Using Factors to Predict and Analyze Landlord-Tenant Decisions to Increase Access to Justice

Hannes Westermann
Cyberjustice Laboratory,
Faculté de droit
Université de Montréal
Montreal Quebec Canada
hannes.westermann@umontreal.ca

Vern R. Walker
Research Laboratory for Law,
Logic and Technology
Maurice A. Deane School of Law
Hofstra University
Hempstead New York USA
vern.r.walker@hofstra.edu

Kevin D. Ashley
Intelligent Systems Program
School of Law
University of Pittsburgh
Pittsburgh Pennsylvania USA
ashley@pitt.edu

Karim Benyekhlef
Cyberjustice Laboratory,
Faculté de droit
Université de Montréal
Montreal Quebec Canada
karim.benyekhlef@umontreal.ca

ABSTRACT

This paper reports results from the JusticeBot Project, in which we analyzed two datasets drawn from 1 million written decisions from the Régie du logement du Québec. Using an empirical methodology, we identified 44 factors that occur in disputes where the tenant seeks a remedy due to problems with the rented apartment, such as the existence of bedbugs, high noise levels or problems with insulation. In the first dataset, we used these factors to tag 149 cases. We found a correlation between how many factors are found in a case and how likely the judge is to award rent reduction to a tenant; the amount of reduction was also higher in cases with more factors. For the second dataset (39 cases with bedbugs, drawn from the first dataset), we developed in-depth factors and used them to tag the cases. We found a number of plausible correlations, such as the average damage award being higher in cases with infestations of high intensity. Finally, in predicting the decision of the judge using the factors present in a case, the results were similar to the baselines or slightly above. We discuss the possible reasons for this, and why the approach shows promise in providing useful information to lay people and lawyers.

CCS CONCEPTS

• **Information systems**~Content analysis and feature selection • **Applied computing**~Law

KEYWORDS

Case Prediction, Access to Justice, Chatbot, Factors, Machine Learning

ICAIL '19, June 17–21, 2019, Montreal, QC, Canada

ACM 978-1-4503-6754-7/19/06…$15.00
https://doi.org/10.1145/3322640.3326732

ACM Reference format:

Hannes Westermann, Vern R. Walker, Kevin D. Ashley and Karim Benyekhlef. 2019. Using Factors to Predict and Analyze Landlord-Tenant Decisions to Increase Access to Justice. In *Proceedings of ICAIL '19, Montreal, Québec, Canada, 10 pages.* *https://doi.org/10.1145/3322640.3326732*

1 Introduction

JusticeBot is a project to use Artificial Intelligence to provide a gateway to law and jurisprudence for lay people. It focuses on the analysis of 1 million case decision texts obtained from the "Régie du logement du Québec" (Régie du Logement, or RDL), which is the tribunal responsible for judging landlord-tenant disputes in Quebec. The goal of the project is to provide a chatbot interface to tenants wanting to know which remedies are available to them. The chatbot will ask a number of questions to determine the tenant's factual situation, and then provide the chatbot user with statistical information about their likelihood of success and averages of awarded damages, as well as displaying previous similar cases. This information could help the tenant get an overview of their legal situation and help inform their decision making.

The area of landlord-tenant disputes was chosen firstly because it is a circumscribed area of law, which is well suited as a starting point for the application of AI to legal decision making. Furthermore, our project in this area of law has a high social relevance and value by making disputes over leases more understandable and accessible to the public. The Régie du Logement is one of the biggest tribunals in Canada, and deals with tens of thousands of cases a year. It is our hope to be able to support the Régie du Logement and the concerned parties using our research.

In this paper, we present an exploration of the data, as well as a method to analyze the data. With the help of student annotators, we analyzed 149 cases where the tenant wanted to obtain a remedy based on an apartment being in a bad state. We then created a taxonomy of factors that we identified in the cases. Such factors could be, for example, the presence of bedbugs, water leakage, heating issues, mold or issues with noise. It is important that they are objective, so that we can ask lay people for the same factors. We then tagged 149 such cases with these factors, and data about the outcome of the case, such as whether rent reduction was awarded or not. For cases with bedbugs, we created an in-depth taxonomy, able to capture nuances in the cases. We used this in-depth taxonomy to tag 39 cases dealing with bedbugs. Finally, we used a number of statistical methods to detect trends in the data.

Section 2 of this paper gives an overview of prior work, and the novelty of this project. Section 3 describes the methodology, including a description of the dataset and how we annotated the cases and analyzed the resulting data. Section 4 shows our results, and section 5 provides some discussion of these. Finally, section 6 details how future work on the project will look.

2 Prior Work

This project builds upon prior work representing legal rules and cases in terms of rule trees and factors, applying legal text analytics to identify factors, predicting case outcomes based on text analysis, and explaining the predictions in terms of legal concepts and case analogies.

In representing legal rules that govern landlords' duty to provide rental units in "good habitable condition", we employ a "rule tree," that is, a tree of authoritative rule conditions, as adapted from Walker's Default Logic Framework (DLF) for modeling evidentiary legal argument [12, 13]. Similar to CATO, IBP, AGATHA, and VJAP [2, 4, 7, 9], we employ a domain model comprising legal concepts such as "good habitable condition", issues concerning essential requirements, sub-issues regarding access to electricity, water, or kitchen appliances and case-indexing factors related to the sub-issues.

The factors connecting legal rule conditions to case facts are stereotypical patterns of facts that strengthen or weaken a side's claim [3, Sec. 3.3.2]. We asked student annotators to identify instances of factors in case texts such as #WarmWaterAccess (insufficient access to warm water) and #BathroomUtilities (some bathroom facilities out of order). Falakmasir and Ashley [8] assembled a corpus of trade secret cases, employed word embeddings, trained an ML algorithm for each factor using a subset of manually-classified cases, and predicted the factors that apply in cases, achieving F1 values of .65, an improvement over the results in [4]. Starting from factor descriptions, Wyner and Peters [15, 16] applied WordNet expansions and expert knowledge to generate factoroids, finer-grained semantic terms related to each factor from which they generated factor-annotating rules in GATE.

In a similar way, using a grounded theory method from empirical legal studies [17], we peruse selected cases to develop an in-depth taxonomy for each factor. For instance, the #Bedbugs factor taxonomy contains such objectively determinable criteria as Duration (1-999 months) and Intensity (1-3), along with a taxonomy for differentiating between levels. Thus, we are attempting to represent quantitative magnitudes of factors [5], and techniques to identify them automatically.

Our goal is to predict outcomes of landlord/tenant cases based on the identified factors and the domain model. Others have predicted outcomes based on automated analysis of case texts. Aletras, et al. [1] predicted outcomes of European Court of Human Rights cases with an accuracy of 79% (although, apparently, without adequately masking text parts which may have biased the results). Sulea, et al. [10] reported accuracy of 92% in predicting outcomes of cases before the French Court of Cassation.

These approaches did not represent substantive features of case scenarios that could be employed in predicting and explaining the predictions. Integrating machine learning (ML) and expert knowledge for purposes of explanation is an important topic of current interest [6]. For example, IBP based its predictions on factors identified automatically in the case texts and explained them in terms of intermediate legal concepts in its domain model and factor-based analogies to similar cases [4]. Similarly, our project will base predictions on automatically identified factors and explain them in terms of legal concepts from the rule trees, as far as we know, a first for a civil law domain. Other factor-based prediction approaches take values into account [7, 9], which we may also be able to incorporate.

3 Methodology

In this section, we first organize and represent the legal rules and issues involved in RDL cases (Section 3.1). Using this organization, we select and describe the dataset for the study we report on in this paper (3.2). Finally, we describe how we classified cases in the dataset, and we describe how we developed our models (3.3).

3.1 Representing the Legal Rules and Issues

One primary task of the JusticeBot Project is to classify cases according to the legal grounds for the decision. For classifying the fact-finding decisions of the Régie du Logement, we began by representing the legal rules that state the conditions under which either the tenant or the landlord is entitled to a legal remedy. Those legal rules are established by the Code Civil du Québec, and to some extent by appellate case law interpreting that code. We used the Default Logic Framework [11] to represent a legal rule as a set of propositions, one of which is the conclusion and the remaining propositions being the rule conditions (i.e., the conditions under which the conclusion is true). An integrated system of such rules that governs a governmental action can be represented as a "rule tree" [14]. Figure 3.1-1 shows a portion of the tenant-oriented rule tree (in English translation) that addresses termination ("résiliation") of a lease due to the dwelling not being in "good habitable condition," formulated from the Code Civil du Québec, articles 1863 and 1910.

This rule tree represents a legal rule by placing the conclusion at the top of an indented list of its conditions, with each condition preceded by the logical connective operating between it and the conclusion. Each condition can function in turn as a conclusion, with its own conditions listed below it. The resulting nested sets of conditions has a tree structure.

The tenant may apply for the resiliation (termination) of the lease.
…
OR A "nonperformance of an obligation" by the lessor "causes" a "serious injury" to the lessee or, in the case of the lease of an immovable, to another occupant. (1863)
AND [1 of 3] There is "nonperformance of an obligation" by the lessor. (1863)
…
OR The lessor does not deliver the dwelling "in good habitable condition" or "maintain it in that condition throughout the term of the lease." (1910)
OR [1 of 2] The lessor does not "deliver the dwelling in good habitable condition." (1910)
OR [2 of 2] The lessor does not "maintain the dwelling in good habitable condition throughout the term of the lease." (1910)
AND [2 of 3] The lessee or another occupant suffers a "serious injury." (1863 al. 1 in fine)
OR [1 of 2] The tenant suffers a "serious injury." (1863 al. 1 in fine)
OR [2 of 2] In the case of a lease of an immovable, another occupant suffers a "serious injury."
AND [1 of 2] It is a case of the lease of an "immovable." (1863 al. 1 in fine)
AND [2 of 2] Another "occupant" suffers a "serious injury." (1863 al. 1 in fine)
AND [3 of 3] The lessor's nonperformance "causes" the serious injury. (1863 al. 1 in fine)

Figure 3.1-1: Partial Rule Tree for Termination of a Lease Due to the Dwelling's Not Being in "Good Habitable Condition" (English Translation)

As Figure 3.1-1 shows, a tenant may apply for the termination of the lease under certain conditions if a nonperformance of an obligation by the lessor causes a serious injury (Code Civil art. 1863). The fact that there are other possibilities for nonperformance is indicated by the logical connective "OR" preceding that condition. Nonperformance of an obligation, however, is only the first of three necessary and jointly sufficient conditions for the ultimate conclusion (as indicated by the three conditions being preceded by "AND [x of 3]"). In addition to nonperformance by providing a dwelling that is not in "good habitable condition," the tenant must prove that the lessee (or another occupant, in the case of the lease of an immovable) suffered a "serious injury" ("un préjudice sérieux"), and that that injury was caused by the lessor's nonperformance.

Use of the rule tree conditions as a method for classifying cases decided by the RDL allows us to organize cases by the legal issues that are decided, and to classify the portions of text within those decisions by legal issue. This enables us to match past cases to future cases, with the unit of analysis (or topic) being the legal rule condition. This matching also enables us to assemble a set of decisions that interpret or apply a single rule condition, so that we can analyze the factors that have come into play in deciding those issues. For example, in this paper we focus on a set of cases that decide what counts as not being in "good habitable condition" (Dataset 1), and particularly those in which the problem is infestation with bedbugs (Dataset 2). From the relevant passages within those decisions, we extract information about the factors that play a role in predicting the outcome - that is, whether the RDL is likely to consider a particular dwelling as not being in "good habitable condition" due to the presence of bedbugs.

3.2 Representing the Legal Rules and Issues

In order to identify the factors relevant for deciding a legal condition in the rule tree, we use a dataset consisting of 149 annotated decisions. It was derived from a set of about 1,000,000 case texts that were given to us by the Régie du Logement du Québec. They are almost exclusively in French. For the purpose of this paper we are focusing on the subset of 38,286 cases decided in 2017. Since these cases were the most recent cases that we have access to, it seems reasonable to assume that they most closely reflect the current practices of the RDL. The case files are in the .doc format, which we converted to .txt files for analysis. They are highly variable in length, with an average of 3,350 characters and a median of 2,540 characters. However, some files are significantly larger, with the maximum length being almost 70,000 characters. In this dataset, 3,103 cases are longer than 6,000 characters. Figure 3.2-1 shows the distribution of the length of cases, where the cases are under 6000 characters long.

Figure 3.2-1: Distribution of the length of cases

For this paper, we decided to focus on cases that deal with remedies sought by a tenant against a landlord, with the tenant arguing that there is something wrong with the rented apartment. Specifically, we focused on cases that dealt with the legal requirement of a landlord to provide an apartment in a "good habitable condition" and to provide "peaceable enjoyment" of the apartment to the tenant. These requirements can be found in Article 1910 and 1854 of the Civil Code of Quebec.

To identify relevant cases, we scanned the dataset for documents containing mention of articles 1910, 1854 or 1864 (the last article requires the landlord to make necessary repairs of the property during the term of the lease). Since these requirements are often applied in parallel, we found it pertinent to include them in the initial analysis in order to capture a broad distribution of facts.

This resulted in a dataset of 594 cases for 2017 which we used during Phase 1 (see below) to establish a matrix of possible facts.
From these 594 cases, we randomly selected 202 for annotation. Of these 202 cases, the annotators considered 26 not relevant. These were usually cases where a landlord sued a tenant, instead of the other way around. A further 27 cases were excluded due to being duplicates annotated to test inter-annotator agreement (see below). This left us with 149 cases for our further analysis (which we call Dataset 1).
Of these 149 cases, 39 dealt with bedbugs. For these 39 cases, we also captured nuanced information such as infestation intensity, see below under Phase 2. We will refer to these 39 cases as Dataset 2. Figure 3.2-2 provides a visual explanation of how the cases were selected.

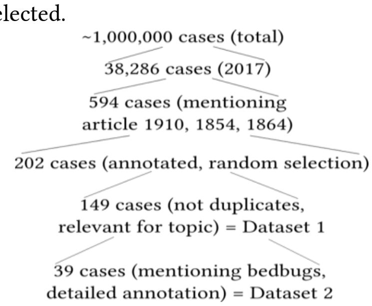

Figure 3.2-2 - Selection of cases for use in this paper

The RDL decision texts in these cases typically follow a certain structure. At the beginning of the decision, the Régie du Logement states the demands of the plaintiff. The decision then lists the facts from the viewpoint of the plaintiff, and then reports the responses of the defendant. After this comes a segment for the analysis, within which the tribunal typically first presents the relevant legislation. Then, the judge applies the law to the factual situation as established by the proof in the case. Based on this, the tribunal comes to conclusions, which they present at the end of the case.
There are different types of remedies available to the tenant. For the purposes of this paper, we focus on the remedy of rent reduction, which is supposed to compensate for the loss of rent value that a tenant experienced. It is an obligation of guarantee, therefore the landlord does not have to be found negligent for the rent reduction to be decided upon. For the analysis in this paper, we consider a case to be "won" by the tenant if the tenant is awarded rent reduction, while a case is considered "lost" if no rent reduction is awarded.

3.3 Case Classification and Model Development

Next, we devised a method to extract and analyze which factors contribute to the decision of a judge, in order to build a model of what "good habitable condition" means in terms of real-world facts. Is the mere presence of bedbugs, for example, always sufficient grounds for deciding that an apartment is not considered to be in "good habitable condition"? Once this model has been constructed, we can then ask real-life tenants for their facts, and compare this to the model, to provide similar cases and statistics on outcomes.

In order to develop a classification system for cases, we use the Grounded Theory Method [17], which does not make any presumptions about the data. This methodology consists of going through qualitative documents, such as court decisions, and creating "memos", or recurring themes in the documents. Through further data analysis, these memos are then refined until they accurately capture the relevant information in the documents.
In our case, this analysis was performed in several stages. These stages are explained below.

Phase 1 – Factor identification

During Phase 1, we established a taxonomy over factors that were recurring in the cases. At first, we simply read cases and took notes on which factors occurred. We then reconvened and together established the taxonomy (hereafter "the taxonomy") and assigned a hashtag to each factor. For example, cases involving moisture would be tagged with #moisture. In creating the factors, we took care to only include factors that are objective, and not dependent upon a value judgement of the judge. This is important so that we can later ask lay people for the factors present in their cases. Asking an individual whether their apartment is in a good habitable state is not possible, since this is a legal fact that the judge has to decide to reach a conclusion. However, it is possible to ask lay people whether there is a water leak in their apartment. As annotations continued, new factors were added based on the feedback of annotators.

Phase 2 – Factor refinement

Once a number of cases had been tagged with the factors from the taxonomy, we delved deeper into a specific factor. For this paper, we decided to focus on bedbugs. We assigned cases from Dataset 2 to the annotators. They were again asked to explore and take notes on the variations of factors involving bedbugs. We then reconvened and established an in-depth taxonomy of the nuances of factors relating to bedbugs (hereafter "the bedbug taxonomy"). We especially focused on ensuring that the variations are objectively derivable from the texts of the cases. We also established clear guidelines on how to assess the variations.

Phase 3 – Case annotation

Once the factors have been refined to a sufficient degree, the annotators were asked to go back to the cases from Dataset 1 to tag them with the taxonomy. We also annotated cases from Dataset 2 with the bedbug taxonomy. At this stage, we classify cases using only the information that the judge considers to be a proven fact, since only such facts help determine the judge's decision. The outcome of the analysis of the judge is also annotated, including outcomes such as damages and rent reduction. The result of this annotation process is to create a "factor vector" for each case - a profile that contains the specified factors of each case as well as the outcome. Essentially, all relevant facts should be extracted from the judgment.

Phase 4 – Model construction

Once enough cases have been annotated, the next step is to use the factor vectors to construct a model of how the factors contribute to a certain decision. The input of the model is the vector of facts, and the output is the decision of the judge on whether to award rent reduction or not. We also used several other analyses to explore the data.

4 Results

In this section, we will present the results of the methodology presented in Section 3.

4.1 Phase 1 - Factor identification

The first session was spent discovering the issues likely to arise in cases. In the end, the taxonomy contained 30 different factors. We used this classification system to "tag" cases with the factors they contain. In line with the Grounded Theory method, any issue in terms of tagging was logged and used to improve the taxonomy. Based on the suggestions of the annotators, the taxonomy was expanded to contain 44 factors.

The factors are structured in a hierarchical way in order to simplify look-up by the annotators. Figure 4.1-1 shows an extract of the taxonomy.

Were any of the following factors present in the case you read?
Other deficiencies
#Dirty - The apartment is dirty.
#Moisture - The apartment has issues with moisture.
#Mold - There is mold in the apartment.
Infestation
#Bedbugs - The apartment is infested with Bedbugs.
#Rats - The apartment is infested with rats.

Figure 4.1-1 - Extract of taxonomy used for classifying case factors

The hashtags are used to provide a common vocabulary to the annotators, and to allow the computer to easily extract the information. At first, the annotators manually inserted the hashtags into a text-field in Gloss, the web-based annotation environment developed by Jaromir Savelka at the University of Pittsburgh [18]. Later we moved to a system where the annotators could simply click on the hashtags in a Google Forms document. This provided more quality assurance over the spelling of factors. Google Forms allows the export of data in .csv format which we used to further analyze the data.

4.2 Phase 2 - Factor Refinement

Once we had created the taxonomy, we created the bedbug taxonomy to capture nuances in cases with bedbugs. The goal here was to identify nuances in the factor that the judge would take into consideration when deciding on the merits of the case. We read through the cases and identified several factors that the judge seems to take into consideration in reaching a decision. Figure 4.2-1 shows the bedbug taxonomy.

We encountered some difficulties in specifying these factors. The decisions usually only discuss an issue if it is raised or if it matters for the decision. In some decisions, for example, it was difficult to determine the duration that the bedbugs were present. For some decisions, the facts did not fit neatly into the categories provided - for example, a landlord might be first helpful and then unhelpful. This shows the difficulty of labelling an often disorganized and unstructured reality. We will come back to this problem in our discussion.

Which intensity of bedbugs were present?
- Low (Few bugs, few bites)
- Medium (if not specified)
- High (Intense infestation, entered furniture)

How long were bedbugs present? (in weeks)
Is it possible to tell who caused the issue?
- Tenant
- Landlord
- Not discussed/ attributed to a third party

How helpful was the landlord in solving the problem?
- Helpful (ordered help when requested)
- Not helpful (responded slowly, did not order professional help)
- Not discussed/applicable

How cooperative was the tenant in helping the extermination?
- Helpful (Prepared apartment)
- Not helpful (Did not prepare apartment, enable access)
- Not discussed/applicable

Figure 4.2-1 - Taxonomy used to classify in-depth factors concerning bedbugs

4.3 Phase 3 - Factor annotation

Once we had created the taxonomy, we used it to annotate cases in Dataset 1. In addition, we annotated a subset of 39 cases using the bedbug taxonomy. Due to the use of the Grounded Theory method, the two phases of discovering categories and of applying them were often overlapping, since we adjusted the taxonomy as we went along. In addition to factors, we also coded other data, such as whether the case was relevant for our purposes (cases filed by landlords, for example, were excluded), the remedies sought by the tenant, the legal rule conditions the judge decided on (such as good habitable state or loss of peaceable enjoyment), and the amounts of rent reduction and damages awarded. We did not annotate individual sentences but assigned labels to the entire case.

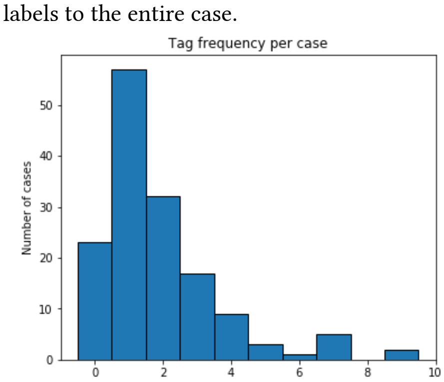

Figure 4.3-1 – Tag frequency distribution

Dataset 1 contained a total of 287 tags. Figure 4.3-1 shows the frequency of how many tags were present in each case.

As seen in Figure 4.3-1, 57 cases have only one tag. Because we are annotating only those factors that are considered proven by the judge, and because it is not always possible to determine which factors were considered "facts" in the case, quite a large number of cases have no tags. Figure 4.3-2 shows a list of the 20 most frequent factors.

Figure 4.3-2. Most frequent factors

Tag	Frequency	Relative frequency
#bedbugs	41	27.50%
#repairsnotconducted	27	18.10%
#mold	16	10.70%
#rainleakage	16	10.70%
#waterleakage	13	8.70%
#heating	11	7.40%
#otherinfestation	11	7.40%
#accesstopremises	10	6.70%
#constantrepairs	10	6.70%
#cockroaches	10	6.70%
#noise	10	6.70%
#intruderprotection	9	6.00%
#landlordunresponsive	9	6.00%
#otheraccessories	8	5.40%
#wallrepair	8	5.40%
#bathroomutilities	8	5.40%
#moisture	7	4.70%
#isolation	7	4.70%
#danger	6	4.00%
#parkingaccess	5	3.40%

Inter-annotator agreement

We had the volunteer annotators annotate 14 cases more than once, so we could determine inter-annotator agreement. Of these, 11 dealt with bedbugs. The results reveal the difficulty of neatly fitting factual occurrences into categories. In only 6 of the cases (43%) did all annotators agree on which tags were present. All of the cases where annotators did agree had either 0 or 1 tags. This difference might stem from the fact that it is sometimes unclear what the judge takes as proven and incorporates into his analysis. Annotator error and vagueness in exactly how to use a factor are also likely to play a role.

Some tags were applied more consistently than others. For example, in 12 of the cases (86%) the annotators agreed on whether there were bedbugs present or not. However, for the more refined factors the reliability of the annotation dropped. The duration of the presence of bedbugs was annotated reliably in only 5 of the 11 cases dealing with bedbugs, representing 45%. It is possible that this is due to the judges often not explicitly stating how long the bedbugs were present. Bedbugs might also disappear and come back, and treatment may continue even while bedbugs are not present, further complicating the issue. The annotators were reliable in the tagging of the intensity of the infestation in 6 of the cases (54%). The other values show similar results.

The annotators consistently tagged whether the tenant was found to have a loss of peaceable enjoyment in 12 of the 14 cases (86%). The amount of rent reduction also showed reasonable consistency, with 11 of the 14 cases agreeing. For moral damage (such as damage awarded for trouble and inconvenience), the annotators agreed in all 14 cases.

Overall, the inter-annotator agreement is not very high. This might be a problem inherent to extracting categorical information from free text, or it might be improved by updating the annotation scheme to match the facts as they come in. In our experiments, we also used a comparatively large number of annotators (10-20) who came in to participate once per week. Results might be more consistent if we used fewer annotators for longer periods of time.

4.4 Phase 4 - Model construction

Finally, we developed models to see the correlations between the factors and the outcome of the cases. First, we analyzed the 149 cases in Dataset 1 (including bedbugs, but without specific bedbug factors such as length of the infestation, etc.). Then, we analyzed the 39 cases in Dataset 2, including the bedbug-specific factors.

For the type of case outcome, we focused on whether the rent was reduced or not. This was chosen since it is sometimes not entirely clear whether a judge found a loss of peaceable enjoyment or not. The same goes for lack of good habitable condition. However, rent reduction can be a legal consequence to compensate for these legal factors. We are therefore using rent reduction as a proxy for whether the judge found the landlord to have breached his obligation of guaranteeing a peaceable enjoyment to the lessee. While this is a simplified version of reality, for the purpose of this section we consider a tenant to have "won" a case if the judge ordered a reduction in rent.

4.4.1 Dataset 1

Dataset 1 contains 149 cases. Out of these, the judge ordered rent reduction in 86 cases, and refused to order rent reduction in 63. In figure 4.4.1-1, we show the number of factors that the judge found proven in a case correlated with the likelihood of the judge deciding on rent reduction. The data shows that judges are more likely to order rent reduction in cases where more factors are seen as proven.

Figure 4.4.1-1 – The percentage of cases where rent reduction was ordered, per number of factors found

Next, we decided to look at the amount of rent reduction. Rent reduction is usually computed as a percentage of monthly rent. For mold, for example, a judge might order a 10% reduction of rent for the 3 months the problem subsisted. For Dataset 1, we only annotated the existence of a factor, and not the length of time that it exists. The unit of measurement for the outcome is therefore how many months of rent reduction the judge ordered in total. The formula used to determine this outcome number ("number of months of rent reduction") is simply the total reduction of rent ordered, divided by the amount of monthly rent for the apartment. For example, if the judge ordered a reduction of rent = $1,000, and the monthly rent was $250, then the number of months of rent reduction would equal 4.

Figure 4.4.1-2 shows the correlation between the number of tags in a case and the number of months of rent reductions. It seems that the amount of rent reduction awarded rises with the number of annotated factors. The blue transparent background shows the 95% confidence interval.

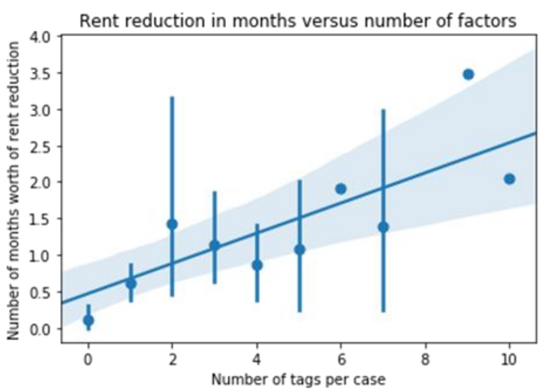

Figure 4.4.1-2 – Rent reduction in months versus number of factors

Figure 4.4.1-3 lists factors used to classify the cases and shows the percentage of wins of the tenant for cases containing each factor. We limited the analysis to factors that were mentioned more than 4 times in the total corpus. The factors have been ordered from the ones most likely to result in rent reduction to those the least likely.

These results seem to show that some factors are stronger indicators of whether rent will be reduced, compared to others. Issues with electricity access, danger in the apartment, issues with insulation and mold seem to be strong factors, where the judge usually ordered a reduction of rent in the cases we looked at. In cases classified as having bedbugs, the rent was reduced in 68% of cases.

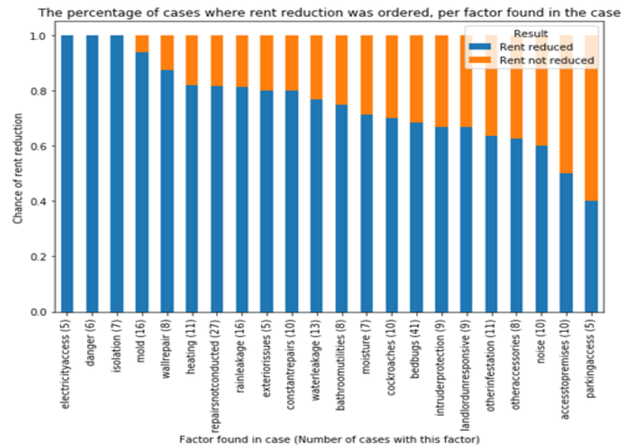

Figure 4.4.1-3 – The percentage of cases where rent reduction was ordered, per factor found in the case

Machine Learning

Classification

We trained a machine learning model to estimate the prediction value of the factors. Our hypothesis was that treating the factors as binary input features (presence or absence of a certain feature), we would be able to create a model able to predict whether a judge will award a rent reduction or not. Instead of using descriptive statistics to discover useful information, as we did before, here we attempted to predict the outcome. We compared the results against a baseline, which always predicts that the rent will be reduced, since this is the most frequent choice. In order to determine reliable results, we used an 8-fold cross validation strategy, where 7/8ths of the 149 cases are used for training and 1/8th for testing. We used a Random Forest Classifier to try to predict the decision of the judge. It is a categorical prediction of whether rent reduction will be awarded, based on the presence or absence of the 44 factors in each case.

The Baseline, which always predicts that the tenant will win, achieves a precision of 57.7%. Our classifier achieves a precision of 66.5%. The classifier thus performs 9 percentage points better than the baseline. It should be noted, however, that this improvement disappears if we only consider cases with at least 1 tag. In this case, the classifier performs equal to or slightly worse than the baseline. We will discuss these results in the conclusion.

Regression

We also tried predicting the number of months of rent reduction based on the factors. The input is the presence or absence of the 44 factors, and the output is the predicted number of months of rent reduction. For this, we used a Random Forest Regressor, and compared it to a baseline of always predicting the average. We used an 8-fold cross validation strategy, and we took the average of the results. The result is expressed in mean squared error, which takes the average difference between the predicted and the real values and squares it.

The mean squared error for the baseline came to 6.55, while our regressor achieved a score of 8.19. Since lower is better, the regressor performs worse than the baseline in this instance. We will discuss this in the conclusion.

4.4.2 Dataset 2

Next, we will look at the selection of cases where bedbugs is one of the factors. There are 39 such cases. These cases have been tagged with the bedbug taxonomy.
Figure 4.4.2-1 shows the quantity of the factors from the bedbug taxonomy, the wins and losses, the win-percentage as well as average rent reduction and moral damage awarded.
Although there might not be enough data points to draw conclusions, some trends are visible. Who caused the problem of bedbugs is mentioned only rarely, therefore little correlation can be drawn from this. There is a clear correlation between the intensity of an infestation and the amount of rent reduction awarded. Where the intensity of an infestation was considered high by annotators, the judge awarded a rent reduction in 100% of cases, with an average reduction of $890. For medium and low intensities of infestation, rent reduction was awarded in about 60% of cases, with an average reduction of $379 for medium or $216 for low intensities.

Tag	Total	Win	Loss	Win % (tenant)	Avg rent reduction	Avg Moral damage
Who was at fault for the infestation of bedbugs?						
Landlord	1	0	1	0.0%	$0	$2,000
Not discussed	36	25	11	69.4%	$481	$270
Tenant	2	2	0	100.0%	$460	$0
How intense was the infestation of bedbugs?						
High	9	9	0	100.0%	$889	$300
Medium	23	14	9	60.9%	$379	$392
Low	7	4	3	57.1%	$216	$0
Was the landlord helpful in exterminating the bedbugs?						
Helpful	23	17	6	73.9%	$489	$183
Not discussed	9	6	3	66.7%	$442	$189
Not helpful	7	4	3	57.1%	$429	$830
Was the tenant cooperative with the exterminators?						
Helpful	11	9	2	81.8%	$790	$437
Not discussed	17	9	8	52.9%	$383	$335
Not helpful	11	9	2	81.8%	$275	$109

Figure 4.4.2-1 – Table showing trends in data based on bedbug taxonomy

The helpfulness of the landlord in exterminating the bedbugs has a slight influence on the win rate and average rent reduction awarded. However, the big difference here is in moral damages awarded: For cases where the landlord was not helpful in exterminating the bedbugs, $830 were awarded on average, compared to under $200 for cases where the landlord was helpful, or the issue was not discussed. Moral damages are sums awarded, for example, for troubles and inconvenience, and are typically only awarded in cases where the landlord is found to be at fault.
The cooperativeness of a tenant does not significantly affect the chances of being awarded rent reduction in these cases but does cause an increase in the amount of rent reduction and moral damages.

Figure 4.4.2-2 demonstrates the correlation between the number of weeks the apartment was affected by bedbugs, and the number of months of rent reductions the tenant was awarded. It should be noted that it was often unclear from the cases how long the bedbugs were present, leaving only 22 cases where the length was annotated. Further, we excluded cases where the infestation lasted more than a year, leaving 18 cases. The data above 1 year was very sparse, and some outliers made the correlation less clear.

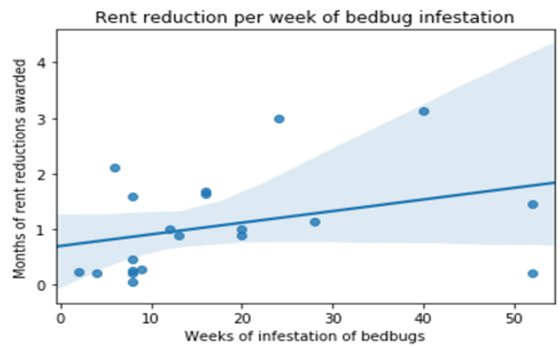

Figure 4.4.2-2 – Rent reduction per week of bedbug infestation

Overall, the correlation is not very clear. It could be expected that an infestation that lasts twice as long as another should also be awarded a rent reduction twice the size. This is not evident from the data. However, 18 cases are likely not enough to obtain significant results.

Machine Learning

As before, we also trained machine learning models to see whether the detailed factors would improve performance in predicting whether rent reduction would occur or not. What we are trying to predict here is whether the judge will order a rent reduction in the cases where bedbugs are present depending on the existence or absence of factors, this time including the in-depth factors specified in phase 3. As such, there are 34 factors considered. These are the new factors, with 12 different factors from the bedbug taxonomy, and 22 factors from the taxonomy, that are also present in the cases with bedbugs. Again, we relied on a 8-fold cross validation to get reliable results. The baseline scored a precision of 70.8%, while our classifier achieved a precision of 70%. The precision is very close to that of the baseline. However, due to the small amount of data, the results may not be evidence of anything. We will get back to this in the discussion section.

5 Conclusion

In this paper, we used the Grounded Theory Method to find 44 factors that occur in legal decisions about the "peaceable enjoyment" and "good habitable condition" of an apartment. We used these factors to tag 149 cases. We further established in-depth factors for cases involving bedbugs. We used these to annotate 39 of the 149 cases. We then tried to use Machine Learning to discover the ways in which the existence or absence of factors influences the decision of the judge to award rent reduction to the tenant, as well as the amount of rent reduction.

The results show the challenges of using computers to help perform legal analysis. Essentially, in extracting categorical information from decision texts, a great deal of information is lost. This might account for the results of our prediction tasks, which barely outperform the baselines, if at all.

5.1 Extracting factors from case decisions

This loss of information starts in the creation of the case decisions. Since the proceedings in front of the Régie du Logement are oral, the judge will analyze a stream of evidence in order to reach a decision. Such evidence includes the testimonies of the plaintiff and defendant about the factual situation (presented with subtle aspects of demeanor such as facial expression, voice and body language) and any documentary evidence presented. The judge must assess the credibility of the witnesses, as well as the trustworthiness of the documents. Necessarily, a large part of this information is lost when the judge writes the case decision, which is the raw text that we have to work with.

Next comes the task of extracting from the case decision any information about a set of factors. In our case, since the goal is to eventually use the data to improve access to justice, we tried to use factors that are as objective as possible, and therefore possible to ask of people without legal education. We established 44 such factors that we used to tag the cases. It is often difficult to fit the RDL decision texts neatly into categories, however. We used the Grounded Theory Method to allow us to maintain a flexible annotation scheme. This, and the large number of factors that might overlap in some instances, caused problems with agreement between annotators. By trying to classify texts using values for categories of tags, we again lose information.

In order to capture some of the more subtle information, we established the bedbug taxonomy for the cases falling into the category of having bedbugs. The idea here is to capture nuances that might influence the decision of the judge. For example, one hypothesis is that the intensity of a bedbug infestation is likely to affect the decision. These nuances cause some analysis problems as well. For example, the evaluation of intensity is to some extent a subjective assessment, and values might differ between judges in determining the facts from case to case, as well as between annotators in tagging the information from decision texts. Even seemingly objective measures, such as weeks of infestation, can be very difficult to assess. What if, for example, an infestation disappears but then comes back, or if the treatment of the infestation continues after the infestation has been exterminated? Even after annotating using the in-depth factors, many factors that might be relevant in the decision are likely to be missed. A judge might change a decision based on a single factor that is unique to the particular case he is ruling on. Further, judges have a certain margin of discretion. Many of these determinants are not mentioned in the tribunal decision and therefore cannot be captured by annotators or serve as data points for Machine Learning.

5.2 Small datasets for analysis

Further, as we have seen, it is difficult to amass datasets large enough to perform quantitative analysis in legal decision making. We are very lucky to be working with a dataset large enough to contain 1 million decisions. However, once a sample is drawn for annotation with factors, and then a subset for in-depth factors, the number of cases in the sample quickly dwindles. Out of some 40,000 cases from 2017, we were able to identify only 39 cases that mentioned the relevant articles from the Code Civil du Québec and dealt with bedbugs. Once we sectioned them by factors, even fewer cases were left in each category, making it very difficult to use quantitative analysis, such as Machine Learning, to establish patterns. To obtain enough cases to fully extract all relevant factors in depth, therefore, seems difficult. Moreover, annotating the resulting samples of cases is a difficult and time-consuming task. It requires the legal analysis of texts, and thus requires both time and competency from the annotators. Finally, if the goal is to predict actual outcomes of actual future cases based on their merits (factors), the tagging data that results from the annotation must be not only reliable (consistently applied among annotators), but also valid (accurately describing the facts in the case).

5.3 Conclusions

However, these challenges do not mean that the analysis of legal decisions is useless. We should remember that the objective and context for the JusticeBot Project is improving access to justice, by using factors extracted from past cases to predict outcomes in new cases. As can be seen from the analysis section, we can draw a number of interesting conclusions that could be useful for practitioners and lay people alike. One interesting result is the fact that cases with more factors present lead to a higher likelihood of the judge reducing the rent. Likewise, the amount of rent reduction awarded also rises with the number of factors tagged in a case. This means that factors do "accumulate" in some way, and together contribute to the decision of the judge. The tables showing percentage success rate per factor and average rent reduced per factor could be useful for users looking for an indication of their chances of success. For example, the fact that users might have a difficult time obtaining a rent reduction for a complaint merely about access to a parking spot could be interesting to people considering a lawsuit.

The in-depth factors for the bedbug category also seem to capture some interesting trends, such as awarded rent reduction being higher for intense infestations and if the tenant is cooperative, and awarded moral damage being higher in cases where landlords are not helpful. There is also a correlation between the duration of an infestation and the rent reduction awarded.

Legal prediction is difficult. Even lawyers have trouble predicting the decisions of courts. This paper shows the difficulty in extracting objective data from decided cases and using it to predict the outcome of further cases. However, it also shows some interesting and plausible trends in the data, despite the comparatively low number of cases analyzed and the challenges of classifying decision texts into discrete factors.

6 Future Work

In the current work, we have provided the proof of concept of the first phase of a longer-term project. We will now continue the work on the project to use the case decisions to enhance access to justice.

In this paper, we have presented a slice of the project, by focusing only on the predictability of court decisions concerning tenants' rights in case of bedbugs in their rental object. Further work will have to be done to create in-depth ontologies for other factors, such as water leakage, noise, temperature and structural integrity. We will continue working with annotators to improve the reliability of annotations and to increase the number of annotated cases, which is likely to contribute to higher accuracy.

In order to expose the information gained to a potential user, the next stage will involve the design of a Dialogue Tree. This will contain questions necessary to gather applicable factors from the user, and analyze their situation using these factors. This Dialogue Tree will be incorporated into a chatbot that asks users for their factors, and uses this information to point them to cases that are similar to theirs, possibly also showing statistics such as win/loss rates and maybe even a range of possible rent reductions.

There are also a number of other approaches that could be used to analyze the data. One is to perform a more qualitative analysis of cases to manually extract the tests judges are likely to apply. This could be incorporated into a traditional expert system. Another approach could be to use Natural Language Processing to try to extract the existence of relevant factors in a court case, which could automatize parts of the prediction. Yet another strategy could be to cluster cases based on the occurrence of words or terms, and compare them to the factors we manually extracted.

ACKNOWLEDGMENTS

We would like to thank the Régie du Logement du Québec for sharing their case decision data with us. We would further like to thank the LexUM Chair on Legal Information, as funding institution and partner. A big thank you also goes to all the student volunteers who participated in the annotation process. Without your support, this paper would not have been possible. Further, we would like to thank Aliaa El-Hage for helping us with her legal expertise of the area, Valentin Callipel for help and guidance, and Florence So for tireless work on the project. I would also like to thank Dirk Westermann for many helpful discussions.

We also thank the Maurice A. Deane School of Law for its financial and other support for the participation of the Research Laboratory for Law, Logic and Technology in this project.

REFERENCES

[1] Nikolaos Aletras, Dimitrios Tsarapatsanis, Daniel Preoţiuc-Pietro and Vasileios Lampos. 2016. Predicting judicial decisions of the European Court of Human Rights: a Natural Language Processing perspective. PeerJ Computer Science 2, e93 (October 2016), 19 pages. DOI: https://doi.org/10.7717/peerj-cs.93.

[2] Vincent Aleven. 2003. Using background knowledge in case-based legal reasoning: A computational model and an intelligent learning environment. Artificial Intelligence 150, 1-2 (November 2003), 183–237. DOI: https://doi.org/10.1016/S0004-3702(03)00105-X

[3] Kevin D. Ashley. 2017. Artificial Intelligence and Legal Analytics: New Tools for Law Practice in the Digital Age. Cambridge University Press, Cambridge, United Kingdom.

[4] Kevin D. Ashley and Stefanie Brüninghaus. 2009. Automatically classifying case texts and predicting outcomes. Artificial Intelligence and Law 17, 2 (2009), 125-165. DOI: https://doi.org/10.1007/s10506-009-9077-9

[5] Katie Atkinson and Trevor Bench-Capon. 2018. Lessons from implementing factors with magnitude. In Legal Knowledge and Information Systems: JURIX 2018: The Thirty-first Annual Conference, ed. Monica Palmirani (Amsterdam: IOS Press, 2018), 11-20. DOI: https://doi.org/10.3233/978-1-61499-935-5-11

[6] Or Biran and Courtenay Cotton. 2017. Explanation and justification in machine learning: A survey (IJCAI 2017). Retrieved January 27, 2019 from http://home.earthlink.net/~dwaha/research/meetings/ijcai17-xai/

[7] Alison Chorley and Trevor Bench-Capon. 2005. AGATHA: Using heuristic search to automate the construction of case law theories. Artificial Intelligence and Law 13, 1 (March 2005), 9-51. DOI: https://doi.org/10.1007/s10506-006-9004-2

[8] Mohammad H. Falakmasir and Kevin D. Ashley. 2017. Utilizing Vector Space Models for Identifying Legal Factors from Text. In Legal Knowledge and Information Systems: JURIX 2017: The Thirtieth Annual Conference, eds. Adam Wyner and Giovanni Casini (Amsterdam: IOS Press, 2017),183-192. DOI: https://doi.org/10.3233/978-1-61499-838-9-183

[9] Matthias Grabmair. 2017. Predicting trade secret case outcomes using argument schemes and learned quantitative value effect tradeoffs. In Proceedings of the 16th edition of the International Conference on Artificial Intelligence and Law (ICAIL '17). ACM, New York, NY, USA, 89-98. DOI: https://doi.org/10.1145/3086512.3086521

[10] Octavia-Maria Sulea, Marcos Zampieri, Mihaela Vela and Josef van Genabith. 2017. Predicting the law area and decisions of french supreme court cases. arXiv:1701.170801681. Retrieved from https://arxiv.org/abs/1708.01681.

[11] Vern R. Walker. 2007. A Default-Logic Paradigm for Legal Fact-Finding. Jurimetrics 47, 2 (2007), 193-243. Retrieved January 28, 2019 from http://www.jstor.org/stable/29762966.

[12] Vern R. Walker, Nathaniel Carie, Courtney C. DeWitt, and Eric Lesh. 2011. A framework for the extraction and modeling of fact-finding reasoning from legal decisions: lessons from the Vaccine/Injury Project Corpus. Artificial Intelligence and Law 19, 4 (August 2011), 291-331. DOI: https://doi.org/10.1007/s10506-006-9004-2

[13] Vern R. Walker, Bernadette C. Lopez, Matthew T. Rutchik, and Julie L. Agris. 2015. Representing the Logic of Statutory Rules in the United States. In Legisprudence Library (Studies on the Theory and Practice of Legislation), vol 2 : Logic in the Theory and Practice of Lawmaking, eds. Michał Araszkiewicz and Krzysztof Płeszka (Cham: Springer, 2015), 357-381. DOI: https://doi.org/10.1007/978-3-319-19575-9_13

[14] Vern R. Walker, Ji Hae Han, Xiang Ni, and Kaneyasu Yoseda. 2017. Semantic types for computational legal reasoning: propositional connectives and sentence roles in the veterans' claims dataset. In Proceedings of the 16th edition of the International Conference on Artificial Intelligence and Law (ICAIL '17). ACM, New York, NY, USA, 217-226. DOI: https://doi.org/10.1145/3086512.3086535

[15] Adam Wyner and Wim Peters. 2010. Towards annotating and extracting textual legal case factors. In Semantic Processing of Legal Texts (SPLeT-2012), 36-45. Retrieved January 28, 2019 from http://www.lrec-conf.org/proceedings/lrec2010/workshops/W23.pdf.

[16] Adam Wyner and Wim Peters. 2012. Semantic Annotations for Legal Text Processing using GATE Teamware. In Semantic Processing of Legal Texts (SPLeT-2012), 34-36. Retrieved January 28, 2019 from http://www.lrec-conf.org/proceedings/lrec2012/workshops/27.LREC%202012%20Workshop%20Proceedings%20SPLeT.pdf.

[17] Lisa Webley. 'Qualitative Approaches to Empirical Legal Research' in The Oxford Handbook of Empirical Legal Research, eds. Peter Cane and Herbert Kritzer (Oxford: Oxford University Press, 2010), 926-950. DOI: https://doi.org/10.1007/3-540-09237-4

[18] Jaromír Šavelka and Kevin D. Ashley. 2018. Segmenting U.S. Court Decisions into Functional and Issue Specific Parts. In Legal Knowledge and Information Systems: JURIX 2018: The Thirty-first Annual Conference, ed. Monica Palmirani (Amsterdam: IOS Press, 2018), 111-120. DOI: https://doi.org/10.3233/978-1-61499-935-5-111