Artificial Intelligence and Predictive Justice: Limitations and Perspectives

Marc Queudot and Marie-Jean Meurs(✉)

Université du Québec à Montréal, Montréal, QC, Canada
meurs.marie-jean@uqam.ca

Abstract. One of the main barriers to effective prediction systems in the legal domain is the very limited availability of relevant data. This paper discusses the particular case of the Federal Court of Canada, and describes some perspectives on how best to overcome these problems. Part of the process involves an automatic annotation system, supervised by a manual annotation process. Several state-of-the-art methods on related tasks are presented, as well as promising approaches leveraging recent advances in natural language processing, such as vector word representations or recurrent neural networks. The insights outlined in the paper will be further explored in a near future, as this work is still an ongoing research.

Keywords: Legal artificial intelligence · Predictive justice
Natural language processing · Machine learning

1 Introduction

The search for techniques to predict justice court decisions has been going on for decades. In 1963, Loevinger described a new science called *Jurimetrics*, concerned among other matters, by "the formulation of a calculus of legal predictability" [14, p. 8]. More recently, Zambrano [26] argued that the ability to accurately predict the probable outcome of a court decision would help lawyers to define their legal strategies, and would also help to relieve congestion in courts since some hopeless cases will be dropped. Our work comes at a time where, according to Richard Tromans in his introduction to Legal AI [23], the legal field begins to adopt software solutions that rely on Machine Learning and Natural Language Processing (NLP). Our work will focus on predictive systems, i.e. predicting the probable outcome of court cases. In this area, many start-ups like Juristat, Lex Predict or Lex Machina, as recensed by Zambrano [26], offer this service to their client. However, the legal field would benefit from more accessible systems that leverage recent advances in sentiment analysis. Access to complete sets of case law is paramount for predictive justice algorithms. Legal decision-making is a process relying heavily on the rule of precedent, even within one case file. The same judge or adjudicator (prothonotary or judge for the Federal Court) may rely on prior decisions, where he or she can find statements

© Springer International Publishing AG, part of Springer Nature 2018
M. Mouhoub et al. (Eds.): IEA/AIE 2018, LNAI 10868, pp. 889–897, 2018.
https://doi.org/10.1007/978-3-319-92058-0_85

of facts and legal reasoning to frame his or her reasons for an order or a judgment. Algorithms can learn to identify common threads and patterns in order to establish a predictive analysis for issues of decisions by adjudicators, provided the precedents constitute a large enough and as complete as possible set of case law.

However, access to complete sets of case law is a thorny issue with the Federal Court of Canada, as its public and free website database is incomplete and not up-to-date. For instance, performing a search for all decisions issued in a case such as Osmose-Pentox, Inc. v. Société Laurentide, Inc., yield roughly 11 decisions, yet when consulting the files history, we find that some decisions dated 2003 and 2005 have been rendered but do not appear in the search. This is illustrative of a much wider problem. Indeed, missing decisions can be accessed through online private and pay-per-use databanks like Westlaw (Law Source) or Quicklaw. Online databanks such as these have proprietary rights on indexed court decisions, and will not give massive access to their databanks. Therefore, access to complete and updated court decisions issued by the Federal Court is tributary of the investment public authorities are willing to put on the website, and in particular the indexing of court decisions in a format readily accessible to the public and researchers. The scope and effectiveness of the predictive algorithm is, as a result, impaired in its capacity to learn and draw inferences with a high probability. Creating our own indexed databases could be a solution, but it would require a staggering amount of resources both financial and in man power, not to mention time-consuming.

The goal of this work is to overcome the difficulties associated with the data access, to provide good quality predictions on justice court decisions. The paper is organized as follows: Sect. 2 presents an overview of various works summarizing the current state-of-the-art in the field, Sect. 3 describes the corpus used for the experiments, Sect. 4 explains our approach while planned experiments are reported in Sect. 5, and finally Sect. 6 concludes.

2 Previous Work

Sulea et al. [21] made predictions on decisions from the French Supreme Court using ensembles of Support Vector Machine (SVM)[7] classifiers. Their dataset contains fields, among others, for the case description and a label of the ruling. They evaluate their approach using the F1-measure. F1-measure is a metric to measure the performance of a system that balances two components. The first one is the number of cases it found from all the ones it was supposed to, while the second one measures the proportion of cases it correctly classified. More details on these metrics will be given in Sect. 4. They achieve an F1-measure of 98,6% in predicting the ruling of cases when such rulings are organized in six classes. This score drops to 95.8% when eight classes are used. In their previous work [22], the authors explained that the labels were derived from the `Conclusion(s)` (outcome) meta-data field of the database they used. In their first experiment, they used the 200 most frequent outcomes, by selecting only its first word. For the

second experiment they used the 200 most frequent outcomes, without filtering on the first word. This process produced respectively 6 and 8 classes, on which they performed their classification experiments. Katz et al. [12] predicted rulings of the United States Supreme Court using Random Forest (RF) [6] classifiers. They found that this algorithm "outperformed [...] support vector machines and feed-forward artificial neural network models" [12, p. 7] in their experiments using the Supreme Court Database[1]. Another benefit of using RF is their ability to perform incremental learning, i.e. improve the model performance when new data are available, without re-training the whole model. They have an interesting approach for evaluating the performance of their system: in addition to using standard metrics, they also develop a baseline algorithm. This baseline "algorithm" always predicts the class most ruled by the Supreme Court judges over the previous years. They found that their approach mostly outperformed the baseline once the training data were of sufficient volume. Aletras et al. [3] classified decisions of the European Court of Human Rights (ECtHR) on matters concerning 3 distinct articles of the European Convention on Human Rights. They obtained good results on their two class classification (either `violating` or `not violating` the convention) by engineering a *topic* features by aggregating semantically close n-grams. Although all these studies produce interesting results, it is important to note that the jurisdictions of the legal systems whose decisions they predict are relatively limited. The United States Supreme Court and the European Court of Human Rights only deal with issues regarding their respective constitutions, and [3] further reduced this scope to only 3 articles. The context is different for the French Supreme Court which only judges on matters of law and its application, as opposed to the study of both facts and law in lower level courts. After Bengio et al. trained word embeddings [5] for the first time in 2003, Mikolov et al. popularized them with the introduction of word2vec [15], a toolkit to train or use pre-trained word-embeddings. Ever since, word embeddings representations pre-trained on big amounts of data such as GloVe [16] have been providing state-of-the-art results in a wide range of semantic tasks [19]. Experiments training cross-lingual embeddings were also successful, allowing to use the huge datasets available in English to improve the performance of models in other languages. In his survey of cross-language embeddings, Ruder et al. [18] described the different approaches developed in recent years. Radford et al. [17] show that even the best vector representations, namely skip-thoughts [13], are still outperformed by supervised models. They reference the work of Dai and Le [8], which fine-tunes a pre-trained unsupervised language model to achieve state-of-the-art performance on some classification datasets, and Dieng et al. [9] that combine unsupervised language modeling with topic modeling and small supervised feature extraction to improve the performance of their model. To deal with inputs of variable length such as text, Recurrent Neural Networks (RNNs), a class of neural networks, have been developed, and in particular Long Short-Term Memory (LSTM) [11] networks which bring performance optimizations. Radford et al. trained a particular kind of LSTM called Multiplicative

[1] United States Supreme Court decisions dataset http://scdb.wustl.edu/.

LSTM (mLSTM), which use a more input-dependent hidden state transition than regular LSTMs. The conclusion of Radford et al.'s work is that language models trained on a big corpus of books cannot be expected to carry enough information to perform well on more specific tasks like review classification, namely sentiment analysis. The next section describes the dataset that we used.

3 Dataset

The dataset consists of all the decisions taken by the Federal Court of Canada (more than 45 000) that were available on the official Federal Court website[2] in August 2017. The decisions can be rendered either in English or French, but have to be traduced upon request and usually are. As a result, most of the decisions are available in both languages. However, 5% of the decisions are only available in English, and 1.2% only in French. Table 1 presents these statistics on the dataset. The rulings are formated as free texts but include the following information before the argument: date and place of ruling, name of the judge or prothonotary, and names of the parties involved in addition to several identifiers. The main part of the decision is the argument, in which the facts are exposed, relevant laws are cited, and reasoning behind the ruling is detailed. Then, one finds the ruling itself, usually in the last paragraph before stating the information found in the header, plus the names of the people who appeared in the trial, and the solicitors of record.

Table 1. Statistics on the dataset

	Occurrences	% of the total number of decisions
Distinct decisions	46 369	100%
English-only decisions	2 329	5%
French-only decisions	602	1.2%
Number of different judges[a]	41	-

[a]This statistic comes from the Federal Court website.

4 Methodology

We gathered the decisions from the website, combined the French and English versions of the same decisions, and indexed them using Apache Lucene[3]. This allowed us to use the Apache Solr[4] search engine in a preliminary exploration phase. The decisions of the Federal Court are not annotated, so we can not apply supervised learning techniques on this corpus as it is. The lack of structure in

[2] http://cas-cdc-www02.cas-satj.gc.ca/fct-cf/.
[3] https://lucene.apache.org/core/.
[4] https://lucene.apache.org/solr/.

the dataset makes it difficult to extract the ruling as categorical variable but the ruling is in the text of the decision, even though there is no established format or way of expressing the ruling. However, a few similar sentences covered a large fraction of the expressed rulings. Using regular expressions, we were able to categorize 12 136 documents. Regular expressions work by matching patterns in text using a particular syntax, and returning the portions of text found this way. We have defined two categories: `granted` and `dismissed`. We select the first one when the judge or prothonotary grants to the plaintiff what they requested, and the other when their request is rejected. We added a third one, "`unknown`", to indicate that we could not choose either class because the regular expressions did not match anything. The specific regular expressions we used are shown in Figs. 1 and 2. The first regular expression in Fig. 1 looks for a portion of text that begins by *orders that* and ends with *dismissed*, regardless of the case. In the second one, we look for a portion of text beginning by *application* followed by none to forty characters, and then *is dismissed*. The third one is the exact part of text it looks for, and the following are equivalent in French. The regular expressions for the `dismissed` class are built in the same way, but the keyword *dismissed* is replaced by *granted*.

```
(?i)orders that(.*)dismissed
application [\w ]{0,40}is dismissed
judicial review is dismissed
(?i)ordonne[ ]?:(.*)rejeté(e)?
demande [\w ]{0,40}est rejetée
```

Fig. 1. Regular expressions for the `dismissed` class

```
(?i)orders that(.*)granted
(?i)application [\w ]{0,40}is granted
judicial review is granted
(?i)ordonne[ ]?:(.*)accueilli(e)?
(?i)demande est accueillie
```

Fig. 2. Regular expressions for the `granted` class

To evaluate the classes extracted by using regular expressions, we plan to manually annotate about 1 000 decisions, randomly selected. Three human experts are currently annotating these documents. We will use the class chosen by the majority and compute the Kappa score [24] to measure the likelihood of these results occurring by chance. This will allow to validate our results in two ways. First, we will compare the classes provided by automatic extraction, with those from the manual annotation process. We will then be able to compute the Precision, Recall and F-measure of our class extraction system. In the context of binary classification, i.e. discriminating positive from negative examples, *precision* is defined as the following: out of the examples that are positive, how

many have been identified as such. With True Positive (TP) being the number of positive examples classified as such, and False Positive (FP) the number of negative examples classified as positive, its formal definition is:

$$precision = \frac{TP}{TP + FP}$$

In the same context, *recall* is, out of all the positive examples, how many have been classified as such. With True Negative (TN) being the number of negative examples correctly classified and False Negative (FN) the number of positive examples classified as negative, the formal definition of recall is:

$$recall = \frac{TN}{TP + FN}$$

F-measure, also called F1-score is used to balance *precision* with *recall*. It does so with an harmonic mean of both:

$$F_1 = 2 \cdot \frac{precision \cdot recall}{precision + recall}$$

The metrics mentioned above will allow us to evaluate both the `dismissed` and `granted` class annotations performed with regular expressions, but also the ones tagged as `unknown`. These will be given a special attention to better analyze their structure and characteristics. This will allow the craft of better regular expressions, to classify a bigger proportion of documents into the `dismissed` and `granted` classes. The corpus of automatically categorized documents shows a strong imbalance: 2 208 siding with the plaintiff (positive) and 9 928 siding with the defendant. Class imbalance is a common problem for classifiers. In their study of the effect of class distribution on classifier learning, Weiss *et al.* [25] shows that "the naturally occurring class distribution often is not best for learning". If this noticed imbalance proves to be similar to the one found in the manually annotated dataset, we will undersample the majority class up to the size of the minority class to obtain a balanced dataset as recommended by Weiss *et al.*, and successfully applied in several studies [4]. The following experiments attempt to classify decisions, which have been stripped from words which are instant give-away of the class. This approach has been proposed by Sulea et al. [21] in an attempt to approximate a more realistic setting where lawyers give a quick introduction to the case as an input to the algorithm.

5 Planned Experiments

We will compare several representations of one document. The baseline classifier uses a Bag Of Words (BOW) to represent documents. In this representation, each word encountered in the corpus is used to build a dictionary. Then, each document is represented by a vector that counts the number of occurrences of each word. It has the advantage of being a simple representation, well suited to

be a baseline. While BOW computes the simplest possible score for a word (the number of occurrences), Term Frequency- Inverse Document Frequency (TF-IDF) [20] computes another score that balances the frequency of a word appearing in the document with its frequency in the corpus. This is to avoid putting the emphasis on words present in too much documents, or too few, which do not help to discriminate documents. Our next step will be to use pre-trained Skip-Thoughts [13] vector representations. We expect these representations to allow our classifiers to outperform the naive ones we tried so far. Radford et al. [17] used mLSTM to classify sentiments from Amazon reviews. As part of the Deep Learning group of algorithms, mLSTM learns the features to describe the documents along with a way to map these features to the class (which is the only part classical Machine Learning algorithms do). Provided we can find a big enough law related dataset, we want to experiment feeding this dataset to a mLSTM similar to the one Randford et al. built, thus creating a language model more adapted to our task.

6 Conclusion

We accessed a public dataset of court decisions which consists of text decisions written in English and in French. We extracted categorical values using a rule-based approach, with the intent of using them as the class of the documents in supervised learning approaches. We will validate this process by manually annotating a portion of the dataset. Then, we will build baseline systems using simple document representations and classical machine learning algorithms. To the best of our knowledge, these are the methods used in state of the art prediction of justice decisions. We will then leverage pre-trained language models to use as documents representations in order to improve the performance of our classification algorithms. We will compare these results with most recently developed mLSTM network architectures on a combination of law related corpora. Doyon [10] argues that case law should be openly accessible to anyone, either law professionals or not. Quebec Court of Appeal aggreed with the editor Wilson & Lafleur Ltée in 2000 [2] when they asked of Société Québécoise d'Information Juridique (SOQUIJ)[5] that they provide them with the all the court decisions ruled in Quebec courts. Canada Supreme Court later ruled [1] that the decisions themselves did not fall under copyright laws, and for this reason, it should be allowed to copy them. In practice, accessibility of legal corpora to the public is limited, which in turn hinders research that could benefit society itself by making law more accessible.

Reproducibility. To ensure full reproducibility and comparisons between systems, our source code will be publicly released as an open source software the following repository: https://github.com/BigMiners.

[5] SOQUIJ website: http://soquij.qc.ca/.

Acknowledgments. We thank José Bonneau for his description of the difficulties in accessing legal court decisions. We also thank Diego Maupomé and Antoine Briand for their valuable comments.

References

1. CCH Canadienne Ltée c. Barreau du Haut-Canada, 2004 CSC 13. https://scc-csc.lexum.com/scc-csc/scc-csc/fr/item/2125/index.do
2. Wilson & Lafleur inc. c. Société Québécoise d'Information Juridique, j.E. 2000–17728 (C.A.)
3. Aletras, N., Tsarapatsanis, D., Preoţiuc-Pietro, D., Lampos, V.: Predicting judicial decisions of the european court of human rights: a natural language processing perspective. PeerJ Comput. Sci. **2**, e93 (2016)
4. Almeida, H., Meurs, M.J., Kosseim, L., Butler, G., Tsang, A.: Machine learning for biomedical literature triage. PLOS ONE **9**(12) (2014)
5. Bengio, Y., Ducharme, R., Vincent, P., Jauvin, C.: A neural probabilistic language model. J. Mach. Learn. Res. **3**(Feb), 1137–1155 (2003)
6. Breiman, L.: Random forests. Mach. Learn. **45**(1), 5–32 (2001)
7. Cortes, C., Vapnik, V.: Support-vector networks. Mach. Learn. **20**(3), 273–297 (1995)
8. Dai, A.M., Le, Q.V.: Semi-supervised sequence learning. In: Advances in Neural Information Processing Systems, pp. 3079–3087 (2015)
9. Dieng, A.B., Wang, C., Gao, J., Paisley, J.: TopicRNN: a recurrent neural network with long-range semantic dependency. arXiv preprint arXiv:1611.01702 (2016)
10. Doyon, J.M.: Accessibilité aux jugements et droit d'auteur. CPI 20(3) (2008). http://www.lescpi.ca/s/2773
11. Hochreiter, S., Schmidhuber, J.: Long short-term memory. Neural Comput. **9**(8), 1735–1780 (1997)
12. Katz, D.M., Bommarito II, M.J., Blackman, J.: A general approach for predicting the behavior of the supreme court of the United States. PLoS ONE **12**(4), e0174698 (2017)
13. Kiros, R., Zhu, Y., Salakhutdinov, R.R., Zemel, R., Urtasun, R., Torralba, A., Fidler, S.: Skip-thought vectors. In: Advances in Neural Information Processing Systems, pp. 3294–3302 (2015)
14. Loevinger, L.: Jurimetrics: the methodology of legal inquiry. Law Contemp. Probl. **28**(1), 5–35 (1963)
15. Mikolov, T., Sutskever, I., Chen, K., Corrado, G.S., Dean, J.: Distributed representations of words and phrases and their compositionality. In: Advances in Neural Information Processing Systems, pp. 3111–3119 (2013)
16. Pennington, J., Socher, R., Manning, C.: Glove: global vectors for word representation. In: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 1532–1543 (2014)
17. Radford, A., Jozefowicz, R., Sutskever, I.: Learning to generate reviews and discovering sentiment. arXiv preprint arXiv:1704.01444 (2017)
18. Ruder, S.: A survey of cross-lingual embedding models. arXiv preprint arXiv:1706.04902 (2017)
19. Schnabel, T., Labutov, I., Mimno, D., Joachims, T.: Evaluation methods for unsupervised word embeddings. In: Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, pp. 298–307 (2015)

20. Spärck Jones, K.: IDF term weighting and IR research lessons. J. Doc. **60**(5), 521–523 (2004)
21. Sulea, O.M., Zampieri, M., Malmasi, S., Vela, M., Dinu, L.P., van Genabith, J.: Exploring the use of text classification in the legal domain. arXiv preprint arXiv:1710.09306 (2017)
22. Sulea, O.M., Zampieri, M., Vela, M., van Genabith, J.: Predicting the law area and decisions of french supreme court cases. arXiv preprint arXiv:1708.01681 (2017)
23. Tromans, R.: Legal AI - A Beginner's Guide. Technical report, Tromans Consulting (2017). https://blogs.thomsonreuters.com/legal-uk/wp-content/uploads/sites/14/2017/02/Legal-AI-a-beginners-guide-web.pdf
24. Viera, A.J., Garrett, J.M., et al.: Understanding interobserver agreement: the kappa statistic. Fam. Med. **37**(5), 360–363 (2005)
25. Weiss, G.M., Provost, F.: The effect of class distribution on classifier learning: an empirical study. Rutgers University (2001). http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.28.9570
26. Zambrano, G.: Précédents et prédictions jurisprudentielles à l'ère des Big Data: parier sur le résultat (probable) d'un procès (2015). https://hal.archives-ouvertes.fr/hal-01496098. Accessed 28 Jan 2018