de travaux entrepris par le propriétaire qui rendent le logement invivable. Les documents-candidats top 1 et top 2 ont été recueillis dans l'ensemble d'entraînement sur la base de la similarité textuelle et thématique. Les informations temporelles et les thèmes sont ici tronqués par nos soins. Les thèmes et labels se recoupant avec ceux de référence sont marqués par le symbole ✓. Dans la section inférieure, le label unique prédit par le modèle FPTCamemBERT correspond à un rejet pur et simple des demandes des locataires tandis que le modèle FPT-CamemBERT_IR_text_topics prédit bien des compensations du locateur aux locataires et un ajustement du loyer.

- la **similarité au niveau du texte et des thèmes** : la similitude entre les instances-requêtes et les documents-candidats sont calculés sur la base des deux critères précédents.

Un exemple d'instance-requête avec ses deux premiers documents-candidats est présenté dans le Tableau 6.1. Les modèles qui y sont mentionnés sont expliqués dans les sections suivantes.

6.2.3. Mise en place d'un indice sur la base de représentations sémantiques

Les différentes approches à la Lucene sont essentiellement basées sur des sacs-de-mots et des thèmes comparables respectivement à des représentations syntaxiques et discrètes (booléennes). Une autre façon d'opérer la recherche d'information consiste à chercher des candidats pertinents sur la base de la proximité entre leur représentation sémantique et celle de l'instance-requête. Dans cette partie, nous reprenons le même *modus operandi* que [**Chalkidis and Kementchedjhieva, 2023**] :

(1) Un classifieur **FPTCamemBERT** est entraîné sur la tâche de prédiction de verdict multilabel (les modalités d'entraînement sont les mêmes que dans le Chapitre 4 à la Section 4.3, mais avec un *batch size* à 16 au lieu de 20). L'entraînement est répété avec cinq chiffres d'amorce[8] différents. Le modèle ayant obtenu le meilleur score en exactitude sur les instances de l'ensemble de validation est conservé pour l'encodage des documents (appelons-le l'encodeur) ;

(2) Dans une phase d'inférence (les paramètres sont gelés), cet encodeur génère la représentation sémantique de chacune des instances d'entraînement, de validation et de test. Cette représentation correspond au premier jeton (c.-à-d. le *token* [CLS]) de la sortie de la douzième et dernière couche (*last hidden state*), soit un vecteur de dimension 768 ;

(3) À l'aide de l'outil Faiss[9] (*Facebook AI Similarity Search*) par [**Johnson et al., 2019**], les 4 documents-candidats les plus similaires à chaque instance sont recherchés. Le calcul de similarité est effectué comme dans le travail de

[8] En anglais, *seed number* ou *random state*.

[9] `https://github.com/facebookresearch/faiss`

[**Chalkidis and Kementchedjhieva, 2023**] avec un produit scalaire (*inner product*[10]) entre la représentation de la requête et celle de chaque candidat. Les résultats proviennent de l'ensemble d'entraînement tandis que les requêtes proviennent des ensembles d'entraînement, de validation et de test. Lorsque l'instance-requête provient de l'ensemble d'entraînement, nous nous assurons qu'elle n'apparaît pas parmi les documents-candidats retenus pour elle.

(4) Pour récapituler, le modèle est entraîné avec des instances de l'ensemble d'entraînement accompagnés de ses plus proches voisins dans l'ensemble d'entraînement. À l'issue de chaque époque d'entraînement, le modèle est évalué avec des instances de l'ensemble de validation accompagnées par leurs plus proches voisins dans l'ensemble d'entraînement. Une fois l'entraînement achevé, le modèle est évalué avec des instances de l'ensemble de test accompagnés par leurs plus proches voisins dans l'ensemble d'entraînement. Ce protocole suit celui de [**Chalkidis and Kementchedjhieva, 2023**] et permet de refléter le fait que, pour chaque instance, le modèle a accès à la jurisprudence passée la plus pertinente.

À l'issue de cette étape, chaque instance comprend désormais le texte des faits (chaîne de caractères), la représentation vectorielle des 4 plus proches documents-candidats provenant de l'ensemble d'entraînement, et les étiquettes cibles pour le verdict.[11]

[10]Dans le détail, le produit scalaire entre deux vecteurs a et b correspond à $|a||b|cos(\theta)$ avec θ étant l'angle entre eux. La similarité cosinus (*cosine similarity*) définie par $cos(\theta)$ est une mesure communément utilisée en TAL pour comparer des plongements lexicaux, mais elle ignore la norme des vecteurs, contrairement au produit scalaire.

[11]Le verdict des documents-candidats recueillis dans le *top 4* n'est pas pris en considération par le modèle pour prédire celui de l'instance-requête. Ceci s'explique par deux points. Premièrement, le verdict de l'instance-requête est indisponible pour le modèle et le calcul de similarité ne prend en compte que le texte des faits et non celui du verdict. Deuxièmement, nous avons voulu reprendre le même système que [**Chalkidis and Kementchedjhieva, 2023**] qui l'avait déployé pour une tâche de *legal judgment prediction* définie par [**Chalkidis et al., 2019a**], et dans laquelle les étiquettes cibles des plus proches voisins sont également indisponibles au modèle.

6.3. Modèles employés

Notre *baseline* est ici **FPTCamemBERT**, déjà présenté dans la Sous-Section 4.3 du Chapitre 4. Ce modèle ayant fait l'objet d'un *further pretraining* sur l'ensemble d'entraînement[12] n'a accès qu'au texte des faits, comme illustré sur la Figure 6.1 (a). Les choix des hyperparamètres pour ce modèle et tous les autres de ce chapitre sont les mêmes que dans la Section 3.2.2 du Chapitre 3. Seul le *batch size* est modifié à 16. Cette même taille est appliquée à tous les modèles étudiés dans le présent chapitre. Une version entraînée de FPTCamemBERT, qui est appelée **FPTCamemBERT′** (prime), est incorporée dans les autres modèles décrits ci-dessous. FPTCamemBERT′ y fait office d'encodeur des documents-candidats pertinents. Les modèles assistés par la modélisation thématique et/ou par la recherche d'information sont des dérivés de FPTCamemBERT ayant accès à d'autres données d'entrée en plus du texte des faits. Ceux assistés par la recherche d'information sont les suivants :

- **FPTCamemBERT_IR_text** ("*IR*" désigne "*information retrieval*") : chaque instance est accompagnée par les 4 documents-candidats les plus similaires sur la base du texte des faits avec la fonction more_like_this de Lucene ;
- **FPTCamemBERT_IR_topics** : chaque instance est accompagnée par les 4 documents-candidats les plus similaires sur la base des thèmes qui sont représentés comme de simples valeurs booléennes ;
- **FPTCamemBERT_IR_text_topics** : il s'agit d'une combinaison des deux approches précédentes. Chaque instance est accompagnée par les 4 documents-candidats les plus similaires sur la base du texte des faits et des thèmes ;
- **FPTCamemBERT_IR_faiss** : les données d'entrée comprennent le texte des faits et les 4 plus proches documents voisins faisant office de jurisprudence. Le calcul de similarité est basé sur un produit scalaire au niveau des représentations sémantiques.

Pour chacun de ces quatre modèles, les documents-candidats les plus pertinents vis-à-vis de l'instance-requête sont retournés par le système de recherche d'information et encodés individuellement par le module FPTCamemBERT′. Ensuite, la représentation sémantique de l'instance-requête (provenant de FPTCamemBERT) et celles des documents-candidats (provenant de FPTCamemBERT′) sont envoyées respectivement en tant que requêtes et

[12]Il s'agit de la tâche non-supervisée de *masked language modeling*.

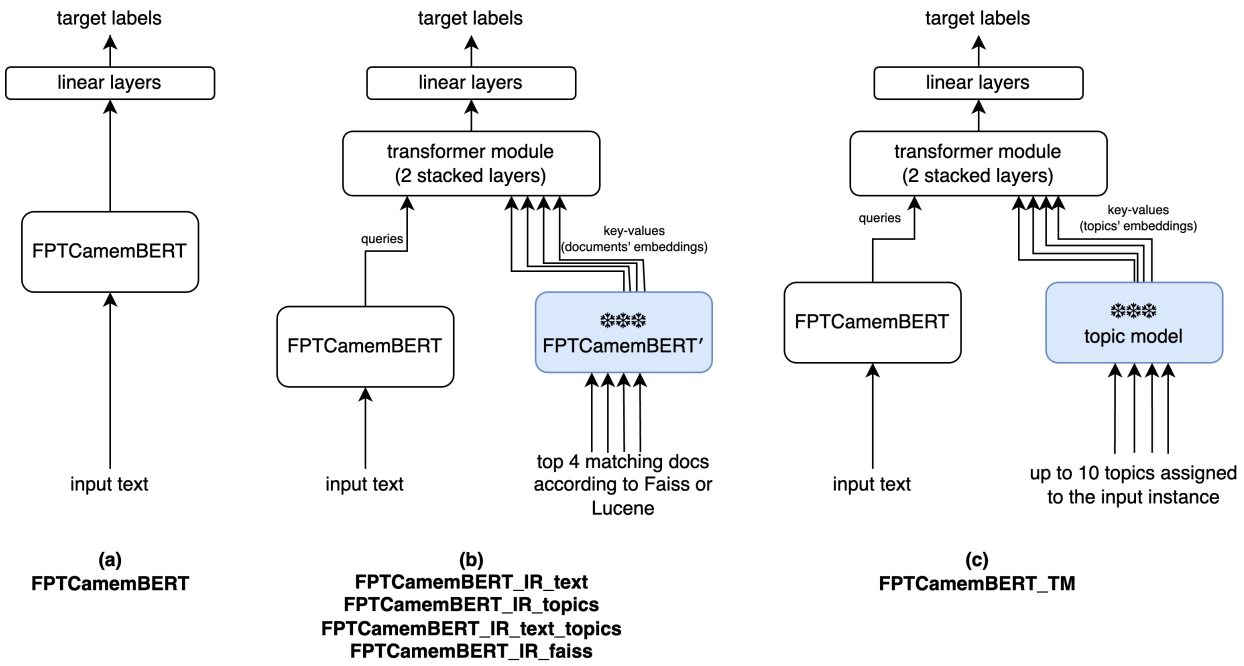

Fig. 6.1. Schéma des différents modèles avec : (a) un modèle FPTCamemBERT classique ayant le texte des faits comme seules données d'entrée ; (b) différents modèles assistés par des systèmes de recherche d'information où les 4 documents les plus proches de l'instance selon différents critères sont encodés par un FPTCamemBERT′ séparé (une version déjà affinée de FPTCamemBERT) aux paramètres gelés (inférence pure) ; (c) un autre modèle où les *embeddings* des *topics* associés à l'instance sont fournis au modèle. Les représentations des thèmes sont obtenues par inférence du modèle thématique dont les paramètres sont gelés. Les paramètres de tous les modules avec un fond blanc sont entraînables. Ceux des modules avec un fond bleu et trois flocons de neige dénotent un gel des paramètres pour une inférence pure.

clés-valeurs dans un module transformeur à deux couches. Ce module est analogue à celui utilisé dans le modèle FPTCamemBERT-TXT dans la Section 4.3 du Chapitre 4. Bien que ces quatre modèles utilisent la même architecture illustrée sur la Figure 6.1 (b), c'est la façon dont la jurisprudence pertinente est récupérée qui varie entre eux. Ainsi, certains modèles s'appuient sur des représentations sémantiques ou syntaxiques des documents dans leurs critères de recherche, tandis que d'autres incorporent des critères liés aux thèmes extraits.

Un dernier modèle non basé sur la recherche d'information mais s'appuyant uniquement sur la modélisation thématique s'intitule **FPTCamemBERT_TM** ("*TM*" signifie ici "*topic model*"). Le modèle a accès jusqu'à un maximum de 10 thèmes[13] qui ont pu être identifiés pour chacune des instances lors de la modélisation thématique décrite à la Section 6.2.1. La représentation de chaque thème correspond à un vecteur de taille 384 qui est la moyenne de toutes les représentations des segments de textes ayant été regroupés dans le même *topic*. La représentation de chaque segment correspond à l'encodage obtenu à la sortie du modèle `sentence-BERT` [**Reimers and Gurevych, 2019**] [14] en inférence. Ainsi, pour chaque instance, c'est une représentation sémantique du texte des faits et une représentation sémantique des thèmes qui sont injectés respectivement comme requêtes et paires de clés-valeurs dans le module transformeur illustré dans la Figure 6.1 (c).

6.4. Résultats

Chaque modèle est exécuté à cinq reprises avec différents chiffres d'amorce. Les résultats sont présentés dans le Tableau 6.2. À première vue, il n'existe pas de modèle excellant pour toutes les métriques utilisées. Le modèle FPTCamemBERT_TM s'appuyant uniquement sur un modèle thématique ne fait pas mieux que la *baseline* FPTCamemBERT, sauf pour le Macro F1 et le score d'exagération. Pour ce qui est des modèles s'appuyant sur un système de recherche d'information (avec l'affixe "IR"), ils peinent à se démarquer de la *baseline* en termes de Micro F1 et d'exactitude. Il faut néanmoins souligner que les variantes FPTCamemBERT_IR_text_topic et FPTCamemBERT_IR_text_topics se distinguent surtout à travers un gain significatif d'environ 4.5 points de pourcentage pour la métrique Macro F1. Le même gain en Macro F1 est aussi observé pour FPTCamemBERT_IR_text et FPTCamemBERT_IR_faiss. Pour ce dernier modèle, il faut aussi souligner le meilleur score d'exagération avec un minimum de 7.66.

[13] Comme dit précédemment, une instance a en moyenne 4.4 thèmes.

[14] Pour rappel, étant donné que notre corpus est en français, c'est un modèle de langage `sentence-transformer` multilingue qui a été employé : `https://huggingface.co/sentence-transformers/paraphrase-multilingual-MiniLM-L12-v2`

Modèle	↑ Micro F1	↑ Macro F1	↑ Exactitude	↓ Score d'exagération
FPTCamemBERT	**88.57 (0.19)**	56.6 (2.94)	**49.9 (0.54)**	8.36 (1.21)
FPTCamemBERT_IR_text	88.19 (0.42)	60.87* (0.48)	48.35* (0.75)	8.03 (0.38)
FPTCamemBERT_IR_topics	88.52 (0.31)	**61.15* (0.66)**	49.30 (0.61)	<u>7.96 (0.55)</u>
FPTCamemBERT_IR_text_topics	88.48 (0.17)	<u>61.10* (0.53)</u>	<u>49.49 (0.74)</u>	8.11 (0.75)
FPTCamemBERT_IR_faiss	<u>88.54 (0.09)</u>	59.84 (1.59)	49.19 (0.35)	**7.66 (0.24)**
FPTCamemBERT_TM	88.44 (0.14)	60.76* (0.56)	49.2 (0.46)	8.16 (0.54)

Tableau 6.2. Scores moyens par modèles sur la base de 5 *runs*. L'écart-type est entre parenthèses. Le meilleur et le deuxième meilleur score de chaque colonne sont respectivement en gras et souligné. Les astérisques dénotent les différences par rapport à FPTCamemBERT pour une valeur p inférieure à 0.05.

De façon générale, l'ajout des seuls thèmes dans les données d'entrée (FPTCamemBERT_TM) n'apporte pas d'amélioration notable par rapport à un modèle ayant uniquement accès au texte des faits, sauf pour le Macro F1. Concernant les modèles assistés par la recherche d'information (avec l'affixe "IR"), les performances stagnent ou diminuent en termes de Micro F1 et d'exactitude, mais sont beaucoup plus importants en termes de Macro F1 et de score d'exagération. Ce phénomène (gains mitigés en Micro F1, mais bien meilleurs en Macro F1) est concordant avec les résultats rapportés par [**Chalkidis and Kementchedjhieva, 2023**] pour les domaines légal et biomédical. Concrètement, cela signifie que le fait de fournir au modèle les 4 documents-voisins les plus proches de l'instance traitée permet d'améliorer significativement les performances pour les labels rares et peu fréquents, ce qui va relever le score Macro F1[15]. Cette amélioration pour les labels dans la traîne de la distribution fait aussi que les modèles ont moins tendance à systématiquement retourner le verdict majoritaire (c.-à-d. pénalités et expulsion contre les locataires ; refus de la demande des locataires par le tribunal), ce qui se traduit par une

[15]Pour rappel, le score Macro F1 est une moyenne non-pondérée entre les scores F1 obtenus pour chacun des labels de verdict. Cela implique qu'un modèle avec des scores F1 très déséquilibrés entre des labels aux distributions très différentes sera pénalisé par cette métrique.

Modèle	↑ Micro F1	↑ Macro F1	↑ Exactitude	↓ Score d'exagération
Propriétaire c. Locataire				
FPTCamemBERT	**89.47**	52.91	**50.90**	7.48
FPTCamemBERT_IR_text	89.06	56.17	49.22	7.48
FPTCamemBERT_IR_topics	89.40	<u>56.79</u>	50.24	<u>7.33</u>
FPTCamemBERT_IR_text_topics	89.36	**56.92**	<u>50.48</u>	7.59
FPTCamemBERT_IR_faiss	<u>89.43</u>	56.01	50.11	**6.78**
FPTCamemBERT_TM	89.33	56.24	50.15	7.54
Locataire c. Propriétaire				
FPTCamemBERT	69.09	30.43	38.13	18.72
FPTCamemBERT_IR_text	69.28	<u>34.66*</u>	38.09	<u>14.57</u>
FPTCamemBERT_IR_topics	**69.48**	**34.88***	<u>38.18</u>	15.37
FPTCamemBERT_IR_text_topics	**69.48**	34.49*	37.88	**14.23**
FPTCamemBERT_IR_faiss	68.85	32.88	**38.31**	18.06
FPTCamemBERT_TM	<u>69.38</u>	34.29*	<u>38.05</u>	15.42

Tableau 6.3. Scores moyens par type de litige et par modèle sur la base de 5 *runs*. Le meilleur et le deuxième meilleur score de chaque colonne sont respectivement en gras et souligné. Les scores avec un astérisque * dénote une différence significative par rapport à FPTCamemBERT.

baisse dans les scores d'exagération. Enfin, il est utile de souligner que l'assistance de la recherche d'information permet de réduire fortement l'écart-type de la performance du modèle en termes de Macro F1 (0.82 en moyenne pour les modèles estampillés "*IR*" contre 2.94 pour FPTCamemBERT) et de score d'exagération (0.48 contre 1.21).

6.4.1. Quels critères de recherche devraient privilégier les systèmes de recherche d'information au moment de collecter la jurisprudence la plus pertinente à chaque instance ?

Les résultats obtenus constituent l'occasion de comparer différents critères de recherche de documents-voisins dans le cas des modèles assistés par la recherche d'information. Sur la base du Macro F1, inclure les *topics* dans les critères de recherche de Lucene avec ou sans le texte des faits FPTCamemBERT_IR_topics, FPTCamemBERT_IR_text_topics) semble donner les meilleurs gains, suivi par le texte seul (FPTCamemBERT_IR_text). Si le critère de proximité sur la base des représentations sémantiques des textes (FPTCamemBERT_IR_faiss) se distingue peu par rapport à un more_like_this de Lucene appliqué sur des sacs-de-mots (FPTCamemBERT_IR_text), il faut souligner que cette option reste compétitive en termes de temps de calcul. En effet, les étapes d'encodage, d'indexation et de recherche des instances dans FPTCamemBERT_IR_faiss prend tout au plus quelques heures. De son côté, Lucene nécessitera plusieurs jours voire une semaine pour ces mêmes étapes[16].

Le détail des scores du Tableau 6.2 par type de litige est présenté dans le Tableau 6.3. L'amélioration significative fu du Macro F1 s'explque en grande partie par Il faut cependant souligner qu'aucun des modèles assistés par des systèmes de recherche d'information ne permet encore de combler l'écart de performance encore très important entre les litiges initiés par les propriétaires et ceux initiés par les locataires.

6.5. Conclusion

Au terme de ce chapitre, il a été vu que les modèles assistés par un système de recherche d'information recueillant la jurisprudence la plus pertinente à chaque instance ont des gains importants en termes de Macro F1 et de scores d'exagération. Cela implique que l'injection de documents voisins similaires aux instances traitées permet d'avoir de meilleures performances pour les labels de verdict rares se trouvant dans la traîne de la distribution. Les critères de

[16]Dans le détail, la recherche des instances avec Lucene est l'étape la plus longue pour laquelle nous avons utilisé une machine avec 64 CPUs afin de traiter les requêtes en parallèle. Malgré cette parallélisation, cette opération reste bien plus longue qu'avec Faiss utilisé pour FPTCamemBERT_faiss.

recherche permettant de tels gains sont la similarité au niveau du texte (sacs-de-mots) et des thèmes attribués à l'instance. Une telle approche apporte cependant des résultats mitigés pour ce qui est du Micro F1 et de l'exactitude. Il faut enfin noter que cette assistance basée sur la recherche d'information, bien que permettant des gains substantiels pour les labels rares, ne permet pas de combler les écarts de performance encore très importants entre les litiges initiés par les propriétaires et ceux initiés par les locataires.

Conclusion

Récapitulatif des chapitres de la thèse

Cette thèse a été l'occasion d'explorer de bout en bout une tâche de TAL appliquée à un corpus de textes juridiques. Le Chapitre 1 présente l'étendue des différents corpora disponibles dans le domaine légal ainsi que différents travaux déjà entrepris au croisement du TAL et du droit. Le Chapitre 2 analyse en détail le corpus du Tribunal administratif du logement du Québec, en soulevant différents biais et certaines particularités métier des documents. Il s'attarde aussi et surtout sur les précautions prises et la difficulté de la mise en forme de telles données pour la tâche de prédiction de verdict sous forme de classification multilabel.

Dans le Chapitre 3, différents modèles neuronaux et non-neuronaux sont employés pour la tâche de prédiction de verdict sous la forme de classification multilabel. Le meilleur modèle s'avère alors être un modèle CamemBERT. Par la suite, le Chapitre 4 s'attarde sur l'amélioration des performances de CamemBERT qui fait préalablement l'objet d'un pré-entraînement non-supervisé étendu (*further pretraining* avec la tâche de *masked language modeling*). Cette partie examine en particulier dans quelle mesure les articles de loi, des connaissances tirées du domaine cible (droit du logement), peuvent aider la prédiction lorsqu'ils sont intégrés comme *input features* additionnelles. En supposant que tous les articles pertinents à une instance soient parfaitement identifiés, il existe des gains de performance substantiels. Ils doivent cependant être relativisés lorsque les articles font l'objet d'une prédiction au lieu d'être fournis par un oracle.

Suite à ces travaux axés sur la classification multilabel, le Chapitre 5 propose une tâche de modélisation thématique permettant de mieux scruter le contenu même des décisions. Une telle tâche permet de dégager rapidement une carte des différentes situations motivant

des locataires à poursuivre leur propriétaire en justice. L'alliance entre un prétraitement fin et l'utilisation de BERTopic s'avère particulièrement utile pour identifier les cas litigieux les plus courants comme les plus rares. Enfin, le Chapitre 6 reprend la meilleure approche de modélisation thématique du Chapitre précédent, et la généralise à l'ensemble du corpus. Chacune des instances est ainsi accompagnée par des *topics* et est incluse dans le répertoire d'un système d'information. Avec un tel dispositif, la tâche de prédiction de verdict est revisitée avec des modèles assistés par des modules de recherche d'information (RI) et de modélisation thématique. Ces modules permettent ainsi de recueillir la jurisprudence la plus pertinente vis-à-vis de chaque litige à traiter (les 4 documents les plus similaires). Les gains de performance sont surtout substantiels en termes de score F1 Macro lorsque les critères de RI s'appuient à la fois sur la similarité au niveau du texte (sacs-de-mots avec Lucene) et des *topics*. Malgré cela, de fortes disparités de performance persistent selon qu'un litige a été initié par un propriétaire ou un locataire.

No data? No task! : De la question sous-estimée de l'accessibilité des données juridiques

Alors que la plupart des travaux de TAL s'appuient sur des corpora relativement faciles d'accès et distribuables, la présente thèse est basée sur un corpus de décisions de justice en français qui ne peut pas faire l'objet d'une diffusion au sein de la communauté. Cette dichotomie mérite un développement concernant les enjeux d'accessibilité des données juridiques qui sont largement sous-estimés.

Les travaux à l'intersection du droit et du TAL existent depuis au moins trois décennies, la première édition de la *International Conference on Artificial Intelligence and Law* remontant à 1987. Il est aussi possible de remonter aux années 1970 avec la mise en place de DATUM, un moteur de recherche juridique mis en place par le Centre de recherche en droit public (CRDP) et le Centre de calcul de l'Université de Montréal.[17] Depuis lors et notamment à partir de 2016, les publications associant droit et informatique ont fortement augmenté en volume depuis 2016, notamment pour l'anglais et le mandarin, comme l'ont constaté [**Katz et al., 2023**]. Ces derniers se réjouissent d'ailleurs de la plus grande diffusion de

[17]Un grand merci à Guy Lapalme pour avoir partagé cette information ainsi que le lien suivant : `https://blogue.soquij.qc.ca/2016/04/07/debut-y-eut-datum/`.

ressources (jeu de données et code) à des fins de reproductibilité des résultats, mais omettent de souligner les difficultés inhérentes à l'accès aux données juridiques. Dans le domaine de l'apprentissage automatique, les praticiens tiennent souvent pour acquis que les corpora sont libres d'accès (absence de péage ou de besoin d'inscription à un registre d'utilisateurs autorisés), et réutilisables à souhait (les données peuvent être rediffusées et redistribuées pour la communauté suite à un traitement préalable). Ces deux hypothèses sont le plus souvent fausses dans le cas des données ayant trait au domaine juridique. Ainsi, dans le cas du corpus utilisé dans cette thèse, les différentes décisions sont accessibles sur internet via un portail dédié[18], mais des barrières techniques (dispositif anti-*scraping*) et légales (l'organisme judiciaire, législatif ou l'éditeur légal garde un monopole exclusif) font que c'est une ressource difficile à exploiter et à diffuser au sein de la communauté de recherche. Un accord entre un centre de recherche à la faculté de droit et le Tribunal administratif du logement a été nécessaire afin de pouvoir accéder au corpus entier de décisions (nous remercions ce laboratoire de nous avoir accordé ce privilège). Ces différentes barrières techniques et légales ont déjà été soulevées par [**Rubinfeld and Gal, 2017**]. La question de l'accessibilité des données juridiques demeure cependant largement sous-estimée dans la communauté alors qu'elle conditionne les possibilités de recherche et de reproductibilité des résultats. Cette situation s'explique par deux principes contradictoires décrits par [**Benatti et al., 2022**] :

- d'une part, les tribunaux et les cours sont tenus de rendre leurs décisions publiques pour garantir la transparence du fonctionnement des institutions ("*state accountability*") et pour garantir le droit à l'information qui est fondamental dans les systèmes démocratiques ;
- de l'autre, la communauté des chercheurs doit s'assurer de produire des travaux reproductibles en redistribuant le corpus utilisé.

Le principal point de friction se situe au niveau du traitement des **données personnelles** contenues dans les documents légaux. Ces données peuvent se retrouver exposées à tous lorsque l'organisme judiciaire rend ses décisions publiques au nom du droit à l'information.

[18] `https://www.canlii.org/fr/qc/qctal/`

Mais elles peuvent aussi rendre des personnes vulnérables encore plus exposées lorsque le corpus les contenant fait l'objet d'une redistribution au sein de la communauté de recherche. Sur cette question, [**Leins et al., 2020**] et [**Benatti et al., 2022**] promeuvent une approche protectrice (certains diront conservatrice) des données personnelles, en suggérant des lignes directrices éthiques (évaluation et atténuation du risque, le cas échéant, avant diffusion ou non-diffusion du corpus). D'autres chercheurs tels que [**Tsarapatsanis and Aletras, 2021**] estiment que cette question éthique touchant au TAL légal ne peut être résolue de par la diversité des différentes normes éthiques et légales à travers la communauté. Ils avancent notamment que ce questionnement éthique, bien que légitime, ne doit pas se transformer en moralisme, ni contrevenir à la liberté du chercher en TAL légal, au risque d'appauvrir le domaine.

Une solution évidente pour répondre à ces problèmes éthiques, et ainsi permettre une plus grande accessibilité des données juridiques pour la communauté des chercheurs, reposerait dans l'**anonymisation** à la source des documents avant publication par les cours et tribunaux. Cependant, un tel procédé est très loin d'être trivial et implique l'utilisation d'importantes ressources. [**Girard-Chanudet, 2023**] avait ainsi montré la difficulté des décisions éditoriales prises par des experts de différents corps de métiers (informaticien, juriste, annotateur) au sein de la Cour de cassation française. À cela, il faut ajouter l'analyse de [**Gruodyte and Milciuviene, 2018**] qui souligne l'hétérogénéité des pratiques entre pays qui peuvent choisir de privilégier davantage le droit à l'information par rapport à la protection des données personnelles ou vice-versa. De telles différences se répercutent sur la disponibilité et l'accessibilité des documents visés, sans compter le fait que les critères de transparence/protection des données personnelles appliqués dans une région pourraient s'opposer à celles appliquées dans une autre.

Il faut enfin souligner un **paradoxe** concernant l'accessibilité des corpora juridiques. Lorsqu'un document provient d'une Cour suprême ou d'un organisme judiciaire qui a pour but de faire respecter des lois placées au plus haut de la hiérarchie des normes[19] (p. ex. la Cour européenne des droits de l'homme, la Cour suprême des États-Unis), il a davantage de

[19]La notion de hiérarchie des normes provient du juriste Hans Kelsen (1881-1973) qui évoquait une pyramide des normes constituée du sommet à la base par : la Constitution, les accords internationaux, les lois votées par les législateurs, les décrets et règlements du pouvoir exécutif, les actes administratifs par les autorités locales. Chaque "étage" doit se conformer à celui qui lui est supérieur.

chances d'être accessible et redistribué par les chercheurs en TAL légal [**Xiao et al., 2018, Aletras et al., 2016, Chalkidis et al., 2019a**]. En effet, plus le niveau juridique du litige est important, plus la décision a de chances d'être publique et accessible en vertu du droit à l'information. C'est ainsi que de nombreuses tâches (dont la prédiction de verdict) sont principalement effectuées sur des décisions concernant des questions de droit constitutionnel ou fondamental. Cependant, l'activité des juges suprêmes est peu susceptible de faire l'objet d'une quelconque automatisation, du fait du caractère inédit (les litiges n'ont pu être résolus de façon satisfaisante par les tribunaux de rang inférieur) et sensible (les décisions des cours supérieures ont une influence sur la pratique des tribunaux de moindre rang) des litiges qui leur sont soumis. Il faut aussi ajouter que ce type de litige est généralement assez complexe pour les citoyens ordinaires[20]. À l'inverse, lorsqu'un document provient d'un tribunal gérant des litiges dans un périmètre légal très délimité et prévisible qui obéit à des lois en bas de la hiérarchie des normes (p. ex. le Tribunal administratif du logement, la Division des petites créances de la Cour du Québec), les documents sont plus difficiles d'accès (ex. : format papier au lieu de numérique, portail en ligne avec consultation au compte-gouttes, besoin de négocier avec les organismes détenteurs des données), et donc moins susceptibles de faire l'objet de travaux et de redistribution au sein de la communauté. Dans le même temps, ces documents-là sont plus susceptibles d'être compréhensibles pour les citoyens car plus proches de préoccupations qui leur sont familières (ex. : droit du logement pour les litiges liés aux baux de location, droit de la consommation pour les questions de garantie).

Il existe ainsi une relation entre le niveau hiérarchique de la cour ayant fourni un corpus juridique et le degré d'accessibilité de ce dernier : plus la cour est importante, plus ses décisions sont complexes, plus celles-ci sont accessibles car plus susceptibles d'être rendues publiques à grande échelle au nom du droit à l'information. À l'inverse, les documents émis par de plus petits tribunaux traitent de questions légales plus simples, en plus grand volume, mais font l'objet d'une diffusion beaucoup plus restreinte. Ce paradoxe revêt des enjeux et conséquences importantes en ceci que les potentialités du TAL juridique sont surtout illustrées par des travaux davantage focalisés sur des tribunaux de haut rang, des questions juridiques complexes, en majorité en langue anglaise, souvent dans le système de *common*

[20]Par exemple, les décisions Roe c. Wade et Dobbs c. Jackson Women's Health Organization de la Cour suprême des États-Unis, bien que très célèbres de par leurs influences sur le droit à l'avortement, comportent des technicités juridiques qui peuvent être difficiles à saisir pour le grand public.

law anglo-saxon (alors que les pays non anglophones s'appuient davantage sur un système de *civil law* ou droit continental), et pour des documents disponibles en format numérique (il existe de nombreuses décisions à l'état de documents imprimés qui font au mieux l'objet d'une numérisation de qualité variable). Sans vouloir encourager un décloisonnement total des données juridiques (éthiquement non souhaitable), il pourrait être bénéfique pour la communauté de TAL juridique d'avoir un meilleur accès aux corpora légaux des tribunaux d'un niveau moindre, de façon à pouvoir développer des modèles sur des questions légales plus simples et plus diversifiées (pas seulement des questions constitutionnelles ou criminelles en anglais), avant de procéder à des travaux sur des litiges juridiques plus complexes (p. ex. relatifs à du droit constitutionnel ou du droit communautaire européen). En ce qui nous concerne, nous avons tenté de combler ce manque de diversité avec d'une part, les publications ayant inspiré la présente thèse, et d'autre part un jeu de données multilingue européen pour la tâche de génération de mots-clés [**Salaün et al., 2024**]. Ce travail fait présentement l'objet d'une soumission.

De l'avenir du TAL juridique : au-delà de la tâche de prédiction de verdict

La thèse s'est principalement focalisée sur la tâche de prédiction de verdict formalisée sous la forme d'une classification multilabel ayant fait l'objet de plusieurs contributions [**Salaün et al., 2020, Salaün et al., 2021, Salaün et al., 2021**]. Différentes connaissances tirées du domaine cible (articles de loi touchant au droit du logement, *topics* extraits avec des modèles thématiques [**Salaün et al., 2022a**]) permettent d'améliorer les performances des classifieurs, notamment pour ce qui est des labels cibles peu fréquents. Elles permettent aussi d'atténuer les amplifications de biais préexistant dans les données et persistant dans les sorties des modèles employés. Il n'en demeure pas moins que ces biais (ex. : locataires ayant moins souvent gain de cause par rapport aux propriétaires) sont amplifiés par les modèles : les classifieurs tendent en effet à prédire des verdicts plus sévères au détriment des locataires, par rapport au verdict humain attendu. De telles observations conduisent à penser qu'il est préférable que de tels prédicteurs de verdict ne soient pas déployés à grande échelle dans le monde réel où ils risqueraient d'amplifier des biais déjà présents. Contrairement à d'autres tâches de prédiction de verdict qui se contentent le plus

souvent de rapporter des métriques de classification binaire (ce qui simplifie la réalité du travail du juge et en donne une vision assez superficielle), notre protocole expérimental a mobilisé des connaissances métier fines qui ont permis de mettre en relief les enjeux sociétaux contenus dans les litiges en droit du logement. Par la même occasion, ils soulignent aussi les enjeux de justice et d'équité qui pèseraient sur des systèmes automatiques devant imiter le raisonnement juridique des magistrats de chair et d'os. À ce jour, les modèles existants délivrent une performance insatisfaisante pour des questions juridiques aussi simples et encadrées que celles relatives au droit du logement. Il est donc peu envisageable qu'ils soient un jour déployés comme assistants juridiques auprès du grand public.

À ce constat d'une performance en deçà de la performance humaine, il faut aussi ajouter que les modèles de langage à la BERT ne permettent pas d'avoir des explications sur le cheminement rationnel aboutissant au verdict, ce qui est pourtant primordial en droit selon [**Bench-Capon, 2021, Bex and Prakken, 2021**]. Aussi, il faut ajouter que les lois changent avec le temps, et qu'un modèle entraîné sur des décisions passées peut cesser d'être pertinent pour des litiges récents avec une législation modifiée[21].

Mis à part cette "déconvenue" du côté de la justice prédictive, l'application de l'apprentissage machine et du TAL reste tout de même pertinente pour accroître l'accessibilité et la lisibilité des documents juridiques. Le Chapitre 5 et notre contribution [**Salaün et al., 2022a**] montrent ainsi que la modélisation thématique permet de déceler des motifs de litiges plus efficacement qu'une lecture manuelle. Pour ce qui est des applications à destination du grand public, telles que le résumé automatique de décision pour les profanes [**Salaün et al., 2022b**], les systèmes actuels nécessitent encore cependant d'être améliorés et éprouvés avant d'envisager un quelconque déploiement à destination des non-experts. D'un autre côté, pour ce qui est des applications destinées aux experts légaux, la génération automatique de mots-clés à partir de documents juridiques volumineux et longs est un cas d'usage pratique mature pour lequel il existe une demande métier concrète, comme l'a montré notre contribution avec la société Lexum [**Cérat et al., 2023**]. Notre travail [**Salaün et al., 2024**], en cours de soumission et relatif à un corpus juridique multilingue

[21] À l'heure d'écrire ces lignes, nous avons ainsi appris que les règles encadrant les cessions de bail sont sur le point d'être modifiées : `https://www.ledevoir.com/politique/quebec/802583/article-projet-loi-31-cessions-bail-bientot-adopte`

abonde aussi en ce sens. Finalement, à défaut d'avoir des systèmes pouvant égaler des magistrats et des juristes dans la capacité de raisonnement juridique[22], le TAL appliqué aux textes légaux reste néanmoins pertinent dans tout ce qui peut aider les professionnels du droit à consulter et traiter plus efficacement le volume de jurisprudence qui croît d'année en année.

[22]Il est à notre humble avis pas plus mal qu'il en soit ainsi, pour éviter de céder à la tentation d'automatiser une activité aussi sensible que la justice. Quelques lectures pertinentes à ce sujet : [**Commission du droit de l'Ontario, 2020**] et [**Kempf, 2023**].

Références bibliographiques

[Aguiar et al., 2022] Aguiar, A., Silveira, R., Furtado, V., Pinheiro, V., and Neto, J. A. M. (2022). Using topic modeling in classification of brazilian lawsuits. In *International Conference on Computational Processing of the Portuguese Language*, pages 233–242. Springer.

[Aletras and Stevenson, 2013] Aletras, N. and Stevenson, M. (2013). Evaluating topic coherence using distributional semantics. In *Proceedings of the 10th International Conference on Computational Semantics (IWCS 2013)–Long Papers*, pages 13–22.

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

[Ali et al., 2021] Ali, B., More, R., Pawar, S., and Palshikar, G. K. (2021). Prior Case Retrieval using Evidence Extraction from Court Judgements. In *Proceedings of the Fifth Workshop on Automated Semantic Analysis of Information in Legal Text (ASAIL 2021), June 25, 2021, São Paulo, Brazil*.

[Alschner, 2019] Alschner, W. (2019). Sense and similarity: Automating legal text comparison. *Computational Legal Studies: The Promise and Challenge of Data-driven Research, Edward Elgar (Forthcoming, 2020)*.

[Ashley and Brüninghaus, 2009] Ashley, K. D. and Brüninghaus, S. (2009). Automatically classifying case texts and predicting outcomes. *Artificial Intelligence and Law*, 17(2):125–165.

[Assemblée nationale, 2016] Assemblée nationale (2016). Loi sur la régie du logement, RLRQ c R-8.1. <`https://canlii.ca/t/69m68`> consulté le 2022-10-30.

[Assemblée nationale, 2018] Assemblée nationale (2018). Code civil du Québec, RLRQ c CCQ-1991. <`https://canlii.ca/t/6b4rq`> consulté le 2022-10-30.

[Barthe, 2017] Barthe, E. (2017). L'intelligence artificielle et le droit. *I2D–Information, données & documents*, 54(2):23–24.

[Benatti et al., 2022] Benatti, R. M., Villarroel, C. M., Avila, S., Colombini, E. L., and Severi, F. (2022). Should i disclose my dataset? caveats between reproducibility and individual data rights. In *Proceedings of the Natural Legal Language Processing Workshop 2022*, pages 228–237.

[Bench-Capon, 2021] Bench-Capon, T. (2021). The need for good old fashioned ai and law. *Jusletter IT*, pages 23–35.

[Benyekhlef and Zhu, 2018] Benyekhlef, K. and Zhu, J. (2018). Intelligence artificielle et justice: justice prédictive, conflits de basse intensité et données massives. *Intelligence*, 30(3).

[Benyekhlef and Zhu, 2020] Benyekhlef, K. and Zhu, J. (2020). At the intersection of odr and artificial intelligence: Traditional justice at the crossroads. *Lex Electronica*, 25:34.

[Bertalan and Ruiz, 2019] Bertalan, V. G. and Ruiz, E. E. S. (2019). Using topic modeling to find main discussion topics in brazilian political websites. In *Proceedings of the 25th Brazilian Symposium on Multimedia and the Web*, pages 245–248.

[Bex and Prakken, 2021] Bex, F. and Prakken, H. (2021). On the relevance of algorithmic decision predictors for judicial decision making. In *Proceedings of the Eighteenth International Conference on Artificial Intelligence and Law*, pages 175–179.

[Bhattacharya et al., 2020] Bhattacharya, P., Ghosh, K., Pal, A., and Ghosh, S. (2020). Methods for computing legal document similarity: A comparative study. *arXiv preprint arXiv:2004.12307*.

[Bird et al., 2009] Bird, S., Klein, E., and Loper, E. (2009). *Natural language processing with Python: analyzing text with the natural language toolkit*. " O'Reilly Media, Inc.".

[Blei et al., 2003] Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022.

[Bouma, 2009] Bouma, G. (2009). Normalized (pointwise) mutual information in collocation extraction. *Proceedings of GSCL*, 30:31–40.

[Branting et al., 2020] Branting, K., Balhana, C., Pfeifer, C., Aberdeen, J. S., and Brown, B. (2020). Judges are from mars, pro se litigants are from venus: Predicting decisions from lay text. In *JURIX*, pages 215–218.

[Carter et al., 2016] Carter, D. J., Brown, J., and Rahmani, A. (2016). Reading the high court at a distance: topic modelling the legal subject matter and judicial activity of the high court of australia, 1903-2015. *University of New South Wales Law Journal*, 39(4):1300–1354.

[Cérat et al., 2023] Cérat, B., Salaün, O., Jillali, N. B., Morissette, M.-A., Pocovnicu, I., Elliot, E., and Harvey, F. (2023). LexKey: A Keyword Generator for Legal Documents. In *Proceedings of the Sixth Workshop on Automated Semantic Analysis of Information in Legal Text (ASAIL 2023)*.

[Chalkidis et al., 2019a] Chalkidis, I., Androutsopoulos, I., and Aletras, N. (2019a). Neural legal judgment prediction in english. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4317–4323.

[Chalkidis et al., 2019b] Chalkidis, I., Fergadiotis, E., Malakasiotis, P., and Androutsopoulos, I. (2019b). Large-scale multi-label text classification on EU legislation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6314–6322, Florence, Italy. Association for Computational Linguistics.

[Chalkidis et al., 2021a] Chalkidis, I., Fergadiotis, M., and Androutsopoulos, I. (2021a). MultiEURLEX - a multi-lingual and multi-label legal document classification dataset for zero-shot cross-lingual transfer. In

Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, pages 6974–6996, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

[Chalkidis et al., 2020] Chalkidis, I., Fergadiotis, M., Malakasiotis, P., Aletras, N., and Androutsopoulos, I. (2020). Legal-bert:"preparing the muppets for court'". In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Findings*, pages 2898–2904.

[Chalkidis et al., 2021b] Chalkidis, I., Jana, A., Hartung, D., Bommarito, M. J., Androutsopoulos, I., Katz, D. M., and Aletras, N. (2021b). Lexglue: A benchmark dataset for legal language understanding in english. *Available at SSRN 3936759*.

[Chalkidis and Kementchedjhieva, 2023] Chalkidis, I. and Kementchedjhieva, Y. (2023). Retrieval-augmented multi-label text classification. *arXiv preprint arXiv:2305.13058*.

[Chang et al., 2009] Chang, J., Gerrish, S., Wang, C., Boyd-Graber, J., and Blei, D. (2009). Reading tea leaves: How humans interpret topic models. *Advances in neural information processing systems*, 22.

[Charmet et al., 2022] Charmet, T., Cherichi, I., Allain, M., Czerwinska, U., Fouret, A., Sagot, B., and Bawden, R. (2022). Complex labelling and similarity prediction in legal texts: Automatic analysis of france's court of cassation rulings. In *LREC 2022-13th Language Resources and Evaluation Conference*.

[Charrow and Charrow, 1979] Charrow, R. P. and Charrow, V. R. (1979). Making legal language understandable: A psycholinguistic study of jury instructions. *Columbia law review*, 79(7):1306–1374.

[Collenette et al., 2020] Collenette, J., Atkinson, K., and Bench-Capon, T. (2020). An explainable approach to deducing outcomes in european court of human rights cases using adfs. In *Computational Models of Argument*, pages 21–32. IOS Press.

[Commission du droit de l'Ontario, 2020] Commission du droit de l'Ontario (2020). Essor et déclin des algorithmes dans la justice pénale des États-unis : quels enseignements pour le canada?

[Dale, 2019] Dale, R. (2019). Law and word order: Nlp in legal tech. *Natural Language Engineering*, 25(1):211–217.

[Devlin et al., 2019] Devlin, J., Chang, M.-W., Lee, K., and 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, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

[Douka et al., 2021] Douka, S., Abdine, H., Vazirgiannis, M., El Hamdani, R., and Restrepo Amariles, D. (2021). JuriBERT: A masked-language model adaptation for French legal text. In *Proceedings of the Natural Legal Language Processing Workshop 2021*, pages 95–101, Punta Cana, Dominican Republic. Association for Computational Linguistics.

[Dragoni et al., 2016] Dragoni, M., Villata, S., Rizzi, W., and Governatori, G. (2016). Combining nlp approaches for rule extraction from legal documents. In *Proceedings of 1st Workshop on MIning and REasoning with Legal texts (MIREL 2016)*.

[Duan et al., 2019] Duan, X., Wang, B., Wang, Z., Ma, W., Cui, Y., Wu, D., Wang, S., Liu, T., Huo, T., Hu, Z., et al. (2019). Cjrc: A reliable human-annotated benchmark dataset for chinese judicial reading comprehension. In *China National Conference on Chinese Computational Linguistics*, pages 439–451. Springer.

[Dumais et al., 1988] Dumais, S. T., Furnas, G. W., Landauer, T. K., Deerwester, S., and Harshman, R. (1988). Using latent semantic analysis to improve access to textual information. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 281–285.

[Dyevre, 2021] Dyevre, A. (2021). Text-mining for lawyers: how machine learning techniques can advance our understanding of legal discourse. *Erasmus L. Rev.*, 14:7.

[Fallah et al., 2023] Fallah, H., Murisasco, E., Bruno, E., and Bellot, P. (2023). Apprentissage de dépendances entre labels pour la classification multi-labels à l'aide de transformeurs. In *18e Conférence en Recherche d'Information et Applications–16e Rencontres Jeunes Chercheurs en RI–30e Conférence sur le Traitement Automatique des Langues Naturelles–25e Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues*, pages 34–40. ATALA.

[Gallié et al., 2016] Gallié, M., Brunet, J., and Laniel, R.-A. (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.

[Garneau et al., 2021] Garneau, N., Gaumond, E., Lamontagne, L., and Déziel, P.-L. (2021). Criminelbart: a french canadian legal language model specialized in criminal law. In *Proceedings of the Eighteenth International Conference on Artificial Intelligence and Law*, pages 256–257.

[Gibbons, 1999] Gibbons, J. (1999). Language and the law. *Annual Review of Applied Linguistics*, 19:156–173.

[Girard-Chanudet, 2023] Girard-Chanudet, C. (2023). Le travail de l'intelligence artificielle: concevoir et entrainer un outil de pseudonymisation automatique à la cour de cassation. *RESET. Recherches en sciences sociales sur Internet*, 12.

[Gonçalves and Quaresma, 2005] Gonçalves, T. and Quaresma, P. (2005). Evaluating preprocessing techniques in a text classification problem. *São Leopoldo, RS, Brasil: SBC-Sociedade Brasileira de Computação*.

[Gong et al., 2018] Gong, H., Sakakini, T., Bhat, S., and Xiong, J. (2018). Document similarity for texts of varying lengths via hidden topics. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2341–2351.

[Goźdź-Roszkowski, 2012] Goźdź-Roszkowski, S. (2012). Legal language. In *The Encyclopedia of Applied Linguistics*. John Wiley & Sons, Ltd.

[Griffiths and Steyvers, 2004] Griffiths, T. L. and Steyvers, M. (2004). Finding scientific topics. *Proceedings of the National Academy of Sciences*, 101(suppl_1):5228–5235.

[Grootendorst, 2022] Grootendorst, M. (2022). Bertopic: Neural topic modeling with a class-based tf-idf procedure. *arXiv preprint arXiv:2203.05794*.

[Grover and Leskovec, 2016] Grover, A. and Leskovec, J. (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.

[Gruodyte and Milciuviene, 2018] Gruodyte, E. and Milciuviene, S. (2018). Anonymization of court decisions in the eu: Actual and comparative issues. *Teises Apzvalga L. Rev.*, 18:60.

[Holzenberger et al., 2020] Holzenberger, N., Blair-Stanek, A., and Van Durme, B. (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*.

[Honnibal et al., 2020] Honnibal, M., Montani, I., Van Landeghem, S., and Boyd, A. (2020). spaCy: Industrial-strength Natural Language Processing in Python.

[Hoyle et al., 2021] Hoyle, A., Goel, P., Hian-Cheong, A., Peskov, D., Boyd-Graber, J., and Resnik, P. (2021). Is automated topic model evaluation broken? the incoherence of coherence. *Advances in Neural Information Processing Systems*, 34:2018–2033.

[Huang et al., 2021] Huang, Y., Giledereli, B., Köksal, A., Özgür, A., and Ozkirimli, E. (2021). Balancing methods for multi-label text classification with long-tailed class distribution. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 8153–8161, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

[Johnson et al., 2019] Johnson, J., Douze, M., and Jégou, H. (2019). Billion-scale similarity search with GPUs. *IEEE Transactions on Big Data*, 7(3):535–547.

[Kano et al., 2018] Kano, Y., Kim, M.-Y., Yoshioka, M., Lu, Y., Rabelo, J., Kiyota, N., Goebel, R., and Satoh, K. (2018). Coliee-2018: Evaluation of the competition on legal information extraction and entailment. In *JSAI International Symposium on Artificial Intelligence*, pages 177–192. Springer.

[Katz et al., 2017] Katz, D. M., Bommarito II, M. J., and Blackman, J. (2017). A general approach for predicting the behavior of the Supreme Court of the United States. *PloS one*, 12(4):e0174698.

[Katz et al., 2023] Katz, D. M., Hartung, D., Gerlach, L., Jana, A., and Bommarito, M. J. (2023). Natural language processing in the legal domain. *Available at SSRN 4336224*.

[Kempf, 2023] Kempf, R. (2023). Calculer et punir, l'essor de la justice algorithmique aux États-unis.

[Kingma and Ba, 2015] Kingma, D. P. and Ba, J. (2015). Adam: A method for stochastic optimization. In *ICLR (Poster)*.

[Koehn, 2005] Koehn, P. (2005). Europarl: A parallel corpus for statistical machine translation. In *Proceedings of Machine Translation Summit X: Papers*, pages 79–86, Phuket, Thailand.

[Krippendorff, 2018] Krippendorff, K. (2018). *Content analysis: An introduction to its methodology*. Sage publications.

[Kumar et al., 2013] Kumar, S., Reddy, P. K., Reddy, V. B., and Suri, M. (2013). Finding similar legal judgements under common law system. In *International workshop on databases in networked information systems*, pages 103–116. Springer.

[Kurzon, 1997] Kurzon, D. (1997). 'Legal language': varieties, genres, registers, discourses. *International Journal of Applied Linguistics*, 7(2):119–139.

[Lam et al., 2020] Lam, J. T., Liang, D., Dahan, S., and Zulkernine, F. (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*.

[Lau et al., 2014] Lau, J. H., Newman, D., and Baldwin, T. (2014). Machine reading tea leaves: Automatically evaluating topic coherence and topic model quality. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pages 530–539.

[Lazaridou et al., 2021] Lazaridou, A., Kuncoro, A., Gribovskaya, E., Agrawal, D., Liska, A., Terzi, T., Gimenez, M., de Masson d'Autume, C., Kocisky, T., Ruder, S., et al. (2021). Mind the gap: Assessing temporal generalization in neural language models. *Advances in Neural Information Processing Systems*, 34:29348–29363.

[Le et al., 2019] Le, H., Vial, L., Frej, J., Segonne, V., Coavoux, M., Lecouteux, B., Allauzen, A., Crabbé, B., Besacier, L., and Schwab, D. (2019). FlauBERT: Unsupervised language model pre-training for French. *arXiv preprint arXiv:1912.05372*.

[Leins et al., 2020] Leins, K., Lau, J. H., and Baldwin, T. (2020). Give me convenience and give her death: Who should decide what uses of NLP are appropriate, and on what basis? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2908–2913, Online. Association for Computational Linguistics.

[Li et al., 2019] Li, X., Michel, P., Anastasopoulos, A., Belinkov, Y., Durrani, N., Firat, O., Koehn, P., Neubig, G., Pino, J., and Sajjad, H. (2019). Findings of the first shared task on machine translation robustness. In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 91–102.

[Lin et al., 2017] Lin, T.-Y., Goyal, P., Girshick, R., He, K., and Dollar, P. (2017). Focal loss for dense object detection. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*.

[Liu et al., 2019] Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., and Stoyanov, V. (2019). Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

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

[Loevinger, 1948] Loevinger, L. (1948). Jurimetrics–the next step forward. *Minnesota Law Review*, 33:455.

[Loevinger, 1963] Loevinger, L. (1963). Jurimetrics: The methodology of legal inquiry. *Law and contemporary problems*, 28(1):5–35.

[Long et al., 2019] Long, S., Tu, C., Liu, Z., and Sun, M. (2019). Automatic judgment prediction via legal reading comprehension. In *China National Conference on Chinese Computational Linguistics*, pages 558–572. Springer.

[Lou et al., 2021] Lou, A., Salaün, O., Westermann, H., and Kosseim, L. (2021). Extracting facts from case rulings through paragraph segmentation of judicial decisions. In *International Conference on Applications of Natural Language to Information Systems*, pages 187–198. Springer.

[Luo et al., 2017] Luo, B., Feng, Y., Xu, J., Zhang, X., and 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*, pages 2727–2736.

[Luz De Araujo and De Campos, 2020] Luz De Araujo, P. H. and De Campos, T. (2020). Topic modelling brazilian supreme court lawsuits. In *Legal Knowledge and Information Systems*, pages 113–122. IOS Press.

[Maley, 2014] Maley, Y. (2014). The language of the law. In Gibbons, J. P., editor, *Language and the Law*, pages 11–50. Routledge.

[Manning et al., 2014] Manning, C. D., Surdeanu, M., Bauer, J., Finkel, J. R., Bethard, S., and McClosky, D. (2014). The stanford corenlp natural language processing toolkit. In *Proceedings of 52nd annual meeting of the association for computational linguistics: system demonstrations*, pages 55–60.

[Martin et al., 2020] Martin, L., Muller, B., Suárez, P. J. O., Dupont, Y., Romary, L., De La Clergerie, É. V., Seddah, D., and Sagot, B. (2020). CamemBERT: a Tasty French Language Model. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7203–7219.

[McHugh, 2012] McHugh, M. L. (2012). Interrater reliability: the kappa statistic. *Biochemia medica*, 22(3):276–282.

[McInnes et al., 2017] McInnes, L., Healy, J., and Astels, S. (2017). hdbscan: Hierarchical density based clustering. *The Journal of Open Source Software*, 2(11):205.

[McInnes et al., 2018] McInnes, L., Healy, J., Saul, N., and Großberger, L. (2018). UMAP: Uniform Manifold Approximation and Projection. *Journal of Open Source Software*, 3(29):861.

[Medvedeva et al., 2021] Medvedeva, M., Üstün, A., Xu, X., Vols, M., and Wieling, M. (2021). Automatic judgement forecasting for pending applications of the european court of human rights. In *ASAIL/LegalAIIA@ ICAIL*.

[Medvedeva et al., 2023] Medvedeva, M., Wieling, M., and Vols, M. (2023). Rethinking the field of automatic prediction of court decisions. *Artificial Intelligence and Law*, 31(1):195–212.

[Mikolov et al., 2013] Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013). Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

[Miller, 1995] Miller, G. A. (1995). Wordnet: a lexical database for english. *Communications of the ACM*, 38(11):39–41.

[Moodley et al., 2019] Moodley, K., Serrano, P. V. H., van Dijck, G., and Dumontier, M. (2019). Similarity and relevance of court decisions: A computational study on cjeu cases. In *JURIX*, pages 63–72.

[Moretti, 2013] Moretti, F. (2013). *Distant reading*. Verso Books.

[Nallapati and Manning, 2008] Nallapati, R. and Manning, C. D. (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.

[Nazarenko and Wyner, 2017] Nazarenko, A. and Wyner, A. (2017). Legal NLP introduction. *Traitement Automatique des Langues*, 58(2):7–19.

[O'Neill et al., 2017] O'Neill, J., Robin, C., O'Brien, L., and Buitelaar, P. (2017). An analysis of topic modelling for legislative texts. In *ASAIL@ICAIL*.

[O'callaghan et al., 2015] O'callaghan, D., Greene, D., Carthy, J., and Cunningham, P. (2015). An analysis of the coherence of descriptors in topic modeling. *Expert Systems with Applications*, 42(13):5645–5657.

[Paley et al., 2021] Paley, A., Zhao, A. L. L., Pack, H., Servantez, S., Adler, R. F., Sterbentz, M., Pah, A., Schwartz, D., Barrie, C., Einarsson, A., et al. (2021). From data to information: automating data science to explore the us court system. In *Proceedings of the Eighteenth International Conference on Artificial Intelligence and Law*, pages 119–128.

[Rabelo et al., 2020] Rabelo, J., Kim, M.-Y., Goebel, R., Yoshioka, M., Kano, Y., and Satoh, K. (2020). A summary of the coliee 2019 competition. In *New Frontiers in Artificial Intelligence: JSAI-isAI International Workshops, JURISIN, AI-Biz, LENLS, Kansei-AI, Yokohama, Japan, November 10–12, 2019, Revised Selected Papers 10*, pages 34–49. Springer.

[Rajpurkar et al., 2016] Rajpurkar, P., Zhang, J., Lopyrev, K., and Liang, P. (2016). Squad: 100, 000+ questions for machine comprehension of text. In *EMNLP*.

[Rasiah et al., 2023] Rasiah, V., Stern, R., Matoshi, V., Stürmer, M., Chalkidis, I., Ho, D. E., and Niklaus, J. (2023). Scale: Scaling up the complexity for advanced language model evaluation. *arXiv preprint arXiv:2306.09237*.

[Řehůřek and Sojka, 2010] Řehůřek, R. and Sojka, P. (2010). Software Framework for Topic Modelling with Large Corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, pages 45–50, Valletta, Malta. ELRA.

[Reimers and Gurevych, 2019] Reimers, N. and Gurevych, I. (2019). Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992.

[Reimers and Gurevych, 2020a] Reimers, N. and Gurevych, I. (2020a). Making monolingual sentence embeddings multilingual using knowledge distillation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

[Reimers and Gurevych, 2020b] Reimers, N. and Gurevych, I. (2020b). Making monolingual sentence embeddings multilingual using knowledge distillation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4512–4525, Online. Association for Computational Linguistics.

[Remmits, 2017] Remmits, Y. (2017). Finding the topics of case law: Latent dirichlet allocation on supreme court decisions.

[Röder et al., 2015] Röder, M., Both, A., and Hinneburg, A. (2015). Exploring the space of topic coherence measures. In *Proceedings of the eighth ACM international conference on Web search and data mining*, pages 399–408.

[Rubinfeld and Gal, 2017] Rubinfeld, D. L. and Gal, M. S. (2017). Access barriers to big data. *Ariz. L. Rev.*, 59:339.

[Salaün et al., 2022a] Salaün, O., Gotti, F., Langlais, P., and Benyekhlef, K. (2022a). Why do tenants sue their landlords? answers from a topic model. In *Legal Knowledge and Information Systems*, pages 113–122. IOS Press.

[Salaün et al., 2021] Salaün, O., Langlais, P., and Benyekhlef, K. (2021). Exploiting domain-specific knowledge for judgment prediction is no panacea. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021)*, pages 1234–1243.

[Salaün et al., 2021] Salaün, O., Langlais, P., and Benyekhlef, K. (2021). Labels distribution matters in performance achieved in legal judgment prediction tasks. In *Proceedings of the Eighteenth International Conference on Artificial Intelligence and Law*, ICAIL '21, page 268–269, São Paulo, SP, Brazil. Association for Computing Machinery.

[Salaün et al., 2020] Salaün, O., Langlais, P., Lou, A., Westermann, H., and Benyekhlef, K. (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.

[Salaün et al., 2024] Salaün, O., Piedboeuf, F., Berre, G. L., Hermelo, D. A., and Langlais, P. (2024). EUROPA: A Legal Multilingual Keyphrase Generation Dataset. *arXiv preprint arXiv:2403.00252*.

[Salaün et al., 2022b] Salaün, O., Troussel, A., Longhais, S., Westermann, H., Langlais, P., and Benyekhlef, K. (2022b). Conditional abstractive summarization of court decisions for laymen and insights from human evaluation. In *Legal Knowledge and Information Systems*, pages 123–132. IOS Press.

[Salton, 1983] Salton, G. (1983). Introduction to modern information retrieval. *McGraw-Hill*.

[Segal, 1984] Segal, J. A. (1984). Predicting Supreme Court cases probabilistically: The search and seizure cases, 1962-1981. *American Political Science Review*, 78(4):891–900.

[Shen et al., 2022] Shen, Z., Lo, K., Yu, L., Dahlberg, N., Schlanger, M., and Downey, D. (2022). Multilexsum: Real-world summaries of civil rights lawsuits at multiple granularities. In *Thirty-sixth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

[Silveira et al., 2021] Silveira, R., Fernandes, C., Neto, J. A. M., Furtado, V., and Pimentel Filho, J. E. (2021). Topic modelling of legal documents via legal-bert. *Proceedings http://ceur-ws org ISSN*, 1613:0073.

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

[Søgaard, 2022] Søgaard, A. (2022). Should we ban english nlp for a year? In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 5254–5260.

[Søgaard et al., 2021] Søgaard, A., Ebert, S., Bastings, J., and Filippova, K. (2021). We need to talk about random splits. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1823–1832, Online. Association for Computational Linguistics.

[Soh et al., 2019] Soh, J., Lim, H. K., and 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*, pages 67–77, Minneapolis, Minnesota. Association for Computational Linguistics.

[Song et al., 2019] Song, C.-W., Jung, H., and Chung, K. (2019). Development of a medical big-data mining process using topic modeling. *Cluster Computing*, 22(1):1949–1958.

[Stevenson, 2018] Stevenson, M. (2018). Assessing risk assessment in action. *Minnesota Law Review*, 103:303.

[Suárez et al., 2019] Suárez, P. J. O., Sagot, B., and Romary, L. (2019). Asynchronous pipeline for processing huge corpora on medium to low resource infrastructures. In *7th Workshop on the Challenges in the Management of Large Corpora (CMLC-7)*. Leibniz-Institut für Deutsche Sprache.

[Şulea et al., 2017] Şulea, O.-M., Zampieri, M., Vela, M., and van Genabith, J. (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.

[Tan et al., 2013] Tan, S., Li, Y., Sun, H., Guan, Z., Yan, X., Bu, J., Chen, C., and He, X. (2013). Interpreting the public sentiment variations on twitter. *IEEE transactions on knowledge and data engineering*, 26(5):1158–1170.

[Thenmozhi et al., 2017] Thenmozhi, D., Kannan, K., and Aravindan, C. (2017). A text similarity approach for precedence retrieval from legal documents. In *FIRE (Working Notes)*, pages 90–91.

[Thomas, 2011] Thomas, S. W. (2011). Mining software repositories using topic models. In *Proceedings of the 33rd International Conference on Software Engineering*, pages 1138–1139.

[Tiedemann, 2012] Tiedemann, J. (2012). Parallel data, tools and interfaces in opus. In *Lrec*, volume 2012, pages 2214–2218. Citeseer.

[Trotman et al., 2014] Trotman, A., Puurula, A., and Burgess, B. (2014). Improvements to bm25 and language models examined. In *Proceedings of the 2014 Australasian Document Computing Symposium*, pages 58–65.

[Tsarapatsanis and Aletras, 2021] Tsarapatsanis, D. and Aletras, N. (2021). On the ethical limits of natural language processing on legal text. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 3590–3599, Online. Association for Computational Linguistics.

[Vacek and Schilder, 2017] Vacek, T. and Schilder, F. (2017). A sequence approach to case outcome detection. In *Proceedings of the 16th edition of the International Conference on Articial Intelligence and Law*, pages 209–215. ACM.

[Vaswani et al., 2017] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. (2017). Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.

[Vīksna et al., 2020] Vīksna, R., Kirikova, M., and Kiopa, D. (2020). Exploring the use of topic analysis in latvian legal documents. In Tagarelli, A., Zumpano, E., Latific, A. K., and Calì, A., editors, *Proceedings of the First International Workshop "CAiSE for Legal Documents" (COUrT 2020) co-located with the 32nd International Conference on Advanced Information Systems Engineering (CAiSE 2020), Grenoble, France, June 9, 2020*, volume 2690 of *CEUR Workshop Proceedings*, pages 39–47. CEUR-WS.org.

[Wagh and Anand, 2020] Wagh, R. S. and Anand, D. (2020). Legal document similarity: a multi-criteria decision-making perspective. *PeerJ Computer Science*, 6:e262.

[Wang et al., 2020] Wang, Y., Gao, J., and Chen, J. (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.

[Westermann et al., 2019] Westermann, H., Walker, V. R., Ashley, K. D., and Benyekhlef, K. (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.

[Wolf et al., 2020] Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., Davison, J., Shleifer, S., von Platen, P., Ma, C., Jernite, Y., Plu, J., Xu, C., Scao, T. L., Gugger, S., Drame, M., Lhoest, Q., and Rush, A. M. (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.

[Wu et al., 2012] Wu, H., Bu, J., Chen, C., Zhu, J., Zhang, L., Liu, H., Wang, C., and Cai, D. (2012). Locally discriminative topic modeling. *Pattern Recognition*, 45(1):617–625.

[Xiao et al., 2018] Xiao, C., Zhong, H., Guo, Z., Tu, C., Liu, Z., Sun, M., Feng, Y., Han, X., Hu, Z., Wang, H., et al. (2018). Cail2018: A large-scale legal dataset for judgment prediction. *arXiv preprint arXiv:1807.02478*.

[Xiao et al., 2019] Xiao, C., Zhong, H., Guo, Z., Tu, C., Liu, Z., Sun, M., Zhang, T., Han, X., Wang, H., Xu, J., et al. (2019). Cail2019-scm: A dataset of similar case matching in legal domain. *arXiv preprint arXiv:1911.08962*.

[Xu et al., 2020] Xu, N., Wang, P., Chen, L., Pan, L., Wang, X., and Zhao, J. (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.

[Zheng et al., 2021] Zheng, L., Guha, N., Anderson, B. R., Henderson, P., and 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*, pages 159–168.

[Zhong et al., 2018] Zhong, H., Guo, Z., Tu, C., Xiao, C., Liu, Z., and Sun, M. (2018). Legal judgment prediction via topological learning. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3540–3549.

[Zhong et al., 2020] Zhong, H., Xiao, C., Tu, C., Zhang, T., Liu, Z., and Sun, M. (2020). Jec-qa: A legal-domain question answering dataset. In *AAAI*, pages 9701–9708.