Extracting Proceedings Data from Court Cases with Machine Learning

Bruno Mathis

stats

MDPI

Article

Extracting Proceedings Data from Court Cases with Machine Learning

Bruno Mathis [1,2]

[1] CHROME Laboratory, Nimes University, 5 Rue du Docteur Georges Salan CS 13019, 30021 Nîmes, France; brunomathis16@gmail.com
[2] European Centre of Law & Economics of ESSEC Business School, 3 Av. Bernard Hirsch, 95000 Cergy, France

Abstract: France is rolling out an open data program for all court cases, but with few metadata attached. Reusers will have to use named-entity recognition (NER) within the text body of the case to extract any value from it. Any court case may include up to 26 variables, or labels, that are related to the proceeding, regardless of the case substance. These labels are from different syntactic types: some of them are rare; others are ubiquitous. This experiment compares different algorithms, namely CRF, SpaCy, Flair and DeLFT, to extract proceedings data and uses the learning model assessment capabilities of Kairntech, an NLP platform. It shows that an NER model can apply to this large and diverse set of labels and extract data of high quality. We achieved an 87.5% F1 measure with Flair trained on more than 27,000 manual annotations. Quality may yet be improved by combining NER models by data type.

Keywords: machine learning; named-entity recognition; information extraction; judicial datae; civil procedur

Citation: Mathis, B. Extracting Proceedings Data from Court Cases with Machine Learning. *Stats* **2022**, *5*, 1305–1320. https://doi.org/10.3390/stats5040079

Academic Editor: Stéphane Mussard

Received: 5 November 2022
Accepted: 5 December 2022
Published: 13 December 2022

1. Introduction

1.1. Context and Objective

France produces between three million and four million court cases a year. A small number of them are supplied with metadata. Appellate cases often include metadata at the beginning of the text but less frequently than supreme court cases do. First-instance cases mostly do not.

Fortunately, French law prescribes that the text of any case indicates:

- The jurisdiction (tribunal or court of appeal) that rendered the case;
- The name of the judges who deliberated on it;
- Its date;
- The name of the representative of the public prosecutor, if any;
- The name of the clerk;
- The first name or denomination of the parties and their domicile or registered office;
- If applicable, the names of the lawyers or of any person who represented or assisted the parties;
- In noncontentious matters, the name of the persons who must be notified.

These data make up the metadata of the case and often show up in the header of the text. The law also says that the judgment must be motivated. It sets out the case in the form of an operative part. The French legislator has decided to have all court cases made available to the public as open data. Though some provisions restrict the content of cases that is eligible to open data, the scope of the final law is unprecedented among developed countries. A first version has been rolled out in September 2021, and the last one is scheduled in December 2025. This project opens up many interesting reuse opportunities.

Predictive justice is the most hyped type of application. Indeed, using machine learning for court cases may help predict the outcome of upcoming cases and help attorneys

advise their customers accordingly. However, using machine learning for court cases may also serve general-interest applications, which have been little talked about until now. This research seeks to pave the way for the design of a database of proceedings data that could be used by social science researchers and legal technology startups ('legaltech'). One possible use case of such a database is to statistically compare different judicial pathways: cases not followed by an appeal, cases followed by an appeal, judgments followed by an appeal in Cassation or by a referral to another appellate court, etc. These different pathways could be compared in terms of case numbers and average duration, according to jurisdictions, areas of law and years of production.

The objective of this research is to show that named-entity recognition (NER) can be applied not only to the traditional named entities (people, places, organizations) but also to the whole set of data of different kinds that are necessary to describe the proceedings of the case, regardless of its substance.

1.2. Related Work

The first experimentation of using machine learning on case law dates back to 2016. Researchers investigated how well the machine could reach the same case as the judges of the European Court of Human Rights (ECHR) on the basis of the facts and the legal proceedings of a dispute [1]. They extracted from Hudoc, ECHR's database, all the judgments of the court, in their English version, in which the judges were asked to rule on Articles 3, 6 and 8 of the European Convention on Human Rights, and divided them to as many datasets. They masked the operative part, then manually annotated in a separate file the expected result of the test with the tag "article violation" or "no violation of the article". They used support vector machine (SVM), a linear model with features based on groups of words and topics to represent legal textual information extracted from the cases. This training led to the same outcome as judges in 79% of cases.

This landmark experience was generalized, in 2019, to all the articles of the European Convention on Human Rights, again using the SVM algorithm [2]. With approximately the same corpus, a comparison between several neural models showed they delivered higher quality than SVM [3]. Again on an ECHR corpus, but with only 30 cases, a comparison between different extraction mechanisms of event dates demonstrated that transformer models achieved the best result and that dates of procedural acts were easier to extract than other dates relative to the circumstances of the case [4]. Where a court database was previously applied a rules-based program to generate metadata, an opportunity arose to conduct training on a large dataset, because no preliminary classification was required. This is what was done for 28,000 cases of the Supreme Court of the US (SCOTUS) [5] and for 130,000 ones of the French Court of Cassation [6]. The aim of the former research was to reproduce the verdict, which it successfully did in 70% of cases, and that of the latter was also to reproduce the subject of law and to estimate the time frame during which the case had been rendered.

Google's 2018 release of its BERT model [7], based on a broad corpus, was followed by sectoral variants. The so-called LEGAL-BERT model [8], shows that pretraining from scratch on legal content is more efficient than pretraining only on the original BERT base supplemented with legal content.

Most experiments on court cases have been carried out on datasets from supreme courts. These have some points in common. The text of the judgments is rather long and well structured: an SVM could provide good results because the structure of the text in different sections (basically: circumstances, facts, procedure, law, outcome) allowed for segregating groups of words by section. The description of the facts and the procedure is exhaustive. The verdict has a binary character (violation or non-violation). The dataset is large and presumably statistically representative. However, because supreme courts rule on the correct application of the law, not on the merits of cases, such experiments can hardly be extrapolated to lower jurisdictions. These experiments use classification, not NER.

In the legal domain, NER can apply not only to names of people, places or organizations but also to names of laws and procedures or to concepts [9]. It can also apply to amounts: a Brazilian team extracted information from cases of the Appellate State Court of Rio de Janeiro (ASCRJ) cases confirming or modifying compensation for moral damage [10]. Amounts were annotated with one out of three mutually exclusive labels, representing either an increase, a decrease or a confirmation of the damage amount granted by the first-instance judge. A more recent NER-based study focused on extracting catchphrases from Indian case law [11].

More research studies have focused on legal and regulatory texts rather than on case law and on the English language rather than on French. One study, applied to court cases, did attempt to generalize extraction to all possible judicial concepts, including facts and pretentions of the parties [12]. This is the first working of this kind on French-language case law. The French Court of Cassation launched a project to pseudonymize court cases [13]. They annotated a dataset with four labels, namely natural person, legal person, address and date of birth, and had these annotations trained by a neuronal network on the basis of a bidirectional long-short term memory (LSTM) algorithm [14]. CamemBERT [15] is the first general-purpose French linguistic model, and JuriBERT [16] is the first one dedicated to the French legal domain.

2. Methodology

2.1. Approach

Within the NLP research flow specializing on the legal domain, this work positions itself among the downstream tasks (Table 1). Like [10,13], it uses NER to address a real-life issue, but with a greater number of labels, of different data types and frequencies: where [10] dealt with labels of roughly balanced annotation numbers, we deal with some labels expected to annotate a few word sequences and other ones expected to annotate many such sequences. Another difference is that this study has involved a significant annotation effort: 1706 cases have been subject to 27,703 manual annotations.

Table 1. Studies on court case data extraction.

Year	Research	Model	Language	Pretrained Corpus	Task	Dataset Origin	Cases	Labels
2016	Aletras	SVM	English		classification	ECHR	584	2
2017	Katz	Random Forest	English		classification	SCOTUS	28,000	3
2018	Sulea	SVM	French		classification	Cassation	126,865	3
2019	Devlin	BERT	English	general				
2019	Medvedeva	SVM	English		classification	ECHR	1942	2
2019	Chalkidis	BERT	English	legal	classification	ECHR	11,500	66
2019	Fernandes	BiLSTM-CRF	Portuguese		NER	ASCRJ	3,022	6
2020	Filtz		German		NER/classification	ECHR	30	
2020	Martin	CamemBERT	French	general				
2020	Chalkidis	Legal-BERT	English	legal				
2020	Ngompe	CRF	French		NER	ECHR	503	11
2021	Mathis	Flair	French		NER	3-degrees mix	1700	26
2022	Douka	JuriBERT	French	legal	NER	Cassation	123,361	151

Proceedings data are common to different areas of law (or types of litigation). There is basically one set of proceedings data for all civil matters and another one for criminal matters. Criminal court cases are subject to more publicity restrictions than civil ones, are available to the public in smaller numbers and will come last in the French government's open data program schedule. That is why this research is limited to civil cases.

We built up a dataset that, though not pretending to be representative of all law areas, focuses on three different cases: road accidents, commercial leases and banking or finance-related litigation. The underlying assumption is that if the quality of extraction is about the same among these three areas of law, then it is likely other ones will provide similar quality.

The research does not intend to devise a new learning algorithm that would deliver a quality higher than that produced by existing ones. Though it does seek which algorithm, among several ones, applies in what context and delivers the best result for court data, its purpose is more practical: the identification of the algorithm, or the set of algorithms, available off the shelf that provides the highest quality for court data. We will not try to optimize further and will keep the default values of hyperparameters proposed by our tool.

Court case files are dirtier than most classical NER datasets [17]. Some legal editors have conducted preliminary data cleansing on cases originated by jurisdictions. They then used their machine-learning experience on court data by pseudonymizing and structuring them, before trying to carry out some more value-added work. Such tasks are off-putting and time-consuming while bringing few scientific lessons. This research therefore seeks to deliver results in one go, directly from raw texts, despite the rather poor quality of their rendition.

2.2. Dataset

For any natural-language processing, a dataset should have sufficient examples to ensure the best balance in the variety of data and language [18]. In our case, the learning of proceedings data should be based on a stock of cases, not only recent cases, and therefore, the learning dataset should be representative of the historically different writing conventions and templates. Our policy was to collect well-distributed cases from the past 15 to 20 years, depending on their level of availability, which varies by jurisdiction. Our dataset is composed of cases from the three jurisdictional degrees and focuses on three law areas (Table 2).

Table 2. Cases breakdown by degree and law area.

Cases	1st Instance	Appeal	Final Appeal	Total
Leases	250	208	63	521
Accidents	80	198	46	324
Bank	250	231	50	531
Other	234	0	96	330
All	814	637	255	1706

One first batch of cases, all on civil responsibility matters, comes from three judicial tribunals: 46 from Chambéry, 119 from Saint-Etienne and 149 from Grenoble, for a total of 314. They had been scanned by each tribunal as PDF images, then OCR'd to PDF text by Université Savoie Mont Blanc (USMB) and then converted to TXT files. As these cases have not been pseudonymized, they will not be published in the context of this research.

A second batch was supplied by Doctrine, a French legal technology. It includes 250 cases on banking law and another 250 on commercial leases. They have been pseudonymized with Doctrine's proprietary ML-based algorithm and randomly selected from their database upon our request. In many cases, parties were not only pseudonymized but actually erased, which damages the intelligibility of the text. USMB and Doctrine are warmly thanked for these corpora.

Three datasets of appellate cases have been downloaded from Légifrance, the portal of Direction de l'Information Légale et Administrative (DILA), the French national gazette editor. These cases have been pseudonymized by DILA and been subject to no subsequent data cleaning. They are published for having been "selected according to the methods specific to each order of jurisdiction", so they are not necessarily statistically representative of the procedural paths. The datasets are fairly well geographically distributed among appellate courts. The Court of Cassation, the supreme court and Légifrance each use their own template for the same case, and each one varies over time. We selected supreme court cases with the help of a page of the Court of Cassation's website that has been clustering cases since 2014, by area of law, such as road accidents, commercial leases and banking. We

downloaded the body of their text from Légifrance, in HTML format, and pasted them to as many TXT files.

Because of legal publicity restrictions and policy variations in the collection and centralization of court cases by supreme courts, the uncertain comprehensiveness of the available corpora makes up the first bias of analysis. French cases raise other issues. Though syntax is generally correct, punctuation hazards, missing blanks and words pseudonymized by error are likely to hinder entity recognition. Many cases are verbose, with redundant arguments or citations, sometimes scattered in different sections. Moreover, court cases are generally not structured. Of appellate cases, 45% do not have an explicit title for all sections [19]. The last chapter, "the outcome" (*le dispositif*), which describes the operative part, is the only section whose starting point is systematically showed, generally by the words "*par ces motifs*".

2.3. Labels

Selected labels describe information related to civil proceedings. About half of them are common to the three jurisdictional orders (see Table 3).

Table 3. Set of labels (alphabetic order). Checkmarks show labels that are applicable per jurisdictional degree and that have been subject to manual annotations.

Label	Label in French	Type	First-Instance Tribunal	Appellate Court	Supreme Court (Cour de Cassation)
Case-law citation	Référence jurisprudentielle	Named entity		✔	✔
Chamber	Formation	Named entity	✔	✔	✔
Costs of instance	Dépens	Amount	✔		
Date of appeal	Date d'appel	Judicial date	✔	✔	
Date of appellate case	Date d'arrêt	Judicial date	✔	✔	✔
Date of bankruptcy act	Date de redressement ou liquidation	Judicial date	✔	✔	✔
Date of birth	Date de naissance	Date	✔	✔	
Date of expertise	Date de rapport d'expertise	Judicial date	✔	✔	✔
Date of injunction	Date de référé	Judicial date	✔	✔	✔
Date of introductory act	Date d'assignation	Judicial date	✔	✔	✔
Date of judgment	Date de jugement	Judicial date	✔	✔	✔
Date of pre-trial	Date de mise en état	Judicial date	✔	✔	✔
Defendant	Défendeur	Named entity (party)	✔	✔	✔
ECLI	ECLI	Decision identifier			✔
Id of appellate case	RG appel	Decision identifier		✔	✔
Id of final appellate case	Numéro de pourvoi	Decision identifier			✔
Id of first-instance case	RG première instance	Decision identifier	✔	✔	

Table 3. *Cont.*

Label	Label in French	Type	First-Instance Tribunal	Appellate Court	Supreme Court (Cour de Cassation)
Jurisdiction	Juridiction	Named entity	✔	✔	✔
Legal citation	Référence juridique	Named entity	✔	✔	✔
Losing party	Condamne	Named entity (party)	✔		
Outcome	Sens de l'arrêt	Boolean		✔	✔
Plaintiff	Demandeur	Named entity (party)	✔	✔	✔
Reference for further explanation	Renvoi pour plus ample exposé	Boolean		✔	✔
Referring jurisdiction	Juridiction de renvoi	Named entity		✔	✔
Type of case	Type de décision	Named entity	✔	✔	✔
Unrecoverable costs	Frais irrépétibles	Amount	✔	✔	✔

Plaintiff, defendant and condemned ("losing") party are three labels that represent one party to a litigation. The plaintiff is the party who initiates a procedure, through a so-called *assignation* or *saisine*, in a first-instance judgment, or the party who, unsatisfied with the solution of the first-instance judge, lodged an appeal. The defendant is the party seeking confirmation of the first-instance judgment, in an appellate judgment, and we also use that label name in the context of a first-instance judgment to describe the party whom the introductory act is aimed at. One might wonder why we should bother to extract the parties to a trial given that names of natural persons are, or are supposed to be, "occulted" (Art. L111-13 of the code of judicial organization). One first answer is that in the absence, most of the time, of a unique identifier of the case preceding an appellate case, every indirect attribute of a case may help the matching. One name of a legal person (ex: "BNP Paribas"), one first name (ex: "Véronique") or one title ('Madame', 'Maître') may be enough to identify a party among many first-instance cases made a given day by a given chamber of a given jurisdiction the case at the origin of a given appellate case. One second answer is to preserve the very intelligibility of the case, as long as cases recognize the parties to a trial by their name, not by their role (essentially plaintiff versus defendant).

The label "date of birth" is proposed as a complementary identifier of a party who is a natural person. It may help disambiguate between different judgments made the same day in the same jurisdiction, when matched with appellate cases to build judicial pathways.

The other date labels define as many judicial milestones. One of these is the date of bankruptcy act. Parties that are legal persons may be subject to acts such as *redressement judiciaire* or *liquidation judiciaire*, which may happen during the litigation timespan. This helps to avoid any such date is being mistaken for a judicial milestone of the current case and helps to avoid getting many false positives, especially under the label "date of judgment". The other dates do refer to the current case. The label "date of introductory act" (*date d'assignation*) refers to the initial seizure of a tribunal by a party. The label "date of judgment" excludes the definition of "date of injunction" (*date de référé*), a type of ordinance that may be ordered by the same first-instance jurisdiction, which we preferred to annotate with a dedicated label. It is also distinct from the "date of appellate case", which is issued following an appeal. The "date of expertise" refers to the date when any expert designated by a judgment hands over their report to the judge. The "date of pre-trial" (*date de mise en état*) marks an update of the litigation case, such as when the case is "joined" with one other case or several other related cases or when it closes the instruction phase. We would expect that if more than one identifier of a first-instance judgment were to appear in the text of a judgment, a pre-trial date would also appear. A long procedure may involve an

assignment, an injunction, the handover of a report by a judicial expert, a first-instance judgment, an appeal, an appellate case, a pre-trial, another appellate case, a final appeal and a final appellate case. Figure 1 shows an example.

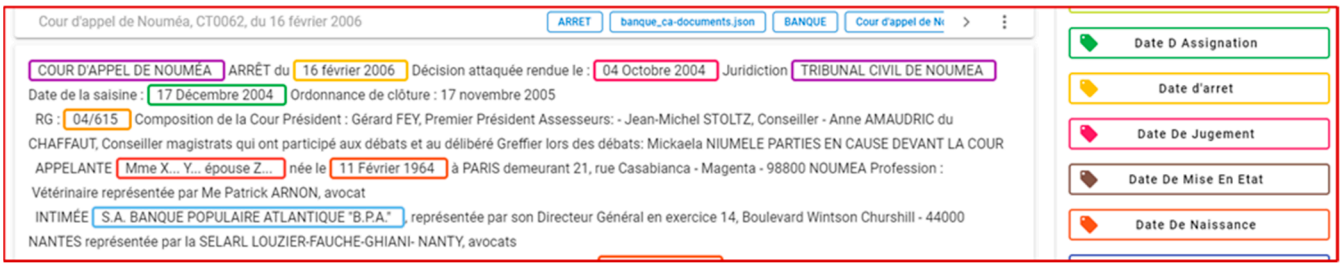

Figure 1. Example of annotations in a court case.

The "date of appeal" had been included before realizing that it had little relevance, insofar as French law states any appeal can be lodged only within the fortnight following the first-instance judgment.

The label "jurisdiction" is defined as the combination of either a tribunal (of first-instance) or a court (of appeal) with a city (in French, a *ressort*), when it has competence in a given territory. Court of Cassation and *Tribunal des Conflits*, which have national competence, are also defined as jurisdictions. We added the "chamber" label to cater to large jurisdictions where cases can be rendered by different courtrooms.

We decided to dedicate a label to the "referring jurisdiction" (*juridiction de renvoi*). The appellate court may hand over the case to a lower jurisdiction or to other magistrates of the same jurisdiction that has been subject to the appeal. As opposed to other annotations of jurisdictions, it may announce an upcoming case and therefore have no date and no chamber associated to it. We believe that this distinction will help reusers separate the past proceedings from the next one in their analysis (see Figure 1).

The label "legal citation" is used to annotate the articles of law invoked by the parties or by the judge, whereas "case-law citation" refers to some case not directly related to the litigation in progress. Many fewer annotations are expected, France having a civil-law tradition, not a common-law system, where previous court cases play a decisive role in every case.

We made the choice to use two separate labels for the Id of the appellate case and the Id of the first-instance case. An analysis of their proximity to the dates of appeal and trial judgment may be sufficient to assign a unified label to either type of proceeding. So that will be considered later.

Though ECLI, the European Case Law identifier, could be generated with even greater quality through some regular expression (regexp) algorithm, "ECLI" has been added as a label because it is a case identifier like three other labels (see Table 1).

For a tribunal, the outcome consists in setting a type of case or naming the losing party (or parties). For an appellate court, it is often a confirmation or a reversal, but the judge may decide on alternative or supplemental provisions. The outcome of a case is binary only at the level of the supreme court. Three labels are therefore needed to describe the operative part of the case. We introduced "type of case" and "losing party" specifically for tribunals, and "outcome", applicable to the other two degrees, is a Boolean that basically classifies the case as a confirmation or a reversal.

Our set of labels include two labels with an amount type: "costs of instance", annotated only for first-instance cases, and "non-recoverable costs". The latter represents the amount dedicated to refund legal fees, mostly paid to lawyers, different from the costs incurred by the proceedings and paid to the jurisdiction, which are fully recoverable by law. Such amounts can be found in cases of every jurisdictional order.

We also adopted "for further account", a Boolean that, if true, invites the reuser to collect and read previous related cases for a comprehensive understanding of the context.

In such a case, this means the case could be unfit for some reuses, such as predictive justice. Finally, we created the label "law area", some metadata available only for Court of Cassation cases, to help us conduct analyses on that axis.

Case categories (*nature d'affaire civile*, or NAC) have not been annotated. They are too seldom mentioned by the judge or difficult to recognize, so learning them would prove ineffective. Nor have the names of attorneys been extracted, but [20] in their own research already studied the extraction of appellant's and appellee's lawyers.

Table 4 shows how manual annotations break down per label.

Table 4. Numbers of annotations.

Label	First-Instance	Appeal	Final Appeal	All Corpora
Number of decisions	**814**	**637**	**255**	**1706**
Case-law citation	6	40	22	68
Chamber	528	589	360	1477
Costs of instance	133	0	0	133
Date of appeal	0	340	0	340
Date of appellate case	58	726	642	1426
Date of bankruptcy act	115	79	23	217
Date of birth	283	407	0	690
Date of expertise	166	136	11	313
Date of injunction	288	124	4	416
Date of introductory act	762	386	47	1195
Date of judgment	832	920	47	1799
Date of pre-trial	130	83	3	216
Defendant	1375	1197	382	2954
ECLI	0	0	177	177
Id of appellate case	0	578	7	585
Id of final appellate case	0	0	307	307
Id of first-instance case	862	242	0	1104
Jurisdiction	1039	1607	664	3310
Legal citation	2355	2071	958	5384
Losing party	833	0	0	833
Outcome	0	513	298	811
Plaintiff	1052	889	370	2311
Reference for further explanation	41	136	0	177
Referring jurisdiction	0	47	121	168
Type of case	114	0	0	114
Type of litigation	0	0	184	184
Unrecoverable costs	548	433	13	994

Table 4. *Cont.*

Label	First-Instance	Appeal	Final Appeal	All Corpora
Total number of annotations	**11520**	**11,543**	**4640**	**27,703**
Number of segments containing annotations	**4331**	**4853**	**1574**	**10,758**
Average number of annotations per decision	**14.15**	**18.12**	**18.20**	**16.24**

2.4. Tool

Kairntech is a French startup specializing in AI applied to text. The company is developing Sherpa [21], a natural-language processing (NLP) platform for software developers. It manages both classification and NER. It includes a graphical interface to annotate and a workbench to test different training algorithms. It caters to French, among several other languages. We use the *segment* (sentence) as the basic learning unit. Sherpa defines as a "dataset" the set of documents or segments that have been subject to at least one manual annotation. It is divided into a training dataset and a test dataset. The application of an experiment to a dataset produces a model, characterized by quality indicators: precision rate, recall rate and f-measure. The precision rate divides the number of correct annotations by that of all returned annotations. The recall rate is the number of correct annotations divided by the number of annotations that should have been returned. The F1 measure is the harmonic mean of precision and recall.

For NER, the workbench proposes to run CRF (conditional random fields), Spacy, Flair or DeLFT. CRF [22], is proposed in five training methods: L-BFGS, L2SGD, PA, AROW and AP. Flair [23], may be used with embeddings, which can be combined: Flair embeddings (contextualized string embeddings), bytecode and transformer embeddings.

DeLFT (deep learning framework for text) is a Keras and TensorFlow framework for text processing, focusing on NER and classification [24]. Flair and DeLFT are both based on recurrent neural networks (RNNs). The DeLFT RNN layer may be followed by a CRF layer. This allows the model to use past and future annotations to set the current annotation [25]. DeLFT can be combined with either ELMo embeddings [26], or BERT embeddings, for English texts, but CamemBERT, its French variant, is not available for DeLFT. We used DeLFT configured with its ELMo embeddings.

The basic experiment uses the CRF algorithm with a training dataset representing 80% of the randomly selected segments in the dataset. Training on the basis of the 80% allows annotations to be projected onto the remaining 20% of segments. The comparison of the manual annotations in this 20% with the projected annotations after training allows for quality indicators to be calculated. We leave this distribution constant in our experimentations.

The tool does not automate cross validation: it is up to the user to multiply runs and to average the results.

2.5. Conduct of Training

To save ourselves the rental costs of a GPU, we sampled one out of 10 cases from our global dataset representing all three jurisdictional degrees. The sampled dataset resulted in 171 cases, 1136 segments and 2857 manual annotations. Each experiment was run on a reshuffled dataset.

We used each algorithm with its default set of hyperparameters. CRF was applied to segments with a maximal size of 1500 characters. Flair was used with the SGD tokenizer. Its learning rate was 0.1, its anneal factor 0.5, batch size 32, patience 3, 100 epochs max, top four layers for embeddings, no storage of embeddings, hidden size 256, 1 LSTM layer, word dropout probability 0.5 and locked dropout probability 0.05. DeLFT used ELMo embeddings. The size of the character embeddings to compute was 25, dropout 0.5, recurrent dropout 0.25, max epochs 50, optimizer was adam, learning rate 0.001, clip

gradients 0.9, patience 5 and model type was bidirectional LSTM cum CRF. The maximum length of character sequence to compute embeddings was 30, the dimensionality of the character LSTM embeddings output space was 25, and the dimensionality of the word LSTM embeddings output space was 100. The batch size was 20, and the maximum number of checkpoints to keep was 5.

Among the different flavors of CRF, we chose CRF-pa, which seemed to deliver the best quality for a comparable model generation time. We looked at relative performances between labels.

Table 5 shows labels whose number of annotations (support) in the test dataset is below 100. The F1 scores are those of the first run and would prove volatile after other runs. A label such as "case-law citation", with only nine annotations to reproduce, understandably provides a fragile result, dependent on the outcome of the split between train and test data.

Table 5. Quality for rare annotations (CRF-pa on all decisions).

Label	Support	%	F1
Date of appeal	62	1.18%	62.5%
Date of birth	75	1.42%	94.7%
Date of expertise	70	1.33%	75.2%
Date of bankruptcy	42	0.80%	19.2%
Date of injunction	97	1.84%	53.9%
Date of pre-trial	40	0.76%	29.9%
Costs of instance	20	0.38%	56.3%
ECLI	42	0.80%	97.6%
Unrecoverable costs	73	1.38%	56.6%
Case-law citation	9	0.17%	40.0%
Outcome	74	1.40%	81.9%
Referring jurisdiction	33	0.63%	81.4%
Reference for further explanation	30	0.57%	78.6%
Other labels	4605	87.35%	
Total	5272	100.00%	

To fine-tune the analyses per jurisdictional degree, law area or label while ensuring statistical representativity, we needed to work on the consolidated dataset. This in turn required us to rent GPUs (one NVIDIA T4 Tensor Core GPUs, eight vCPUs, 32 GB RAM) to get acceptable training times.

3. Results

3.1. By Type of Algorithm

Our first analysis was to rank algorithm performances irrespective of the type of jurisdiction and the law area.

This cycle of runs shows that DeLFT is the best-performing model and CRF-pa (passive/aggressive) is the best-performing algorithm for the shortest training time. SpaCy (with 30 iterations) appears to be a good compromise between quality and training time. Table 6 sorts the performance outcomes of algorithms in different setups by ascending F1 score average over five runs.

Table 6. Quality, training time and run time by type of algorithm.

Algorithm	Average f1 on 5 Reshuffled Runs	Time to Train the Model on a CPU	Time to Train the Model on a GPU	Seconds to Annotate One Decision *
CRF—passive/aggressive	64.0%	1′35″	1′35″	0.32
Spacy—20 iterations	65.1%	7 mn	5 mn	0.24
Spacy—30 iterations	67.1%	9 mn	7 mn	0.24
Flair without embeddings (GRU)	69.4%	2 h	37 mn	0.31
Spacy—60 iterations	69.9%	19 mn	14 mn	0.25
Flair without embeddings (LSTM)	70.4%	3 h	1 h	0.36
Flair with Flair embeddings (LSTM)	75.3%	11 h	1 h	0.78
Flair with Flair, bytecode, transformer embeddings (LSTM)	75.3%	15 h	1 h	8.96
DELFT w. ELMo/bid. LSTM	76.0%	6 h	44 mn	16.04

(*): simulations conducted on a CPU server with decision RG 2016 000472 of Tribunal de Commerce de Thonon-les-Bains, 24 November 2016, from the Doctrine corpus. There may have been other computing activity on the Kairntech server upon running.

That step is not extended to transformers and to ensemble learning, which are not available in Kairntech at the time of writing.

Because CRF-pa was the configuration that performed best among the CRF Suite, we decided to keep it in our further research when comparing CRF with other families of algorithms.

3.2. By Degree of Jurisdiction

We wondered whether the nine labels that are common to the three jurisdictional levels (Table 2) should be trained together or be subject to degree-specific training. In this experimentation, on the whole dataset (1706 cases), we ran the same four classes of algorithms on each degree-specific dataset, with the same hyperparameters, and again on a cross-degree dataset. The algorithms were:

- CRF-pa;
- SpaCy with 30 iterations;
- Flair with its embeddings;
- DeLFT.

For practical reasons, the multijurisdictional dataset was capped at a sample of 4000 segments among the 10,750 of the aggregated segments. This represents a size comparable to the first-instance dataset (4331). Table 7 shows that the ranking between algorithms remains unchanged at every jurisdictional level. Flair outperforms DeLFT by a tiny margin. It is not clear-cut whether jurisdictional-level models outperform a single cross-jurisdictional-level one. We will therefore pursue investigations at both levels.

Table 7. Quality by jurisdictional degree.

Jurisdictional Degree [1]	Algorithm	F1-Score [2]
Tribunals	CRF-PA	78.3%
	Spacy	83.2%
	Flair with its embeddings	87.2%
	DeLFT with ELMo	86.9%
Appellate Courts	CRF-PA	76.3%
	Spacy	80.9%
	Flair with its embeddings	84.5%
	DeLFT with ELMo	83.1%
Court of Cassation	CRF-PA	80.2%
	Spacy	81.7%
	Flair with its embeddings	87.3%
	DeLFT with ELMo	85.7%
All jurisdictions (4000 segments)	CRF-PA	75.0%
	Spacy	80.4%
	Flair with its embeddings	85.7%
	DeLFT with ELMo	83.6%

(1) labels: date of introductory act, date of judgment, date of appellate decision, plaintiff, defendant, jurisdiction, chamber, legal reference, non-recoverable costs. (2) average on 5 runs after re-shuffle of training dataset.

3.3. By Law Area

Given that the purpose of this research is to assess machine-learning quality on the extraction of civil proceedings–related information, it is necessary to verify whether the law area could be a factor of that quality. Because supreme court cases refer to the legality of the procedure, not to the merits of the litigation, we excluded Court of Cassation cases from the scope of experiences conducted on a specific law area.

The F1 score rises by some 2% or 3% under any algorithm if learning is separately applied to each law area (Table 8). This contradicts our initial assumption. It seems the law area weighs more than the jurisdictional degree as a quality factor. Though it is true that in large jurisdictions with specialized chambers, judges may repeat writing patterns, in small ones, which make up the majority in our dataset, the same judge is likely to write cases in any of our three chosen law areas. So this alone cannot explain why a consolidated model produces lesser quality. Some next steps could be to extend the experimentation to other types of litigation.

Table 8. Quality by law area: lower-court cases.

F1-Score, All Labels, 1000 Segments, avg./ Runs	Leases	Accidents	Banking	Weighted avg. in 3 Law Areas	Consolidated 3 Law Areas
CRF-pa	65.5%	68.1%	62.7%	65.0%	62.8%
Spacy	68.2%	75.2%	65.8%	68.9%	66.7%
Flair with its embed.	77.9%	81.3%	74.8%	77.4%	75.2%
Delft	76.0%	78.7%	73.3%	75.6%	72.4%

Table 8. *Cont.*

F1-Score, All Labels, 1000 Segments, avg./ Runs	Leases	Accidents	Banking	Weighted avg. in 3 Law Areas	Consolidated 3 Law Areas
Nb documents	458	278	481	1217	1217
Nb segments in dataset	2995	1872	2638	7505	7505
Nb annotations	6187	5328	7060	18,575	18,575

3.4. By Type of Label

Any multilabel learning strategy must navigate between two pitfalls. Learning all labels together would produce unreliable figures for small-population ones. Learning them all separately would result in many word sequences annotated with conflicting labels. It is worthwhile to determine whether a model mix is possible, a mix where complementary models would each run on their best-performing algorithm.

We decided to divide the labels into different groups according to their numerical weight or their syntactic nature: identifiers, which are regular expressions (Group 1); dates (Group 2); parties, which are oft-repeated, named and partly pseudonymized but not truly taxonomic entities (Group 3); legal citations, which are oft-repeated, high-volume, compoundable regular expressions (Group 4); amounts (Group 5); and jurisdictions and chambers, which are taxonomic named entities (Group 6). The rest make up Group 7, the last group. Each group is therefore subject to 12 types of training.

Table 9 shows F1 scores of random runs for each group, in each type of algorithm and for each jurisdictional degree, on the full dataset. The best scores for each label are shown in red. Figures must be welcomed with prudence since these (intensive) calculations, for practical reasons, were launched only once.

Table 9. Quality by type of label (first run).

Group of Learning	Label	Tribunals				Appellate Courts				Court of Cassation			
		CRF-PA	Spacy	Flair	Delft	CRF-PA	Spacy	Flair	Delft	CRF-PA	Spacy	Flair	Delft
1	Appeal id	N/A				92.9%	92.6%	95.4%	94.2%	0.0%	93.6%	95.9%	94.9%
	First instance id	95.2%	93.6%	95.9%	94.9%	71.0%	81.9%	88.4%	90.6%	N/A			
	Final appeal id	N/A				N/A				98.3%	95.7%	96.7%	96.0%
	ECLI	N/A				N/A				96.1%	98.6%	98.6%	98.8%
2	Date of appeal	N/A				76.5%	74.8%	93.9%	80.0%	N/A			
	Date of appellate case	33.3%	50.0%	100.0%	92.3%	81.7%	91.0%	93.0%	91.6%	93.5%	91.8%	96.7%	94.3%
	Date of introductory act	79.0%	87.4%	75.4%	81.4%	69.1%	70.7%	83.4%	88.3%	0.0%	27.6%	15.4%	40.0%
	Date of judgment	75.1%	78.1%	80.0%	82.4%	78.5%	83.1%	82.7%	87.7%	28.6%	26.7%	72.7%	35.3%
	Date of pre-trial	34.3%	57.1%	57.1%	64.0%	34.8%	43.3%	90.9%	92.3%	N/A			
	Date of bankruptcy	12.1%	26.7%	83.3%	63.2%	0.0%	26.1%	36.4%	62.5%	0.0%	0.0%	83.3%	0,0%
	Date of injunction	67.9%	74.4%	88.9%	83.9%	38.1%	50.0%	60.0%	77.3%	0.0%	0.0%	0.0%	0.0%
	Date of expertise	69.4%	89.2%	57.1%	81.0%	68.2%	81.4%	87.5%	89.2%	66.7%	22.2%	66.7%	0.0%
	Date of birth	96.5%	96.5%	96.2%	97.0%	98.3%	95.7%	96.7%	93.9%	N/A			

Table 9. *Cont.*

Group of Learning	Label	Tribunals				Appellate Courts				Court of Cassation			
		CRF-PA	Spacy	Flair	Delft	CRF-PA	Spacy	Flair	Delft	CRF-PA	Spacy	Flair	Delft
3	Plaintiff (pseudon.)	63.2%	82.3%	77.4%	80.4%	54.8%	67.8%	72.1%	42.4%	53.7%	64.6%	88.9%	79.6%
	Defendant (pseudon.)	73.6%	77.5%	85.5%	83.7%	52.0%	66.1%	76.6%	60.7%	50.4%	69.4%	90.5%	83.0%
	Losing party (pseudon.)	49.6%	63.1%	75.8%	68.7%	N/A				N/A			
4	Legal citation	91.6%	90.4%	92.7%	91.5%	93.2%	88.4%	91.8%	88.5%	82.9%	78.6%	81.7%	81.0%
5	Non-recoverable costs	84.5%	86.7%	88.0%	92.0%	78.7%	83.6%	84.7%	35.3%	100.0%	80.0%	0.0%	50.0%
	Costs of instance	73.2%	69.0%	0.0%	69.2%	N/A				N/A			
6	Jurisdiction	94.5%	93.5%	91.3%	92.1%	95.0%	94.2%	91.3%	93.2%	97.6%	95.0%	92.1%	91.8%
	Chamber	94.9%	93.1%	92.1%	93.8%	87.9%	89.3%	98.2%	95.3%	95.3%	89.1%	86.4%	90.5%
7	Case-law citation	0.0%	0.0%	0.0%	0.0%	25.0%	54.5%	20.0%	36.4%	40.0%	50.0%	26.7%	40.0%
	Type of litigation	N/A				N/A				96.0%	100.0%	87.9%	89.4%
	For further account	93.3%	88.9%	100.0%	100.0%	79.2%	75.9%	76.7%	74.4%	N/A			
	Outcome	N/A				89.0%	88.1%	91.7%	93.5%	96.5%	96.1%	95.3%	96.8%
	Type of case	70.6%	58.5%	28.6%	84.2%	N/A				N/A			

Flair still ranks highest in a majority of labels, but not in all of them. It comes on top in neither degree for legal citations. DeLFT outperforms it for the date of birth and the outcome, and CRF for the type of case. DeLFT performs equally well globally for tribunals and appellate courts. If Flair maintains a slight advantage over DeLFT, it is at the cost of significantly longer training times. CRF is the best for legal citations in appellate and first-instance cases and still provides excellent results for other labels, as [Ngompe, 2019] already showed, especially with respect to its learning time.

Results are fairly good: five from those models give a quality higher than 95%. Only one gives a quality below 75%, and it involves parties (Group 3, appellate courts).

4. Conclusions and Next Steps

4.1. Conclusions

Our best overall performance was 87.5%, with Flair. It showed that building a database of proceedings data was feasible despite the variety of label types and the numerical weights of their annotations. In principle, results will be even better once the civil-justice open data program has been completed, allowing for a fully random sampling process to create the right-fitting, statistically representative training dataset.

It was not clear whether models built on the basis of a specific jurisdictional degree provided higher performance outcomes. A model dedicated to identifiers can extract them with a quality ranging from 90% to 98% (Table 10). Judicial dates can be extracted with a quality between 60% (date of injunction for appellate cases) and 96% (date of appellate case for Cassation). Jurisdictions and chambers obtained a quality near 95%. Scores of parties ranged from 72% to 90%, depending on the jurisdictional degree, and a preliminary rules-based propagation barely enhanced them.

Table 10. Model mix (first run).

Full Dataset	Tribunals		Courts of Appeal		Supreme Court	
Group	**Best Model**	**F1**	**Best Model**	**F1**	**Best Model**	**F1**
1	Flair	95.9%	Flair	93.4%	Flair	96.9%
2	Flair	82.1%	Flair	89.0%	Flair	89.6%
3	Flair	81.5%	Flair	74.4%	Flair	89.7%
4	Flair	92.7%	CRF-pa	93.2%	CRF-pa	82.9%
5	Delft	80.6%	Flair	84.7%	CRF-pa	100.0%
6	CRF-pa	94.7%	Flair	94.8%	CRF-pa	96.8%
7	Delft	92.1%	Delft	84.2%	Spacy	96.8%

We obtained poor results on the labels "type of case" and "losing party" used for first-degree cases. This may reflect confusing the "losing party" label with the labels "plaintiff" and "defendant"—like that between jurisdiction and referring jurisdiction, seen above. Additional research will have to be conducted to determine the outcome for first-instance cases, possibly using a taxonomy of case types. On the other hand, we obtained very good results from the outcome on appellate courts (94.7%) and on Cassation (96,8%), possibly because this outcome was often treated as metadata and embedded in the text header. Contrary to most labels, DeLFT seemed to be the best-performing algorithm for it.

Our best F1 score (87.5%) was achieved thanks to "brute force": some 27,000 annotations and one GPU. This is not optimal. Machine-learning practitioners seek the right balance between quality, annotation workload and computing power requirements.

The type of label and the numerical weight of annotations in each label appear to be key quality factors—and more clearly so than the jurisdictional degree and the law area: all date annotations should be trained together, as well as all amounts and annotations, to mitigate the risk of conflicting annotations. Another common property is the volume of annotations: jurisdictions and chambers, for instance, can be trained together because their respective numbers for annotations are roughly similar. Conversely, case-law references should be trained separately from legal citations, because their occurrences are comparatively much less frequent: a much larger dataset is required to reach comparable quality.

4.2. Next Steps

As already implied above, another step could be to complement the dataset with cases from other areas of civil law, to assess whether this is a factor of learning quality. A second one would be to reproduce the experimentation on a dataset of French criminal justice or administrative justice cases.

A third one could be to reapply to this dataset and this set of 26 labels the experiment of [27], who compared performances of a CamemBERT model refined in legal content (Judicial CamemBERT) with an ensemble method.

A fourth one would be to test a pretraining using the JuriBERT model.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Data are available on https://github.com/Bruno-Mathis/Court-decisions.

Conflicts of Interest: The author declares no conflict of interest.

References

1. Aletras, N.; Tsarapatsanis, D.; Preotiuc-Petro, D.; Lampos, V. Predicting judicial decisions of the European Court of Human Rights: A Natural Language Processing perspective. *Peer J. Comput. Sci.* **2016**, *2*, e93. [CrossRef]

2. Medvedeva, M.; Vols, M.; Wieling, M. Using machine learning to predict decisions of the European Court of Human Rights. *Artif. Intell. Law* **2020**, *28*, 237–266. [CrossRef]
3. Chalkidis, I.; Fergadiotis, M.; Malakasiotis, P.; Aletras, N.; Androutsopoulos, I. Neural Legal Judgment Prediction in English. *arXiv* **2019**, arXiv:1906.02059.
4. Filtz, E.; Navas-Loro, M.; Santos, C.; Polleres, A.; Kirrane, S. Events Matter: Extraction of Events from Court Decisions. In *Legal Knowledge and Information Systems*; IOS Press: Amsterdam, The Netherlands, 2020; ISBN 978-1-64368-151-1.
5. Katz, D.M.; Bommarito, M.J.; Blackman, J. A general approach for predicting the behavior of the Supreme Court of the United States. *PLoS ONE* **2017**, *12*, e0174698. [CrossRef] [PubMed]
6. Şulea, O.-M.; Zampieri, M.; Vela, M.; van Genabith, J. Predicting the Law Area and Decisions of French Supreme Court Cases. *arXiv* **2017**, arXiv:1708.01681.
7. Devlin, J.; Chang, M.-W.; Lee, K.; Toutanova, K. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv* **2018**, arXiv:1810.04805.
8. Chalkidis, I.; Fergadiotis, M.; Malakasiotis, P.; Aletras, N.; Androutsopoulos, I. LEGAL-BERT: The Muppets straight out of Law School. In *Findings of the Association for Computational Linguistics: EMNLP 2020*; Association for Computational Linguistics: Stroudsburg, PA, USA, 2020; pp. 2898–2904.
9. Cardellino, C.; Teruel, M.; Alonso Alemany, L.; Villata, S. A Low-cost, High coverage Legal Named Entity Recognizer, Classifier And Linker. In Proceedings of the 16th edition of the International Conference on Artificial Intelligence and Law, London, UK, 12–16 June 2017; pp. 9–18.
10. Fernandes, W.P.D.; Silva, L.J.S.; Frajhof, I.Z.; Konder, C.N.; Nasser, R.B.; de Carvalho, G.R.; Almeida, G.F.C.F.; Barbosa, S.D.J.; Lopes, H.C.V. Appellate Court Modifications Extraction for Portuguese. *Artif. Intell. Law* **2019**, *28*, 1–34. [CrossRef]
11. Mandal, A.; Ghosh, K.; Ghosh, S.; Mandal, S. A sequence labeling model for catchphrase identification from legal case documents. *Artif. Intell. Law* **2021**, *30*, 325–358. [CrossRef]
12. Ngompe, G.T.; Harispe, S.; Zambrano, G.; Montmain, J.; Mussard, S. Detecting sections and entities in court decisions using HMM and CRF graphical models. In *Advances in Knowledge Discovery and Management*; Springer: Berlin/Heidelberg, Germany, 2019; pp. 61–86.
13. Barrière, V.; Fouret, A. May I Check Again? A simple but efficient way to generate and use contextual dictionaries for Named Entity Recognition. Application to French Legal Texts. *arXiv* **2019**, arXiv:1909.03453.
14. Hochreiter, S.; Schmidhuber, J. Long Short-Term Memory. *Neural Comput.* **1997**, *9*, 1735–1780. [CrossRef] [PubMed]
15. Martin, L.; Muller, B.; Ortiz Suárez, P.J.; Dupont, Y.; Romary, L.; de La Clergerie, É.V.; Seddah, D.; Sagot, B. CamemBERT: A Tasty French Language Model. *arXiv* **2020**, arXiv:1911.03894v3.
16. Douka, S.; Abdine, H.; Vazirgiannis, M.; Hamdani, R.E.; Restrepo, D. Juribert: A masked-language model adaptation for French legal text. *arXiv* **2021**, arXiv:2110.01485.
17. Benesty, M. Why We Switched from Spacy to Flair to Anonymize French Case Law. Available online: towardsdatascience.com (accessed on 4 December 2022).
18. Xiao, R. Corpus Creation. In *Handbook of Natural Language Processing*, 2nd ed.; Indurkhya, N., Damerau, F.J., Eds.; Chapman and Hall: London, UK, 2010; pp. 146–165.
19. Miribel, A.; Chavallard, P. Structuring Legal Documents with Deep Learning. 2019. Available online: https://medium.com/doctrine/structuring-legal-documents-with-deep-learning-4ad9b03fb19 (accessed on 4 December 2022).
20. Boniol, P.; Panagopoulos, G.; Xypolopoulos, C.; Rajaa El Hamdani, R.; Restrepo Amariles, D.; Vazirgiannis, M. Performance in the Courtroom: Automated Processing and Visualization of Appeal Court Decisions in France. *arXiv* **2020**, arXiv:2006.06251.
21. Geissler, S. The Kairntech Sherpa—An ML Platform and API for the Enrichment of (not only) Scientific Content. In Proceedings of the 1st International Workshop on Language Technology Platforms, Language Resources and Evaluation Conference, Marseille, France, 11–16 May 2020; pp. 54–58.
22. Lafferty, J.; McCallum, A.; Pereira, F. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In Proceedings of the Eighteenth International Conference on Machine Learning, San Francisco, CA, USA, 28 June–1 July 2001; pp. 282–289.
23. Akbik, A.; Bergman, T.; Blythe, D.; Rasul, K.; Schweter, S.; Vollgraf, R. FLAIR: An Easy-to-Use Framework for State-of-the-Art NLP. In Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics (demonstrations), Minneapolis, MN, USA, 2–7 July 2019; pp. 54–59.
24. Lopez, P. Deep Learning Framework for Text. Available online: https://github.com/kermitt2/delft (accessed on 4 December 2022).
25. Huang, Z.; Xu, W.; Yu, K. Bidirectional LSTM-CRF models for sequence tagging. *arXiv* **2015**, arXiv:150801991.
26. Peters, M.E.; Neumann, M.; Iyyer, M.; Gardner, M.; Clark, C.; Lee, K.; Zettlemoyer, L. Deep contextualized word representations. *arXiv* **2018**, arXiv:1802.05365.
27. Mahmoudi, S.-A.; Condevaux, C.; Mathis, B.; Zambrano, G.; Mussard, S. NER sur décisions judiciaires françaises: CamemBERT Judiciaire ou méthode ensembliste? In Proceedings of the Extraction et Gestion des connaissances, Blois, France, 24–28 January 2022.