Information Extraction for Planning Court Cases

Drish Mali
The University of Edinburgh
D.Mali@sms.ed.ac.uk

Rubash Mali
Himalaya College of Engineering
rubash@hcoe.edu.np

Claire Barale
The University of Edinburgh
claire.barale@ed.ac.uk

Abstract

Legal documents are often long and unstructured, making them challenging and time-consuming to apprehend. An automatic system that can identify relevant entities and labels within legal documents, would significantly reduce the legal research time. We developed a system to streamline legal case analysis from planning courts by extracting key information from XML files using Named Entity Recognition (NER) and multi-label classification models to convert them into structured form. This research contributes three novel datasets for the Planning Court cases: a NER dataset, a multi-label dataset fully annotated by humans, and newly re-annotated multi-label datasets partially annotated using LLMs. We experimented with various general-purpose and legal domain-specific models with different maximum sequence lengths. It was noted that incorporating paragraph position information improved the performance of models for the multi-label classification task. Our research highlighted the importance of domain-specific models, with LegalRoBERTa and LexLM demonstrating the best performance.

1 Introduction

The use of Artificial Intelligence (AI) techniques within the legal domain has been rapidly growing, transforming the way legal professionals handle their complex tasks (Jacey and Yuniarti, 2023). The advancement in Natural Language Processing (NLP) in legal informatics (Krasadakis et al., 2024; Quevedo et al., 2023) has significantly enhanced tasks such as question-answering, judgment predictions, and information extraction from legal text (Zhong et al., 2020; Barale et al., 2023; Licari et al., 2023). For countries with common-law jurisdictions like the UK and the US, legal research needs to be consistent with referencing similar past cases (Shulayeva et al., 2017). However, legal research is extremely time-consuming due to the extensive length of legal texts (Vági, 2023) and the need for domain expertise to navigate the specialised vocabulary and legal jargon (Cemri et al., 2022). Additionally, the unstructured nature of legal documents, such as court hearings, adds to the complexity (Li and Li, 2021). To address these challenges, an NLP-based technique that can automatically extract relevant information from unstructured legal cases into a structured format would be highly beneficial. The primary task is twofold: (1) structuring the raw document formats (PDF and XML) of these Planning Court cases, and (2) curating a novel dataset to support future research efforts. To achieve this, we apply Named Entity Recognition (NER) and multi-label classification techniques, which are effective at organizing and categorizing legal information.

Large Language Models (LLMs) have demonstrated strong capabilities in the legal sector (Fei et al., 2023a), but they require large amounts of domain-specific, accurate data (Lai et al., 2023).We choose to use traditional extractive methods, such as named-entity recognition, which are well suited to the need for precision in the legal workflow and do not yield hallucinations. We opt to build upon those methods that are widely used by legal search softwares and propose to improve them using LLMs. This research focuses on cases from the Planning Court, part of the Administrative Court of England and Wales [1], provided by the Find Case Law service of The National Archives UK. It addresses two main issues: structuring the initial document format (PDF and XML) of Planning Court cases using LLMs, and curating a novel dataset for future research. We have employed NER and multilabel classification to bring structure to it. The project will benefit:

- **Legal Professionals:** For legal professionals

[1] https://www.judiciary.uk/courts-and-tribunals/high-court/administrative-court/planning-court

Proceedings of the Natural Legal Language Processing Workshop 2024, pages 97–114

this system will make the searching of similar cases easier and streamline the legal research process as compared to the traditional manual searching approach (Vági, 2023). It would improve efficiency in legal research, enabling professionals to make more consistent decisions and prepare better for new cases (Barale et al., 2023).

- **Legal NLP Researchers:** The generated structured data will be a valuable asset for various research areas such as judgment prediction, summarisation, drafting, and content selection tasks. The availability of such data facilitates the exploration of new research questions, reducing the challenge of finding high-quality human-labeled legal domain-specific datasets (Song et al., 2022). This study also contributes to filling knowledge gaps in research on Planning Court cases.

The primary research questions guiding the project are as follows:

1. **RQ1:** Can language models accurately extract legal entities such as court name, location, citation, judges, and date from legal cases?

2. **RQ2:** Can language models comprehend legal text and classify it as introduction, factual text, citations to other cases, and judgment?

3. **RQ3:** Do transformer-based models pretrained in the legal domain perform better than general-purpose models in legal entity extraction and multi-label classification?

To address these research questions, this study investigated the utility of language models in extracting information from legal documents specific to the Planning Court. Our contributions are as follows: (1) we create a novel dataset of Planning Court cases specifically curated for NER and multi-label classification, (2) we propose an end-to-end pipeline to extract and structure this data using NER and multi-label classification to analyse those cases automatically by extracting legally relevant entities and paragraphs, (3) lastly we create a structured database from our results, allowing for a quick and efficient search based on the extracted entities.

2 Background and Related Work

2.1 Legal Named Entity Recognition (NER)

NER is a foundational task of Natural Language Processing (NLP), where algorithms are trained to detect and classify entities like location, date or person in the given text (Yu et al., 2020). NER models perform token classification. Research on NER approaches has been ongoing for decades, utilizing methods like graph-based dependency parsing, LSTM, maximum entropy (Yu et al., 2020; Chieu and Ng, 2003; Chiu and Nichols, 2015). With the current advancements in transformer-based models, the performance of NER tasks has been improving significantly. Models like T5 and XLM-RoBERTa have achieved state-of-the-art results (Tavan and Najafi, 2022; Pu et al., 2022). However, the challenge with legal texts is their length and complexity (Mamakas et al., 2022a). They are often difficult to understand due to their complex language, ambiguities, cross-references, frequent amendments, and the specialized legal jargon involved, which requires domain-specific knowledge(Cemri et al., 2022; Ganguly et al.; Otto and Antón, 2009). Additionally, the domain-specific entities like courts, judges, statutes, and articles make the general NER models incompatible with legal documents (Zhao et al., 2023). Transformer-based models have shown promising results even for legal NER tasks (Kalamkar et al., 2022; Barale et al., 2023; Li et al., 2022; Bernsohn et al., 2024). These models perform well across various languages and legal systems (Kalamkar et al., 2022; Păis et al., 2023; Luz de Araujo et al., 2018; Smădu et al., 2022).

For evaluation of the Legal NER systems previous research has used a macro-average F1 score, as there can be an imbalance in the distribution of entities in legal texts (Barale et al., 2023; Keshavarz et al., 2022; Skylaki et al., 2020). Precision and recall are "also crucial for advancing future research and meeting the needs of potential legal end users" (Barale et al., 2023). High precision means the model identifies mostly correct entities, while high recall ensures it finds most of the relevant entities.

2.2 Multi-label classification in legal context

Multi-label classification is a supervised learning method where a single instance of input, such as text, image, or sound, can have multiple labels from a predefined set (Pant et al., 2018; de Leon Ferreira de Carvalho and Freitas, 2009). Compared

to simple multi-class classification problems, multi-label classification is more complex as labels are not mutually exclusive leading to challenges such as label space dimensionality, label drifting, data imbalance, and label dependency (Pant et al., 2018). Multi-label problems can be addressed using multi-class algorithms with a Binary Relevance transformation (Pereira et al., 2018). But it would be extremely slow as for N labels we would require N number of binary classification models (one model for each class) which would not be feasible. Another issue in multi-label classification is choosing evaluation metrics, which can be label-based or instance-based. Popular metrics for such tasks include hamming loss, exact match, AUC PR score, precision, recall, and F1 score (Pereira et al., 2018; Riyanto et al., 2023).

For legal text, the main problem with the dataset is the imbalance of labels, as some labels occur frequently while others are rare. To tackle this problem, F1 score and hamming loss are good candidate metrics (ster et al., 2024; Pereira et al., 2018). Hamming loss evaluates the fraction of incorrectly predicted labels relative to the total number of labels, and the F1 score considers both precision and recall, providing a balance between them. Domain-specific encoder-based models like LEGAL-BERT and LegalRoBERTa (ster et al., 2024; Geng et al., 2021) have shown impressive performance, but as noted, the length of legal texts is large. Therefore, larger models like Longformer and BIGBIRD, which support a larger maximum sequence length, may be needed (Mamakas et al., 2022b). Recent advancements in legal research have led to even larger domain-specific models like LexLM, which offer both larger max sequences and domain knowledge (Chalkidis* et al., 2023).

2.3 Prompting and Few-shot learning

Prompting is the task of providing input instructions to large language models (LLMs) such that these pre-trained models generate output through analogical learning (Bhandari, 2024; Chang et al., 2024). Advancements in LLMs have made prompting a standard approach for various NLP tasks (Chang et al., 2024). However, such models are extremely resource-intensive and require significant effort from the human side to design effective prompts as each model has their prompt format.

While paid services like ChatGPT offer powerful options, cost-effective alternatives like open-source models such as LLaMA (Touvron et al., 2023) also exist. To mitigate the computational demands of open-source models, techniques like Post-Training Quantization can be applied (Zhang et al., 2023) where the size of weights of a neural network are reduced without any retraining. This approach can reduce the computing resource requirements but may also diminish the model's capabilities, creating a performance versus resource trade-off. One solution to this challenge is to use 4-bit quantization (Jin et al., 2024) along with the NF4 Quantization scheme(Dettmers et al., 2023), and use bfloat16 format for performing computations, which aims to balance both the accuracy and efficiency of LLMs. Acquiring adequate amounts of labeled data is quite difficult (Bahrami et al., 2023) in today's day and age, especially with legal data being complex, unstructured, and rare to find. One of the boons of the emerging research in LLMs is their ability to learn patterns and perform specific tasks with few examples, a method called few-shot learning. Few-shot learning involves providing tasks based on a few particular examples in the prompt, allowing LLMs to understand the task, analyse the given examples, and infer accordingly (Brown et al., 2020). This technique has shown promising results in various NLP tasks, including text classification, and sentiment analysis (Min et al., 2021) with larger models like GPT4 and LLaMA performing well in the legal domain (Fei et al., 2023b). However, using few-shot examples alone is not always efficient, especially for complex domain-specific tasks (Naguib et al., 2024; Jayakumar et al., 2023), for such cases domain-specific models are required.

3 Data Collection and Exploratory Data Analysis (EDA)

3.1 Case data collection

We filtered data for Planning Court cases using the keyword search "planning court" on the Case Law service of The National Archives UK, yielding 845 cases. These documents were available in both PDF and XML formats, and we chose XML to avoid data inaccuracies associated with OCR processing of PDFs. These cases can be divided into two sections: the cover section (which contains the initial page of case with typical information as the neutral citation of the case, judges involved, date of the judgment.), the main section (includes the hearing cases from introduction to judgment). By analysing the XML document structure and using the National Archives of the LegalDocM-

L/Akoma Ntoso XML format (Palmirani and Vitali, 2011), it was identified that the cover section was located within the header tag and the main section within the <judgmentBody>tag. We extracted case-wise cover section data by retrieving text inside the <p>tags within the header tag, and the main section data by retrieving text inside the <p>tags within the <judgmentBody>tag. The cover text data was used to train the NER model, while the main section data was used to train the multi-label classification model as illustrated in Figure 1.

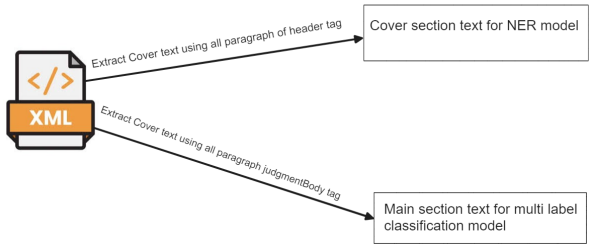


Figure 1: Overall workflow of the data extraction from XML file.

3.2 Cover Section Data Annotation and EDA

We obtained 845 cover sections and used the IOB (Inside, Outside, Beginning) format for annotating entities (Krishnan and Ganapathy, 2005). The entities we extracted using the NER model are 'Citation', 'Court', 'Judge', 'Location' and 'Date' (further descriptions and examples of these entities are listed in the appendix section). These categories were chosen for their crucial role in legal search workflows. 'Citation' aids in linking relevant cases, while 'Court' and 'Judge' allow filtering by jurisdiction or authority. 'Location' helps with regional relevance, and 'Date' enables chronological tracking of cases. For data labeling and creating the NER dataset, we utilised the UBIAI platform.

To understand the data better, we examined the word count of each cover section and the number of labeled entities present in this NER dataset. The descriptive statistics for the word count of all cases are detailed in Table 1. From observing the word counts, it became clear that models with at least a maximum token capacity of 2048 are required. Seventy-five percent of documents have cover sections with fewer words than 1339, about 1741 tokens (1 word is about 1.3 tokens [2]). Further analysis revealed that about 85% of cases have a cover section with fewer than 1500 words (about 2000 tokens), reinforcing the need for models with a 2048 maximum token capacity.

Table 1: Descriptive Statistics of Word Count for Cover Section Data

Statistic	Value
Average Word Count	1199.421
Minimum Word Count	94
Q1 (25%) of Word Count	960
Median (50%) of Word Count	1132
Q3 (75%) of Word Count	1339
Maximum Word Count	5877

The bar plot depicted in Figure 2 shows the counts of various entities within an NER dataset, highlighting the distribution of different entity types. The entity DATE has the highest frequency, appearing 1645 times, while CITATION is the least frequent with 971 occurrences. This visualisation underscores the prevalence of DATE entities in the dataset compared to others and indicates that the dataset is not balanced.

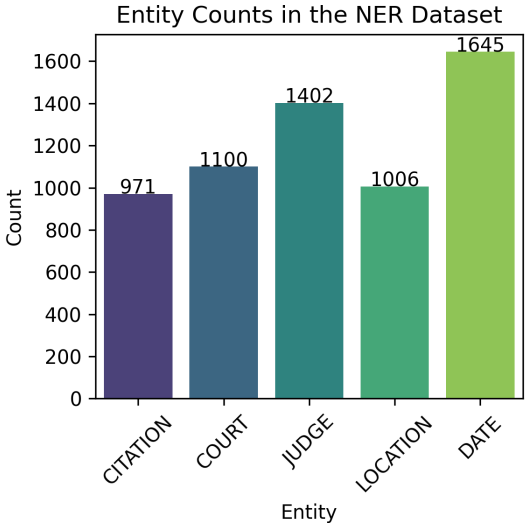


Figure 2: Distribution of Entities in NER Dataset

3.3 Main Section Data Annotation and EDA

We collected 140,377 paragraphs from 845 cases and decided to use four labels: introduction, fact, citation, and judgment for the multi-label classification task. These labels were chosen after discussions and suggestions from domain experts. The core motivation for selecting these four labels was to identify text segments that are important to legal professionals and to ensure that the annotation could be done without requiring specialised legal domain expertise. Due to time constraints and the

[2] https://platform.openai.com/tokenizer.

labor-intensive nature of manual annotation, we decided to annotate 400 out of 845 cases, resulting in 59,302 annotated paragraphs. Descriptive statistics for the word count of paragraphs in the main section are provided in Table 2, detailed descriptions with examples of each label are included in the Appendix. We initially grouped all the paragraphs

Table 2: Descriptive Statistics for Main Section Data

Statistic	Value
Total Number of paragraphs	140377
Average Word Count	359.85
Minimum Word Count	4
Q1 (25%) of Word Count	94
Median (50%) of Word Count	255
Q3 (75%) of Word Count	516
Maximum Word Count	4408

according to cases and manually labeled them. To visualise the data we plotted a bar chart showing the count of paragraphs and their respective labels, as illustrated in Figure 3. The distribution is imbalanced, with the 'fact' label having the highest count (15,511 paragraphs), while the 'introduction' and 'judgment' labels have the lowest counts (1,792 and 422 paragraphs, respectively). This imbalance is expected, as a case usually has a single paragraph for the conclusion, a few for the introduction, but many paragraphs presenting facts.

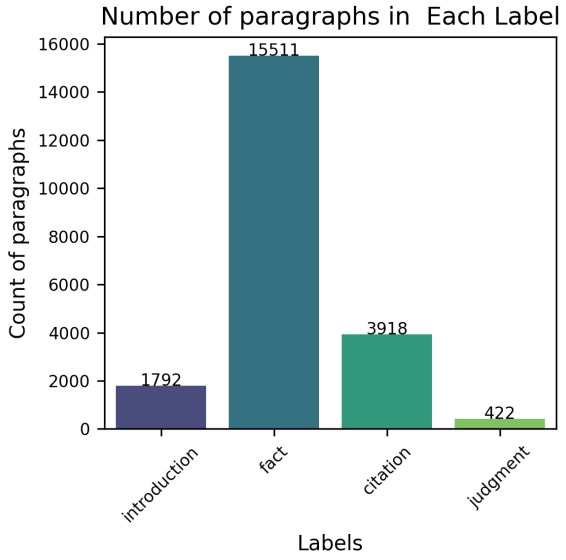


Figure 3: Distribution of label in multi-label Dataset

3.4 Main Section Data re-annotation and EDA

We re-annotated the data after identifying that separating paragraphs using tags disrupted the context needed for accurate labeling. An example of such a case is presented in Figure 4 where the context of the first paragraph is crucial to understanding the second paragraph. The first paragraph mentions the rule CPR, and the second paragraph elaborates on it. If examined individually, the first paragraph should have a fact label of 1 and the second paragraph a fact label of 0. However, as the second paragraph continues from the first, both should be labeled with a fact label of 1. To address this, we restructured and re-annotated the data.

Paragraph 1 with Civil Procedure Rules (CPR)

CPR 54.5 sets out the time limits for filing a claim form in claims for judicial review and statutory review. CPR 54.5(5) specifies that:

Paragraph 2 with continuation statement of above rule

"Where the application for judicial review relates to a decision made by the Secretary of State or local planning authority under the planning acts, the claim form must be filed not later than six weeks after the grounds to make the claim first arose"

Figure 4: Example of paragraphs needing context from preceding paragraph

We improved paragraph extraction from the XML files by using the <num>tag with a style attribute, which allowed connected paragraphs to be treated as a single and made paragraphs longer. This approach resolved the previous issue as mentioned in Figure 4, enabling the extraction of 69,881 paragraphs from 845 cases. Similar to above approach, we looked into the descriptive statistics of the word count of paragraphs presented in Table 3 and 85% of the paragraphs contained fewer than 1,200 words, indicating that models with a maximum token size of around 1,536 tokens would be appropriate for this task.

Table 3: Descriptive Statistics for Re-annotation Data

Statistic	Value
Total Number of paragraphs	69,881
Average Word Count	766.75
Minimum Word Count	4
Q1 (25%) of Word Count	341
Median (50%) of Word Count	579
Q3 (75%)of Word Count	938
Maximum of Word Count	46,559

For the re-annotation process, we utilised both manual and automated methods. We manually la-

beled the 'introduction' and 'judgment' categories by reviewing the initial and final paragraphs of each case. For the 'fact' and 'citation' labels, we employed large language models (LLMs) due to the need for detailed paragraph analysis. We employed the LLaMA 3 70B model with 18 few-shot examples to predict whether paragraphs contained 'citation', achieving 86% accuracy (345 correct out of 400 randomly sampled paragraphs), the prompt for this task is presented in Figure 8. For the 'fact' label, LLaMA 3's performance was unsatisfactory, so we used ChatGPT 3.5, which accurately labeled 241 out of 300 randomly sampled paragraphs (about 80% accuracy) using five examples for a few-shot classification. We used a combination of 4-bit quantization along with double quantization, utilizing the NF4 quantization scheme, and performing computations in bfloat16 to achieve efficient and accurate LLaMA 3 model inference. As shown in Figure 5, the re-annotated data remains highly imbalanced, with the 'fact' label dominating at 45,774 paragraphs, while 'introduction' and 'judgment' labels have 3,429 and 948 paragraphs, respectively.

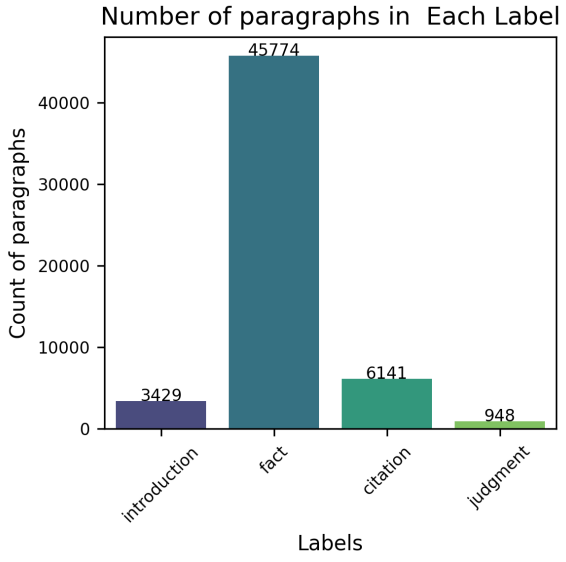


Figure 5: Distribution of label for re-annotated multi-label Dataset

4 Experimental Setup and Modeling

Experimental Setup for NER task: Literature suggests that models like LEGAL-BERT (Chalkidis et al., 2020) have been performing very well for NER tasks in the legal domain (Barale et al., 2023; Keshavarz et al., 2022; Kalamkar et al., 2022). However, the maximum sequence size of such BERT (Devlin et al., 2019) based models is just 512 tokens. Our EDA of NER data (Section 3.2) indicated that we would need models with larger maximum sequence lengths than 512 tokens, making it necessary to explore models with larger maximum sequence lengths. For legal text, such a smaller token size can be restrictive (Mamakas et al., 2022c). In our search for other legal domain-specific models, we identified two additional options: LegalRoBERTa (Geng et al., 2021) that still had 512 tokens as max sequence size limit, and LexLM supports 4096 tokens (Chalkidis et al., 2023). Further research led us to Google's BIGBIRD (Zaheer et al., 2021) model that also supports 4096 tokens. Given these findings, we decided to experiment with these four models which were available in the huggingface platform (Wolf et al., 2020). This selection allows us to evaluate both general-purpose and legal domain-specific models, and also compare the performance of models with smaller (512 tokens) and larger (4096 tokens) maximum sequence lengths. We had a total of 845 cases from which we got 845 cover sections, we split the data in a 70:15:15 ratio for train, test, and validation splits respectively. The models were trained on an NVIDIA A100 GPU with the following training configurations: Learning Rate: 1×10^{-5}, Number of Epochs: 200, Weight Decay: 0.01, Per Device Train Batch Size: 16, Per Device Eval Batch Size: 16, LR Scheduler Type: Cosine, Warmup Ratio: 0.1, Evaluation Strategy: Epoch, Save Strategy: Epoch, and Early Stopping Patience: 30.

As illustrated in Figure 2, the dataset is slightly imbalanced. To account for this, we report the Area Under the Precision-Recall Curve (AUC-PRC) score, which "measures the fraction of true positives among positive predictions" and varies with the ratio of positives to negatives (Saito and Rehmsmeier, 2015). Although the imbalance was not severe with the least frequent entity, 'citation,' still occurring 971 times, the variation in class frequencies was significant enough to warrant consideration. Therefore, we chose to report the F1 score, precision, and recall for each model.

Experimental Setup for multi-label classification: Multi-label classification in the legal domain is challenging due to severe label imbalance and complex label co-occurrence patterns (Forster et al., 2024). Models like LEGAL-BERT, DistilBERT (Sanh et al., 2020), LegalRoBERTa, and LexLM (Forster et al., 2024; Chalkidis et al., 2023; Wei

et al., 2023; Geng et al., 2021) have shown impressive performance in legal multi-label text classification task. From our EDA in Section 3.3, it became clear that we would need models with sequence sizes of about 1,000 tokens. Similar to our NER experiment setup, we explored variations using legal domain-specific and general-purpose models with different maximum sequence sizes. To add another model with a larger maximum sequence size, we again selected the BIGBIRD model. Hence, we decided to use the five models mentioned above. The models were trained on three NVIDIA A100 Tensor Core GPUs using accelerate package [3]. Similar to the NER task, we had 845 cases with 59,302 paragraphs. We opted to split the data case-wise rather than label-wise to maintain the distribution of labels as they would appear in actual case documents. The data was divided into a 70:15:15 ratio for training, testing, and validation splits. The configuration for training is as follows: Per Device Train Batch Size: 16, Per Device Eval Batch Size: 16, Number of Epochs: 30, Evaluation Strategy: Epoch, Save Strategy: Epoch, Checkpoint Limit: 2, and Early Stopping Patience: 15.

Figure 3 highlighted the significant imbalance in the label distribution. To address this issue, we decided to report the F1 score as one of our primary evaluation metrics. The AUC-ROC score was selected because it balances precision and recall, while the Hamming loss was chosen for its sensitivity to class imbalance, capturing errors across all labels, including the rare ones. To check an individual model performance across each label, we assessed performance by reporting the F1 score, AUC-ROC score, recall, precision, and accuracy.

Experimental Setup for re-annotated multi-label classification: In evaluating the re-annotated multi-label data, we used the same four models: LEGAL-BERT, LegalRoBERTa, and LexLM, excluding DistilBERT due to its poor performance with the multi-label data. The evaluation metrics and train test evaluation splits ratio were consistent with those used for the original multi-label data. The only change in the training configuration was a reduction in batch size to 8, which was necessary to manage the increased memory requirements. This adjustment was made because the re-annotated data contained more words, as shown in Table 3, resulting in more tokens per paragraph.

[3] https://github.com/huggingface/accelerate.

5 Evaluation

5.1 Named Entity Recognition (NER) Evaluation

We evaluated these models using these metrics: average precision, recall, F1 score, and AUC-PR score. The results are presented in Table 4. The results clearly show that the LexLM model performed best in terms of precision, and AUC-PR. Meanwhile, LegalRoBERTa excelled in recall and F1 score. As anticipated, the general-purpose Google BIGBIRD model performed the worst among the models tested.

Table 4: Evaluation metrics for different legal models for the NER task

Model	Precision	Recall	F1	AUC PR
LexLM	**0.802**	0.795	0.798	**0.943**
Legal-BERT	0.799	0.804	**0.802**	**0.943**
Legal-RoBERTa	0.791	**0.813**	**0.802**	0.939
Google BigBird	0.731	0.724	0.727	0.926

5.2 Multi-label classification task Evaluation

During data annotation, we observed that "Introduction" typically appears in earlier paragraphs, while labels like "Judgment" appear towards the end. Based on this observation, we decided to test models with and without paragraph position information. The paragraph information was added to the text by explicitly mentioning the paragraph number before the paragraph content. Table 5 and Table 6 show the overall performance metrics for each model, including ROC AUC score, Hamming loss, and F1 score. Including paragraph position information in the models significantly improved their performance across all metrics. Without this information, LegalRoBERTa consistently outperformed other models in most metrics except recall, where it lagged slightly. With the inclusion of this information, the performance differences among models became more balanced. This suggested that models can better interpret and classify legal text with this additional contextual information. Additionally, the models exhibited varied strengths across different labels, indicating that no single model was universally superior. LegalRoBERTa and LexLM were particularly effective, demonstrating strong adaptability and consistent performance enhancements with the added paragraph position context.

Table 5: Evaluation metrics for different models for the multi-label dataset

Model	ROC AUC	Hamming Loss	F1 Score
DistilBERT	0.803	**0.048**	0.675
LexLM	0.800	0.053	0.643
LEGAL-BERT	0.820	**0.048**	0.669
LegalRoBERTa	**0.849**	**0.048**	**0.707**
Google BigBird	0.739	0.053	0.538

Table 6: Evaluation metrics for different models for the multi-label dataset with paragraph information

Model	ROC AUC	Hamming Loss	F1 Score
DistilBERT	0.825	0.046	0.721
LexLM	**0.847**	0.046	0.665
LEGAL-BERT	0.840	0.051	0.734
LegalRoBERTa	0.843	**0.042**	**0.745**
Google BigBird	0.812	0.049	0.654

5.3 Re-annotation Multi-label Task Evaluation

We experimented with three models: LegalRoBERTa, LexLM, and Google BIGBIRD for the re-annotated data. LEGAL-BERT was not used it was not performing best in evaluation criteria for multi-label classification task with paragraph information as illustrated in Table 6. We reported the average ROC AUC score, F1 score and Hamming loss as presented in Table 7. LexLM achieved the highest overall F1 score (0.851), indicating a strong balance between precision and recall, though it had the worst ROC AUC score, and tied for the lowest Hamming loss (0.063) with Google BIGBIRD. LegalRoBERTa demonstrated the highest ROC AUC score (0.877), highlighting its effectiveness in class separation. Its F1 score (0.850) was impressive, just slightly behind LexLM. Google BIGBIRD, while having the lowest F1 score (0.829), excelled in minimising Hamming loss (0.063).

Table 7: Evaluation metrics for different legal models for multi-label classification on re-annotated data

Method	ROC AUC	Hamming Loss	F1 Score
LegalRoBERTa	**0.877**	0.065	0.850
Google BigBird	0.866	**0.063**	0.829
LexLM	0.837	**0.063**	**0.851**

6 Discussion and Conclusion

Our study focused on planning court cases of the Administrative Court of England and Wales, where we designed and experimented with various models to extract important entities and label paragraphs. We added significant contributions to the legal research domain by creating a novel Named Entity Recognition (NER) dataset and a multi-label paragraph dataset which were both fully annotated by humans. Additionally, we developed another multi-label dataset with improved paragraph separation. We applied few-shot learning techniques using state-of-the-art models such as ChatGPT-3.5 and LLaMA 3 70B instruct model to generate two labels: 'fact' and 'citation' respectively.
For the NER task, it became clear that legal domain-specific models performed reasonably well even with smaller maximum sequence sizes. Notably, LegalRoBERTa achieved the highest recall of 0.813 and an F1 score of 0.802. This strong performance was likely because the entities often appeared early in the text, as we observed during the annotation process. LexLM model also excelled in various evaluation criteria and achieved the highest scores in precision (0.802), and AUC PR score (0.943). While Google BIGBIRD (a general-purpose model with a large maximum sequence length) performed the worst across all evaluation metrics. The success of LegalRoBERTa and LexLM highlights the importance of using specialised models for domain-specific applications. Conversely, the poor performance of the general-purpose Google BIGBIRD model reinforces the need for tailored approaches in legal text analysis and research.

In the multi-label task, incorporating paragraph position information had increased the model's performance. For the fully human-annotated dataset, LegalRoBERTa had the best performance. However, when paragraph information was added we found that there was no single superior model; both LegalRoBERTa and LexLM performed well in various metrics. As expected, Google BIGBIRD did not perform on par with the other models.

For re-annotated data, we tested only LegalRoBERTa, Google BIGBIRD, and LexLM. The LegalRoBERTa performed well in various metrics. However, the general-purpose model with a large max sequence like Google BIGBIRD, did not perform well further reinforcing the importance of domain-specific models. These findings from both the NER and multi-label classification tasks underscore the importance of using specialized models tailored to the legal domain to achieve superior performance, advancing legal research in this area.

7 Limitations and Future Work

One limitation of this research is that the re-extracted data was not fully annotated by humans. Due to time and cost constraints, we used LLMs to annotate the 'citation' and 'fact' labels. Future studies can leverage newer state-of-the-art models like LLaMA 3.1 and ChatGPT 4 for more accurate annotation or even fully human annotation can also be done. Another limitation is the dependency on powerful GPUs for fine-tuning and inferencing transformer-based models, which may not always be available in legal or academic settings. Additionally, the generalization performance of our methods has not been tested on other similar datasets.

For future work, we could explore more advanced models for annotation and extract paragraphs and cover section text from all cases within the Administrative Court to build a larger corpus. Additionally, testing our methods on other similar datasets and reporting their metrics would help assess the generalization of our approach.

8 Ethics Statement

The curated dataset contains sensitive information, including the names of claimants and appellants. Our research utilizes data that is already publicly available and not anonymous. We have obtained permission to use this data under the Open Justice Licence provided by the Find Case Law service, which allows us to copy, publish, distribute, and transmit the information. Our primary task is to transform this semi-structured data into a structured format. While we acknowledge the potential concerns regarding dual use, we focus on streamlining the analysis of legal cases, making the likelihood of such concerns minimal.

References

Morteza Bahrami, Muharram Mansoorizadeh, and Hassan Khotanlou. 2023. Few-shot learning with prompting methods. *2023 6th International Conference on Pattern Recognition and Image Analysis (IPRIA)*, pages 1–5.

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.

Dor Bernsohn, Gil Semo, Yaron Vazana, Gila Hayat, Ben Hagag, Joel Niklaus, Rohit Saha, and Kyryl Truskovskyi. 2024. LegalLens: Leveraging LLMs for legal violation identification in unstructured text. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2129–2145, St. Julian's, Malta. Association for Computational Linguistics.

Prabin Bhandari. 2024. A survey on prompting techniques in llms. *arXiv.org*.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. *Preprint*, arXiv:2005.14165.

Mert Cemri, Tolga Çukur, and Aykut Koç. 2022. Unsupervised simplification of legal texts. *arXiv.org*.

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*, Nicolas Garneau*, Catalina Goanta, Daniel Martin 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*, Toronto, Canada. Association for Computational Linguistics.

Ilias Chalkidis, Nicolas Garneau, Catalina Goanta, Daniel Martin Katz, and Anders Søgaard. 2023. Lexfiles and legallama: Facilitating english multinational legal language model development. *Preprint*, arXiv:2305.07507.

Kaiyan Chang, Songcheng Xu, Chenglong Wang, Yingfeng Luo, Xiao Tong, and Jingbo Zhu. 2024. Efficient prompting methods for large language models: A survey. *arXiv.org*.

Hai Leong Chieu and Hwee Tou Ng. 2003. Named entity recognition with a maximum entropy approach. In *Conference on Computational Natural Language Learning*.

Jason P. C. Chiu and Eric Nichols. 2015. Named entity recognition with bidirectional lstm-cnns. *Transactions of the Association for Computational Linguistics*, 4:357–370.

André Carlos Ponce de Leon Ferreira de Carvalho and Alex Alves Freitas. 2009. A tutorial on multi-label

classification techniques. In *IEEE Symposium on Foundations of Computational Intelligence*.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. Qlora: Efficient finetuning of quantized llms. *Preprint*, arXiv:2305.14314.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. *Preprint*, arXiv:1810.04805.

Zhiwei Fei, Xiaoyu Shen, Dawei Zhu, Fengzhe Zhou, Zhuo Han, Songyang Zhang, Kai Chen, Zongwen Shen, and Jidong Ge. 2023a. Lawbench: Benchmarking legal knowledge of large language models. *arXiv.org*.

Zhiwei Fei, Xiaoyu Shen, Dawei Zhu, Fengzhe Zhou, Zhuo Han, Songyang Zhang, Kai Chen, Zongwen Shen, and Jidong Ge. 2023b. Lawbench: Benchmarking legal knowledge of large language models. *Preprint*, arXiv:2309.16289.

Martina Forster, Claudia Schulz, Prudhvi Nokku, Melicaalsadat Mirsafian, Jaykumar Kasundra, and Stavroula Skylaki. 2024. The right model for the job: An evaluation of legal multi-label classification baselines. *Preprint*, arXiv:2401.11852.

Debasis Ganguly, Jack G. Conrad, Kripabandhu Ghosh, Saptarshi Ghosh, Pawan Goyal, Paheli Bhattacharya, Shubham Kumar Nigam, and Shounak Paul. Legal ir and nlp: The history, challenges, and state-of-the-art. In *Advances in Information Retrieval*, Lecture Notes in Computer Science, pages 331–340. Springer Nature Switzerland, Cham.

Saibo Geng, Rémi Lebret, and Karl Aberer. 2021. Legal transformer models may not always help. *arXiv.org*.

Precia Jacey and Siti Yuniarti. 2023. Artificial intelligence: Implementation in legal services (comparative study on china, united stated and indonesia). *Proceedings of the International Conference on Industrial Engineering and Operations Management*.

Thanmay Jayakumar, Fauzan Farooqui, and Luqman Farooqui. 2023. Large language models are legal but they are not: Making the case for a powerful legalllm. *ArXiv*, abs/2311.08890.

Renren Jin, Jiangcun Du, Wuwei Huang, Wei Liu, Jian Luan, Bin Wang, and Deyi Xiong. 2024. A comprehensive evaluation of quantization strategies for large language models. *Preprint*, arXiv:2402.16775.

Prathamesh Kalamkar, Astha Agarwal, Aman Tiwari, Smita Gupta, Saurabh Karn, and Vivek Raghavan. 2022. Named entity recognition in indian court judgments. *arXiv.org*.

Hossein Keshavarz, Zografoula Vagena, Pigi Kouki, Ilias Fountalis, Mehdi Mabrouki, Aziz Belaweid, and Nikolaos Vasiloglou. 2022. Named entity recognition in long documents: An end-to-end case study in the legal domain. In *2022 IEEE International Conference on Big Data (Big Data)*, pages 2024–2033.

Panteleimon Krasadakis, Evangelos Sakkopoulos, and Vassilios S. Verykios. 2024. A survey on challenges and advances in natural language processing with a focus on legal informatics and low-resource languages. *Electronics (Basel)*, 13(3):648–.

Vijay Krishnan and Vignesh Ganapathy. 2005. Named entity recognition. https://cs229.stanford.edu/proj2005/KrishnanGanapathy-NamedEntityRecognition.pdf. [Accessed 13-04-2024].

Jinqi Lai, Wensheng Gan, Jiayang Wu, Zhenlian Qi, and Philip S Yu. 2023. Large language models in law: A survey. *arXiv.org*.

Jing Li, Aixin Sun, Jianglei Han, and Chenliang Li. 2022. A survey on deep learning for named entity recognition. *IEEE transactions on knowledge and data engineering*, 34(1):50–70.

Xiao Li and Jing Li. 2021. Law - net: A new method for legal text mining.

Daniele Licari, Praveen Bushipaka, Gabriele Marino, Giovanni Comandé, and Tommaso Cucinotta. 2023. Legal holding extraction from italian case documents using italian-legal-bert text summarization. In *Proceedings of the Nineteenth International Conference on Artificial Intelligence and Law*, ICAIL '23, page 148–156, New York, NY, USA. Association for Computing Machinery.

Pedro Henrique Luz de Araujo, Teófilo E. de Campos, Renato R. R. de Oliveira, Matheus Stauffer, Samuel Couto, and Paulo Bermejo. 2018. Lener-br: A dataset for named entity recognition in brazilian legal text. In *Computational Processing of the Portuguese Language*, pages 313–323, Cham. Springer International Publishing.

Dimitris Mamakas, Petros Tsotsi, Ion Androutsopoulos, and Ilias Chalkidis. 2022a. Processing long legal documents with pre-trained transformers: Modding legalbert and longformer. *arXiv.org*.

Dimitris Mamakas, Petros Tsotsi, Ion Androutsopoulos, and Ilias Chalkidis. 2022b. Processing long legal documents with pre-trained transformers: Modding LegalBERT and longformer. In *Proceedings of the Natural Legal Language Processing Workshop 2022*, pages 130–142, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Dimitris Mamakas, Petros Tsotsi, Ion Androutsopoulos, and Ilias Chalkidis. 2022c. Processing long legal documents with pre-trained transformers: Modding legalbert and longformer. *Preprint*, arXiv:2211.00974.

Sewon Min, Michael Lewis, Hannaneh Hajishirzi, and Luke Zettlemoyer. 2021. Noisy channel language model prompting for few-shot text classification. *ArXiv*, abs/2108.04106.

Marco Naguib, Xavier Tannier, and Aur'elie N'ev'eol. 2024. Few shot clinical entity recognition in three languages: Masked language models outperform llm prompting. *ArXiv*, abs/2402.12801.

Paul N. Otto and Annie I. Antón. 2009. Managing legal texts in requirements engineering.

Monica Palmirani and Fabio Vitali. 2011. Akoma-ntoso for legal documents. In *Legislative XML for the Semantic Web*, Law, Governance and Technology Series, pages 75–100. Springer Netherlands, Dordrecht.

Pooja Pant, A. Sai Sabitha, Tanupriya Choudhury, and Prince Dhingra. 2018. Multi-label classification trending challenges and approaches. In *Emerging Trends in Expert Applications and Security*, volume 841 of *Advances in Intelligent Systems and Computing*, pages 433–444. Springer Singapore Pte. Limited, Singapore.

Rafael B. Pereira, Alexandre Plastino, Bianca Zadrozny, and Luiz H.C. Merschmann. 2018. Correlation analysis of performance measures for multi-label classification. *Information processing management*, 54(3):359–369.

Keyu Pu, Hongyi Liu, Yixiao Yang, Jiangzhou Ji, Wenyi Lv, and Yaohan He. 2022. Cmb ai lab at semeval-2022 task 11: A two-stage approach for complex named entity recognition via span boundary detection and span classification. In *International Workshop on Semantic Evaluation*.

Vasile Păis, Maria Mitrofan, Carol Luca Gasan, Alexandru Ianov, Corvin Ghit, ă, Vlad Silviu Coneschi, and Andrei Onut. 2023. Legalnero: A linked corpus for named entity recognition in the romanian legal domain. *Semantic Web*, pages 1–14.

Ernesto Quevedo, Tomas Cerny, Alejandro Rodriguez, Pablo Rivas, Jorge Yero, Korn Sooksatra, Alibek Zhakubayev, and Davide Taibi. 2023. Legal natural language processing from 2015-2022: A comprehensive systematic mapping study of advances and applications. *IEEE access*, pages 1–1.

Slamet Riyanto, Imas Sukaesih Sitanggang, Taufik Djatna, and Tika Dewi Atikah. 2023. Comparative analysis using various performance metrics in imbalanced data for multi-class text classification. *International Journal of Advanced Computer Science and Applications*.

Takaya Saito and Marc Rehmsmeier. 2015. The precision-recall plot is more informative than the roc plot when evaluating binary classifiers on imbalanced datasets. *PloS one*, 10(3):e0118432–e0118432.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2020. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *Preprint*, arXiv:1910.01108.

Olga Shulayeva, Advaith Siddharthan, and Adam Wyner. 2017. Recognizing cited facts and principles in legal judgements. *Artificial intelligence and law*, 25(1):107–126.

Stavroula Skylaki, Ali Oskooei, Omar Bari, Nadja Herger, and Zac Kriegman. 2020. Named entity recognition in the legal domain using a pointer generator network. *Preprint*, arXiv:2012.09936.

Răzvan-Alexandru Smădu, Ion-Robert Dinică, Andrei-Marius Avram, Dumitru-Clementin Cercel, Florin Pop, and Mihaela-Claudia Cercel. 2022. Legal named entity recognition with multi-task domain adaptation. In *Proceedings of the Natural Legal Language Processing Workshop 2022*, pages 305–321, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Dezhao Song, Andrew Vold, Kanika Madan, and Frank Schilder. 2022. Multi-label legal document classification: A deep learning-based approach with label-attention and domain-specific pre-training. *Information systems (Oxford)*, 106:101718–.

Martina ster, Claudia Schulz, Prudhvi Nokku, Melicaal-sadat Mirsafian, Jaykumar Kasundra, and Stavroula Skylaki. 2024. The right model for the job: An evaluation of legal multi-label classification baselines. *arXiv.org*.

Ehsan Tavan and Mary Najafi. 2022. Marsan at semeval-2022 task 11: Multilingual complex named entity recognition using t5 and transformer encoder. In *International Workshop on Semantic Evaluation*.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. Llama: Open and efficient foundation language models. *Preprint*, arXiv:2302.13971.

Renátó Vági. 2023. How could semantic processing and other nlp tools improve online legal databases? *TalTech Journal of European Studies*, 13(2):138–151.

Fusheng Wei, Robert Keeling, Nathaniel Huber-Fliflet, Jianping Zhang, Adam Dabrowski, Jingchao Yang, Qiang Mao, and Han Qin. 2023. Empirical study of llm fine-tuning for text classification in legal document review. *2023 IEEE International Conference on Big Data (BigData)*, pages 2786–2792.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2020. Huggingface's transformers: State-of-the-art natural language processing. *Preprint*, arXiv:1910.03771.

Juntao Yu, Bernd Bohnet, and Massimo Poesio. 2020. Named entity recognition as dependency parsing. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6470–6476, Online. Association for Computational Linguistics.

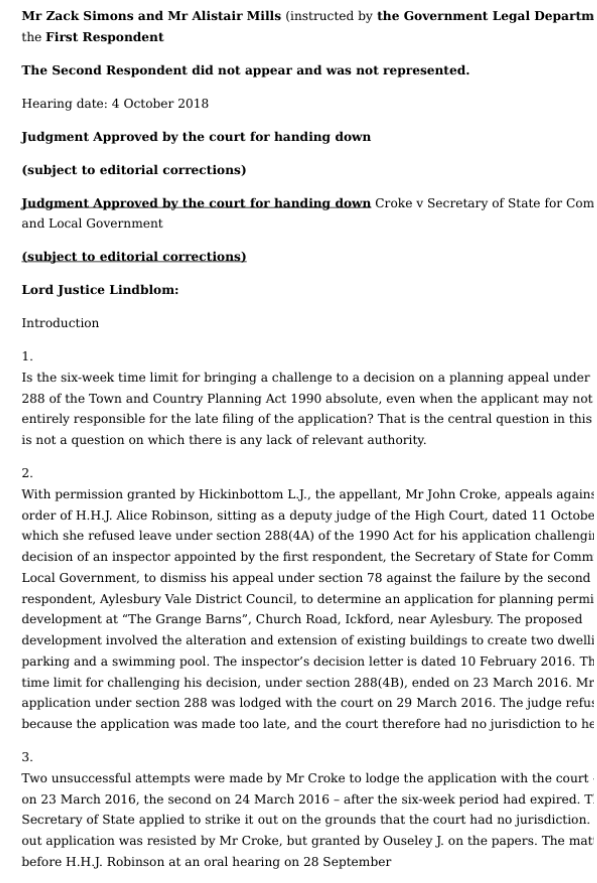
Manzil Zaheer, Guru Guruganesh, Avinava Dubey, Joshua Ainslie, Chris Alberti, Santiago Ontanon, Philip Pham, Anirudh Ravula, Qifan Wang, Li Yang, and Amr Ahmed. 2021. Big bird: Transformers for longer sequences. *Preprint*, arXiv:2007.14062.

Jinjie Zhang, Yixuan Zhou, and Rayan Saab. 2023. Post-training quantization for neural networks with provable guarantees. *SIAM journal on mathematics of data science*, 5(2):373–399.

Junzhe Zhao, Yingxi Wang, Nicolay Rusnachenko, and Huizhi Liang. 2023. Legal_try at semeval-2023 task 6: Voting heterogeneous models for entities identification in legal documents. In *International Workshop on Semantic Evaluation*.

Haoxi Zhong, Chaojun Xiao, Cunchao Tu, Tianyang Zhang, Zhiyuan Liu, and Maosong Sun. 2020. How does nlp benefit legal system: A summary of legal artificial intelligence. *arXiv.org*.

A Appendix A: Document layout



Mr Zack Simons and Mr Alistair Mills (instructed by **the Government Legal Departm** the **First Respondent**

The Second Respondent did not appear and was not represented.

Hearing date: 4 October 2018

Judgment Approved by the court for handing down

(subject to editorial corrections)

Judgment Approved by the court for handing down Croke v Secretary of State for Com and Local Government

(subject to editorial corrections)

Lord Justice Lindblom:

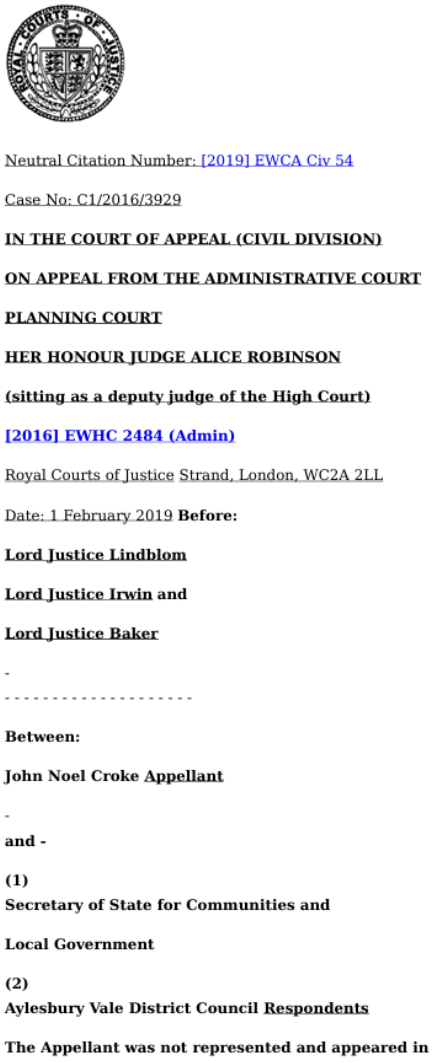
Introduction

1.
Is the six-week time limit for bringing a challenge to a decision on a planning appeal under 288 of the Town and Country Planning Act 1990 absolute, even when the applicant may not entirely responsible for the late filing of the application? That is the central question in this is not a question on which there is any lack of relevant authority.

2.
With permission granted by Hickinbottom L.J., the appellant, Mr John Croke, appeals agains order of H.H.J. Alice Robinson, sitting as a deputy judge of the High Court, dated 11 Octobe which she refused leave under section 288(4A) of the 1990 Act for his application challengi decision of an inspector appointed by the first respondent, the Secretary of State for Comm Local Government, to dismiss his appeal under section 78 against the failure by the second respondent, Aylesbury Vale District Council, to determine an application for planning permi development at "The Grange Barns", Church Road, Ickford, near Aylesbury. The proposed development involved the alteration and extension of existing buildings to create two dwelli parking and a swimming pool. The inspector's decision letter is dated 10 February 2016. Th time limit for challenging his decision, under section 288(4B), ended on 23 March 2016. Mr application under section 288 was lodged with the court on 29 March 2016. The judge refus because the application was made too late, and the court therefore had no jurisdiction to he

3.
Two unsuccessful attempts were made by Mr Croke to lodge the application with the court on 23 March 2016, the second on 24 March 2016 – after the six-week period had expired. T Secretary of State applied to strike it out on the grounds that the court had no jurisdiction. out application was resisted by Mr Croke, but granted by Ouseley J. on the papers. The mat before H.H.J. Robinson at an oral hearing on 28 September

Figure 6: Snapshot of an example of main section



Neutral Citation Number: [2019] EWCA Civ 54

Case No: C1/2016/3929

IN THE COURT OF APPEAL (CIVIL DIVISION)

ON APPEAL FROM THE ADMINISTRATIVE COURT

PLANNING COURT

HER HONOUR JUDGE ALICE ROBINSON

(sitting as a deputy judge of the High Court)

[2016] EWHC 2484 (Admin)

Royal Courts of Justice Strand, London, WC2A 2LL

Date: 1 February 2019 **Before:**

Lord Justice Lindblom

Lord Justice Irwin and

Lord Justice Baker

-
- - - - - - - - - - - - - - - - - - -

Between:

John Noel Croke Appellant

-
and -

(1)
Secretary of State for Communities and

Local Government

(2)
Aylesbury Vale District Council Respondents

The Appellant was not represented and appeared in person.

Figure 7: Snapshot of an example of main section

B Appendix B: Links to Models and Platforms Used

- **LEGAL-BERT:** https://huggingface.co/nlpaueb/legal-bert-base-uncased
- **LegalRoBERTa:** https://huggingface.co/Saibo-creator/legal-roberta-base
- **LexLM:** https://huggingface.co/lexlms/legal-longformer-large
- **BIGBIRD:** https://huggingface.co/google/bigbird-roberta-base
- **DistilBERT:** https://huggingface.co/distilbert/distilbert-base-uncased
- **GitHub Repository:** https://tinyurl.com/d434zc34

C Appendix C: Individual metrics for classes for re-annotated data

Table 8: Evaluation metrics for different legal models across various labels for re-annotated data.

method	label	f1	roc auc	precision	recall	accuracy
google BIGBIRD	introduction	0.781	0.897	0.754	0.811	0.975
	fact	0.844	**0.731**	**0.783**	0.916	**0.783**
	citation	0.974	0.992	0.960	0.983	0.995
	judgment	0.719	0.845	0.747	0.694	0.992
LexLM	introduction	**0.799**	0.906	**0.775**	0.824	**0.977**
	fact	**0.846**	0.719	0.777	0.935	0.781
	citation	0.977	**0.993**	0.964	0.986	0.996
	judgment	0.785	**0.904**	0.759	**0.812**	**0.993**
LegalRoBERTa	introduction	0.788	**0.913**	0.740	**0.844**	0.975
	fact	0.840	0.705	0.761	**0.938**	0.771
	citation	**0.980**	**0.993**	**0.970**	**0.991**	**0.998**
	judgment	**0.793**	0.897	**0.787**	0.798	**0.993**

D Appendix D: Individual metrics for classes for multi-label classification

Table 9: Evaluation metrics for different models across various labels for multi-label dataset

method	label	f1 score	roc auc	precision	recall	accuracy
DistilBERT	introduction	0.388	0.657	**0.472**	.329	**0.960**
	fact	0.759	0.831	0.766	0.752	0.866
	citation	0.851	0.929	0.832	0.870	0.979
	judgment	0.704	0.797	**0.864**	0.594	**0.996**
LexLM	introduction	0.351	0.642	0.424	0.300	0.957
	fact	0.756	0.833	0.742	0.770	0.860
	citation	0.832	0.917	0.817	0.848	0.977
	judgment	0.635	0.811	0.645	0.625	0.995
LEGAL-BERT	introduction	0.357	0.656	0.385	0.332	0.954
	fact	0.757	0.829	**0.770**	0.745	0.866
	citation	0.856	0.931	**0.840**	0.874	**0.980**
	judgment	0.707	0.866	0.681	0.734	**0.996**
LegalRoBERTa	introduction	**0.425**	**0.687**	0.465	**0.392**	0.959
	fact	**0.788**	**0.865**	0.734	**0.850**	**0.871**
	citation	**0.862**	**0.950**	0.816	**0.914**	**0.980**
	judgment	**0.756**	**0.897**	0.718	**0.797**	**0.996**
google BIGBIRD	introduction	0.285	0.598	0.466	0.205	**0.960**
	fact	0.764	0.852	0.694	**0.850**	0.853
	citation	0.859	0.938	0.831	0.889	**0.980**
	judgment	0.247	0.570	**1.000**	0.141	0.994

Table 10: Evaluation metrics for different legal models across various labels for multi-label dataset with paragraph information.

method	label	f1	roc_auc	precision	recall	accuracy
DistilBERT	introduction	0.631	0.767	**0.762**	0.541	0.978
	fact	0.757	0.834	0.733	0.780	0.859
	citation	**0.868**	0.929	0.863	0.867	0.982
	judgment	0.631	0.772	0.745	0.594	0.993
LexLM	introduction	0.621	0.774	0.701	0.558	0.974
	fact	0.748	0.827	0.739	0.757	0.857
	citation	0.850	0.920	0.817	0.870	0.978
	judgment	0.444	**0.869**	0.316	0.251	**0.996**
LegalRoBERTa	introduction	**0.654**	0.787	0.749	0.518	**0.979**
	fact	0.804	0.833	0.767	0.804	0.858
	citation	0.867	**0.941**	**0.937**	**0.894**	**0.984**
	judgment	0.656	0.811	0.685	0.625	0.992
LEGAL-BERT	introduction	0.630	0.784	0.699	0.572	**0.979**
	fact	0.752	0.836	0.707	0.803	0.859
	citation	0.837	0.936	0.802	0.871	0.981
	judgment	**0.719**	0.820	0.790	**0.656**	0.995
google BIGBIRD	introduction	0.638	**0.798**	0.670	**0.603**	0.964
	fact	**0.866**	**0.940**	**0.842**	0.834	**0.980**
	citation	0.865	0.940	0.842	**0.834**	0.980
	judgment	0.247	0.570	**1.000**	0.140	0.993

E Appendix E: Individual metrics for classes for NER task

Table 11: Performance metrics for different legal models for each class of NER dataset.

	LexLM			LEGAL-BERT			LegalRoBERTa			google BIGBIRD		
class	precision	recall	f1-score	precision	recall	f1-score	precision	recall	f1-score	precision	recall	f1-score
O	0.97	**0.99**	**0.98**	**0.98**	0.98	**0.98**	**0.98**	0.98	**0.98**	**0.98**	0.97	**0.98**
B-CITATION	0.98	0.95	**0.96**	0.98	0.95	**0.96**	**1**	0.93	**0.96**	0.93	**0.97**	0.95
I-CITATION	0.99	0.98	**0.98**	0.99	**0.99**	**0.98**	**1**	0.96	**0.98**	0.98	**0.99**	**0.98**
B-DATE	**0.97**	0.96	0.96	0.96	0.96	0.96	**0.97**	**0.97**	**0.97**	**0.97**	0.95	0.96
I-DATE	0.93	0.98	**0.96**	**0.96**	0.91	0.93	0.92	**1**	0.95	0.95	0.94	**0**.95
B-JUDGE	0.68	0.77	0.72	**0.72**	0.71	0.71	0.69	**0.79**	**0.74**	0.68	0.61	0.65
I-JUDGE	0.90	0.87	0.89	0.89	0.86	0.86	**0.91**	**0.90**	**0.90**	0.90	0.81	0.84
B-LOCATION	0.72	0.61	0.66	0.66	0.64	0.74	0.68	**0.67**	**0.77**	**0.77**	**0.67**	0.62
I-LOCATION	**0.95**	0.82	0.88	0.88	**0.88**	**0.92**	0.90	0.84	0.88	0.89	0.62	0.86
B-COURT	**0.92**	0.73	**0.81**	0.81	**0.91**	0.75	0.82	0.65	0.74	0.74	0.62	0.75
I-COURT	**0.98**	0.82	**0.89**	0.89	**0.98**	0.86	0.92	0.82	**0.89**	0.89	0.80	0.87

F Appendix F: Description of Labels used in multi-label classification with examples

Table 12: Labels used in multi-label classification with description and examples.

Labels	Description with example
Introduction	Text containing the topic of discussion in court, usually preceding facts, history, and background. *Example:* What is the scope of the "presumption in favour of sustainable development" in the National Planning Policy Framework ("the NPPF")? That is the basic question in this appeal. Judges in the Planning Court have differed in their answers to it.
Fact	Text containing rules, facts, or references such as section 10, s.10, S 10, article 10, CPR (Civil Procedure Rule), regulations, etc. *Example:* Section 70(2) of the 1990 Act requires that, in dealing with an application for planning permission, a local planning authority must have regard to the provisions of the development plan, so far as is material to the application, and any "other material considerations".
Citation	Text containing references to cases, including neutral citations of different cases. *Example:* Time starts to run on the day after the date of the decision letter itself, not the day on which it is received by the applicant (see Griffiths v Secretary of State for the Environment [1983] 1 All E.R. 439).
Judgment	Text consisting of outcomes of cases and appeal (successful or dismissed). *Example:* For the reasons I have given, I would dismiss this appeal.

G Appendix G: Description of Entities used in NER with examples

Table 13: Entities used for NER model with description and examples.

Entities	Description with Examples
CITATION	A unique identifier for cases consisting of the year, jurisdiction, court, and case number. *Examples:* [2023] EWHC 2629 (KB), [2018] EWCA Civ 2532, [2011] UKSC 7
JUDGE	Name of the judges involved. *Examples:* Lord Justice Lindblom, MR JUSTICE JAY, MR JUSTICE HOLGATE
COURT	Name of the court where the case is heard. *Examples:* High Court (Administrative Court), Court of Appeal (Civil Division), High Court (Planning Court)
LOCATION	Location where the case was heard. *Examples:* Bristol Civil Justice Centre, Strand, London, WC2A 2LL, Manchester Civil Justice Centre
DATE	Date when the case was heard. *Examples:* 14 November 2018, 20/10/2023, 12/10/2015

H Appendix H: Few shot prompt example labeling the fact using ChatGPT

Here is the prompt used in the labeling the 'fact' using ChatGPT 3.5, the prompt is similar for LLaMA 3 70B which has 18 examples (used to label 'citation').

```
{
    "role": "system",
    "content": f"""Task:** For the given text and determine whether it contains a fact
content. Assign the label 1 only when you are extremely certain that the fact content is present
(as mostly don't have fact content); otherwise, assign the label 0. You must generate an label for
every given input and ouput the label only. The description of text with fact content is delimited
by << >>.
    <<
    Fact content: Text containing rules, facts like section (like section 10, s 10 ,S10), Act,
article, Amendment, rule, policy,local plan, paragrapgh from NPPF, CPR (civil preceding rule)
etc.

    >>
    The few shot examples are delimited by ``` ```
    ```
    1) Input text : For all the above reasons, these second applications to reconsider must
fail. They fail to meet any – let alone all – of the criteria set out in CPR 52.30. These are not
exceptional cases. There has been no injustice to the applicant. There is no probability of a
different result. There was never any tenable basis for an appeal, for the reasons given by both
the judge and Lewison L.J.. We consider that neither application for reconsideration was
justifiable. The applications before us are therefore dismissed.

    Output: 1

    2) Input text: In Lawal v Circle 33 Housing Trust [2014] EWCA Civ 1514, [2015] 1 P. &
C.R. 12, Sir TerenceEtherton, then the Chancellor of the High Court, said at paragraph 65 that
the paradigm case for reopening "is where the litigation process has been corrupted, such as by
fraud or bias or where the judge read the wrong papers". He reiterated that the broad principle
was that "for an appeal to be reopened, the injustice that would be perpetrated if the appeal is
not reopened must be so grave as to overbear the pressing claim of finality in litigation". Finally,
he said:
    "It also follows that the fact that a wrong result was reached earlier, or that there is
fresh evidence, or that the amounts in issue are very large, or that the point in issue is very
important to one or more of the parties or is of general importance is not of itself sufficient to
displace the fundamental public importance of the need for finality."

    Output: 0

    3) Input text : These and other statements of principle were brought together in the
judgment of this court inGoring-on-Thames Parish Council, to which we have already referred.
Importantly, at paragraph 15,emphasis was placed on the requirement that "there must be a
powerful probability that the decision in question would have been different if the integrity of the
earlier proceedings had not been critically undermined". More recently, the scope of the
jurisdiction was summarised byHickinbottom L.J.in Balwinder Singh v Secretary of State for the
Home Department [2019] EWCA Civ 1504, paragraph 3, in terms with which we entirely
agree:"This is an exceptional jurisdiction, to be exercised rarely: "the injustice that would be
perpetrated if the appeal is not reopened must be so grave as to overbear the pressing claim of
finality in litigation"(Lawal v Circle 33 Housing Trust [2014] EWCA Civ 1514; [2015] HLR 9 at
[65] per Sir TerenceEtherton VC (as he then was)). The jurisdiction will therefore not be
exercised simply because the determination was wrong, but only where it can be demonstrated
that the integrity of the earlier proceedings has been "critically undermined" (R
(Goring-on-Thames Parish Council) v SouthOxfordshire District Council [2018] EWCA Civ 860;
[2018] 1 WLR 5161 at [10]-[11]."

    Output: 0

    4) Input : There were several delays prior to the Secretary of State's decision owing to
additional consultations,which included a consultation on the Court of Appeal's decision in East
Northamptonshire DistrictCouncil v Secretary of State for Communities and Local Government
(the Barnwell Manor case)[2014] EWCA Civ 137. The Court of Appeal interpreted section 66(1)
of the Planning (Listed Buildings And Conservation Areas) Act 1990 as requiring the
decision-maker to give "the desirability of preserving the [relevant] building or its setting" not
merely careful consideration but considerable importance and weight when balancing the
advantages of the proposed development against the harmit might do.
    Output: 1

    5) Input: For these reasons, which are somewhat different from those of the judge, I
would dismiss this appeal. I agree.
    Output: 0
    ```
    **Note the output should be 1 or 0. So return only 0 or 1 accordingly**

The legal text is  {paragraph}"""
    }
```

Figure 8: prompt for labeling fact

I Appendix I: Screen shot of the Data generated by the models

NEUTRAL CITATION	COURT	LOCATION	DATE	JUDGE	introduction	judgement	facts	citation
[2023] EWHC 171 (Admin)	THE HIGH COURT OF JUSTICE	2 Redcliff Street, Bristol BS	2023, 8 & 9 November 2022	MR JUSTICE LANE, MR JUSTICE LANE	['1. Climate change, with its consequences f...	['257. This judicial review is dismissed.']	['1. Climate change, with its consequences f...	['13. DL 78 to 82 are headed "Climate change...
[2022] EWHC 2406 (Admin)	THE HIGH COURT OF JUSTICE	Strand , London	2022, 7 July 2022	MR C M G OCKELTON , VICE PRESIDENT OF THE UPP...	['1. The Claimant is the registered propriet...	['31. I therefore hold that on its true cons...	['1. The Claimant is the registered propriet...	['5. The issue to be decided is therefore a ...
[2022] EWHC 2991 (Admin)	THE HIGH COURT OF JUSTICE	Strand , London	25 November 2022, 10 November 2022	JUDGE JARMAN KC	["1. The second interested party in these pr...	['38. In my judgment, each of the grounds of...	["1. The second interested party in these pr...	['27. He points out that the inspector did n...
[2022] EWHC 3317 (Admin)	THE HIGH COURT OF JUSTICE, COURT	Strand , London	2022, 8 December 2022	JUDGE JARMAN KC	['1. The claimant challenges a decision of a...	['37. I would therefore quash the decision a...	['1. The claimant challenges a decision of a...	['12. In contrast, the GPDO specifies those ...

Figure 9: Screenshot of data generated by the Named Entity Recognition (NER) and multi-label classification models. The NER model extracts key entities such as Neutral Citation, Court, Location, Date, and Judge. Meanwhile, the multi-label classification model generates relevant sections including Introduction, Judgment, Facts, and Citations.