Author/s:
Vihikan, WO;MISTICA, M;Levy, I;Christie, A;Baldwin, T;Mistica, M

Title:
Automatic Resolution of Domain Name Disputes

Date:
2021

Citation:
Vihikan, W. O., MISTICA, M., Levy, I., Christie, A., Baldwin, T. & Mistica, M. (2021). Automatic Resolution of Domain Name Disputes. Natural Legal Language Processing Workshop 2021, pp.228-238. Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.nllp-1.24.

Persistent Link:
https://hdl.handle.net/11343/291996

Automatic Resolution of Domain Name Disputes

Wayan Oger Vihikan[♦] **Meladel Mistica**[♥]
Inbar Levy[♣] **Andrew F. Christie**[♣] **Timothy Baldwin**[♦]
[♦] Computing and Information Systems, The University of Melbourne
[♥] School of Languages and Cultures, The University of Queensland
[♣] Melbourne Law School, The University of Melbourne
wvihikan@student.unimelb.edu.au, m.mistica@uq.edu.au
{a.christie, inbar.levy}@unimelb.edu.au, tb@ldwin.net

Abstract

We introduce the new task of domain name dispute resolution (DNDR), that predicts the outcome of a process for resolving disputes about legal entitlement to a domain name. The ICANN UDRP establishes a mandatory arbitration process for a dispute between a trademark owner and a domain name registrant pertaining to a generic Top-Level Domain (gTLD) name (one ending in .COM, .ORG, .NET, etc). The nature of the problem leads to a very skewed data set, which stems from being able to register a domain name with extreme ease, very little expense, and no need to prove an entitlement to it. In this paper, we describe the task and associated data set. We also present benchmarking results based on a range of models, which show that simple baselines are in general difficult to beat due to the skewed data distribution, but in the specific case of the respondent having submitted a response, a fine-tuned BERT model offers considerable improvements over a majority-class model.

1 Introduction

Domain name abuse, or cybersquatting, is the act of registering and using a domain name in bad faith. Examples of such behaviour include registering a generic Top-Level Domain (gTLD) name (i.e. one ending in .COM, .ORG, .NET, etc) that incorporates another's trademark, with the intent to on-sell it to the trademark owner; to dupe the public into erroneously assuming that the owner of the domain name is the trademark owner; or to interfere with the trademark owner's business, to name a few. A victim of cybersquatting has a means to remedy the situation. In 1999, the agency responsible for the internet's naming system, the Internet Corporation for Assigned Names and Numbers (ICANN), introduced a fast and low-cost mandatory arbitration system as an alternative to the slow and expensive litigation of cybersquatting disputes — the Uniform Domain Name Dispute Resolution Policy (UDRP).[1] Since its establishment, approximately 80,000 complaints have been resolved through the process (Christie, 2020).

UDRP cases, over 90% of which are in English, are decided by an arbitration panel comprising either one or three experienced trademark lawyers. The typical arbitrator's fee of USD 1,000 is small relative to the time expended and the usual hourly rate for legal services of this type. The performance of the task is, in effect, cross-subsidized by the arbitrator's day job. We introduce this task as a means to assist the service providers and arbitrators by automatically pre-assessing complaints. The ultimate aim is to identify those complaints whose characteristics are predictive of success. Doing so has the potential to aid in the optimal allocation of finite resources — in this case, the time and expertise of the arbitrator — because such cases should be relatively easier for the arbitrator to decide. By identifying the easier cases, service providers and arbitrators can direct more resources towards the remainder of complaints, with the result that the cases most in need of resources are given a greater share of time. That should improve the quality of decision-making, benefiting all stakeholders in the domain name system.

The distribution of outcomes under the UDRP is highly skewed. Of the complaints that proceed to a panel determination, 87% succeed. However, the distribution of cases is also highly skewed by whether or not a response to the complaint is filed. In the arbitration process, the domain name registrant has the right to file a rebuttal (a "response") to the complainant's claim. For nearly three-quarters of all complaints, no response is filed, in which case the complaint almost always (94% of the time) succeeds. Where a response is filed, however, the success rate is only 66% (Christie, 2014). It follows that cases where a response is filed are of particular

[1] https://www.icann.org/resources/pages/policy-2012-02-25-en

Natural Legal Language Processing Workshop 2021, pages 228–238

interest to the task.

While we are by no means the first to propose the task of legal judgment prediction (Aletras et al., 2016; Zhong et al., 2018a; Chalkidis et al., 2019; Zhong et al., 2020), previous work has focused almost exclusively on criminal or human rights cases, which we argue have sensitivities and ethical dimensions that are yet to be fully quantified (Leins et al., 2020; Tsarapatsanis and Aletras, 2021). Part of our intention with this paper is to introduce a "lower-stakes" task/data set in terms of its direct and indirect implications for individuals and civil liberties, which is far from "solved" in terms of current NLP capabilities.

In the introduction of this new task, we outline in Section 2 the process of how a complaint is dealt with once it is submitted to an ICANN-accredited service provider, and how a decision on the complaint is reached by the arbitrator appointed to decide it. In this study, we ultimately aim to ascertain if providing predicted outcomes to the arbitrators will assist them in the dispute resolution process — not as a means to make the decision for them, but as a way to pre-sort the difficult cases from the straightforward ones. In Section 3, we describe how we develop our corpus for the DNDR data set, and Section 4 details how we prepare this data for our experiments as well as the set-up of the fine-tuning experiments. Given that the data set is highly skewed towards one class, we expect that the reporting of the overall binary task to not impressively exceed a majority-class baseline. However, the results from the important partitioning of the data between cases with and without responses in Section 5 shows that this task has the potential to aid arbitrators in performing their duties. In Section 6, we discuss how we foresee AI to be of use in legal processes and assisting legal professionals in their decision-making, and why it's important to define best practice when it comes to employing artificial intelligence to assist in the legal domain.

2 Background

A domain name dispute commences when a trademark owner ("complainant") files a complaint with an ICANN-accredited service provider, seeking a remedy against a domain name registrant ("respondent"). Of the six bodies accredited by ICANN as UDRP service providers,[2] the largest by far is the World Intellectual Property Organisation (WIPO). Over 52,000 cases — more than one half all UDRP disputes — have been filed with WIPO (Christie, 2020). Its case-load is growing, with filings in 2020 topping 4,000.[3]

To obtain a remedy, the complainant must show that four conditions are satisfied: (1) the complainant owns a trademark; (2) the domain name is identical or confusingly similar to that trademark; (3) the respondent has no rights or legitimate interests in the domain name; and (4) the respondent registered and has used the domain name in bad faith. The dominant service provider, WIPO, has a pro forma Model Complaint form. As a result, complaints filed with WIPO are highly standardized in terms of structure and form.

Upon the filing of a complaint, WIPO checks it for compliance with administrative requirements, obtains certain information about the domain name and the registrant of it, then notifies the complaint to the respondent. The respondent has 20 business days in which to file a response. Once the deadline for filing a response has passed, WIPO appoints a panel and supplies it with the case file — essentially, the complaint and any response. The panel has 14 days in which to decide the case and to write a decision justifying its outcome, which are both publicly disclosed (Christie, 2014).

A panel decision will generally come to one of two outcomes: either the complaint succeeds, or the complaint fails (is "denied"). In a small number of cases, where the parties settle the dispute after commencement but prior to a decision being rendered, the panel will "terminate" the complaint. Where the complaint succeeds, the panel will order one of two remedies, based on the desires of the complainant: either cancellation of the domain name, or transfer of the domain name to the complainant. The registrar of the domain name will give effect to the order (i.e. either cancel or transfer the domain name) unless the respondent challenges the decision by filing a court case in the appropriate jurisdiction (Christie, 2002), something which almost never happens. Thus, in practice, the panel's decision is final and binding. Ideally the task would be undertaken on the complaint and any response that was filed. Unfortunately, however, neither of those documents are published. Instead we use the panelist's decision, which is publicly

[2]For the list of accredited service providers, see https://www.icann.org/resources/pages/providers-6d-2012-02-25-en

[3]For full statistics, see https://www.wipo.int/amc/en/domains/statistics/outcome.jsp

WIPO Arbitration and Mediation Center

ADMINISTRATIVE PANEL DECISION

Groupon, Inc. v. Domain Admin, Whois Privacy Corp.

Case No. D2020-0001

1. The Parties

The Complainant is Groupon, Inc., United States of America ("United States"), represented by Greenberg Traurig, LLP, United States.

The Respondent is Domain Admin, Whois Privacy Corp., Bahamas.

2. The Domain Name and Registrar

The disputed domain name <ggroupon.com> is registered with Internet Domain Service BS Corp (the "Registrar").

3. Procedural History

The Complaint was filed with the WIPO Arbitration and Mediation Center (the "Center") on January 1, 2020. On January 2, 2020, the Center transmitted by email to the Registrar a request for registrar verification in connection with the disputed domain name. On January 3, 2020, the Registrar transmitted by email to the Center its verification response confirming that the Respondent is listed as the registrant and providing the contact details.

Figure 1: Excerpt from a panel decision document showing the first sections, from https://www.wipo.int/amc/en/domains/search/text.jsp?case=D2020-0001.

available. WIPO decisions are highly standardized in structure and form. They typically contain seven sections, with the following headings: (1) "The Parties"; (2) "The Domain Name and Registrar"; (3) "Procedural History"; (4) "Factual Background"; (5) "Parties' Contentions"; (6) "Discussion and Findings"; and (7) "Decision". Sections (4) and (5) are the panel's summaries of the facts, and the arguments about the legal import of those facts, respectively, that are contained in the complaint and response. They are, therefore, a proxy for the information that is contained in the complaint and response, which are not publicly available.

3 Corpus Creation

The corpus was derived by the scraping the text from the published decision of each gTLD domain name case[4] for the years 2000 to mid-2020. Figure 1 shows the first sections of a sample decision. Each section of the decision is extracted and stored in JSON format, as shown in Figure 2, with the text of each section labeled according to its heading.

For the period between January 2000 and August 2020, we extracted over 33k decisions in total. We found WIPO decisions in a number of languages, including Swedish, Spanish, French, and Chinese. However, the number of non-English cases were too few (a little less than 3k), and therefore we opted to create an English-only corpus for our initial release of the data set. After filtering on the English documents using saffsd/langid library (Lui and Baldwin, 2012), we arrived at a distribution of decision outcomes shown in Table 1.

For this preliminary work, we decided to only keep the outcome labels that had enough instances to be able to reasonably train a model, which is why we only kept the top 3 labels: TRANSFER, COMPLAINT DENIED, and CANCELLATION (which account for over 99% of the data), as shown in Table 2.

The purpose of the task is to predict the outcome of the case based on the complainant's grievance and the respondent's counter to it. As detailed in Section 2, the grievance in the form in which it was filed with the service provider (the complaint) is *not* published. This is also the case for the respondent's counter to the grievance (the response). However, the contents of both the complaint and the response are summarized by the panel in the sections on "Factual Background" and "Parties' Contentions". We retain those sections, as a proxy for the facts alleged by complainant and the respondent, and for their legal arguments pertaining to those facts.

We also retain the three preceding sections of the decision. The first two of these are "The Parties" and "The Domain Name and Registrar", in

[4] https://www.wipo.int/amc/en/domains/decisionsx/index.html

```
{
    "case-0": {
        "status": "transfer",
        "url": "https://www.wipo.int/amc/en/domains/search/text.jsp?case=D2020-0001",
        "no": "D2020-0001",
        "complainant": "Groupon, Inc.",
        "respondent": "Domain Admin, Whois Privacy Corp.",
        "title": "",
        "text": {
            "section-3": {
                "h3": "3. Procedural History",
                "p": [
                    "The Complaint was filed with the WIPO Arbitration and Mediation Center...
                ],
                "a": []
            },
            "section-4": {
                "h3": "4. Factual Background",
                "p": [
                    "The Complainant is a major player in local commerce that offers consumers a...
                ],
                "a": []
            },
            "section-5": {
                "h3": "5. Parties' Contentions",
                "p": [],
                "a": []
            },
            "section-6": {
                "h4": "A. Complainant",
                "p": [
                    "The Complainant contends that the disputed domain name is confusingly similar to...
                ],
                "a": []
            },
            "section-7": {
                "h4": "B. Respondent",
                "p": [
                    "The Respondent did not reply to the Complainant's contentions."
                ],
                "a": []
            }
        }
    }
}
```


Figure 2: The panel decision document in json format showing the sections of the data from Figure 1.

which the complainant, the respondent, the domain name, and the registrar are named. The third is the "Procedural History", in which is recorded the dates on which the complaint and response were filed, the date on which the panel was appointed, and the name of the panelist(s).

The final two sections of the decision are "Discussion and Findings", in which the panel analyzes whether or not the complainant has proven its claim; and "Decision", in which the outcome is formally recorded. We remove those sections, as they contain the outcome of the case, which is the very thing the task is seeking to predict.

In total there were 33,841 cases at the time of scraping. However, not all of them could be collected. There are four cases that are not available, and 67 cases cannot be scraped due to inconsistencies in the HTML tags. On top of that, there are 2,757 non-English cases. So, after the filtering, 31,013 cases remained. We filtered the data further by selecting cases that have both Procedural History and Discussion and Findings sections. Moreover, we only chose cases that have transfer, complaint denied, and cancellation decision. In the end we obtained 30,311 cases. This data set only contains Procedural History, Factual Background, and Parties' Contentions sections from each cases. The vocabulary size of the data set is roughly 245K unique tokens. The data set is split into training, development, and testing set. Their length on average is 1,131, 940, and 939 words respectively. However, these numbers are reduced after performing text processing before training the model.

We provide a Data Statement (Bender and Friedman, 2018) for the data set in the Appendix.

4 Experiments

We conducted a series of experiments on our new data set. We developed 2 RNN models, 3 BERT models, as well as 2 non-neural models. Given that the data set is highly imbalanced, we also provide a majority-class baseline.

TRANSFER	29,565
COMPLAINT DENIED	3,397
CANCELLATION	530
TRANSFER, DENIED IN PART	110
COMPLAINT DENIED WITH DISSENTING OPINION	71
TRANSFER WITH DISSENTING OPINION	65
COMPLAINT DENIED, TRANSFER IN PART	32
TERMINATED BY PANEL (ORDER PUBLISHED)	11
COMPLAINT DENIED WITH DISSENTING AND CONCURRING OPINION	5
TRANSFER WITH CONCURRING OPINION	5
CANCELLATION, TRANSFER IN PART	4
TRANSFER, DENIED IN PART WITH DISSENTING OPINION	4
CANCELLATION, DENIED IN PART	3
COMPLAINT DENIED, TRANSFER IN PART WITH DISSENTING OPINION	2

Table 1: Document count per decision category after filtering for English documents.

TRANSFER	26,865
COMPLAINT DENIED	2,970
CANCELLATION	476

Table 2: The three largest categories after removing cases with HTML issues and further clean-up

	Label	Train	Dev	Test
1	ACCEPT	21,645	2,848	2,848
0	DENY	2,604	183	183

Table 3: WIPO data set: division of training, development and testing data for cases that accept or deny the complainant's claims.

4.1 Data Preparation

Of three labels in our data set, TRANSFER and CANCELLATION largely amount to the same result of the case for the complainant being upheld, while COMPLAINT DENIED is a rejection of the complainant's claims. As such, we merge the first two classes into ACCEPT to define a binary classification task.

Table 3 shows the number of instances per class based on an 80/10/10 split of the data into training/development/testing partitions, determined chronologically to reflect the reality of applying the model to future data (with the training instances being the earliest data, followed by the development, and finally the test data).

Each case is represented as a JSON file with seven fields: (1) the (binary) outcome decision; (2) the URL of the original case description; (3) the WIPO case number; (4) the name of the complainant; (5) the name of the respondent; (6) the title of the case; and (7) the textual content of the case, structured based on headings and section boundaries. All of the data described as well as the code is available for download.[5]

4.2 Method

We first developed two RNN-based models: (1) BIGRU (a bidirectional GRU: Cho et al. (2014)); and (2) BILSTM (a bidirectional LSTM: Hochreiter and Schmidhuber (1997)). For both of these models we use GLOVE pretrained word vectors (Pennington et al., 2014), derived from 42 billion tokens of Common Crawl[6] with 300-dimension vectors.[7] For both models the pretrained GLOVE embeddings are not updated in the training phase (i.e. they are set to 'untrainable').

The BIGRU and BILSTM models are built using the layers module in Tensorflow's Keras API, with the pretrained embeddings being passed through a SpatialDropout layer with value 0.2, followed by either an LSTM or GRU layer with 128 units and 0.2 dropout. It is then followed by the second (reverse-direction) layer but with 64 units, a 0.2 dropout layer, and finally a dense layer that has 3 units as output, with softmax activation. The loss is calculated using categorical cross-entropy,

[5] Data: https://people.eng.unimelb.edu.au/tbaldwin/resources/wipo/; Code: https://github.com/vihikan/automatic-resolution-of-domain-name-disputes

[6] https://commoncrawl.org

[7] https://nlp.stanford.edu/projects/glove/

trained using the Adam optimizer (Kingma and Ba, 2017) with a learning rate of 0.001. We iterate through 20 epochs with a batch size of 64.

We next train 3 variants of BERT (Devlin et al., 2019): (1) OOBERT (out-of-the-box BERT); (2) FTBERT (fine-tuned BERT); and (3) LEGAL-BERT, a BERT model pre-trained on legal documents (Chalkidis et al., 2020) which we fine-tune over our data. The difference between OOBERT and FTBERT is that we additionally pre-train FTBERT over the non-test data using the masked language model objective. By default, the BERT base model has 12 transformer layers, and in this experiment, we freeze the first eight layers, in addition to the embeddings layer. The classification layer has 64 units followed by a dropout layer with value of 0.2, and an output layer with 3 units. The loss is once again calculated using cross-entropy loss with the Adam optimizer and a learning rate of 1e-5.

The BERT models are trained for 20 epochs. To save on training time, an early stopping algorithm is implemented by monitoring the macro-averaged F1 score with a patience of 3 and minimum delta of 0.0001.

To counter the effects of randomization, we perform all experiments with the RNN and BERT models 5 times with different random seeds, and average over the runs.

As non-neural baselines, we train a logistic regression (LOGREG) and linear-kernel SVM (SVM) model over TFIDF-weighted feature vectors.

In terms of document representation, for BIGRU and BILSTM we truncate the document to the first 5,000 tokens, for the BERT models we truncate to the first 512 tokens, and for LOGREG and SVM we represent the full document as a (TF-IDF weighted) bag of words without truncation.

4.3 Evaluation

We evaluate based on accuracy, macro-averaged precision, recall and f1-score. In addition to the overall results, we also break down results across cases that have a response and those that do not.

Although the "Procedural History" section of the decision states in plain language whether or not a response was filed, there is no convenient machine-readable category that discretely labels a case as having a response or not, and this information needs to be extracted from the decision. In the "Parties' Contentions" section, the panelist generally states whether or not the respondent exercised their right to respond. For example, the panelist can report that *The Respondent did not reply to the Complainant's contentions.* Other examples of a negative response are *The Respondent, having been duly notified of the Complaint and of these proceedings, did not reply to the Complainant's contentions or take any part in these proceedings* and *The Respondent expressed his willingness to transfer the disputed domain name to the Complainant, but the parties were not able to reach a settlement. The Respondent did not reply to the Complainant's contentions substantively.* The variation in these examples show that automatically detecting a non-response isn't entirely straightforward.

In determining whether a substantive response was made to the complaint for the purpose of partitioning cases based on whether there was a response or not, we develop the following set of heuristics for the "Parties' Contentions" section, where 'RESPONSE = 1' means there was a response, and 'RESPONSE = 0' means there was not.

- IF the "Respondent" section is empty
  - THEN, RESPONSE = 0
- IF the "Respondent" section is non-empty
  - IF text includes *did not*, *has not*, or *no response*
    - IF the text length is ≤ 100 words
      - THEN, RESPONSE = 0
    - OTHERWISE, RESPONSE = 1
  - OTHERWISE, RESPONSE = 1

We set the threshold of 100 words because if there was any prose in the respondent's section of the report that was less than 100 words, it usually only indicated that the respondent did not reply in any substantive way. Employing the above heuristics, we found that of the 3,031 instances in the test set, 441 had responses and 2,590 did not.

5 Results

Table 4 reports the accuracy (ACC), and macro-average precision (P), recall (R) and f1-score (F).

As can be seen, the Majority-class baseline is high due to natural skew in the data set.

In terms of accuracy, all systems exceed the majority-class baseline, with little difference between the best-performing BERT models and the

Model	Acc	Macro-average		
		P	R	F
Majority	0.940	0.470	0.500	0.484
LogReg	0.959	**0.888**	0.710	0.769
SVM	0.960	0.869	0.733	0.783
BiGRU	0.953	0.815	0.731	0.763
BiLSTM	0.951	0.800	0.725	0.755
ooBert	0.956	0.833	0.732	0.769
ftBert	0.960	0.837	**0.792**	**0.813**
LegalBert	**0.961**	0.848	0.773	0.805

Table 4: Overall results for DNDR

traditional machine learning models (LogReg and SVM). However, LegalBert slightly improves over the other models in terms of raw accuracy. For the macro-average precision, recall, and f1-score, all systems far exceed the baseline, with the overall best-performing model being ftBert with an f1-score of 0.813 ($\sigma = 0.009$), primarily due to the high recall of 0.792. LegalBert is a close second with an f1-score of 0.805 ($\sigma = 0.010$).

Of all the models, LogReg performs best in terms of precision with 0.888, but unfortunately its low recall of 0.710 diminishes its overall performance. The RNN models (BiGRU and BiLSTM) perform the worst of the trained models, in terms of accuracy, precision, and f1-score, and are omitted for the remainder of the paper.

Although all systems outperform the majority-class baseline, it is difficult to ascertain the benefit of these systems in terms of real-world utility, and their ability to indeed assist a panelist in their task. If they do not perform substantially above the majority-class baseline, then our systems are not better the simple (and, in practical terms, useless) majority-class baseline heuristic of labeling every case as Transfer.

We know from previous analyses that cases in which no response is filed will almost always succeed, as it is almost certain that such cases are straightforward instances of illegitimate cybersquatting. Conversely, where the complaint is not clear-cut, there is more incentive for a respondent to contest it. Thus, we expect that it is only the more difficult cases that have had responses filed.

Table 5 shows the breakdown of the results split into the two cases of with a response vs. without a response. Looking first at the 'No Response' column, in terms of accuracy all systems are on par with the majority-class baseline. However, LogReg far exceeds all the other methods in terms of macro-averaged precision, recall, and f1-score. While this result may at first glance seem surprising, it is important to bear in mind that LogReg and SVM have a representation of the entire document, where the Bert models only capture the first 512 tokens. As such, there is a tradeoff between the expressiveness of the model (higher for the Bert models) and how holistic the document representation is (better for LogReg and SVM), which we see play out in these results.

Looking next to the 'Response' column of Table 5, we see that the trained models generally perform well above baseline, with all trained models other than ooBert far exceeding the majority class baseline, and ftBert and LegalBert achieving almost identical precision, recall, and f-score. It is interesting to observe that LegalBert has no advantage over ftBert in this setting, that is extensive pre-trained over legal texts of different types prior to pre-training over the WIPO data, is almost identical to simply pre-training over the WIPO data.

6 Related Work and Discussion

This is the first work to propose predicting judgments of IP cases, in the low-stakes (but technically challenging, as evidenced by the empirical results) setting of World Intellectual Property Organisation domain name resolution cases. While the legal context for this research and resulting data set are novel, we are in no way the first to propose judgment prediction as a task for Legal NLP (Kort, 1957; Lawlor, 1963; Aletras et al., 2016; Zhong et al., 2018a; Chalkidis et al., 2019; Zhong et al., 2020, inter alia). Perhaps the most popular judgment-related data set is the Chinese AI and Law challenge data set ("CAIL 2018"), where the task is to predict which of 202 charges the accused is guilty of as a multi-label classification task, based on more than 2.6m criminal cases published by the Supreme People's Court of China in Mandarin Chinese (Xiao et al., 2018). The original competition attracted over 200 submitting teams (Zhong et al., 2018b), and included two other tasks: (1) law articles: identify criminal law articles that are cited as being relevant to a given case; and (2) sentencing: predict the jail sentence, in months, for the guilty party. The data set has since been used

Model	Response				No Response			
	Acc	P	R	F	Acc	P	R	F
Majority	0.692	0.346	0.500	0.409	0.982	0.491	0.500	0.495
LogReg	0.828	**0.827**	0.753	0.775	0.982	**0.888**	**0.710**	**0.769**
SVM	0.828	0.820	0.759	0.779	0.982	0.759	0.584	0.625
ooBERT	0.692	0.770	0.738	0.750	0.982	0.769	0.586	0.628
ftBERT	0.831	0.808	**0.806**	**0.803**	0.982	0.763	0.598	0.642
LegalBERT	**0.834**	0.808	0.798	0.802	0.982	0.753	0.573	0.611

Table 5: Breakdown of results based on response vs. no response

widely as a benchmark data set for reasoning and legal NLP (Yang et al., 2019; Li et al., 2019; Xu et al., 2020; Zhong et al., 2020).

For English, the most popular data set is perhaps that of Chalkidis et al. (2019) based on over 11k European Court of Human Rights ("ECtHR") cases, in the form of two tasks: (1) a binary classification task: does the case violate any of the 66 articles and protocols of the ECtHR; and (2) a multi-label classification task: which, if any, of the 66 articles and protocols does a given case violate.[8] Other data sets include a French data set of 126k cases from the French Supreme Court (Şulea et al., 2017), an English data set of 5k cases from the UK Court of Appeal (Strickson and De La Iglesia, 2020), an English data set of 28k cases from the Supreme Court of the United States (Katz et al., 2017), and a German data set of around 6k tax law appeal cases from the German Fiscal Courts (Waltl et al., 2017).

While the majority of this work has focused solely on the task of predicting the case judgment, Chalkidis et al. (2019) equally focus on using the trained models to analyze the fairness of judicial decisions, and more generally as an analytic tool for enhancing understanding of how judicial decisions are made.

Methodologically speaking, much of the past work has approached judgment prediction using off-the-shelf topic classification methods (e.g. using TF-IDF-weighted bag-of-words document representations with a linear-kernel SVM or naive Bayes model, similar to our LogReg and SVM), although recently the data sets of Xiao et al. (2018) and Chalkidis et al. (2019) have equally been used as benchmark data sets to evaluate longer-context pre-trained LMs over (based on the fact that the average document length tends to be much longer than the standard 512 token limitation of conventional BERT models).

As alluded to in the introduction, Leins et al. (2020) recently questioned the appropriateness of data sets such as Xiao et al. (2018), on the grounds of issues such as there being only superficial anonymisation of the data (leading to risk of the model learning demographic biases based on names and place mentions), privacy concerns (e.g. LMs fine-tuned on the data would potentially generate outputs containing sensitive information regarding defendants mentioned in the data), the data set being developed without engagement with the relevant legal authorities, and the appropriateness of the criminal domain for the deployment of automatic decision-making tools, given the high personal stakes for defendants. While noting some of the counter-examples to these concerns of Tsarapatsanis and Aletras (2021) and the need for further debate regarding the ethics of the applications of legal NLP, we return to our earlier claim that legal areas such as domain name disputes are a more appropriate short-term application domain for judgment prediction, in the sense that there are very few personal sensitivities which would have implications on the later life of the individuals represented in the data.

WIPO documents are relatively consistent in formatting and content between cases, similarly to the ECtHR documents, e.g., which follow a rigid format. The WIPO documents are generally more consistent than other legal cases, in which there is a significant issue for NLP techniques, as discussed by Higgins et al. (2020).

7 Conclusion

We have introduced a new large-scale data set for legal judgment prediction, derived from domain

[8] There is also a third task of case importance regression, on a scale of 1–4, where no explicit judgment prediction is made.

name dispute cases from the World Intellectual Property Organisation where a complainant has applied for a domain name to be transferred or cancelled on the grounds of trademark infringement. The data set is released with a data statement as defined by Bender and Friedman (2018) in the Appendix.

In the context of binary classification of the complaint being accepted or denied, we provided results for a number of benchmark NLP methods. In terms of macro-averaged f-score (to counter the effects of class imbalance), we showed that BERT with fine-tuning performed the best, including over the subset of cases where a response has been submitted. However, we equally showed that there is considerable room for improvement in the results, and plenty of room for further work on this task.

In terms of future work, we aim to further build on the models presented in this paper, and test their utility as decision support models. Our initial design was to simply experiment with the threshold of the model's output as a means to trade off precision for recall, and rely on the probability score as a proxy for model confidence and a measure of reliability. However, pretrained models such as BERT and LEGALBERT can suffer from severe miscalibration due to over-parameterisation. (Desai and Durrett, 2020; Kong et al., 2020).

As a first step, we aim to perform model calibration experiments to address the over-confidence issues commonly seen in these fine-tuned models. Further to this step, we will validate the accuracy of these calibrated models by revisiting the errors made by the system and verifying the original judgments by the panelists with a fresh set of legal eyes.

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.

Emily M. Bender and Batya Friedman. 2018. Data statements for natural language processing: Toward mitigating system bias and enabling better science. *Transactions of the Association for Computational Linguistics*, 6:587–604.

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.

Kyunghyun Cho, Bart van Merriënboer, Dzmitry Bahdanau, and Yoshua Bengio. 2014. On the properties of neural machine translation: Encoder–decoder approaches. In *Proceedings of the Eighth Workshop on Syntax, Semantics and Structure in Statistical Translation (SSST-8)*.

Andrew F. Christie. 2002. The ICANN domain-name dispute resolution system as a model for resolving other intellectual property disputes on the internet. *Journal of World Intellectual Property*, 5(1).

Andrew F. Christie. 2014. Online dispute resolution — the phenomenon of the UDRP. In Paul Torremans, editor, *Research Handbook on Cross-Border Enforcement of Intellectual Property*, chapter 16. Edward Elgar, Cheltenham, UK.

Andrew F. Christie. 2020. WIPO and IP Dispute Resolution. In Sam Ricketson, editor, *The Elgar Companion to the World Intellectual Property Organization*, chapter 14. Edward Elgar, Cheltenham, UK.

Shrey Desai and Greg Durrett. 2020. Calibration of pre-trained transformers. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 295–302, Online. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186.

Andrew Higgins, Inbar Levy, and Thibaut Lienart. 2020. The bright but modest potential of algorithms in the courtroom. In Rabeea Assy and Andrew Higgins, editors, *Principles, Procedure, and Justice: Essays in honour of Adrian Zuckerman*, pages 113–132. Oxford University Press.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural Computation*, 9(8):1735–1780.

Daniel Martin Katz, Michael J Bommarito, and Josh Blackman. 2017. A general approach for predicting the behavior of the Supreme Court of the United States. *PloS one*, 12(4):e0174698.

Diederik P. Kingma and Jimmy Ba. 2017. Adam: A method for stochastic optimization. arXiv cs.LG 1412.6980.

Lingkai Kong, Haoming Jiang, Yuchen Zhuang, Jie Lyu, Tuo Zhao, and Chao Zhang. 2020. Calibrated language model fine-tuning for in- and out-of-distribution data. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1326–1340, Online. Association for Computational Linguistics.

Fred Kort. 1957. Predicting supreme court decisions mathematically: A quantitative analysis of the "right to counsel" cases. *American Political Science Review*, 51(1):1–12.

Reed C Lawlor. 1963. What computers can do: Analysis and prediction of judicial decisions. *American Bar Association Journal*, pages 337–344.

Kobi Leins, Jey Han Lau, and Timothy Baldwin. 2020. Give me convenience and give her death: Who should decide what uses of NLP are appropriate, and on what basis? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2908–2913, Online. Association for Computational Linguistics.

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.

Marco Lui and Timothy Baldwin. 2012. langid.py: An off-the-shelf language identification tool. In *Proceedings of the ACL 2012 System Demonstrations*, pages 25–30, Jeju Island, Korea. Association for Computational Linguistics.

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.

Benjamin Strickson and Beatriz De La Iglesia. 2020. Legal judgement prediction for UK courts. In *Proceedings of the 2020 The 3rd International Conference on Information Science and System*, pages 204–209.

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.

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.

Bernhard Waltl, Georg Bonczek, Elena Scepankova, Jörg Landthaler, and Florian Matthes. 2017. Predicting the outcome of appeal decisions in Germany's tax law. In *International Conference on Electronic Participation*, pages 89–99.

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 preprint arXiv:1807.02478*.

Nuo Xu, Pinghui Wang, Long Chen, Li Pan, Xiaoyan Wang, and Junzhou Zhao. 2020. Distinguish confusing law articles for legal judgment prediction. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3086–3095.

Wenmian Yang, Weijia Jia, Xiaojie Zhou, and Yutao Luo. 2019. Legal judgment prediction via multi-perspective bi-feedback network. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI 2019*, pages 4085–4091. ijcai.org.

Haoxi Zhong, Zhipeng Guo, Cunchao Tu, Chaojun Xiao, Zhiyuan Liu, and Maosong Sun. 2018a. 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, Yuzhong Wang, Cunchao Tu, Tianyang Zhang, Zhiyuan Liu, and Maosong Sun. 2020. Iteratively questioning and answering for interpretable legal judgment prediction. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 1250–1257.

Haoxi Zhong, Chaojun Xiao, Zhipeng Guo, Cunchao Tu, Zhiyuan Liu, Maosong Sun, Yansong Feng, Xianpei Han, Zhen Hu, Heng Wang, and Jianfeng Xu. 2018b. Overview of CAIL2018: Legal judgment prediction competition. *ArXiv*, abs/1810.05851.

A Appendix — Data Statement

A.1 Curation Rationale

Goal: This data set was created with the goal of building models to automatically categorize domain name disputes pertaining to generic Top-Level Domains (gTLDs, i.e. domain names ending in .com, .org, .net, etc), in close collaboration with lawyers (who are also co-creators of the data set), including a panelist who has processed a large number of cases through WIPO.

Background: Domain name disputes can be lodged with an ICANN-accredited service provider, such as the World Intellectual Property Organisation (WIPO) for arbitration. The 'complainant' lodges a complaint, and the 'respondent' can in turn file a response. A panel then decides on the case and writes up a justification of their decision. These reports are publicly disclosed, and form the basis of this data set.

A.2 Language Variety

BCP47: en

Background: The data set was derived from WIPO panelists' written reports as part of the arbitration process in settling domain name disputes. These are formal written reports by 1 or 3 panelists and follow a highly structured format (expected set of headings). The collection of reports are written in a variety of languages, however a vast majority (over 90%) are written in English. This collection only includes English reports.

Further Information: The variety of English cannot be further specified because panelists can be from various countries/regions.

A.3 Demographic Information

Age: Unknown

Gender: Unknown

Race/Ethnicity: Unknown

Native Language: Unknown

Socioeconomic Status: All registered or practising lawyers

Number of different speakers represented: Unknown

Further Information: The current list of panelists can be found here: https://www.wipo.int/amc/en/domains/panel/panelists.jsp

A.4 Annotation Information

Age: N/A

Gender: N/A

Race/Ethnicity: N/A

Native Language: N/A

Socioeconomic Status: N/A

Number of different speakers represented: N/A

Further Information: The 'annotation' was derived from the categories published on WIPO's decision page: https://www.wipo.int/amc/en/domains/decisionsx/index.html. These categories were binarized for the purposes of the domain name dispute resolution (DNDR) described in this paper.

A.5 Text Characteristics

Modality: Written text in a report format

A.6 Other Information

Original Format/Encoding: HTML

Method of Collection: The text was scraped from the WIPO site from custom-built tools using bs4 (BeautifulSoup).

A.7 Provenance Appendix

Provenance: All text derived from https://www.wipo.int/amc/en/domains/decisionsx/index.html

Date: Collection dates from January 2000 to August 2020.