Swiss-Judgment-Prediction: A Multilingual Legal Judgment Prediction Benchmark

Joel Niklaus
Research Center for Digital Sustainability
Institute of Computer Science
University of Bern
Bern, Switzerland
`joel.niklaus@inf.unibe.ch`

Ilias Chalkidis
Coastal NLP Group
Department of Computer Science
University of Copenhagen
Copenhagen, Denmark
`ilias.chalkidis@di.ku.dk`

Matthias Stürmer
Research Center for Digital Sustainability
Institute of Computer Science
University of Bern
Bern, Switzerland
`matthias.stuermer@inf.unibe.ch`

Abstract

In many jurisdictions, the excessive workload of courts leads to high delays. Suitable predictive AI models can assist legal professionals in their work, and thus enhance and speed up the process. So far, Legal Judgment Prediction (LJP) datasets have been released in English, French, and Chinese. We publicly release a multilingual (German, French, and Italian), diachronic (2000-2020) corpus of 85K cases from the Federal Supreme Court of Switzerland (FSCS). We evaluate state-of-the-art BERT-based methods including two variants of BERT that overcome the BERT input (text) length limitation (up to 512 tokens). Hierarchical BERT has the best performance (approx. 68-70% Macro-F1-Score in German and French). Furthermore, we study how several factors (canton of origin, year of publication, text length, legal area) affect performance. We release both the benchmark dataset and our code to accelerate future research and ensure reproducibility.

1 Introduction

Frequently, legal information is available in textual form (e.g. court decisions, laws, legal articles or commentaries, contracts). With the abundance of legal texts comes the possibility of applying Natural Language Processing (NLP) techniques to tackle challenging tasks (Chalkidis and Kampas, 2018; Zhong et al., 2020; Chalkidis et al., 2021b). In this work, we study the task of Legal Judgment Prediction (LJP) where the goal is to predict the outcome (verdict) of a decision given its facts (Aletras et al., 2016; Şulea et al., 2017; Luo et al., 2017; Zhong et al., 2018; Hu et al., 2018; Chalkidis et al., 2019). Many relevant applications and tasks, such as court opinion generation (Ye et al., 2018) and analysis (Wang et al., 2012) have been also studied, while there is also work aiming to interpret (explain) the decisions of particular courts (Ye et al., 2018; Chalkidis et al., 2021a).

Models developed for LJP and relevant supportive tasks may assist both lawyers, e.g., help them prepare their arguments by identifying their strengths and weaknesses, and judges and clerks, e.g., review or prioritize cases, thus speeding up judicial processes and improving their quality. Especially in areas with many pending cases such as Indian[1] and Brazilian[2] jurisdictions or US immigration cases[3] the deployment of such models may drastically shorten the backlog. Such models can also help legal scholars to study case law (Katz, 2012) and help sociologists and research ethicists to expose irresponsible use of AI in the justice system (Angwin et al., 2016; Dressel and Farid, 2018). So far, LJP datasets have been released for English (Katz et al., 2017; Medvedeva et al., 2018; Chalkidis et al., 2019), French (Şulea et al., 2017) and Chinese (Xiao et al., 2018; Long et al., 2019).

---

[1] `https://tinyurl.com/mjy2uf9a`
[2] `https://tinyurl.com/2uttucmn`
[3] `https://tinyurl.com/4ybhhff8`

We introduce a new multilingual, diachronic LJP dataset of FSCS cases, which spans 21 years (from 2000 to 2020) containing over 85K (50K German, 31K French and 4K Italian) cases. To the best of our knowledge, it is the only publicly available multilingual LJP dataset to date. Additionally, it is annotated with publication years, legal areas and cantons of origin; thus it can be used also as testbed for fairness and robustness in the critical application of NLP to law (Wang et al., 2021).

Rogers (2021) argues that the NLP community is investing many more resources in the development of models rather than data. As a result, there are not enough challenging, high-quality and well curated benchmarks available. Rogers assumes that the main reason for this imbalance is that the "data work" is considered less prestigious and top conferences are more likely to reject resource (dataset) papers. With our work (and the associated code and data) we hope to make a valuable contribution to the legal NLP field, where there are not many ready-to-use benchmarks available.

Contributions

The contributions of this paper are threefold:

- We publicly release a large, high quality, curated, multilingual, diachronic dataset of 85K Swiss Federal Supreme Court (FSCS) cases annotated with the respective binarized judgment outcome (*approval/dismissal*), posing a challenging text classification task. We also provide additional metadata, i.e., the publication year, the legal area and the canton of origin per case, to promote robustness and fairness studies on the critical area of legal NLP (Wang et al., 2021).

- We provide experimental results with strong baselines representing the current state-of-the-art in NLP. Since the average length of the facts (850 tokens in the French part) is longer than the 512 tokens limit by BERT (Devlin et al., 2019), special methods are needed to cope with that. We show results comparing standard BERT models (up to 512 tokens) with two variants (hierarchical and prolonged BERT) that use up to 2048 tokens.

- We analyze the results of the German dataset in terms of diachronicity (publication year), legal area and input (text) length and the French dataset by canton of origin. We find that performance deteriorates as cases are getting more complex (longer facts), while also performance

varies across legal areas. There is no sign of performance fluctuation across years.

2 Related Work

European Court of Human Rights (ECtHR)

Aletras et al. (2016) introduced a dataset of 584 ECtHR cases concerning the violation or not of three articles of the European Convention of Human Rights (ECHR). They used a Support Vector Machine (SVM) (Cortes and Vapnik, 1995) with Bag-of-Words (BoW) (n-grams) and topical features on a simplified binarized LJP. In contrast to our work, they evaluated with random 10-fold cross-validation instead of the more realistic temporal split based on the date (Søgaard et al., 2021). Medvedeva et al. (2018) extended the ECtHR dataset to include 9 instead of 3 Articles resulting in a total of approx. 11.5K cases. They also experimented with an SVM operating on n-grams on the LJP task. Chalkidis et al. (2019) experimented on a similarly sized dataset using neural methods. On the binary LJP task, they improve the state-of-the-art using a hierarchical version of BERT. Additionally, they experimented with a multi-label LJP task predicting for each of the 66 ECHR Articles whether it is violated or not.

Supreme Court of the United States (SCOTUS)

Katz et al. (2017) experimented on LJP with 28K cases from the SCOTUS spanning almost two centuries. They trained a Random Forest (Breiman, 2001) classifier using extensive feature engineering with many non textual features. Kaufman et al. (2019) improved results using an ADABoost (Freund and Schapire, 1997) classifier, while also incorporating more textual information (i.e., statements made by the court judges during oral arguments).

French Supreme Court (Court of Cassation)

Şulea et al. (2017) studied the LJP task on a dataset of approx. 127K French Supreme Court cases. They experimented on a 6-class and a 8-class setting using an SVM with BoW features. They reported very high scores, which they claim are justified by the high predictability of the French Supreme Court. Although they used as input the entire case description and not only the facts, thus there is a strong possibility of label information leak. They also used 10-fold stratified cross-validation selecting the test part at random.

German Courts

Urchs et al. (2021) present a corpus of over 32K German court decisions from 131 Bavarian courts. The corpus is annotated with rich metadata including, among others, facts and judgment outcome needed for the LJP task. They present sample experiments predicting the type of the decision (judgment, resolution or other) and detecting conclusion, definition and subsumption in a subset of 200 randomly chosen and manually annotated decisions. They used traditional Machine Learning (ML) methods such as Logistic Regression (LR) on unigrams (BoW features) and SVM on Term Frequency - Inverse Document Frequency (TF-IDF) features.

Supreme People's Court of China (SPC)

Luo et al. (2017) experimented with the Hierarchical Attention Network (Yang et al., 2016) on Chinese criminal cases. They trained a model jointly on charge prediction, a form of LJP, and the relevant criminal law article extraction task using the relevant articles as support for the charge prediction. Xiao et al. (2018) introduced a large-scale LJP dataset of more than 2.6M Chinese criminal cases from the SPC. Their dataset is annotated with extensive metadata such as applicable law articles, charges, and prison terms. Zhong et al. (2018) viewed the dependencies between the different subtasks of LJP as a Directed Acyclic Graph (DAG) and apply a topological multitask learning framework. They work on three different datasets each containing Chinese criminal cases. Long et al. (2019) studied the LJP task on 100K Chinese divorce proceedings considering three types of information as input: applicable law articles, fact description, and plaintiffs' pleas. Li et al. (2019) use a multichannel attentive neural network on four datasets containing Chinese criminal cases. They considered all three subtasks of the Chinese LJP datasets: charges, law articles and prison term. Yang et al. (2019) apply a recurrent attention network on three Chinese LJP datasets.

3 Data Description

3.1 Dataset Construction

The decisions were downloaded from the platform entscheidsuche.ch and have been pre-processed by the means of HTML parsers and Regular Expressions (RegExps). The dataset contains more than 85K decisions from the FSCS written in three languages (50K German, 31K French, 4K Italian) from the years 2000 to 2020.[4] The FSCS is the last level of appeal in Switzerland and hears only the most controversial cases which could not have been sufficiently well solved by (up to two) lower courts. In their decisions, they often focus only on small parts of previous decision, where they discuss possible wrong reasoning by the lower court. This makes these cases particularly challenging.

In order to fight the reproducibility crisis (Britz, 2020), we release the Swiss-Judgment-Prediction dataset on Zenodo[5] and on Hugging Face[6], while also open-sourcing the complete code used for constructing the dataset[7] as well as for running the experiments[8] on GitHub.

3.2 Structure of Court Decisions

A typical Swiss court decision is made up of the following four main sections: *rubrum*, *facts*, *considerations* and *rulings*.[9] The *rubrum* (introduction) contains the date and chamber, mentions the involved judge(s) and parties and finally states the topic of the decision. The *facts* describe what happened in the case and form the basis for the considerations of the court. The higher the level of appeal, the more general and summarized the facts. The *considerations* reflect the formal legal reasoning which form the basis for the final ruling. Here the court cites laws and other influential rulings. The *rulings*, constituting the final section, are an enumeration of the binding decisions made by the court. This section is normally rather short and summarizes the considerations.

3.2.1 Use of Facts instead of Considerations

We deliberately did not consider the considerations as input to the model, unlike Aletras et al. (2016) for the following reasons.The facts are the section which is most similar to a general description of the case, which may be more widely available, while

---

[4]The dataset is not parallel, all cases are unique and decision are written only in a single language.

[5]https://zenodo.org/record/5529712

[6]https://huggingface.co/datasets/swiss_judgment_prediction

[7]https://github.com/JoelNiklaus/SwissCourtRulingCorpus

[8]https://github.com/JoelNiklaus/SwissJudgementPrediction

[9]See examples in Figures 5 and 6 of Appendix B

Split	de			fr			it		
	approval	dismissal	total	approval	dismissal	total	approval	dismissal	total
train	8369 (24%)	27003 (76%)	35452	5197 (25%)	15982 (75%)	21179	625 (20%)	2447 (80%)	3072
val	959 (20%)	3746 (80%)	4705	649 (21%)	2446 (79%)	3095	65 (16%)	343 (84%)	408
test	1915 (20%)	7810 (80%)	9725	1264 (19%)	5556 (81%)	6820	152 (19%)	660 (81%)	812
all	11243 (23%)	38639 (77%)	49882	7110 (23%)	23984 (77%)	31094	842 (20%)	3450 (80%)	4292

Table 1: The number of cases per label (*approval*, *dismissal*) in each language subset.

being less biased.[10] Additionally, the facts do not change that much from one to the next level of appeal (apart from being more concise and summarized in the higher levels of appeal). According to estimations from several court clerks we consulted, the facts take approximately 10% of the time for drafting a decision while the considerations take 85% and the outcome 5% (45%, 50% and 5% in penal law respectively). So, most of the work being done by the judges and clerks results in the legal considerations. Therefore, we would expect the model to perform better if it had access to the considerations. But on the other hand, the value of the model would be far smaller, since most of the work is already done, once the considerations are written. Thus, to create a more realistic and challenging scenario, we consider only the facts as input for the predictive models.

3.3 The Binarized LJP Task - Verdict Labeling Simplification

The cases have been originally labeled with 6 labels: *approval*, *partial approval*, *dismissal*, *partial dismissal*, *inadmissible* and *write off*. The first four are judged on the basis of the facts (merits) and the last two for formal reasons. A case is considered *inadmissible*, if there are formal deficiencies with the appeal or if the court is not responsible to rule the case. A court rules *write off* if the case has become redundant so there is no reason for the proceeding anymore. This can be for several reasons, such as an out-of-court settlement or procedural association (two proceedings are unified). *Approval* and *partial approval* mean that the request is deemed valid or partially valid respectively. *Dismissal* and *partial dismissal* mean that the request is denied or partially denied respectively. A *partial* decision is usually ruled in parallel with a decision of the opposite kind or with *inadmissible*.

In practice, court decisions may have multiple requests (questions), where each can be judged indi-

vidually. Since the structure of the outcomes in the decisions is non-standard, parsing them automatically is very challenging. Therefore, we decided to focus on the main request only and discard all side (secondary) requests. Even the main request sometimes contains multiple judgments referring to different parts of the main request, with some more important than others (it is very hard to automatically detect their criticality). So, to simplify the task and make it more concise, we transform the document labeling from a list of partial judgments into a single judgment, as follows:

1. We excluded all cases that have been ruled with both an approval and a dismissal in the main request, since that could be rather confusing.
2. We excluded cases ruled with *write off* outcomes since these cases are rejected for formal reasons that are not written (described) in the facts. Therefore, a model has no chance of inferring it correctly. We also excluded cases with *inadmissible* outcomes for similar reasons.
3. Since *partial* approvals/dismissals are very hard to distinguish from *full* approvals/dismissals respectively, we converted all the partial ones to full ones. Thus, the final labeling includes two possible outcomes, approvals and dismissals (i.e., the court "leans" positive or negative to the request).

By implementing these simplifications, we made the dataset more feasible (solvable) and semantically coherent targeting the core ruling process (see Section 5). Table 2 shows the numbers of decisions after each processing step. Note that we reduced the dataset with these preprocessing steps significantly (from over 141K to close to 85K decisions) to achieve higher quality. We also made the task structurally simpler by converting it from a multi-label to a binary classification task.[11]

The dataset is highly imbalanced containing more than $\frac{3}{4}$ dismissed cases (see Table 1 for de-

---

[10]Note however, that the facts are drafted together with the considerations and are often formulated in a way to support the reasoning in the considerations.

[11]Although, we look forward to recover at least part of the complexity in the future, if we have the appropriate resources to manually extract per-request judgments, introducing a new multi-task (multi-question) LJP dataset.
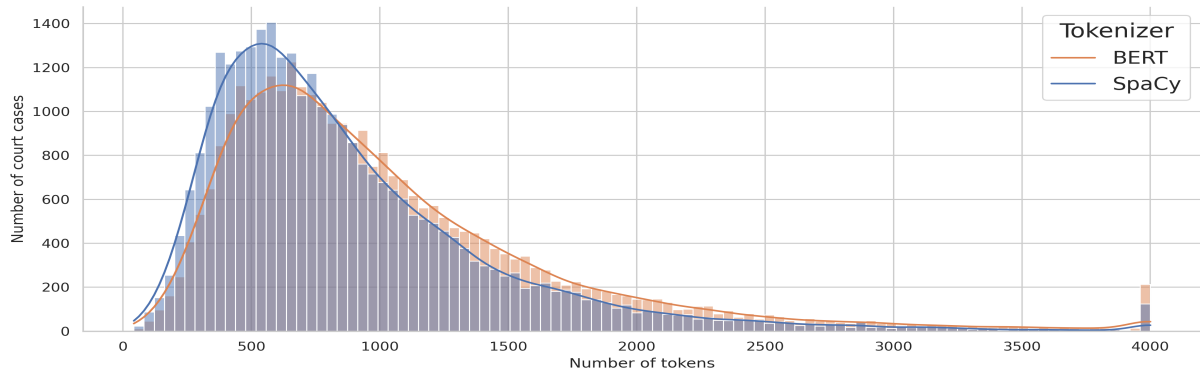
Figure 1: The distribution of the document (the facts of a case) length for French decisions. The blue histogram shows the document (case) length distribution in regular words (using the spacy tokenizer (Honnibal et al., 2020)). It is useful for a human estimation of the length and for methods building upon word embeddings (Mikolov et al., 2013; Pennington et al., 2014). The orange histogram shows the distribution in sub-word units (generated by the SentencePiece tokenizer (Kudo and Richardson, 2018) used in BERT). It is useful e.g. for estimating the maximum sequence length of a BERT-like model. Decisions with length over 4000 tokens have been grouped in the last bin.

tails). The label skewness makes the classification task quite hard and beating dummy baselines, e.g., predicting always the majority class, on micro-averaged measures (e.g., Micro-F1) is challenging. In our opinion, macro-averaged measures (e.g., Macro-F1) are more suitable in this setting, since they consider both outcomes (classes); they can also better discriminate better methods. In other words, they favor models that can actually learn the task (discriminate the two classes) and they do not always predict the majority class, i.e., *dismissal*, regardless of the facts.

Language	Total	2000-2020	Rulings	Judgments	Binarized
de	96337	95449	95273	84083	49882
fr	52278	51748	49132	49083	31094
it	8784	8643	8457	8441	4292
all	157399	155840	152862	141607	85268

Table 2: *Rulings* is the number of cases where rulings could be extracted. *Judgments* is the number of cases where we could extract any judgment types described in Section 3.3. *Binarized* is the number of cases considered in the final dataset after removing decisions containing labels other than *approval* or *dismissal*.

3.4 Case Distribution

This Section presents statistics about the distribution of cases according to different metadata like input (text) length, legal area and origin cantons.

3.4.1 The Curse of Long Documents

Figure 1 shows the distribution of the document (facts of the case) length of French cases.[12] We see that there are very few decisions with more

than 2K tokens in German (very similar for Italian). The French decisions are more evenly distributed, including a large portion of decisions with more than 4K tokens. For all languages, there is a considerable portion of decisions (50%+) containing more than 512 sub-word units (BERTs maximum sequence length) posing a fundamental challenge for standard BERT models.

3.4.2 Legal Areas

Table 3 presents the distribution of legal areas across languages. The legal areas are derived from the chambers where the decisions were heard. The website of the FSCS[13] describes in detail what kinds of cases the different chambers hear.

Legal Area	de	fr	it
public law	12182 (24%)	8514 (27%)	1583 (37%)
penal law	10942 (22%)	8039 (26%)	692 (16%)
social law	10742 (22%)	4048 (13%)	673 (16%)
civil law	8208 (16%)	7348 (24%)	763 (18%)
insurance law	7625 (15%)	2950 (9%)	561 (13%)
other	183 (0.4%)	195 (0.6%)	20 (0.5%)

Table 3: The distribution of legal areas in each language subset.

3.4.3 Origin Cantons

To study robustness and fairness in terms of geographical (regional) groups, we extracted the canton of origin from the decisions. As we observe in Table 4, most of the cantons (e.g., Zürich, Ticino) are monolingual and the distribution of the cases across cantons is very skewed with 1-2 cantons per language covering a large portion of the total cases.

---

[12]See Figures 7 and 8 in Appendix C for the German and Italian cases, respectively.

[13]https://tinyurl.com/52a4x8yz (in German)

Canton of Origin	de	fr	it
Zürich (ZH)	**12749 (25%)**	-	-
Berne (BE)	4705 (9%)	469 (2%)	-
Lucerne (LU)	3124 (6%)	-	-
Uri (UR)	248 (0.5%)	-	-
Schwyz (SZ)	1408 (3%)	-	-
Obwalden (OW)	190 (0.4%)	-	-
Nidwalden (NW)	364 (0.7%)	-	-
Glarus (GL)	363 (0.7%)	-	-
Zug (ZG)	1321 (3%)	-	-
Fribourg (FR)	487 (1%)	1826 (6%)	-
Soleure (SO)	2022 (4%)	-	-
Basel-City (BS)	1651 (3%)	-	-
Basel-Country (BL)	1578 (3%)	-	-
Schaffhausen (SH)	591 (1%)	-	-
Appenzell Outer-Rhodes (AR)	73 (0.2%)	-	-
Appenzell Inner-Rhodes (AI)	103 (0.2%)	-	-
St. Gall (SG)	3188 (6%)	-	-
Grisons (GR)	1300 (3%)	-	85 (2%)
Argovia (AG)	5494 (11%)	-	-
Thurgovia (TG)	2066 (4%)	-	-
Ticino (TI)	-	-	**3302 (77%)**
Vaud (VD)	-	**8926 (29%)**	-
Valais (VS)	502 (1%)	2095 (7%)	-
Neuchâtel (NE)	-	1732 (6%)	-
Genève (GE)	-	**9320 (30%)**	-
Jura (JU)	-	630 (2%)	-
Swiss Confederation (CH)	1854 (4%)	348 (1%)	83 (2%)
uncategorized	4488 (9%)	5742 (18%)	818 (19%)

Table 4: The distribution of cantons of origin in each language subset. No entry means that this language is not spoken in that canton. The cantons are ordered in the official order determined by the Swiss Confederation (mostly based on the date of entry into the confederation). High-resource cantons (> 20% of decisions per language) are marked in bold. Low-resource cantons (< 5% of decisions per language) are underlined.

4 Methods

4.1 Baselines

We first experiment with three baselines. The first one is a *majority* baseline that selects the majority (*dismissal*) class always across cases. The *stratified* baseline predicts labels randomly, respecting the training distribution. The last baseline is a *linear* classifier relying on TF-IDF features for the 35K most frequent n-grams in the training set.

4.2 BERT-based methods

BERT (Devlin et al., 2019) and its variants (Yang et al., 2020; Liu et al., 2019; Lan et al., 2020), inter alia, dominate NLP as state-of-the-art in many tasks (Wang et al., 2018, 2019). Hence, we examine an arsenal of BERT-based methods.

Standard BERT We experimented with monolingual BERT models for German (Chan et al., 2019), French (Martin et al., 2020) and Italian (Parisi et al., 2020) and also the multilingual BERT of (Devlin et al., 2019). Since the facts are often longer than 512 tokens (see Section 3 for details), there is a need to adapt the models to long textual input.

Long BERT is an extension of the standard BERT models, where we extend the maximum sequence length by introducing additional positional embeddings. In our case, the additional positional encodings have been initialized by replicating the original pre-trained 512 ones 4 times (2048 in total). While Long BERT can process the full text in the majority of the cases, its extension leads to longer processing time and higher memory requirements.

Hierarchical BERT, similar to the one presented in Chalkidis et al. (2019), uses a shared standard BERT encoder processing segments up to 512 tokens to encode each segment independently. To aggregate all (in our case 4) segment encodings, we pass them through an additional Bidirectional Long Short-Term Memory (BiLSTM) encoder and concatenate the final LSTM output states to form a single document representation for classification.

5 Experiments

In this Section, we describe the conducted experiments alongside the presentation of the results and an analysis of the results of the German dataset in terms of diachronicity (judgment year), legal area, input (text) length and canton of origin.

5.1 Experimental SetUp

During training, we over-sample the cases representing the minority class (*approval*).[14] Across BERT-based methods, we use Early Stopping on development data, an initial learning rate of 3e-5 and batch size 64 across experiments. The standard BERT models have been trained and evaluated with maximum sequence length 512 and the two variants of BERT with maximum sequence length 2048. The 2048 input length has been chosen based on a balance between memory and compute restrictions and the statistics of the length of facts (see Section 3.4.1), where we see that the vast majority of cases contains less than 2K tokens. Additionally, this gives us the possibility to investigate differences by input (text) length (see Section 5.3.2). We report both micro- and macro-averaged F1-score on the test set. Micro-F1 is averaged across samples whereas Macro-F1 is averaged across samples inside each class and then across the classes. Therefore, a test example in

---

[14]In preliminary experiments, we find that this sampling methodology outperforms both the standard Empirical Risk Minimization (ERM) and the class-wise weighting of the loss penalty, i.e., considering each class loss 50-50.

Model	de		fr		it	
	Micro-F1↑	Macro-F1↑	Micro-F1↑	Macro-F1↑	Micro-F1↑	Macro-F1↑
baselines						
Majority	**80.3**	44.5	**81.5**	44.9	**81.3**	44.8
Stratified	66.7 ± 0.3	50.0 ± 0.4	66.3 ± 0.2	50.0 ± 0.4	69.9 ± 1.8	48.8 ± 2.4
Linear (BoW)	65.4 ± 0.2	52.6 ± 0.1	71.2 ± 0.1	56.6 ± 0.2	67.4 ± 0.5	53.9 ± 0.6
standard (up to 512 tokens)						
Native BERT	74.0 ± 4.0	63.7 ± 1.7	74.7 ± 1.8	58.6 ± 0.9	76.1 ± 3.7	55.2 ± 3.7
Multilingual BERT	68.4 ± 5.1	58.2 ± 4.8	71.3 ± 4.3	55.0 ± 0.8	77.6 ± 2.4	53.0 ± 1.1
long (up to 2048 tokens)						
Native BERT	76.5 ± 3.7	67.9 ± 1.8	77.2 ± 3.4	68.0 ± 1.8	77.1 ± 3.9	**59.8 ± 4.6**
Multilingual BERT	75.9 ± 1.6	66.5 ± 0.8	73.3 ± 1.9	64.3 ± 1.5	76.0 ± 2.6	58.4 ± 3.5
hierarchical (two-tier 4× 512 tokens)						
Native BERT	77.1 ± 3.7	**68.5 ± 1.6**	80.2 ± 2.0	**70.2 ± 1.1**	75.8 ± 3.5	57.1 ± 6.1
Multilingual BERT	76.8 ± 3.2	57.1 ± 0.8	76.3 ± 4.1	67.2 ± 2.9	72.4 ± 16.6	55.5 ± 9.5

Table 5: All the models have been trained and evaluated in the same language. With *Native BERT* we mean the BERT model pre-trained in the respective language. The best scores for each language are in bold. Given the high class imbalance, BERT-based methods under-perform in Micro-F1 compared to the *Majority* baseline, while being substantially better in Macro-F1.

a minority class has a higher weight in Macro-F1 than an example from the majority class. In classification problems with imbalanced class distributions (such as the one we examine), Macro-F1 is more realistic than Micro-F1 given that we are equally interested in both classes. Each experiment has been run with 5 different random seeds. We report the average score and standard deviation across experiments. The experiments have been performed on a single GeForce RTX 3090 GPU with mixed precision and gradient accumulation. We used the Hugging Face Transformers library (Wolf et al., 2020) and the BERT models available from `https://huggingface.co/models`.

5.2 Main Results

Table 5 shows the results across methods for all language subsets. We observe that the native BERT models outperform their multi-lingual counterpart; while not being domain-specific, these models can still better model the case facts. Given the high class imbalance, all BERT-based methods underperform in Micro-F1, being biased towards *dismissal* performance compared to the naive Majority baseline, while doing substantially better in Macro-F1. Hierarchical and Long BERT-based methods consistently out-perform the linear classifiers across languages (+10% in Macro-F1), while standard BERT is comparable or better than linear models, although it considers only up to 512 tokens. While performance of BERT-based methods is quite comparable between the German and French subsets with 35K and 21K training samples respectively, it is far worse in the Italian subset, where there are only 3K training samples. In two out of three languages (German and French with 20K+ training samples) hierarchical BERT has borderline better performance compared to long BERT (+1.6-2.2% in Macro-F1), but in both cases the difference is very close to the error margin (standard deviation). We would like to remark that the results of Hierarchical BERT could possibly be improved considering a finer segmentation of the text into sentences or paragraphs.[15] We leave the investigation for alternative segmentation schemes for future work.

5.3 Discussion - Bivariate Analysis

In this section, we analyze the results in relation to specific attributes (publication year, input (text) length, legal area and canton of origin) in order to evaluate the model robustness and identify how specific aspects affect the model performance.

---

[15]Currently, we segment the text into chunks of 512 tokens to avoid excessive padding that will further increase the needed number of segments and will lead to even higher time and memory demands.

Legal Area			standard		long		hierarchical	
Legal Area	# cases	approval rate	Micro-F1↑	Macro-F1↑	Micro-F1↑	Macro-F1↑	Micro-F1↑	Macro-F1↑
public law	2587	20.6%	66.6 ± 6.2	53.1 ± 1.8	64.6 ± 6.7	53.8 ± 2.1	64.8 ± 8.1	53.7 ± 3.0
penal law	2900	21.0%	83.6 ± 1.8	74.8 ± 1.5	87.6 ± 1.6	81.1 ± 2.3	88.4 ± 1.0	82.6 ± 2.5
social law	661	19.3%	71.1 ± 4.3	65.2 ± 2.6	74.8 ± 4.0	69.1 ± 2.8	75.4 ± 3.9	69.4 ± 2.5
civil law	1574	16.5%	73.6 ± 4.8	55.5 ± 1.0	79.0 ± 3.4	65.1 ± 2.4	78.9 ± 3.8	65.9 ± 2.8

Table 6: We used the German native BERT model pre-trained and evaluated on the German data. In the German test set there are no insurance law cases and only 3 cases with other legal areas. The area where models perform best is in bold and the area where they perform worst is underlined.
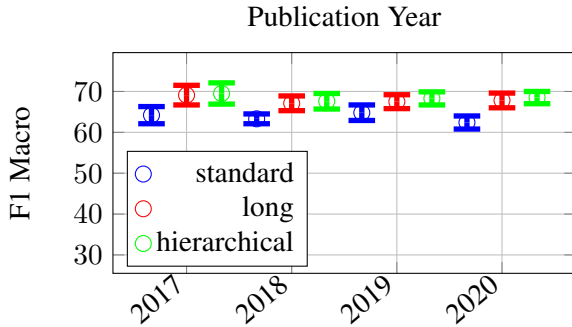
Figure 2: This table compares the different BERT types on cases from different years. We used the native German BERT model.
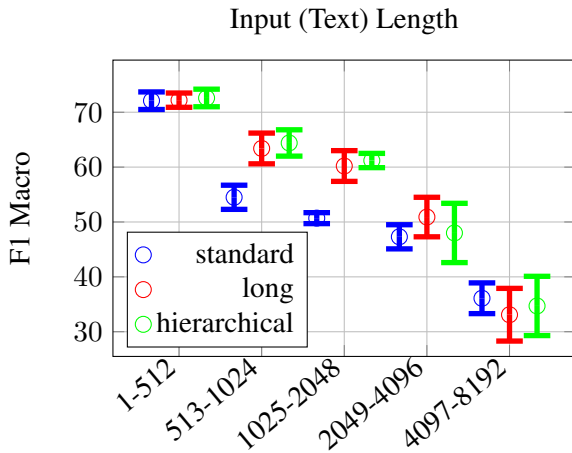

Figure 3: This table compares the different long BERT types on different input (text) lengths. We used the native German BERT model.

5.3.1 Diachronicity

In Figure 2, we present the results grouped by years in the test set (2017-2020). We cannot identify a notable fluctuation in performance across years as there is a very small decrease in performance (approx. -2% in Macro-F1); most probably because the testing time-frame is really short (4 years). Comparing the performance between the validation (2015-2016) and the test (2017-2020) set (approx. 70% vs. 68.5%), again we do not observe an exceptional fluctuation time-wise.

5.3.2 Input (Text) Length

In Figure 3, we observe that model performance deteriorates as input (text) length increases, i.e., there is an absolute negative correlation between performance and input (text) length. The two variants of BERT improve results, especially in cases with 512 to 2048 tokens. Since the two variants of BERT have a maximum length of 2048 they perform similar to the standard BERT type in cases longer than 2048 tokens.

5.3.3 Legal Area

In Table 6, we observe that the models do not equally perform across legal areas. All models seem to be much more accurate in penal law cases, while the performance is much worse (approx. 30%) in public law cases. According to the experts, the jurisprudence in penal law is more united and aligned in Switzerland and outlier judgments are rarer making the task more predictable. Additionally, in the case of not enough evidence the principle of *"in dubio pro reo"* (reasonable doubt) is applied. [16] Another possible reason for the higher performance in penal law could be the increased work performed by the legal clerks in drafting the facts of the case (see Section 3.2.1), thus including more useful information relevant to the task.

5.3.4 Canton of Origin

In Figure 4, we observe a performance disparity across cantons, although this is neither correlated with the number of cases per canton, nor with the dismissal/approval rate per canton. Thus, the disparity is either purely coincidental and has to do with the difficulty of particular cases in some cantons or there are other factors (e.g., societal, economics) worth considering in future work.

---

[16]The principle of *"in dubio pro reo"*, i.e., "When in doubt, in favor of the defendant.", is only applicable in penal law cases.
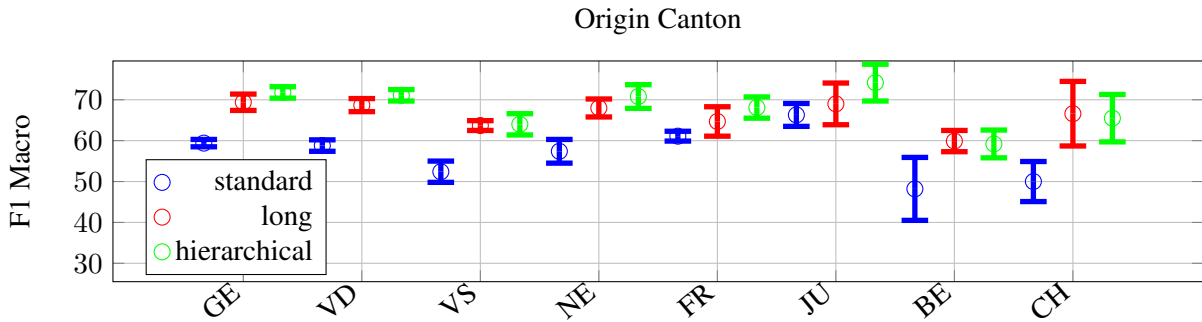
Figure 4: This table compares the different long BERT types on different origin cantons. We used the native French BERT model. The cantons are sorted by the number of cases in the training set descending.

6 Conclusions & Future Work

We introduced a new multilingual, diachronic dataset of 85K Swiss Federal Supreme Court (FSCS) cases, including cases in German, French, and Italian. We presented results considering three alternative BERT-based methods, including methods that can process up to 2048 tokens and thus can read the entirety of the facts in most cases. We found that these methods outperform the standard BERT models and have the best results in Macro-F1, while the naive majority classifier has the best overall results in Micro-F1 due to the high class imbalance of the dataset (more than $\frac{3}{4}$ of the cases are dismissed). Further on, we presented a bivariate analysis between performance and multiple factors (diachronicity, input (text) length, legal area, and canton of origin). The analysis showed that performance deteriorates as input (text) length increases, while the results in cases from different legal areas or cantons vary raising questions on models' robustness under different attributes.

In future work, we would like to investigate the application of cross-lingual transfer learning techniques, for example the use of Adapters (Houlsby et al., 2019; Pfeiffer et al., 2020). In this case, we could possibly improve the poor performance in the Italian subset, where approx. 3K cases exists, by training a multilingual model across all languages, thus exploiting all available resources, ignoring the traditional language barrier. In the same direction, we could also exploit and transfer knowledge from other annotated datasets that aim at the LJP task (e.g., ECtHR and SCOTUS).

More in depth analysis on robustness is also an interesting future avenue. In this direction, we would like to explore distributional robust optimization (DRO) techniques (Koh et al., 2021; Wang et al., 2021) that aim to mitigate disparities across groups of interest, i.e., labels, cantons and/or legal areas could be both considered in this framework.

Another interesting direction is a deeper analysis with models handling long textual input (Beltagy et al., 2020; Zaheer et al., 2020) using alternative attention schemes (window-based, dilated, etc.). Furthermore, none of the examined pre-trained models is legal-oriented, thus pre-training and evaluating such specialized models is also needed, similarly to the English Legal-BERT of Chalkidis et al. (2020).

Ethics Statement

The scope of this work is not to produce a robot lawyer, but rather to study LJP in order to broaden the discussion and help practitioners to build assisting technology for legal professionals. We believe that this is an important application field, where research should be conducted (Tsarapatsanis and Aletras, 2021) to improve legal services and democratize law, while also highlight (inform the audience on) the various multi-aspect shortcomings seeking a responsible and ethical (fair) deployment of technology. In this direction, we provide a well-documented public resource for three languages (German, French, and Italian) that are underrepresented in legal NLP literature. We also provide annotations for several attributes (year of publication, legal area, canton/region) and provide a bivariate analysis discussing the shortcomings to further promote new studies in terms of fairness and robustness (Wang et al., 2021), a critical part of NLP application in law. All decisions (original material) are publicly available on the entscheid-suche.ch platform and the names of the parties have been redacted (See Figures 5 and 6) by the court according to its official guidelines[17].

---

[17]https://tinyurl.com/mtu23szy (In German)

Acknowledgements

References

Nikolaos Aletras, Dimitrios Tsarapatsanis, Daniel Preoţiuc-Pietro, and Vasileios Lampos. 2016. Predicting judicial decisions of the European Court of Human Rights: a Natural Language Processing perspective. *PeerJ Computer Science*, 2:e93. Publisher: PeerJ Inc.

Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner. 2016. Machine bias: There's software used across the country to predict future criminals. and it's biased against blacks. *ProPublica*.

Iz Beltagy, Matthew E. Peters, and Arman Cohan. 2020. Longformer: The Long-Document Transformer. *arXiv:2004.05150 [cs]*. ArXiv: 2004.05150.

Leo Breiman. 2001. Random forests. *Machine Learning*, 45(1):5–32.

Denny Britz. 2020. AI Research, Replicability and Incentives.

Ilias Chalkidis, Ion Androutsopoulos, and Nikolaos Aletras. 2019. Neural Legal Judgment Prediction in English. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4317–4323, Florence, Italy. Association for Computational Linguistics.

Ilias Chalkidis, Manos Fergadiotis, Prodromos Malakasiotis, Nikolaos Aletras, and Ion Androutsopoulos. 2020. LEGAL-BERT: The muppets straight out of law school. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2898–2904, Online. Association for Computational Linguistics.

Ilias Chalkidis, Manos Fergadiotis, Dimitrios Tsarapatsanis, Nikolaos Aletras, Ion Androutsopoulos, and Prodromos Malakasiotis. 2021a. Paragraph-level rationale extraction through regularization: A case study on european court of human rights cases. In *Proceedings of the Annual Conference of the North American Chapter of the Association for Computational Linguistics*, online.

Ilias Chalkidis, Abhik Jana, Dirk Hartung, Michael Bommarito, Ion Androutsopoulos, Daniel Martin Katz, and Nikolaos Aletras. 2021b. LexGLUE: A Benchmark Dataset for Legal Language Understanding in English.

Ilias Chalkidis and Dimitrios Kampas. 2018. Deep learning in law: early adaptation and legal word embeddings trained on large corpora. *Artificial Intelligence and Law*, 27:171–198.

Branden Chan, Timo Möller, Malte Pietsch, Tanay Soni, and Chin Man Yeung. 2019. deepset - Open Sourcing German BERT.

C. Cortes and V. Vapnik. 1995. Support vector networks. *Machine Learning*, 20:273–297.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *arXiv:1810.04805 [cs]*. ArXiv: 1810.04805.

Julia Dressel and Hany Farid. 2018. The accuracy, fairness, and limits of predicting recidivism. *Science Advances*, 4(10).

Yoav Freund and Robert E Schapire. 1997. A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of Computer and System Sciences*, 55(1):119–139.

Matthew Honnibal, Ines Montani, Sofie Van Landeghem, and Adriane Boyd. 2020. spaCy: Industrial-strength Natural Language Processing in Python.

Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin de Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. 2019. Parameter-efficient transfer learning for nlp. In *Proceedings of the 36th International Conference on Machine Learning (ICML)*, Long Beach, CA, USA.

Zikun Hu, Xiang Li, Cunchao Tu, Zhiyuan Liu, and Maosong Sun. 2018. Few-Shot Charge Prediction with Discriminative Legal Attributes. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 487–498, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Daniel Martin Katz. 2012. Quantitative legal prediction-or-how I learned to stop worrying and start preparing for the data-driven future of the legal services industry. *Emory Law Journal*, 62:909.

Daniel Martin Katz, Michael J. Bommarito Ii, and Josh Blackman. 2017. A general approach for predicting the behavior of the Supreme Court of the United States. *PLOS ONE*, 12(4):e0174698. Publisher: Public Library of Science.

Aaron Russell Kaufman, Peter Kraft, and Maya Sen. 2019. Improving supreme court forecasting using boosted decision trees. *Political Analysis*, 27(3):381–387.

Pang Wei Koh, Shiori Sagawa, Henrik Marklund, Sang Michael Xie, Marvin Zhang, Akshay Balsubramani, Weihua Hu, Michihiro Yasunaga, Richard Lanas Phillips, Irena Gao, Tony Lee, Etienne David, Ian Stavness, Wei Guo, Berton A. Earnshaw, Imran S. Haque, Sara Beery, Jure Leskovec, Anshul Kundaje, Emma Pierson, Sergey Levine, Chelsea Finn, and Percy Liang. 2021. WILDS: A benchmark of in-the-wild distribution shifts. In *International Conference on Machine Learning (ICML)*.

Taku Kudo and John Richardson. 2018. SentencePiece: A simple and language independent subword tokenizer and detokenizer for Neural Text Processing. *arXiv:1808.06226 [cs]*. ArXiv: 1808.06226.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2020. ALBERT: A Lite BERT for Self-supervised Learning of Language Representations. *arXiv:1909.11942 [cs]*. ArXiv: 1909.11942.

Shang Li, Hongli Zhang, Lin Ye, Xiaoding Guo, and Binxing Fang. 2019. MANN: A Multichannel Attentive Neural Network for Legal Judgment Prediction. *IEEE Access*, 7:151144–151155. Conference Name: IEEE Access.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach. *arXiv:1907.11692 [cs]*. ArXiv: 1907.11692.

Shangbang Long, Cunchao Tu, Zhiyuan Liu, and Maosong Sun. 2019. Automatic Judgment Prediction via Legal Reading Comprehension. In *Chinese Computational Linguistics*, Lecture Notes in Computer Science, pages 558–572, Cham. Springer International Publishing.

Bingfeng Luo, Yansong Feng, Jianbo Xu, Xiang Zhang, and Dongyan Zhao. 2017. Learning to Predict Charges for Criminal Cases with Legal Basis. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2727–2736, Copenhagen, Denmark. Association for Computational Linguistics.

Louis Martin, Benjamin Muller, Pedro Javier Ortiz Suárez, Yoann Dupont, Laurent Romary, Éric Villemonte De La Clergerie, Djamé Seddah, and Benoît Sagot. 2020. CamemBERT: a Tasty French Language Model. pages 7203–7219.

Masha Medvedeva, Michel Vols, and Martijn Wieling. 2018. Judicial decisions of the European Court of Human Rights: Looking into the crystal ball. In *Proceedings of the Conference on Empirical Legal Studies*, page 24.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient Estimation of Word Representations in Vector Space. *arXiv:1301.3781 [cs]*. ArXiv: 1301.3781.

Loreto Parisi, Simone Francia, and Paolo Magnani. 2020. UmBERTo: an Italian Language Model trained with Whole Word Masking. Original-date: 2020-01-10T09:55:31Z.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. Glove: Global Vectors for Word Representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar. Association for Computational Linguistics.

Jonas Pfeiffer, Ivan Vulić, Iryna Gurevych, and Sebastian Ruder. 2020. MAD-X: An Adapter-Based Framework for Multi-Task Cross-Lingual Transfer. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7654–7673, Online.

Anna Rogers. 2021. Changing the World by Changing the Data. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 2182–2194, Online. Association for Computational Linguistics.

Anders Søgaard, Sebastian Ebert, Jasmijn Bastings, and Katja Filippova. 2021. We need to talk about random splits. In *Proceedings of the 2021 Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, Online.

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

Stefanie Urchs, Jelena Mitrović, and Michael Granitzer. 2021. Design and Implementation of German Legal Decision Corpora:. In *Proceedings of the 13th International Conference on Agents and Artificial Intelligence*, pages 515–521, Online Streaming, — Select a Country —. SCITEPRESS - Science and Technology Publications.

Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R Bowman. 2019. SuperGLUE: A Stickier Benchmark for General-Purpose Language Understanding Systems. page 30.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2018. GLUE: A Multi-Task Benchmark and Analysis Platform for Natural Language Understanding. In

Proceedings of the 2018 EMNLP Workshop Black-boxNLP: Analyzing and Interpreting Neural Networks for NLP, pages 353–355, Brussels, Belgium. Association for Computational Linguistics.

William Yang Wang, Elijah Mayfield, Suresh Naidu, and Jeremiah Dittmar. 2012. Historical Analysis of Legal Opinions with a Sparse Mixed-Effects Latent Variable Model. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 740–749, Jeju Island, Korea. Association for Computational Linguistics.

Yuzhong Wang, Chaojun Xiao, Shirong Ma, Haoxi Zhong, Cunchao Tu, Tianyang Zhang, Zhiyuan Liu, and Maosong Sun. 2021. Equality before the law: Legal judgment consistency analysis for fairness. *Science China - Information Sciences*, abs/2103.13868.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-Art Natural Language Processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Chaojun Xiao, Haoxi Zhong, Zhipeng Guo, Cunchao Tu, Zhiyuan Liu, Maosong Sun, Yansong Feng, Xianpei Han, Zhen Hu, Heng Wang, and Jianfeng Xu. 2018. CAIL2018: A Large-Scale Legal Dataset for Judgment Prediction. *arXiv:1807.02478 [cs]*. ArXiv: 1807.02478.

Ze Yang, Pengfei Wang, Lei Zhang, Linjun Shou, and Wenwen Xu. 2019. A Recurrent Attention Network for Judgment Prediction. In *Artificial Neural Networks and Machine Learning – ICANN 2019: Text and Time Series*, Lecture Notes in Computer Science, pages 253–266, Cham. Springer International Publishing.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Ruslan Salakhutdinov, and Quoc V. Le. 2020. XLNet: Generalized Autoregressive Pretraining for Language Understanding. *arXiv:1906.08237 [cs]*. ArXiv: 1906.08237.

Zichao Yang, Diyi Yang, Chris Dyer, Xiaodong He, Alex Smola, and Eduard Hovy. 2016. Hierarchical attention networks for document classification. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1480–1489, San Diego, California. Association for Computational Linguistics.

Hai Ye, Xin Jiang, Zhunchen Luo, and Wenhan Chao. 2018. Interpretable Charge Predictions for Criminal Cases: Learning to Generate Court Views from Fact Descriptions. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1854–1864, New Orleans, Louisiana. Association for Computational Linguistics.

Manzil Zaheer, Guru Guruganesh, Kumar Avinava Dubey, Joshua Ainslie, Chris Alberti, Santiago Ontanon, Philip Pham, Anirudh Ravula, Qifan Wang, Li Yang, et al. 2020. Big bird: Transformers for longer sequences. *Advances in Neural Information Processing Systems*, 33.

Haoxi Zhong, Zhipeng Guo, Cunchao Tu, Chaojun Xiao, Zhiyuan Liu, and Maosong Sun. 2018. Legal Judgment Prediction via Topological Learning. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3540–3549, Brussels, Belgium. Association for Computational Linguistics.

Haoxi Zhong, Chaojun Xiao, Cunchao Tu, Tianyang Zhang, Zhiyuan Liu, and Maosong Sun. 2020. How does NLP benefit legal system: A summary of legal artificial intelligence. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5218–5230, Online. Association for Computational Linguistics.

Octavia-Maria Şulea, Marcos Zampieri, Mihaela Vela, and Josef van Genabith. 2017. Predicting the Law Area and Decisions of French Supreme Court Cases. In *Proceedings of the International Conference Recent Advances in Natural Language Processing, RANLP 2017*, pages 716–722, Varna, Bulgaria. INCOMA Ltd.

A Training Effort

Type	BERT	RoBERTa
standard	3.377E+11	3.398E+11
long	1.365E+12	1.374E+12
hierarchical	1.476E+12	1.477E+12

Table 7: This table shows the total floating point operations per epoch per training example used for training each type. Each model has been trained for 2 to 4 epochs (variable because of early stopping). This table can be used to choose a suitable model with limited resources. Additionally, it can be used to measure the environmental impact.

Table 7 shows the training effort required for finetuning each type. Training one of the types capable of handling long input results in 4 to 5 times more training operations compared to the standard model. This seems justifiable since the gain from the longer models in terms of F1 score is considerable. Also, the entire cost of finetuning is relatively small.

B Examples

In this appendix we show some examples of court decisions with their respective labels. Figure 5 shows an example of a dismissed decision and Figure 6 an example of an approved decision. Both decisions are relatively short, but still contain all sections (rubrum, facts, considerations and judgments). They are both very recent, dating from 2019 and 2017 respectively.

C Input Length Distribution

In this appendix we show the input length distributions for the German (Figure 7) and Italian (Figure 8) datasets. We observe that the average Italian decision is longer than the average German decision. Additionally, there is also a higher density in moderately long decisions (over 1000 tokens) and there are many more decisions over 4000 tokens. Apart from the availability of more training data in the German dataset, the shorter decisions may also be an important factor in the better performance we see in most models trained on the German dataset in comparison to the Italian case and to some extent the French case (see Table 5).

D Tables to Plots

In this appendix, we show tables belonging to plots in the main paper to show the exact numbers. Table 8 shows the results regarding the different input lengths. Table 9 shows the results regarding different years in the test set. Table 10 shows the model performance across different cantons.

E Training with Class Weights

In this appendix we show the results of training the models with class weights instead of oversampling. Table 11 shows the training results. We notice, that for many configurations (especially with XLM-R), the model only learns the majority classifier. This leads to a very low Macro-F1 score. We also experimented with undersampling as an alternative to oversampling, but saw similar results to the training with class weights.

F Classifier Confidence

In this appendix, we discuss the reliability of the confidence scores of the classifier output alongside the predictions. The confidence scores are computed by taking the softmax on the classifier outputs, so that we get a probability (confidence) score of a given class between 0 and 100. The hierarchical and long BERT types show an increase in both the confidence in the correct predictions and the incorrect predictions compared to the standard BERT type (with the increase in the correct predictions being more pronounced). This finding holds across all three languages.

Model	standard		long		hierarchical	
	Micro-F1↑	Macro-F1↑	Micro-F1↑	Macro-F1↑	Micro-F1↑	Macro-F1↑
1-512 (5479 decisions)	81.1 ± 2.7	72.1 ± 1.6	80.8 ± 2.5	72.2 ± 1.3	39.3 ± 37.2	25.1 ± 17.4
513-1024 (3364 decisions)	65.3 ± 6.2	65.3 ± 6.2	71.8 ± 5.4	63.4 ± 2.8	43.3 ± 30.8	30.5 ± 13.2
1025-2048 (788 decisions)	63.8 ± 4.9	50.7 ± 1.0	69.1 ± 5.4	60.2 ± 2.8	54.9 ± 26.7	37.2 ± 15.3
2049-4096 (82 decisions)	64.9 ± 6.7	47.3 ± 2.2	65.1 ± 9.2	50.9 ± 3.6	60.2 ± 13.3	48.0 ± 5.4
4097-8192 (12 decisions)	56.7 ± 7.0	36.1 ± 2.8	50.0 ± 10.2	33.1 ± 4.8	50.0 ± 11.8	34.7 ± 5.4

Table 8: Results on the German data grouped by text length. Performance deteriorates as text length is increased.

Model	standard		long		hierarchical	
	Micro-F1↑	Macro-F1↑	Micro-F1↑	Macro-F1↑	Micro-F1↑	Macro-F1↑
2017	73.9 ± 4.2	64.2 ± 2.1	77.1 ± 3.9	69.1 ± 2.4	77.4 ± 3.9	69.5 ± 2.6
2018	74.2 ± 3.8	63.3 ± 1.2	76.6 ± 3.7	67.1 ± 1.8	76.7 ± 4.0	67.6 ± 1.9
2019	74.5 ± 4.0	64.8 ± 1.9	76.0 ± 3.7	67.5 ± 1.7	76.9 ± 3.8	68.3 ± 1.6
2020	73.5 ± 4.2	62.4 ± 1.6	76.6 ± 3.4	67.8 ± 1.8	77.4 ± 3.1	68.5 ± 1.5

Table 9: We used the German native BERT model pretrained and evaluated on the German data.

Canton	# cases	approval rate	standard		long		hierarchical	
			Micro-F1↑	Macro-F1↑	Micro-F1↑	Macro-F1↑	Micro-F1↑	Macro-F1↑
Berne (BE)	332	9.5%	79.4 ± 4.6	48.2 ± 7.7	78.7 ± 4.7	59.9 ± 2.6	78.5 ± 2.7	59.2 ± 3.4
Fribourg (FR)	1121	14.7%	76.7 ± 3.1	61.1 ± 1.2	75.8 ± 5.2	64.7 ± 3.6	79.5 ± 3.4	68.1 ± 2.6
Vaud (VD)	5684	17.0%	76.0 ± 1.8	58.8 ± 1.4	78.9 ± 3.0	68.7 ± 1.6	82.5 ± 1.7	71.1 ± 1.4
Valais (VS)	1399	20.6%	75.1 ± 1.0	52.4 ± 2.6	75.0 ± 2.6	63.7 ± 1.2	76.1 ± 3.3	64.0 ± 2.6
Neuchâtel (NE)	1226	14.9%	76.2 ± 3.6	57.4 ± 2.9	79.0 ± 3.9	68.0 ± 2.2	82.3 ± 2.7	70.8 ± 2.9
Genève (GE)	6017	21.8%	72.0 ± 3.1	59.4 ± 0.9	76.0 ± 3.3	69.4 ± 2.0	79.4 ± 2.3	71.8 ± 1.7
Jura (JU)	425	15.7%	80.1 ± 3.2	66.3 ± 2.8	78.9 ± 5.8	69.0 ± 5.1	83.8 ± 4.3	74.2 ± 4.5
Swiss Confederation (CH)	227	26.7%	70.0 ± 2.7	50.0 ± 4.9	72.0 ± 8.7	66.6 ± 7.9	73.3 ± 4.4	65.5 ± 5.8

Table 10: We used the French native BERT model pretrained and evaluated on the French data. The number of cases is counted on the training set per canton. The approval rate is calculated on the test set.

Model	de		fr		it	
	Micro-F1↑	Macro-F1↑	Micro-F1↑	Macro-F1↑	Micro-F1↑	Macro-F1↑
baselines						
Most Frequent	80.3	44.5	81.5	44.9	81.3	44.8
Stratified	66.7 ± 0.3	50 ± 0.4	66.3 ± 0.2	50 ± 0.4	69.9 ± 1.8	48.8 ± 2.4
Uniform	50 ± 0.3	44.8 ± 0.4	50 ± 0.6	44.5 ± 0.5	49.7 ± 2.4	44 ± 2.3
standard						
Native BERT	71.1 ± 3.3	62.6 ± 1.6	72.8 ± 5.5	58.2 ± 1.2	67 ± 13.1	49.4 ± 5.1
XLM-RoBERTa	77.8 ± 6.3	47.3 ± 6.3	76.1 ± 7.4	48.4 ± 4.9	80.4 ± 1.9	44.7 ± 0.4
long						
Native BERT	81.9 ± 1.2	**69.5 ± 0.9**	**81.8 ± 1.5**	69.4 ± 1.7	80.2 ± 1.4	46.1 ± 2.2
XLM-RoBERTa	**81.5 ± 0.7**	59.4 ± 9.6	81.5 ± 0.5	51.3 ± 8.8	81.3	44.8
hierarchical						
Native BERT	78.6 ± 2.1	69.2 ± 0.6	79.3 ± 0.8	**70 ± 0.7**	80.6 ± 1.1	**50.5 ± 6.5**
XLM-RoBERTa	80.3	44.5	80.3 ± 1.8	49.6 ± 9.8	81.3	44.8

Table 11: All the models have been trained and evaluated in the same language. With *Native BERT* we mean the BERT model pretrained in the respective language. The *Most Frequent* baseline just selects the majority class always. The *Stratified* baseline predicts randomly, respecting the training distribution. The best scores for each language are in bold. To combat label imbalance, we weighted the minority class samples more in the loss function.

Model	de		fr		it	
	Correct↑	Incorrect↓	Correct↑	Incorrect↓	Correct↑	Incorrect↓
standard	75.8 ± 13.6	64.7 ± 10.6	71.9 ± 12.2	64.4 ± 9.8	77.6 ± 12.2	68.3 ± 11.3
long	78.9 ± 12.2	65.8 ± 10.9	78.3 ± 11.6	67.8 ± 11.0	81.2 ± 11.2	68.4 ± 10.5
hierarchical	86.6 ± 15.9	69.3 ± 13.6	85.9 ± 15.2	70.8 ± 13.9	88.7 ± 14.7	71.4 ± 13.4

Table 12: This table shows the average confidence scores (0-100) of the different types of multilingual BERT models on the test set for correct and incorrect predictions respectively. Both the mean and standard deviation are averaged over 5 random seeds. The model has been finetuned on the entire dataset (all languages) and evaluated on the respective language.

Rubrum

5F_5/2019

Urteil vom 28. Mai 2019

II. zivilrechtliche Abteilung

Besetzung
Bundesrichter Herrmann, Präsident,
Bundesrichter von Werdt, Bovey,
Gerichtsschreiber Möckli.

Verfahrensbeteiligte
A._____,
Gesuchsteller.

Gegenstand
Gesuch um Revision des bundesgerichtlichen Urteils 5A_89/2018 vom 5. Februar 2018.

Sachverhalt: (facts)
Gegen die KESB Pfäffikon, diverse Sozialdienste, die Postfinance, Sozialversicherungen, eine Einwohnergemeinde, verschiedene Versicherungsgesellschaften, Banken und weitere Gesellschaften sowie mehrere Scientology Kirchen und andere religiöse Vereinigungen erhebt A._____ eine als subsidiäre Verfassungsbeschwerde betitelte Eingabe, mit welcher er die Revision des bundesgerichtlichen Urteils 5A_89/2018 vom 5. Februar 2018 verlangt. Ferner verlangt er die Verurteilung der Gegenparteien wegen Persönlichkeitsverletzung, die Aufhebung eines Erbscheines, die Überweisung von IV-Renten, die Erstellung von Abschlussrechnungen und vieles mehr.

Erwägungen: (considerations)

1.
Der Gesuchsteller zählt zwar verschiedene Revisionsgründe auf. Indes begründet er mit keinem Wort, inwiefern ein Revisionsgrund vorliegen soll. Ebenso wenig äussert er sich zur Einhaltung der Fristen (Art. 124 Abs. 1 BGG). Somit ist auf das Revisionsgesuch nicht einzutreten (Art. 42 Abs. 2 BGG).

2.
Soweit die Eingabe auch einen Beschwerdecharakter haben sollte, wäre darauf ebenfalls nicht einzutreten. Es ist zwar die Rede von "Urteilen des Obergerichts des Kantons Zürich" und solche könnten grundsätzlich Anfechtungsobjekt sein (Art. 75 Abs. 1 BGG). Jedoch werden diese nicht näher bezeichnet und es liegt auch keine Beschwerdebegründung im Sinn von Art. 42 Abs. 2 BGG vor, wenn festgehalten wird, die Beschwerde schütze die körperliche und geistige Unversehrtheit vor ausländischen Vereinen und Folgedelikten der ausserordentlichen Gerichte, um entsprechend den Beweisen, welche sich jeden Tag mit Völkermord beweisen, die Rechtsgleichheit und Unversehrtheit entsprechend der Mehrheit zu beweisen.

3.
Die Urteile des Bundesgerichtes erfolgen grundsätzlich im schriftlichen Verfahren (zu den Ausnahmen vgl. Art. 57 und 58 BGG), weshalb der Antrag auf eine öffentliche Verhandlung abzuweisen ist.

4.
Die Gerichtskosten sind dem Gesuchsteller aufzuerlegen (Art. 66 Abs. 1 BGG).

Demnach erkennt das Bundesgericht: (rulings)

1.
Der Antrag auf öffentliche Verhandlung wird abgewiesen.

2.
Auf das Revisionsgesuch wird nicht eingetreten.

3.
Soweit eine Beschwerde erhoben werden sollte, wird auf diese nicht eingetreten.

4.
Die Gerichtskosten von Fr. 1'000.-- werden dem Gesuchsteller auferlegt.

5.
Dieses Urteil wird dem Gesuchsteller schriftlich mitgeteilt.

Lausanne, 28. Mai 2019

Im Namen der II. zivilrechtlichen Abteilung des Schweizerischen Bundesgerichts

Der Präsident: Herrmann

Der Gerichtsschreiber: Möckli

Figure 5: This is an example of a dismissed decision: https://tinyurl.com/n44hathc

---

Rubrum

9C_502/2017

Urteil vom 21. September 2017

II. sozialrechtliche Abteilung

Besetzung
Bundesrichterin Pfiffner, Präsidentin,
Bundesrichterin Glanzmann, Bundesrichter Parrino,
Gerichtsschreiberin Oswald.

Verfahrensbeteiligte
A._____,
vertreten durch Rechtsanwalt Jan Herrmann,
Beschwerdeführerin,

gegen

IV-Stelle Basel-Stadt,
Lange Gasse 7, 4052 Basel,
Beschwerdegegnerin.

Gegenstand
Invalidenversicherung
(vorinstanzliches Verfahren; Prozessvoraussetzung),

Beschwerde gegen den Entscheid des Sozialversicherungsgerichts des Kantons Basel-Stadt vom 10. Juni 2017 (IV.2016.186).

Nach Einsicht (facts)
in die Beschwerde vom 18. Juli 2017 (Poststempel) gegen den Entscheid des Sozialversicherungsgerichts des Kantons Basel-Stadt vom 10. Juni 2017, mit welchem auf das Gesuch vom 21. November 2016 um Revision des Entscheids vom 11. Oktober 2016 nicht eingetreten wurde,

in Erwägung, (considerations)
dass das kantonale Gericht erkannte, das einschlägige Prozessrecht (§ 18 Abs. 1 lit. a des Gesetzes über das Sozialversicherungsgericht des Kantons Basel-Stadt und über das Schiedsgericht in Sozialversicherungssachen vom 9. Mai 2001 [Sozialversicherungsgerichtsgesetz, SVGG/BS SGS 154.200]) sehe die Revision unter anderem bei Entdeckung neuer erheblicher Tatsachen oder Beweismittel vor (Art. 61 lit. i ATSG),
dass es erwog, beim Revisionsgesuch handle es sich um ein ausserordentliches Rechtsmittel, das sich gegen einen rechtskräftigen Beschwerdeentscheid richte, ein solcher jedoch vorliegend nicht

gegeben sei, da die Gesuchstellerin gegen den Entscheid des Sozialversicherungsgerichts Basel-Stadt vom 11. Oktober 2016 Beschwerde beim Bundesgericht erhoben habe,
dass diese Beschwerde vom 21. November 2016 beim Bundesgericht noch hängig ist (9C_782/2016),
dass eine Vorinstanz des Bundesgerichts auf ein Revisionsgesuch nicht einzig mit der Begründung nicht eintreten darf, gegen den zu revidierenden Entscheid sei Beschwerde beim Bundesgericht erhoben worden (**BGE 138 II 386** E. 6 S. 389 ff.; Urteil 8C_921/2014 vom 12. Mai 2015 E. 2.3),
dass die Beschwerde damit offensichtlich begründet und deshalb im Verfahren nach Art. 109 Abs. 2 lit. b BGG mit summarischer Begründung (Art. 109 Abs. 3 Satz 1 BGG) gutzuheissen ist,
dass umständehalber auf die Erhebung von Gerichtskosten zu verzichten ist (Art. 66 Abs. 1 Satz 2 BGG),

Demnach erkennt das Bundesgericht:

1.
Die Beschwerde wird gutgeheissen. Der Entscheid des Sozialversicherungsgerichts des Kantons Basel-Stadt vom 10. Juni 2017 wird aufgehoben. Die Sache wird an die Vorinstanz zurückgewiesen, damit sie über die übrigen Eintretensvoraussetzungen bezüglich des Gesuchs vom 21. November 2016 entscheide und dieses gegebenenfalls materiell behandle.

2.
Es werden keine Gerichtskosten erhoben.

3.
Die Beschwerdegegnerin hat die Beschwerdeführerin für das bundesgerichtliche Verfahren mit Fr. 2'000.- zu entschädigen.

4.
Dieses Urteil wird den Parteien, dem Sozialversicherungsgericht des Kantons Basel-Stadt und dem Bundesamt für Sozialversicherungen schriftlich mitgeteilt.

Luzern, 21. September 2017

Im Namen der II. sozialrechtlichen Abteilung des Schweizerischen Bundesgerichts

Die Präsidentin: Pfiffner

Die Gerichtsschreiberin: Oswald

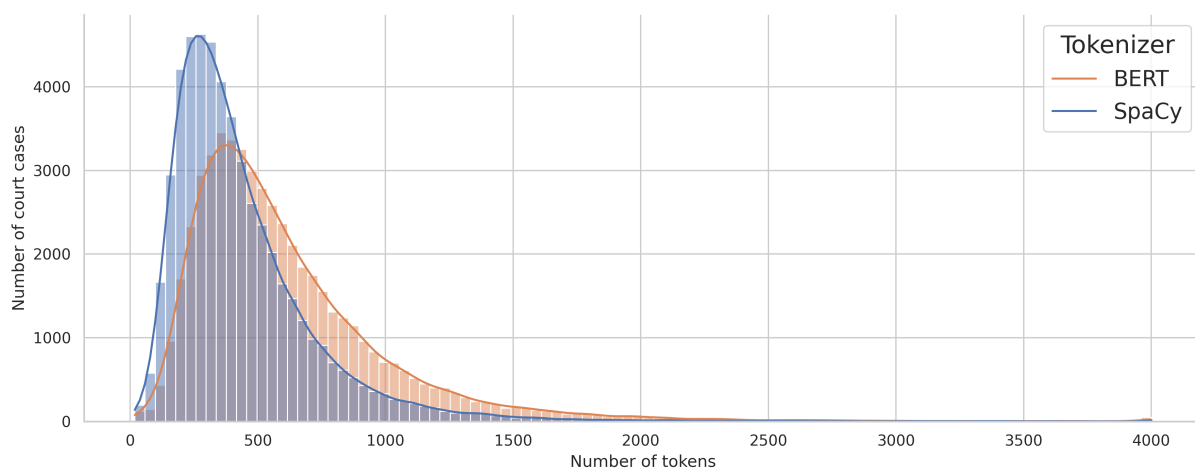Figure 6: This is an example of an approved decision: https://tinyurl.com/mjxfjn65

Figure 7: This histogram shows the distribution of the input length for German decisions. The blue histogram is generated from tokens generated by the spacy tokenizer (regular words). The orange histogram is generated from tokens generated by the SentencePiece tokenizer used in BERT (subword units). Decisions with length over 4000 tokens are grouped in the last bin (before 4000).
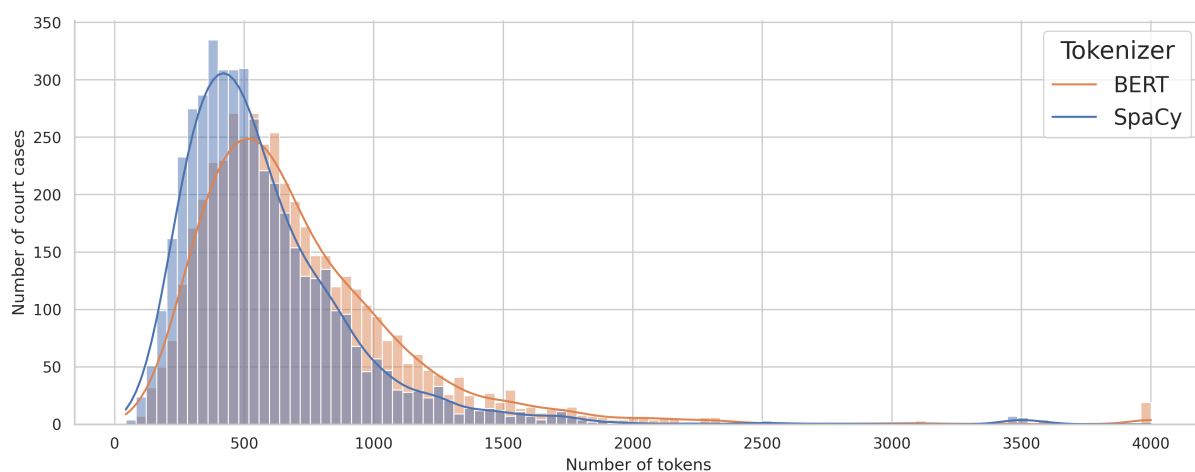

Figure 8: This histogram shows the distribution of the input length for Italian decisions. The blue histogram is generated from tokens generated by the spacy tokenizer (regular words). The orange histogram is generated from tokens generated by the SentencePiece tokenizer used in BERT (subword units). Decisions with length over 4000 tokens are grouped in the last bin (before 4000).