AsyLex: A Dataset for Legal Language Processing of Refugee Claims

Claire Barale[1] **Mark Klaisoongnoen**[2] **Pasquale Minervini**[1]
Michael Rovatsos[1] **Nehal Bhuta**[3]
[1]School of Informatics, The University of Edinburgh
[2]EPCC, The University of Edinburgh [3]School of Law, The University of Edinburgh
claire.barale@ed.ac.uk

Abstract

Advancements in natural language processing (NLP) and language models have demonstrated immense potential in the legal domain, enabling automated analysis and comprehension of legal texts. However, developing robust models in Legal NLP is significantly challenged by the scarcity of resources. This paper presents AsyLex, the first dataset specifically designed for Refugee Law applications to address this gap. The dataset introduces 59,112 documents on refugee status determination in Canada from 1996 to 2022, providing researchers and practitioners with essential material for training and evaluating NLP models for legal research and case review. Case review is defined as entity extraction and outcome prediction tasks. The dataset includes 19,115 gold-standard human-labeled annotations for 20 legally relevant entity types curated with the help of legal experts and 1,682 gold-standard labeled documents for the case outcome. Furthermore, we supply the corresponding trained entity extraction models and the resulting labeled entities generated through the inference process on AsyLex. Four supplementary features are obtained through rule-based extraction. We demonstrate the usefulness of our dataset on the legal judgment prediction task to predict the binary outcome and test a set of baselines using the text of the documents and our annotations. We observe that models pretrained on similar legal documents reach better scores, suggesting that acquiring more datasets for specialized domains such as law is crucial. The dataset is available at https://huggingface.co/datasets/clairebarale/AsyLex.

1 Introduction

While large language models (LLMs) have gained significant attention in NLP, many real-world applications have yet to leverage their capabilities fully. One of the main challenges lies in the fact that these models heavily rely on training or fine-tuning models with specific datasets to effectively transfer their capabilities to real-world and specialized applications. However, collecting such datasets is often time-consuming and expensive, primarily due to the need for human annotation. This challenge becomes even more pronounced in specialized domains where human annotators require domain expertise. Collecting thousands of annotations is a barrier to developing advanced NLP tools that support researchers and practitioners in the legal domain.

	Main text	Case cover	Case outcome
Documents	59,112	45,882	32,627
Sentences	4,946,438	-	53,977
Paragraphs	1,781,240	-	-
Labels	16	8	3
Labeled (human)	16,628	2,487	1,682
Labeled (rule-based)	-	57,408	-
Labeled (inferred)	6,154,226	123,802	30,944

Table 1: Overview of AsyLex

Refugee Law is a specific area of law that currently lacks publicly available datasets and benchmarks to evaluate NLP applications and compare their performance with standard baselines. Lawyers specializing in Refugee Law devote a significant amount of time to reviewing past cases to prepare for new ones. This process of legal research and case review is not only time-consuming but also expensive. Given the limited resources and funding available in Refugee Law, where a majority of claimants rely on legal aid, it becomes even more critical to find ways to free up time and speed up the case review process.

The availability of a comprehensive dataset in Refugee Law would have significant benefits. It would enable lawyers to efficiently search and analyze relevant cases, extract valuable insights, and apply them to new cases. By streamlining the case review process, lawyers could save time, reduce

Proceedings of the Natural Legal Language Processing Workshop 2023, pages 244–257

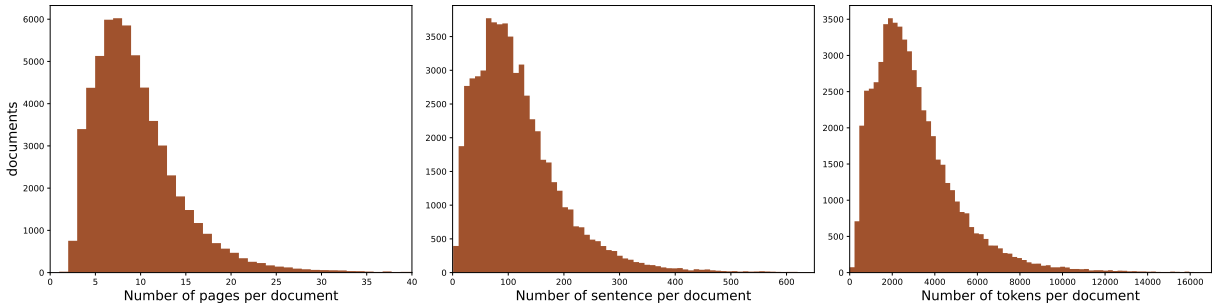

Figure 1: Distribution of text input length, measured in pages, sentences, and tokens

costs, and ultimately provide more efficient legal assistance to those seeking refugee protection. The absence of a Refugee Law dataset affects lawyers and claimants, as Refugee Status Determination (RSD) plays a pivotal role in determining refugee protection. Refugee protection decisions directly impact the lives of the approximately 4.6 million asylum seekers worldwide as of mid-2022[1]. To provide a concrete example, 48,014 new claims and 10,055 appeals were filed[2] in Canada alone in 2021. Processing times of refugee claims range from a few months to several years.

This variation in processing times exacerbates the uncertainty and instability experienced by asylum seekers. By enhancing access to data, we aim to promote fairness, effectiveness, and improved outcomes for individuals seeking refuge globally. Automating aspects of case review and providing high-quality data can facilitate better access to legal counseling and support for claimants.

Our dataset is particularly valuable because we collected a total of 76,523 gold annotations (both from human labeling and rule-based) and trained specific information extraction models to gain insights and structure the dataset (section 3) to facilitate two tasks: (1) entity extraction for legal search and (2) legal judgment prediction. Both tasks, according to lawyers we consulted, have the potential to help draft new claims for refugee protection. We experiment with judgment prediction on different baseline masked language models to showcase the dataset's potential (section 4).

AsyLex contributes directly to applying NLP in high-stakes legal domains and as a novel research asset to the NLP research community. Specifically, our contributions are as follows:

1. We provide the anonymized raw text of the decision documents by case, by paragraphs, and by sentences;

2. We implement a state-of-the-art methodology for annotating our dataset, with a primary emphasis on speed and replicability for other datasets;

3. We provide documents labeled with their decision outcome for the task of legal judgment prediction;

4. We provide the set of extracted sentences that directly indicates the determination of each case with gold-standard annotations on the outcome;

5. We provide gold-labeled data for relevant legal entity types for entity extraction; and

6. We use state-of-the-art Transformer-based models for text classification to accurately predict the outcome of cases and evaluate the presented data.

In addition to the data, the code of the experiments is public and can be found at: https://github.com/clairebarale/AsyLex.

2 Related Work

Legal NLP Legal NLP is an active and promising field of research. Legal information is predominantly conveyed through text, with crucial details typically documented in written form. In principle, this makes the law an ideal domain for leveraging natural language processing techniques. However, applying NLP in the legal domain brings about notable challenges. These challenges arise from the distinct structure, specialized vocabulary, contextual nuances, importance of legal citations, and high stakes in legal applications. A wide range of tasks and functionalities have been explored in

[1]United Nations High Commissioner for Refugees report: https://www.unhcr.org/global-trends-report-2022

[2]https://irb.gc.ca/en/statistics/Pages/index.aspx

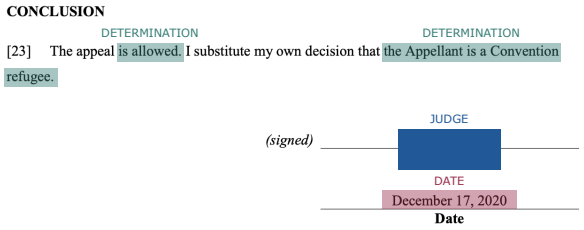

Figure 2: Example of the end of a case document containing the determination sentence, the name of the judge, and the date of the decision

legal NLP (Zhong et al., 2018), such as legal summarization (Elaraby and Litman, 2022; Aumiller et al., 2022), legal information extraction and retrieval (Bommarito II et al., 2021; Brüninghaus and Ashley, 2001; Chalkidis et al., 2021), legal question answering, automatic text generation, text classification (Chalkidis et al., 2022b, 2019c), legal judgment prediction (Katz et al., 2017; Branting et al., 2019, 2018; Chalkidis et al., 2019a) or multi-task benchmarks (Chalkidis et al., 2022a).

With this dataset, we target two tasks for **Case Review**: **(i) Entity Extraction** and **(ii) Legal Judgment Prediction**. Legal judgment prediction is a common task in legal AI. Work has been conducted on legal datasets gathering decisions from the Supreme Court of the United States (Katz et al., 2017; Martin et al., 2004; Ruger et al., 2004a; Undavia et al., 2018) or the European Court of human rights (Medvedeva et al., 2020; Kaur and Bozic, 2019). Similar studies have been conducted on asylum decision data sets (Chen and Eagel, 2017; Dunn et al., 2017; Rehaag, 2012). However, most of these works rely on statistical data rather than text documents. One reason is the lack of specialized language resources and datasets available in law. We recognize this gap and aim to address it by providing a new dataset of legal texts, thereby contributing to legal NLP research. Legal information extraction and classification are difficult tasks because approaches are mostly supervised and, therefore, require precisely labeled datasets.

Legal NLP Datasets While large language models have recently made natural language tasks more accessible, their utilization in the legal domain has been relatively limited. One of the reasons behind this is the need to adapt these models to legal applications (Gururangan et al., 2020). These are typically trained on general corpora and may not fully capture the intricacies of legal text, as shown by previous studies (Barale et al., 2023). In our information extraction pipeline, we observed that LegalBERT outperformed RoBERTa in accuracy, highlighting the significance of fine-tuning models specifically for legal tasks.

A range of datasets for legal NLP tasks in English exists in Tax Law (Holzenberger et al., 2020), European Legislation and the European Court of Human Rights (Chalkidis et al., 2019b,a), Corporate and Contract Law (Hendrycks et al., 2021; Tuggener et al., 2020), Supreme Court cases and US court cases (Ruger et al., 2004b; Zheng et al., 2021). LeXFiles is the most comprehensive to date (Chalkidis et al., 2023). However, no such dataset is available for refugee decisions. To further investigate how well large language models perform on specialized tasks and to ensure advances also benefit the field of Refugee Law, we curated the first-ever dataset of refugee status determination cases in English. This dataset will serve as a valuable resource to train and evaluate models tailored to the unique challenges and requirements of Refugee Law.

3 AsyLex: a Case Review Dataset

Case Review Our dataset focuses on facilitating the legal application of case review. Case review includes similar past case retrieval and case analysis and is an essential part of legal research. Lawyers face the challenge of dealing with an overwhelming number of cases under significant time constraints. Additionally, they require valuable insights to guide them in identifying critical elements for new application drafting. The anticipated benefit of our research is to expedite the legal case review process while offering supplementary insights to enhance efficiency.

The dataset contains labeled data suited for two NLP tasks: **(1) Entity extraction** and **(2) Legal Judgment Prediction**. For the first task of interest, we offer a total of 19,115 human-labeled annotated samples, data spanning across 22 categories, an additional 57,408 annotations using rule-based extraction on the *case cover*, along with 6,278,028 silver-standard labeled samples inferred through a named-entity recognition model presented in details in Barale et al. (2023). For the *case cover*, which constitutes the first page of each case and contains meta-data about each document, we annotated 346 documents. Details of the number of labeled samples per category are shown in Table 3 for the *case cover* and in Table 4 for the full body

	Case #123259 – *Initial* Case Cover *shown in Annex*	**Case #100049**
extracted_dates	'april 04 , 2019', '2020', 'december 17, 2020', 'december 15, 2020'	'august 14, 2013', 'august 15, 2013', '2013'
loc_hearing	['toronto', 'on']	['ottawa', 'ontario', 'montreal', 'quebec']
tribunal	['refugee appeal division']	['immigration appeal division']
public_private_hearing	-	-
in_chamber_virtual	-	['videoconference']
judge	-	dana kean
date_decision	-	-
text_case_cover	immigration and refugee board...	immigration and refugee board...

Table 2: Example of two case documents' *Case Cover* and the additional structured information provided in AsyLex

Label	**gold**	**inferred**
extracted_dates	1,219	45,884
loc_hearing	871	43,715
tribunal	278	34,203
public_private_hearing	307	-
in_chamber_virtual	7,224	-
judge_name	18,691	-
date_decision	31,186	-
person	119	-

Table 3: Entities Annotations on the *Case Cover*

Label	**gold**	**inferred**
CLAIMANT_EVENT	3,730	-
CLAIMANT_INFO	687	209,623
GPE	1,545	1,027,918
NORP	206	206,927
ORG	1,041	748,612
PROCEDURE	1,788	617,659
CREDIBILITY	1,020	464,504
DETERMINATION	342	116,489
DOC_EVIDENCE	1,878	861,357
EXPLANATION	1,274	362,973
DATE	1,474	975,625
LAW	874	317,277
LAW_CASE	211	125,497
LAW_REPORT	47	30,892
LEGAL_GROUND	158	88,873
PERSON	353	-

Table 4: Entities Annotations on the *Main Text*

text of the documents.

For the second task of interest, we provide a comprehensive collection of data for legal judgment prediction, which comprises a test set with 1,682 gold-standard classified documents and a train set with 30,944 classified documents. In addition, we provide the raw text of the document cases in full text and split it into sentences, each presented with their corresponding case number (`decisionID`).

3.1 Dataset Assembly Pipeline

Dataset Source We retrieved 59,112 historic decision documents (dated 1996 to 2022, as shown in Figure 4) from the online services of the Canadian Legal Information Institute (CanLII) to curate a collection of federal refugee cases. The documents are initially collected both in PDF and HTML and are all available online. Our automated retrieval process is more efficient and error-resistant than manual retrieval, covers all available cases to date, and is superior to human-based manual retrieval in terms of error proneness and processing time. We obtain two sets: (1) a set of *Case Covers* that consists of semi-structured data and displays meta-information (Appendix A) and (2) a *Main Text* set that contains the body of each case, in full text.

Anonymization When retrieved, the data had already undergone a partial anonymization process to protect sensitive information such as names and locations. However, certain cases still contained identifiable details. To reinforce privacy protection, we took additional steps to anonymize all documents using *Microsoft Presidio* tool[3] and a sequence-labeling task. This involved utilizing our trained NER model to identify personal names (labeled as `PERSON`) and replacing them with the placeholder `X`. Our fine-tuned RoBERTa transformer achieved an F1 score of 85.71% on this label.

Choice of Categories The labels have been defined and decided upon with the help of experienced refugee lawyers. We chose labels that represent characteristics reflective of similarity among different cases to facilitate future legal searches. While each case exhibits individual characteristics, legal practitioners typically search for similarities based on elements such as the constitution of the panel, the country of origin and the characteristics of the claimant, the year the claim was made in relation to a particular geopolitical situation, the legal procedures involved, the grounds for the decision, relevant legislation, as well as other cases or reports that are cited.

[3] https://github.com/microsoft/presidio

Flagged Phrase	# of Occurrences
male	5493
canadian citizen	4508
female appellant	4126
woman	2844
roma	2758
female claimant	2329
male appellant	2209
male claimant	1912
female	1688
minor appellant	1655
minor appellants	1257
community	1227
canadian citizens	1018
citizen of china	917
homosexual	833
women	827
lesbian	811
associate appellant	782
citizen of haiti	709
man	677

Table 5: The 20 most frequently flagged phrases with the label `CLAIMANT_INFO`, using LegalBERT fine-tuned on our dataset (best run evaluated with F1 score is reported here).

Collecting gold-standard annotations Recognizing the importance of collecting annotations of high quality in the domain of Refugee Law, we first collect manually labeled annotations, as explained in Barale et al. (2023). After consulting with refugee lawyers to determine appropriate annotation guidelines, we annotated the data ourselves. We chose to annotate the text split by sentences, as we have found that annotating by paragraph takes more time for the annotator.

For the **human annotation** task, we used state-of-the-art semi-automatic annotation tools: the `Prodigy` annotation tool[4] was used under an academic research license in order to speed up and improve the manual labeling work in terms of consistency and accuracy of annotations. To collect annotated samples on traditional NER labels (`DATE`, `ORG`, `GPE`, `PERSON`, `NORP`, `LAW`), we use suggestions from general purpose pretrained embeddings. For the rest of the labels (`CLAIMANT_INFO`, `CLAIMANT_EVENT`, `PROCEDURE`, `DOC_EVIDENCE`, `EXPLANATION`, `DETERMINATION`, `CREDIBILITY`) and with the aim of improving consistency of annotation, we created a terminology base with the help of lawyers based on `word2vec` (Mikolov et al., 2013). At annotation time, patterns are matched with sentences considered for annotation; the human annotator only corrects them, creating a gold-annotated set of sentences and considerably speeding up the labeling task.

Second, recognizing the different kinds of target information, we also performed a complementary **rule-based extraction and labeling** for the following categories of interest: the name of the judge, the date of the decision, and whether the hearing was in-person, virtual, private, or public. This mostly consisted of searching the text for keywords that we had predefined in advance in our terminology base. For `judge_name` and `date_decision` in the *Case Cover*, we rely on cross-checking information between the names and dates extracted from the first page and the name and date present at the very end of each case in a format presented in Figure 2. Similarly, we determine the outcome of 1,682 cases by keyword search on keywords that are present at the end of a case in some documents to indicate the final decision, *positive* of *negative*. We then manually review the assigned label and this allows us to determine the final outcome with a high degree of certainty and low ambiguity.

Generating Silver-standard Annotated Data To gather more labeled cases, we used a transformer-based text classification model to generate silver-standard labels that indicate the outcome of each case. We first used our NER model to extract the sentences pertaining to the decision outcomes, which were flagged with the label `DETERMINATION`. The extracted sentences contain on average 5.01 tokens. We then trained a classifier on gold-labeled sentences with positive or negative outcomes which for future research are destined to be used as a test set. Because of the imbalance of the available data we oversample our set of gold-standard labels on the extracted sentences labeled *granted*. Since there may be multiple extracted sentences per case, we employed a majority vote mechanism to determine the outcome of each case. Sentences that could not be confidently classified as positive or negative (i.e. those which returned a weight between 0.4 and 0.6) were categorized as *Uncertain* and account for 19.09% of the analyzed sentences. The results using different pretraining are presented in Table 6 with a best-achieved accuracy of 99.34%. We further comment on those results in section 4. In total, we classify 52,234 sentences into three categories: granted (1), rejected (0), and uncertain (2). Thereby we obtain the deci-

[4] `Prodigy`: https://prodi.gy/docs

DETERMINATION Extracted Sentences							
	BERT	**RoBERTa**	**DeBERTa**	**LegalBERT**	**CaseHOLD**	**PoL**	**LexLM**
Accuracy	**99.34**	99.15	98.94	98.73	99.15	98.52	88.98
Macro F1	**98.82**	98.43	98.02	97.67	98.43	97.30	82.59
Weighted F1	**99.36**	99.15	98.94	98.73	99.15	98.53	89.74

Table 6: Results of the sentence classification task, on the best-achieved run, after oversampling (Appendix C)

Label	**Precision**	**Recall**	**F1**
CLAIMANT_EVENT	66.29	64.80	65.54
CLAIMANT_INFO	88.64	88.64	88.64
CREDIBILITY	80.43	77.08	78.72
DATE	85.19	93.88	89.32
DETERMINATION	72.41	67.74	70.00
DOC_EVIDENCE	82.95	86.90	84.88
EXPLANATION	81.43	61.29	69.94
GPE	96.24	98.90	97.55
LAW	51.61	60.38	55.65
LAW_CASE	62.50	71.43	66.67
NORP	90.00	85.71	87.80
ORG	89.00	90.82	89.90
PROCEDURE	69.51	65.52	67.46
LEGAL_GROUND	60.00	42.86	50.00

Table 7: Evaluation metrics for the LegalBERT-based fine-tuned entity extraction model

sion outcome for a total of 30,944 case documents (to use as the training split) of which 21.68% are positive decisions as detailed in Table 8.

Similarly, we infer from our previously trained entity extraction model, on the remaining sentences extracted from each document. This allows us to collect silver-standard annotations for the whole dataset of the *main text*. For inference, we use our best-performing trained NER model, which relies on LegalBERT and for which the achieved performance numbers on each concerned category can be found in Table 7 (Chalkidis et al., 2020; Barale et al., 2023).

Outcome \ **Split**	**Test** (Gold)	**Train** (Inferred)
Granted	311	6,709
Rejected	1,371	18,264
Uncertain	-	5,971
Total	**1,682**	**30,944**

Table 8: Outcome distribution of the cases

3.2 Dataset Description and Statistics

From the *Case Covers* entity extraction, we are able to detect the tribunal in 23,022 documents. 92.19% of the decision in this sample are appeal decisions rendered by the Refugee Appeal Division, and the rest (7.80%) is rendered by the Refugee Protection Division in the first instance (flagged with `tribunal`). In terms of location of the hearings (`loc_hearing`), the most commonly found is Toronto with 55.31% of the documents analyzed, second is Montreal with 21.54%, followed by Vancouver with 13.57%, Ottawa with 5.09%, Calgary with 2.57%, Edmonton 1.18%, Winnipeg 0.48%, Halifax 0.25% (figures on 28,127 documents for which the hearing location was detected). We find that 981 documents indicate that the hearing was held by video conference (`in_chamber_virtual`, a field that was mostly added since 2020 during the covid-19 pandemic), the rest defaults to in-chambers hearing. We flagged only 3 private hearings on the whole dataset (`public_private_hearing`).

From the *main text*, we choose to detail two of the extracted categories of entities: `GPE` and `CLAIMANT_INFO` which are respectively described in Figure 3 and Table 5. We believe the latter is a valuable foundation for further document classification and analysis of potential biases across cases. The former category extracts mentions of locations and in particular countries. This can refer to country of origin or country of transit. It is also to note that there can be multiple occurrences of one or several countries per case.

4 Experiments: Text Classification for Outcome Prediction

This task serves multiple purposes. Firstly, it is of interest to lawyers and the claimant as they seek to measure the probability of winning the case. Additionally, they aim to understand the grounds on which cases can be won and identify the key factors that require attention to enhance the chances

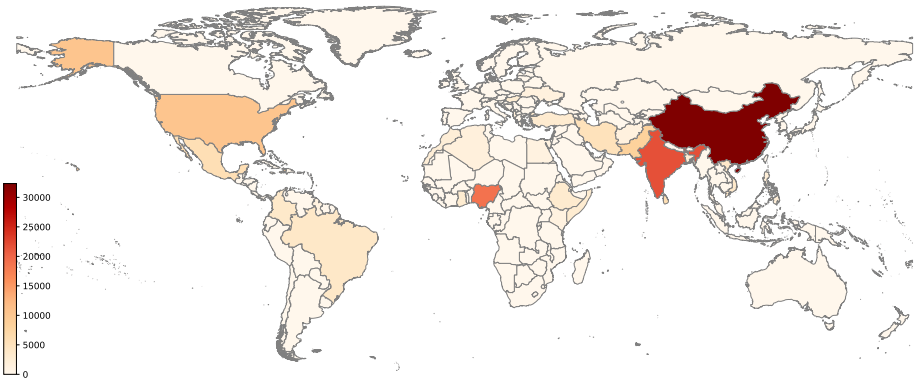

Figure 3: Occurrences of countries on the whole dataset text

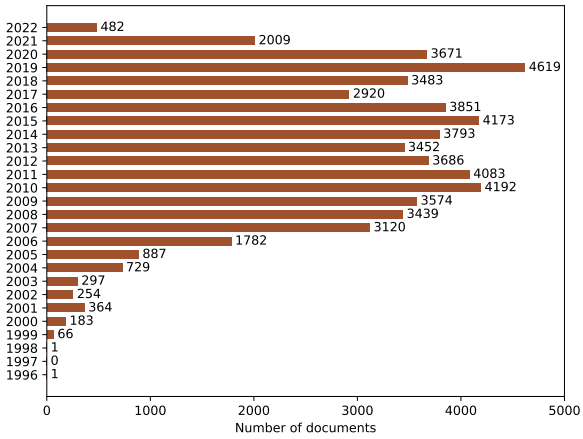

Figure 4: Distribution of the documents over the years by the date of the decision

of success. For judges, engaging in this task offers the potential to analyze their decisions, leading to potential improvements and increased consistency. Furthermore, researchers can utilize this task to assess the fairness of Refugee Status Determination (RSD) decisions and uncover potential biases.

Task: Judgment Prediction The task consists of predicting whether the decision outcome was positive (refugee status granted) or negative (the claim was rejected).

Data The task is performed on two sets of data: the entities extracted from the *case cover* and on the sentences flagged as `determination` by our entity extraction model as described in section 3.1. After cleaning the dataset of *case covers*, we obtained a total of 25,232 documents, 6,985 of which are cases with a positive outcome, i.e. 27.68% of our dataset of *case covers*. After splitting 80-20%, we obtain a train set of 20,186 documents and a test set of 5,046 documents (Appendix D).

We concatenate the entities to form one string per document, separating each type of entity with a separation token. An input string example: *"2010 april 28 2010[SEP]toronto ontario[SEP]in chambers[SEP]april 28 2010"*.

Metrics and Evaluation We evaluate the accuracy, macro F1, and weighted F1 to take into account the initial imbalance in the classes of our dataset, where a majority of cases have a negative decision outcome. The macro average F1 score treats the two classes equally regardless of the number of examples in the set, while the weighted F1 score considers the number of examples in each class, accounting for the class distribution. We present our evaluation on a test set of unseen sequences. For the *Case Cover* entities classifier, we compare to a majority baseline, a trivial baseline that predicts the majority class ("rejected") for all the samples.

Baseline and Models Used We fine-tune our binary classifier on different baseline models with a language modeling objective, both pretrained on general-purpose data (BERT, RoBERTa, and DeBERTa-V3 (Devlin et al., 2019; Liu et al., 2019; He et al., 2023)) and on legal documents (Legal-BERT, CaseHOLD (Chalkidis et al., 2020; Zheng et al., 2021), Pile of Law (Henderson et al., 2022) and LexLM (Chalkidis et al., 2023)). We exclusively employ BERT-based architectures for several reasons. Firstly, BERT's masked language modeling objective aligns well with our language understanding goals. Secondly, BERT's fine-tuning capabilities allow us to fully leverage our dataset, making it particularly suitable for our benchmark task. Additionally, we utilize BERT as a baseline due to its proven robustness. Prior research has

Case Cover Entities								
	Majority	**BERT**	**RoBERTa**	**DeBERTa**	**LegalBERT**	**CaseHOLD**	**PoL**	**LexLM**
Accuracy	72.36	72.49	74.45	**74.79**	74.57	74.61	74.36	73.96
Macro F1	42.00	66.20	64.26	66.02	65.97	**67.78**	67.53	65.79
Weighted F1	60.77	72.80	72.77	73.91	73.82	**74.39**	74.18	73.26

Table 9: Results of the legal judgment prediction task on the best-achieved run. *Majority* is the majority baseline

demonstrated that BERT-based architectures excel in contextual understanding compared to autoregressive language models. We leave a detailed comparison to future work.

RoBERTa and DeBERTav3 are built upon a larger corpus (160GB) compared to BERT's 16 GB corpus. Pile of Law's corpus is the largest with 256GB, followed by LexLM with 175GB. DeBERTav3 holds fewer parameters (86M) than RoBERTa and improves upon it by using a disentangled attention mechanism. CaseHOLD consists of 37GB of American case law while LegalBERT has the smallest corpus with 12 GB. LegalBERT (US and EU), Pile of Law (US, Canada, EU) CaseHOLD and LexLM (US, Canada, EU, UK, India) are the corpora that include human rights and asylum texts which is the closest domain match to AsyLex.

Experimental Set Up We fine-tune the pretrained language models on AsyLex. We perform a hyperparameter search using the Optuna library[5]. The details of the hyperparameter used for each implementation can be found in Appendix E.

Experimental Results The results of the two experiments conducted on the task of Legal Judgment Prediction are presented in Table 6 and 9. They demonstrate the improved performance of models pretrained on legal documents and relevant legal domains, compared to general-purpose models.

Firstly, one must note that the task of determination sentences judgment prediction is much easier than the prediction task on the dataset of *case covers entities*. On the determination sentences classification task, BERT reaches the best level of accuracy with 99.34%, the second best being achieved by the RoBERTa and CaseHOLD models. On the concatenated entities classification, DeBERTa achieves the best accuracy, very closely with Pile of Law, LegalBERT, and CaseHOLD. As the classes are imbalanced, it is worth noting that the best macro F1 score is in fact achieved by CaseHOLD with 67.78%. Similarly, in the first experiment, BERT achieved the best macro F1 score with 99.82%. All models exhibit closely comparable scores (within one percent), suggesting that the difference in performance between them is not significant. LexLM is the only model showing a significantly lower score. One conjecture is that the English legal pretraining corpus's magnitude exacerbates the model's difficulty in retrieving legal information.

As expected, the classifier directly fined-tune on determination sentences achieves superior performance across all evaluation metrics. However, the classifier trained on entities extracted from the case also demonstrates promising results. This highlights the possibilities for leveraging extracted entities in subsequent studies.

These findings collectively demonstrate the significant advantages of utilizing pretrained models specifically designed for legal documents, with a close match in the legal domain, in the context of Judgment Prediction and therefore the importance of gathering data for legal NLP. The superiority of CaseHOLD and Pile of Law showcases the potential of domain-specific language models to enhance the performance of legal NLP tasks.

5 Conclusion

We introduce a high-quality dataset of annotated refugee claims to streamline case review procedures and contribute to the overall effectiveness of legal practitioners' workflows. Furthermore, we explore the performance of generic models in specialized domains, particularly in the context of Refugee Law. This dataset aims to not only enable advancements in legal NLP but also specifically address the scarcity of resources in Refugee Law. By providing researchers and practitioners access to high-quality labeled data, we hope to foster further research and innovation in this crucial area, ultimately facilitating the development of intelligent legal systems with applications in Refugee Law and beyond.

[5] https://github.com/optuna/optuna

Limitations

In this section, we enumerate the limitations and shortcomings of our work:

- The need to train Transformer-based architectures to perform inference on a large dataset is a limitation as the process typically requires parallelization efforts across GPUs and CPUs to finish within acceptable time frames of multiple hours to several days.

- The manual annotation process for entity extraction and text classification is a weakness: while it results in gold-standard annotations, it is very time-consuming. One of the proposed solutions of this work is to use human labeling in combination with rule-based extraction. However, in future work, it would be interesting to look at methods of indirect supervision and automated annotation generation.

- The dataset contains text in English only and is therefore destined for further research on English-speaking jurisdictions only.

Ethics Statement

Firstly, the proposed dataset contains sensitive personal data. The documents were already available online with partial anonymization. We perform further anonymization of all cases. Second, the task of outcome prediction presented in this paper is not intended to be used for automating the final decision. It is presented in this paper to provide a baseline for future work, with the goal of further analyzing the decision-making process, possible biases, and inconsistencies in the features leading to a decision outcome.

AsyLex is intented to be used for research purposes only.

References

Dennis Aumiller, Ashish Chouhan, and Michael Gertz. 2022. EUR-lex-sum: A multi- and cross-lingual dataset for long-form summarization in the legal domain. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 7626–7639, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Claire Barale, Michael Rovatsos, and Nehal Bhuta. 2023. Automated refugee case analysis: A NLP pipeline for supporting legal practitioners. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 2992–3005, Toronto, Canada. Association for Computational Linguistics.

Michael J Bommarito II, Daniel Martin Katz, and Eric M Detterman. 2021. Lexnlp: Natural language processing and information extraction for legal and regulatory texts. In *Research Handbook on Big Data Law*, pages 216–227. Edward Elgar Publishing.

K. Branting, B. Weiss, B. Brown, C. Pfeifer, A. Chakraborty, L. Ferro, M. Pfaff, and A. Yeh. 2019. Semi-Supervised Methods for Explainable Legal Prediction. In *Proceedings of the Seventeenth International Conference on Artificial Intelligence and Law*, pages 22–31, Montreal QC Canada. ACM.

L Karl Branting, Alexander Yeh, Brandy Weiss, Elizabeth Merkhofer, and Bradford Brown. 2018. Inducing predictive models for decision support in administrative adjudication. In *AI Approaches to the Complexity of Legal Systems: AICOL International Workshops 2015-2017: AICOL-VI@ JURIX 2015, AICOL-VII@ EKAW 2016, AICOL-VIII@ JURIX 2016, AICOL-IX@ ICAIL 2017, and AICOL-X@ JURIX 2017, Revised Selected Papers 6*, pages 465–477. Springer.

Stefanie Brüninghaus and Kevin D Ashley. 2001. Improving the representation of legal case texts with information extraction methods. In *Proceedings of the 8th international conference on Artificial Intelligence and Law*, pages 42–51.

Ilias Chalkidis, Ion Androutsopoulos, and Nikolaos Aletras. 2019a. 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, Emmanouil Fergadiotis, Prodromos Malakasiotis, Nikolaos Aletras, and Ion Androutsopoulos. 2019b. Extreme multi-label legal text classification: A case study in EU legislation. In *Proceedings of the Natural Legal Language Processing Workshop 2019*, pages 78–87, Minneapolis, Minnesota. Association for Computational Linguistics.

Ilias Chalkidis, Emmanouil Fergadiotis, Prodromos Malakasiotis, and Ion Androutsopoulos. 2019c. Large-scale multi-label text classification on EU legislation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6314–6322, 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. 2021. Paragraph-level rationale extraction through regularization: A case study on European court of human rights cases. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 226–241, Online. Association for Computational Linguistics.

Ilias Chalkidis, Nicolas Garneau, Catalina Goanta, Daniel Katz, and Anders Søgaard. 2023. LeXFiles and LegalLAMA: Facilitating English multinational legal language model development. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15513–15535, Toronto, Canada. Association for Computational Linguistics.

Ilias Chalkidis, Abhik Jana, Dirk Hartung, Michael Bommarito, Ion Androutsopoulos, Daniel Katz, and Nikolaos Aletras. 2022a. LexGLUE: A benchmark dataset for legal language understanding in English. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4310–4330, Dublin, Ireland. Association for Computational Linguistics.

Ilias Chalkidis, Tommaso Pasini, Sheng Zhang, Letizia Tomada, Sebastian Schwemer, and Anders Søgaard. 2022b. FairLex: A multilingual benchmark for evaluating fairness in legal text processing. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4389–4406, Dublin, Ireland. Association for Computational Linguistics.

Daniel L. Chen and Jess Eagel. 2017. Can machine learning help predict the outcome of asylum adjudications? In *Proceedings of the 16th edition of the International Conference on Articial Intelligence and Law*, pages 237–240, London United Kingdom. ACM.

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, Minneapolis, Minnesota. Association for Computational Linguistics.

Matt Dunn, Levent Sagun, Hale Şirin, and Daniel Chen. 2017. Early predictability of asylum court decisions. In *Proceedings of the 16th edition of the International Conference on Articial Intelligence and Law*, pages 233–236, London United Kingdom. ACM.

Mohamed Elaraby and Diane Litman. 2022. ArgLegalSumm: Improving abstractive summarization of legal documents with argument mining. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 6187–6194, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A. Smith. 2020. Don't stop pretraining: Adapt language models to domains and tasks. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8342–8360, Online. Association for Computational Linguistics.

Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2023. DeBERTav3: Improving deBERTa using ELECTRA-style pre-training with gradient-disentangled embedding sharing. In *The Eleventh International Conference on Learning Representations*.

Peter Henderson, Mark S. Krass, Lucia Zheng, Neel Guha, Christopher D. Manning, Dan Jurafsky, and Daniel E. Ho. 2022. Pile of law: Learning responsible data filtering from the law and a 256gb open-source legal dataset.

Dan Hendrycks, Collin Burns, Anya Chen, and Spencer Ball. 2021. Cuad: An expert-annotated nlp dataset for legal contract review. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, volume 1. Curran.

Nils Holzenberger, Andrew Blair-Stanek, and Benjamin Van Durme. 2020. A dataset for statutory reasoning in tax law entailment and question answering. In *Proceedings of the Natural Legal Language Processing Workshop 2020 co-located with the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (KDD 2020), Virtual Workshop, August 24, 2020*, volume 2645 of *CEUR Workshop Proceedings*, pages 31–38. CEUR-WS.org.

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.

Arshdeep Kaur and Bojan Bozic. 2019. Convolutional neural network-based automatic prediction of judgments of the european court of human rights. In *AICS*, pages 458–469.

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. *CoRR*, abs/1907.11692.

Andrew D Martin, Kevin M Quinn, Theodore W Ruger, and Pauline T Kim. 2004. Competing approaches to predicting supreme court decision making. *Perspectives on Politics*, 2(4):761–767.

Masha Medvedeva, Michel Vols, and Martijn Wieling. 2020. Using machine learning to predict decisions of the european court of human rights. *Artificial Intelligence and Law*, 28(2):237–266.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Sean Rehaag. 2012. Judicial review of refugee determinations: The luck of the draw. *Queen's LJ*, 38:1.

Theodore W. Ruger, Pauline T. Kim, Andrew D. Martin, and Kevin M. Quinn. 2004a. The Supreme Court Forecasting Project: Legal and Political Science Approaches to Predicting Supreme Court Decisionmaking. *Columbia Law Review*, 104(4):1150.

Theodore W Ruger, Pauline T Kim, Andrew D Martin, and Kevin M Quinn. 2004b. The supreme court forecasting project: Legal and political science approaches to predicting supreme court decisionmaking. *Columbia law review*, pages 1150–1210.

Don Tuggener, Pius von Däniken, Thomas Peetz, and Mark Cieliebak. 2020. LEDGAR: A large-scale multi-label corpus for text classification of legal provisions in contracts. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 1235–1241, Marseille, France. European Language Resources Association.

Samir Undavia, Adam Meyers, and John Ortega. 2018. A Comparative Study of Classifying Legal Documents with Neural Networks. pages 515–522.

Lucia Zheng, Neel Guha, Brandon R. Anderson, Peter Henderson, and Daniel E. Ho. 2021. When does pretraining help? assessing self-supervised learning for law and the casehold dataset. In *Proceedings of the 18th International Conference on Artificial Intelligence and Law*. Association for Computing Machinery.

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.

A Appendix A: Example of a *Case Cover* (Case #123259)

Immigration and Refugee Board of Canada

Refugee Appeal Division

ORG

Commission de l'immigration et du statut de réfugié du Canada

Section d'appel des réfugiés

RAD File / Dossier de la SAR : TB8-17005

Private Proceeding / Huis clos

Private/Public

Reasons and decision – Motifs et décision

Person who is the subject of the appeal	PERSON XXXX XXXX	**Personne en cause**
Date(s) of hearing	DATE December 15, 2020	**Date(s) de l'audience**
Appeal considered / heard at	LOCATION Toronto, ON	**Appel instruit / entendu à**
Date of decision	DATE December 17, 2020	**Date de la décision**
Panel	PERSON [redacted]	**Tribunal**
Counsel for the person who is the subject of the appeal	[redacted]	**Conseil de la personne en cause**
Designated representative	N/A	**Représentant(e) désigné(e)**
Counsel for the Minister	N/A	**Conseil du ministre**

RAD.25.02 (April 04, 2019)
Disponible en français

Canada

B Appendix B: Legal Entity Types Description

Type	Description	Examples
LOCATION	cities, countries, regions	"toronto, ontario"
DATE	absolute or relative dates or periods	"june, 4th 1996", "two years"
NORP	adjectives of nationalities, religious, political or ethnic groups or communities	"hutu", "nigerian", "christian"
ORG	tribunals, companies, NGOs	"immigration appeal division", "human rights watch"
CREDIBILITY	mentions of credibility	"lack of evidence", "inconsistencies"
DETERMINATION	outcome of the decision (accept/reject)	"appeal is dismissed", "not a convention refugee"
CLAIMANT_INFO	age, gender, citizenship, occupation	"28 year old", "citizen of Iran", "female"
PROCEDURE	steps in the claim and legal procedure events	"removal order", "sponsorship for application"
DOC_EVIDENCE	pieces of evidence, proofs, supporting documents	"passport", "medical record", "marriage certificate"
EXPLANATION	reasons given by the panel for the determination	"fear of persecution", "no protection by the state"
LEGAL_GROUND	referring to the Convention, refugee status is granted for reasons of race, religion, nationality, membership of a particular social group or political opinion	"homosexual", "christian"
LAW	citations: legislation and international conventions	"section 1(a) of the convention"
LAW_CASE	citations: case law and past decided cases	"xxx v. minister of canada, 1994"
LAW_REPORT	country reports written by NGOs or the United Nations	"amnesty international: police and military torture of women in mexico, 2016"

C Appendix C: Determination Sentences Dataset Split

Because of the dataset imbalance, we chose to oversample the minority class for training. The split is shown in the table below.

Outcome \ Split	Train	Test
Granted	1,064	76
Rejected	1,685	396
Total	**2,749**	**472**

Table 10: Split for determination sentences labeling in number of sentences after oversampling the minority class (granted), used for training – batch size=16, Adam optimizer, 1 epoch

D Appendix D: *Case Cover* Entities Dataset Split

Outcome \ Split	Train	Test
Granted	5,590	1,395
Rejected	14,596	3,651
Total	**20,186**	**5,046**

Table 11: Split for the *Case Cover* Dataset with Outcomes of the Judgments as Labels (Binary Classification Task).

E Appendix E: Hyperparameters used for the Task of Legal Judgment Prediction

We perform experiments on a single GPU and single node/multi-GPU settings. We use a cluster including 38 GPU nodes available, each comprising two 20-core Intel Xeon Cascade Lake CPUs running at 2.4 Ghz, four Nvidia Tesla V100-SMX2-16GB GPUs (640 Tensor cores each), 384 GB of memory per node, and Infiniband interconnect at 100 Gbit/s. Cirrus is an EPSRC UK-Tier 2 HPC supercomputer hosted by EPCC.

For the task of Legal Judgment Prediction on the Determination sentences presented in Table 6, we use the same hyperparameters with all tested models. We use a learning rate of 2.0×10^{-5}, a batch size of 16, a weight decay of 1.8×10^{-2} on one epoch.

For the second experiment on the *Case Cover* Entities (presented in Table 9) the hyperparameters used are detailed in the table below.

Model	LR	Batch	W decay
BERT	1.0×10^{-4}	32	1.6×10^{-2}
RoBERTa	1.0×10^{-3}	8	1.6×10^{-2}
DeBERTav3	5.0×10^{-6}	16	1.8×10^{-2}
LegalBERT	1.1×10^{-5}	4	1.6×10^{-2}
CaseHOLD	2.5×10^{-5}	16	1.0×10^{-2}
Pol	1.0×10^{-6}	16	1.5×10^{-2}
LexLM	1.1×10^{-5}	16	1.8×10^{-2}

Table 12: Hyper-parameters used for training on the *Case Cover* Entities, after performing hyper-parameter search. All implementations use an Adam optimizer and are trained on 3 epochs, except LexLM which is trained on 2 epochs. *LR* is the Learning rate, *Batch* the Batch size, *W decay* the Weight decay.