Automatically Identifying Eviction Cases and Outcomes within Case Law of Dutch Courts of First Instance

Masha MEDVEDEVA [a,b,1], Thijmen DAM [a,b], Martijn WIELING [b] and Michel VOLS [a]
[a] *Department of Legal Methods, University of Groningen (UG), The Netherlands*
[b] *Center for Language and Cognition Groningen, UG, The Netherlands*

Abstract. In this paper we attempt to identify eviction judgements within all case law published by Dutch courts in order to automate data collection, previously conducted manually. To do so we performed two experiments. The first focused on identifying judgements related to eviction, while the second focused on identifying the outcome of the cases in the judgements (eviction vs. dismissal of the landlord's claim). In the process of conducting the experiments for this study, we have created a manually annotated dataset with eviction-related judgements and their outcomes.

Keywords. outcome identification, case law, machine learning, judicial decision

1. Introduction

Legal scholars and practitioners are confronted with an enormous and expanding body of case law. For example, in the Netherlands the judiciary dealt with over 1.3 million cases in 2020 alone.[2] Many of these cases involve bulk cases on, for example, family law or labour law. Another area that results in a considerable number of bulk cases is landlord-tenant law. It is estimated that courts have to decide whether or not a tenant needs to be evicted in nearly 20.000 cases every year (1). The Dutch judiciary does not publish all judgements online, but a significant number of cases can be found online in the *Open Data van de Rechtspraak* dataset.[3] Traditionally, legal scholars and practitioners collect and analyse these cases manually (2). Of course, this is time-consuming and will become impossible due to the increasing amount of published case law online.[4]

In this paper we are trying to solve this legal research problem. Specifically, we want to identify judgements concerning eviction within all judgements published by the Dutch judiciary and extract their outcome from the text (i.e. eviction/non-eviction). This work builds upon existing research that until now has been done manually (3), and our goal is to test how much of the data collection and outcome extraction can be automated. Some

[1] Corresponding Author. E-mail: m.medvedeva@rug.nl

[2] `https://www.rechtspraak.nl/Organisatie-en-contact/Rechtspraak-in-Nederland/Rechtspraak-in-cijfers` (Dutch)

[3] `https://www.rechtspraak.nl/Uitspraken/Paginas/Selectiecriteria.aspx` (Dutch)

[4] `https://www.volkskrant.nl/nieuws-achtergrond/raad-voor-de-rechtspraak-meer-vonnissen-online-publicerenb̃f045df7` (Dutch)

of the case law under review has already been annotated by hand and can be used for training machine learning systems.

In this paper, we use 'judgement' to denote the text of a published judgement. The word 'outcome', and its synonyms 'verdict' and 'decision', are used to define a specific closed class (i.e. a limited number) of labels for verdicts. An example of an outcomes is *eviction* or *non-eviction* in the landlord-tenant law context.

2. Related work

This paper deals with the identification of the topic of a judgement (i.e. an eviction case or a non-eviction case). To our knowledge, the number of publications dealing with automatically identifying the topic of a case for dataset creation is limited. Some similar work involves topic modelling techniques that allow one to identify and cluster multiple topics at once (4; 5; 6), and using document similarity to find the documents that deal with similar issues (7; 8), both of which can be particularly hard to evaluate.

Besides the identification of the topic of a judgement, this paper concerns outcome identification (i.e. the extraction of the outcome from the full text). This identification task can be useful in itself, for instance if one wants to know the statistics of cases concerning eviction actually ending in a tenant being evicted.

Depending on the court, identifying the outcome can be more or less complicated. Some courts publish their judgements with meta-data stating the outcome (e.g., the European Court of Human Rights). As a result, one just needs to extract this information in order to get the outcomes. While in other judgements, the wording of the outcome may be standardised and therefore easy to extract (e.g., "The Court of Appeal therefore affirms the decision of the Court of First Instance."), the majority of courts seem to formulate their decisions in free-form natural language, making the task of extracting a specific outcome a more complex task.

There is a small number of studies focusing specifically on identifying the outcome within the judgements. Recent papers extracted outcomes from Appellate Decisions in US State courts (9; F1-score: 0.82), US federal court dockets (10; recall up to 0.96) as well as French courts (11; F1-scores: 0.8-1.0) using various machine learning methods. In this paper we compare the performance of a simpler keyword-search approach (not requiring annotated data) to a simple machine learning system.

3. Data

For our dataset we rely upon the *Open Data van de Rechtspraak*,[5] an official, publicly available, database of the judiciary of the Netherlands. Not all Dutch case law is published online, but merely a subset of judgements that *De Rechtspraak* allows for publication. Unfortunately, exact criteria are not available to the public, though some guidance is provided by a dedicated page on their website.[6] The *Open Data van de Rechtspraak* dataset can be downloaded as one large file archive (>4GB) of XML files containing the texts of the judgements as well as some basic meta-data (e.g., court, date).

[5] `https://www.rechtspraak.nl` (Dutch)

[6] `https://www.rechtspraak.nl/Uitspraken/Paginas/Selectiecriteria.aspx` (Dutch)

For this paper, we are specifically interested in the cases of the courts of first instance (*rechtbanken*). A collection of 591 eviction cases between 2000 and 2020 (manually collected and annotated, including the verdicts: eviction or non-eviction) from the courts of first instance was already available (based on existing research from our lab). This dataset was compiled in such a way that it should include the large majority of all published eviction cases between 2000 and 2020. As this dataset only contains a relatively limited number of eviction cases, and no non-eviction cases, we aimed to supplement it by including both cases about eviction, and cases on other topics, but still somewhat related to the subject matter. This was to ensure the task was useful and not trivial, and that we had a larger dataset to train the system.

To increase the likelihood of identifying eviction cases, we used the following procedure. We extracted all (2,641,946) judgements from the *Open Data van de Rechtspraak* dataset between the years 2000 and 2020. From this set, we only included judgements from the courts of first instance, and furthermore selected the judgements that contain at least one of the following words *huurovereenkomst* (rental agreement), *ontruiming* (eviction) or *woning* (home). Subsequently, we narrowed down the selection by only retaining judgements with the label "private law", which is the appropriate label for eviction cases. These relatively simple filters allowed us to reduce the amount of judgements to 24,268 cases. Unfortunately, this number was still rather large. Consequently, we made a further reduction by only including cases from 2016-2018, and excluding cases already included in the original set of 591 cases of the original set, yielding a set of 4,795 judgements. From this set, we randomly sampled 69 judgements (1 hour of manual annotation) to assess the proportion of cases related to eviction. A manual inspection showed that more than half of the judgements (37) were eviction cases. This suggests that our manually curated dataset of 591 eviction cases was missing a substantial amount of eviction-related cases.

To increase the amount of data, we took all 591 manually annotated eviction judgements, and we again randomly sampled from the 2016-2018 judgements, extracting twice the amount (1182) of manually annotated eviction judgements. We then built a simple three-fold cross-validation support vector machine (SVM) only using 1-3 n-grams (i.e. sequences of one to three words from the text of the judgement) as features. When training the model, we treated the 591 judgements as eviction cases and the 1182 judgements as non-eviction cases.[7] Of course this is a sub-optimal class distinction, as potentially many of the 1182 judgements may, in fact, be eviction cases. Consequently, from all cases that were classified as non-eviction cases, we only retained those which were (when included in the test set during the three-fold cross-validation procedure) assigned the non-eviction label with over 99% confidence (using Platt Scaling; 13). This reduced the number of non-eviction judgements in our training set to 809. We then trained the system again (using 809 non-eviction cases and 591 eviction cases) and evaluated it by using the rest of the judgements from between 2000 and 2020. Out of 22,868 judgements, 3,277 (14%) were predicted as eviction-related.

Of course, not all predictions will be correct. To supplement our final correct training dataset, however, we did not use these predictions, but instead used these simply to guide two subsequent manual annotation rounds (the first annotation round included the 69 aforementioned cases). Specifically, we asked two legal experts to spend eight hours

[7] For a more detailed explanation of machine learning classification and its evaluation (i.e. precision, recall, f1-score, accuracy) applied to legal texts, see (12) and (4).

in total on annotating judgements that our model predicted as eviction-related in the second annotation round (under the assumption that many would *not* be eviction related), and an additional four hours in total in a third annotation round focusing on the judgements that our predicted as *non-eviction*. The annotators were provided the full text of a randomly selected judgement and they had to simply identify whether the judgement was concerning an eviction or not. In the allocated time, 716 judgements were annotated. Out of predicted eviction judgements 298 (55%) turned out to be eviction related, while 243 judgements (45%) were not. In addition, and the vast majority of non-eviction cases 161 out 175 (92%) turned indeed out to be non-eviction related. This left us with a dataset with 940 eviction judgements, and 436 non-eviction judgements. Table 1 provides an overview of our final dataset.

Table 1. Number of available data in the initial dataset and after three rounds of annotation.

	Eviction	Non-eviction
Initial dataset	591	0
First annotation round	37	32
Second annotation round (predicted as eviction)	298	243
Third annotation round (predicted as non-eviction)	14	161
Total	940	436

Once we identify the eviction-related judgements, we are also interested in their outcome. In the judgements concerning evictions, the courts of first instance can decide to evict the resident and/or cancel the lease (labelled as *eviction*) or reject the property owner's claim (labelled as *non-eviction*). The cases are decided on by a single judge. All of the eviction cases in the court of first instance are property owner vs. resident, with the latter being the defendant.

4. Experiment I: Identifying Eviction-Related Judgements

4.1. Methodology

From the final dataset (see Table 1), we used 200 judgements (100 eviction-related and 100 non-related) to test and evaluate the model, which left us with 840 eviction and 336 non-eviction judgements to train and fine-tune the system. We then balanced this dataset for training, leaving us with 336 eviction-related judgements and the same number of non-related judgements. We used three-fold cross-validation to fine-tune the parameters and ended up using a linear support vector machine, using as features the frequencies of 1-6 character n-grams (i.e. sequences of one to six characters).[8] The results of the best model can be found in Tables 2 and 3.

[8]The following command, showing all used parameters, was used to fit our final model: `CountVectorizer(analyzer = 'char', ngram_range = (1,6), max_features=None, max_df = 0.7, lowercase=False, binary=True); LinearSVC(C=0.01)`. For more details on each parameter see the `sklearn` documentation available at `https://scikit-learn.org/`. The full set of parameters we experimented with can be found in our code and data available at `https://github.com/masha-medvedeva/EVICT`.

Table 2. Results (precision, recall, f1-score and accuracy) for identifying eviction-related judgements using three-fold cross-validation.

	Precision	Recall	F1-score	Support
Non-eviction	0.90	0.88	0.89	336
Eviction	0.88	0.90	0.89	336
Accuracy			0.89	672
Macro avg.	0.89	0.89	0.89	672
Weighted avg.	0.89	0.89	0.89	672

Table 3. Results (confusion matrix) for identifying eviction-related judgements using three-fold cross-validation.

		Actual topic	
		Non-eviction	Eviction
Predicted topic	Non-eviction	294	42
	Eviction	32	304

4.2. Results

Our final results, when evaluating our model on the held-out test set, are shown in Tables 4 and 5.

Table 4. Results (precision, recall, f1-score and accuracy) on the test set for identifying eviction-related judgements.

	Precision	Recall	F1-score	Support
Non-eviction	0.92	0.81	0.87	100
Eviction	0.83	0.95	0.89	100
Accuracy			0.88	200
Macro avg.	0.89	0.88	0.88	200
Weighted avg.	0.89	0.88	0.88	200

Table 5. Results (confusion matrix) on the test set for identifying eviction-related judgements.

		Actual topic	
		Non-eviction	Eviction
Predicted topic	Non-eviction	81	19
	Eviction	5	95

The results suggest that when having a reasonable amount of annotated data, it is possible to identify eviction-related cases with a relatively high accuracy of about 88%. Consequently, this automatic procedure can be suitably used to speed up the process of finding relevant (eviction-related) case law.

When we evaluated the model on all (filtered) judgements published between 2000 and 2020 not included in our dataset, a total of 3,248 out 22,872 cases (all original judgements between 2000 and 2020, excluding all annotated judgements) were classified as eviction-related judgements. With an estimated precision of 83%, we expect about 2,695 cases to be actual eviction-related judgements. Similarly, with an estimated precision of 92% in identifying non-eviction related judgements, we expect an additional 8% of these (i.e. 1569 judgements) to be eviction-related.

5. Experiment II: Identifying the Outcome

5.1. Methodology

Once we identified the eviction-related judgements, we were interested in identifying how many of them actually resulted in the eviction of a resident. Identifying the verdict should not necessarily always be a machine learning task. A simple keyword search could potentially be sufficient. Therefore, we first tried determining words that may be characteristic of a specific outcome. While judgements of the courts of first instance do not have a clear structure, they could potentially use the same wording for the verdict itself. We then compare these results to a more sophisticated machine learning system which is able to take more advanced features into account.

For this experiment we used the full set of 940 eviction-related cases shown in Table 1. Except for the cases included in the initial dataset which already included an annotated outcome, we asked two legal experts to annotate the outcome of each case: *eviction* or *non-eviction*. We excluded 28 cases that had other types of verdicts, such as only cancellation of the lease, but no eviction, etc. The final dataset for this task contained 912 judgements (620 having an eviction outcome, whereas 292 had a non-eviction outcome).

5.1.1. Keyword-Based System

For the keyword-based system, we determined (via manual inspection of several cases) a number of terms that relate to each specific outcome. We then automatically searched for these terms in the *decision* section of the published judgement, and in cases where the decision section was not specified, in the bottom part (2500 characters) of the text. The keywords that we chose for being representative of an *eviction* outcome were (including different forms of the same words): *ontbinding* (cancellation), *ontruiming* (eviction), and *verlaten* (leave). If none of these words were found, our keyword-based system determined that no eviction had been ordered by the court.

We tested the method on all 912 judgements that we had labels for. The results of this system can be found in Tables 6 and 7.

Table 6. Results (precision, recall, f1-score and accuracy) for identifying the outcome of eviction cases using keyword extraction.

	Precision	Recall	F1-score	Support
Non-eviction	0.88	0.65	0.75	292
Eviction	0.85	0.96	0.90	620
Accuracy			0.86	912
Macro avg.	0.87	0.80	0.82	912
Weighted avg.	0.86	0.86	0.85	912

Table 7. Results (confusion matrix) on the test set of identifying the outcome of eviction cases using keyword extraction.

		Actual outcome	
		Non-eviction	Eviction
Predicted outcome	Non-eviction	189	103
	Eviction	25	595

This simple system achieved reasonably good results, although we can see that non-eviction is not categorised very well, 103 (35%) out 299 non-eviction cases were misclassified. However, the issue with a keyword-based system, is that it is very hard to improve upon, unless one can come up with more specific keywords. Moreover, if the keywords from one type of outcome are found in the judgement with a different outcome, this is hard to correct. For instance, a judgement can contain the phrase "at this point, eviction is not necessary". While the word 'eviction' is present in this judgement, the case clearly resulted in no eviction. However since we are just dealing with individual words, it is hard to incorporate all possible nuances.

Nonetheless, as opposed to a system using machine learning, which we will discuss in the next subsection, this system does not require any prior annotation, other than determining the keywords.

5.1.2. Machine Learning System

During the keyword-based experiment, we determined that the outcome usually appears within the last 2500 characters of the judgement. While we experimented with shorter and longer fragments, this subset seemed to work best for both of the experiments in identifying the verdict. We used the initial dataset for training and cases from the first, second and third rounds of annotation for testing. We have built a Linear SVC that uses character (1-7) n-grams, and optimised it for a number of other parameters.[9] The results of the model during the cross-validation stage can be found in Tables 8 and 9.

Table 8. Results (precision, recall, f1-score and accuracy) for identifying the outcome of eviction cases using three-fold cross-validation.

	Precision	Recall	F1-score	Support
Non-eviction	0.97	0.96	0.96	183
Eviction	0.98	0.99	0.98	379
Accuracy			0.98	562
Macro avg.	0.98	0.97	0.97	562
Weighted avg.	0.98	0.98	0.98	562

Table 9. Results (confusion matrix) for identifying the outcome of eviction cases using three-fold cross-validation.

		Actual outcome	
		Non-eviction	Eviction
Predicted outcome	Non-eviction	175	8
	Eviction	5	374

5.2. Results

We then tested the model on the cases that we were able to extract in the previous experiment. The performance on the test set can be found in Tables 10 and 11.

[9]The following command, showing all used parameters, was used to fit our final model: `CountVectorizer(analyzer = 'char', ngram_range = (1,7), max_features=2000, max_df = 0.9, lowercase=True, binary=True); LinearSVC(C=0.001)` The full set of parameters we experimented with can be found in our code and data available at `https://github.com/masha-medvedeva/EVICT`.

Table 10. Results (precision, recall, f1-score and accuracy) on a test set for identifying the verdict of eviction cases.

	Precision	Recall	F1-score	Support
Non-eviction	0.82	0.94	0.88	109
Eviction	0.97	0.91	0.94	241
Accuracy			0.92	350
Macro avg.	0.90	0.92	0.91	350
Weighted avg.	0.92	0.92	0.92	350

Table 11. Results (confusion matrix) on a test set for identifying the verdict of eviction cases.

		Actual outcome	
		Non-eviction	Eviction
Predicted outcome	Non-eviction	102	7
	Eviction	22	219

As we can see from the results, we were able to achieve a very high performance, especially for the eviction class. When inspecting the cases manually, it is clear that the phrasing of the judgement outcomes varies to a large extent from case to case. Similar as in many other natural language processing tasks, the best-performing model included not word n-grams, but character n-grams (14; 15). While we did try using word n-grams for this experiment, in the hope of identifying additional keywords for the keyword-based approach, we did not identify any additional unique words for both outcomes. The performance of the machine learning approach was much higher than the performance of the keyword-based approach. However, whereas the machine learning approach requires annotated data, the keyword-based method does not.

6. Discussion and Conclusion

In this paper we have presented two experiments, one to identify case law having a certain topic, specifically judgements concerning evictions, and one to subsequently identify the outcomes of these eviction judgements. For both tasks, we have shown a high performance, being able to identify eviction-related cases with 88% accuracy, and correctly identifying the outcome in eviction-related cases in 92% of cases. While identifying this type of information may seem easy (as the information is available when reading the document), a keyword-based approach showed it is not straightforward when the information is provided as natural text. While in this paper we were not able to identify *all* eviction cases perfectly, our machine learning approach can suitably be used to identify cases which are *likely* to be eviction cases. Manually checking this smaller set of cases (at a rate of about 1 case per minute) is feasible, whereas checking the full set is not. With relatively little effort, a new database containing thousands of cases can therefore easily be created.

Such a more restricted subject-specific database is also useful in the context of increasing research focusing on categorising or forecasting court decisions (12; 16; 17; 18; 19). This type of research is mostly limited to only a few courts, such as the US Supreme Court or the European Court of Human Rights. This is partly due to the courts' publishing policies, even though more and more courts publish their case law. The dom-

inant focus on a few courts, however, is also caused by the relative large diversity of uncategorised cases in other courts. Therefore narrowing down the task, as we have done here, will likely help subject-specific machine learning systems for these courts (e.g., distinguishing between bankruptcy cases) to be developed.

References

[1] Vols M. Evictions in the Netherlands. In: Loss of Homes and Evictions across Europe. Edward Elgar Publishing; 2018. p. 214–238.

[2] Vol M. Legal Research. One Hundred Questions and Answers. Eleven; 2021.

[3] Vols M. European Law and Evictions: Property, Proportionality and Vulnerable People. European Review of Private Law. 2019;27(4).

[4] Dyevre A. Text-mining for Lawyers: How Machine Learning Techniques Can Advance our Understanding of Legal Discourse. Available at SSRN 3734430. 2020.

[5] Silveira R, Fernandes CG, Neto JAM, Furtado V, Pimentel Filho JE. Topic Modelling of Legal Documents via LEGAL-BERT. Proceedings http://ceur-ws org ISSN. 2021;1613:0073.

[6] Remmits Y. Finding the topics of case law: Latent dirichlet allocation on supreme court decisions [Bachelor's thesis]; 2017.

[7] Novotná T, et al. Document Similarity of Czech Supreme Court Decisions. Masaryk University Journal of Law and Technology. 2020;14(1):105-22.

[8] Barco Ranera LT, Solano GA, Oco N. Retrieval of Semantically Similar Philippine Supreme Court Case Decisions using Doc2Vec. In: 2019 International Symposium on Multimedia and Communication Technology (ISMAC); 2019. p. 1-6.

[9] Petrova A, Armour J, Lukasiewicz T. Extracting Outcomes from Appellate Decisions in US State Courts. In: Legal Knowledge and Information Systems: JURIX 2020: The Thirty-third Annual Conference, Brno, Czech Republic, December 9-11, 2020. vol. 334. IOS Press; 2020. p. 133.

[10] Vacek T, Schilder F. A sequence approach to case outcome detection. In: Proceedings of the 16th edition of the International Conference on Artificial Intelligence and Law; 2017. p. 209-15.

[11] Tagny-Ngompé G, Mussard S, Zambrano G, Harispe S, Montmain J. Identification of Judicial Outcomes in Judgments: A Generalized Gini-PLS Approach. Stats. 2020;3(4):427-43.

[12] Medvedeva M, Vols M, Wieling M. Using machine learning to predict decisions of the European Court of Human Rights. Artificial Intelligence and Law. 2020;28(2):237-66.

[13] Platt J, et al. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. Advances in large margin classifiers. 1999;10(3):61-74.

[14] Basile A, Dwyer G, Medvedeva M, Rawee J, Haagsma H, Nissim M. N-GrAM: New Groningen Author-profiling Model—Notebook for PAN at CLEF 2017. In: CEUR Workshop Proceedings. vol. 1866; 2017. .

[15] Medvedeva M, Kroon M, Plank B. When sparse traditional models outperform dense neural networks: the curious case of discriminating between similar languages. In: Proceedings of the Fourth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial); 2017. p. 156-63.

[16] Chalkidis I, Androutsopoulos I, Aletras N. Neural Legal Judgment Prediction in English. In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. Florence, Italy: Association for Computational Linguistics; 2019. p. 4317-23. Available from: https://www.aclweb.org/anthology/P19-1424.

[17] Katz DM, Bommarito II MJ, Blackman J. A General Approach for Predicting the Behavior of the Supreme Court of the United States. PloS one. 2017;12(4).

[18] Waltl B, Bonczek G, Scepankova E, Landthaler J, Matthes F. Predicting the outcome of appeal decisions in Germany's tax law. In: International Conference on Electronic Participation. Springer; 2017. p. 89-99.

[19] Strickson B, De La Iglesia B. Legal Judgement Prediction for UK Courts. In: Proceedings of the 2020 The 3rd International Conference on Information Science and System; 2020. p. 204-9.