CourtPressGER: A Dataset of German Court Decisions and Press Releases with Baseline Models

TECHNICAL UNIVERSITY OF MUNICH, School of Computation, Information and Technology, Germany

Legal text generation is gaining importance as a research area in the intersection of legal technology and natural language processing. However, the scarcity of high-quality, domain-specific datasets remains a challenge, particularly for languages other than English. In this paper, we present CourtPressGER, a novel dataset containing approximately 6,500 German court decisions paired with their official press releases. This dataset enables research in automatic summarization, controlled text generation, and evaluation of language models in the legal domain. We describe the data collection, cleaning processes, and present baseline results using various language models for generating press releases from court decisions. Additionally, we provide a comprehensive evaluation framework featuring both automatic metrics and human assessments. CourtPressGER serves as a valuable resource for researchers and practitioners working on legal NLP applications in German, addressing the gap in language-specific legal datasets and enabling new approaches to legal text generation and analysis.

CCS Concepts: • **Information systems** → **Data sets**; • **Human-centered computing** → *Natural language interfaces*; • **Applied computing** → *Law*; • **Computing methodologies** → *Natural language generation*.

Additional Key Words and Phrases: Legal NLP, German language resources, court decisions, press releases, dataset, baseline models, text generation

ACM Reference Format:
Technical University of Munich. 2024. CourtPressGER: A Dataset of German Court Decisions and Press Releases with Baseline Models. In *Proceedings of ACM Conference on Research Data (Dataset 2024).* ACM, New York, NY, USA, 6 pages. https://doi.org/XX.XXXX/XXXXXXX.XXXXXXX

1 Introduction

Natural Language Processing (NLP) applications in the legal domain have gained significant attention in recent years, with tasks such as legal document classification, information extraction, and text summarization becoming increasingly important for legal professionals and researchers [3]. However, the development of effective legal NLP tools is often hindered by the scarcity of high-quality, domain-specific datasets, particularly for languages other than English [2].

To address this limitation, we present CourtPressGER, a novel dataset containing approximately 6,500 German court decisions paired with their corresponding official press releases. This unique resource offers several advantages for researchers:

- A substantial collection of professional summaries created by legal experts
- Paired texts that demonstrate how complex legal reasoning is condensed into concise, accessible language
- Rich metadata including court information, decision dates, and case references
- A clean, processed dataset ready for use in various NLP tasks

Author's Contact Information: Technical University of Munich, contact@tum.de, School of Computation, Information and Technology, Munich, Germany.

Manuscript submitted to ACM

Furthermore, we provide baseline results for generating press releases from court decisions using various language models, ranging from large proprietary models (like GPT-4o) to smaller open-source models fine-tuned for German legal text (such as Teuken-7B and EuroLLM-9B). We evaluate these baselines using both automatic metrics and human assessments to establish benchmark performance on this dataset.

The paper is structured as follows: Section 2 discusses related work in legal NLP datasets and text generation. Section 3 describes the dataset collection and cleaning process. Section 4 presents our baseline models and generation approach. Section 5 details the evaluation framework and results. Finally, Section 6 concludes with implications and future work.

2 Related Work

2.1 Legal NLP Datasets

Legal NLP has emerged as a specialized field within natural language processing, focusing on applications tailored to legal documents and processes. Several legal datasets have been released in recent years, but most focus on English language materials, such as the Legal BERT corpus [3], which contains over 12 million legal documents from various sources, or the SCOTUS dataset [2], which includes U.S. Supreme Court opinions.

For languages other than English, resources are considerably scarcer. Notable exceptions include CAIL2018 [8], a Chinese dataset for legal judgment prediction, and JurisCorpusFR [7], a corpus of French legal documents. For German legal texts, LegalBERT-German [6] was introduced, but it focuses on model training rather than providing a comprehensive dataset for specific tasks.

2.2 Text Generation in the Legal Domain

Text generation in the legal domain presents unique challenges due to the need for factual accuracy, adherence to legal principles, and clear reasoning. Previous work has explored various applications such as contract generation [4], legal document summarization [5], and judgment prediction [1].

However, the generation of press releases from court decisions represents a relatively unexplored area, particularly for non-English languages. This task is complex as it requires not only summarizing the decision but also adapting the language for public consumption while preserving legal accuracy.

3 Dataset Description

3.1 Data Collection

The CourtPressGER dataset was collected from publicly available sources, primarily from the official websites of German courts at various levels, including:

- Federal Constitutional Court (Bundesverfassungsgericht)
- Federal Court of Justice (Bundesgerichtshof)
- Federal Administrative Court (Bundesverwaltungsgericht)
- Federal Labor Court (Bundesarbeitsgericht)
- Federal Social Court (Bundessozialgericht)
- Federal Finance Court (Bundesfinanzhof)

The initial raw dataset contained over 7,000 entries, each consisting of a court decision and its corresponding press release published by the court. The data spans multiple years and covers a wide range of legal areas, including constitutional law, civil law, criminal law, administrative law, labor law, social security law, and tax law.

3.2 Data Cleaning and Processing

We implemented a robust cleaning pipeline to ensure data quality and consistency. The cleaning process involved several steps:

(1) Removal of duplicate entries and entries with missing decision texts or press releases
(2) Normalization of text formatting, including removal of excessive whitespace, standardization of paragraph breaks, and proper handling of special characters
(3) Rule-based filtering to exclude entries with problematic content (e.g., extremely short or truncated texts)
(4) Semantic similarity analysis to validate the pairing between decisions and press releases
(5) Extraction and standardization of metadata from decision headers and footers

After cleaning, the final dataset contains approximately 6,500 high-quality pairs of court decisions and press releases, with comprehensive metadata for each entry.

3.3 Dataset Structure

Each entry in the CourtPressGER dataset contains the following elements:

- **Decision text:** The full text of the court decision, often ranging from 5,000 to 50,000 words
- **Press release:** The official press release published by the court, typically 500-1,500 words
- **Metadata:**
  - Court name and level
  - Decision date
  - Case reference number
  - Legal area
  - Panel/chamber information
  - Publication date of press release

The dataset is provided in multiple formats:

- JSON files with complete entries
- CSV files for tabular analysis
- Hugging Face dataset for easy integration with NLP pipelines

3.4 Dataset Statistics

The CourtPressGER dataset exhibits the following key statistics:

Table 1. Dataset Statistics

Statistic	Value
Total number of document pairs	6,489
Average decision length (words)	7,842
Average press release length (words)	754
Average compression ratio	10.4:1
Covered legal areas	18
Covered time period	1998-2023

4 Baseline Models and Generation

To establish baseline performance for the task of generating press releases from court decisions, we implemented a comprehensive generation pipeline using various language models.

4.1 Models

We evaluated three categories of language models:

(1) **Large proprietary models:**
   - GPT-4o (OpenAI)
   - Llama-3-70B (via DeepInfra API)

(2) **Medium-sized open models:**
   - Llama-3-8B
   - Teuken-7B (specialized for German)

(3) **Small open models:**
   - EuroLLM-9B (multilingual European model)

4.2 Synthetic Prompts

For each model, we developed synthetic prompts following various strategies:

- **Basic prompting:** Direct instruction to generate a press release
- **Role-based prompting:** Framing the task as being a court press officer
- **Format-guided prompting:** Providing examples of desired output format
- **Context enhancement:** Including metadata in the prompt

Example of a synthetic prompt:

```
Du bist ein Pressesprecher eines deutschen Gerichts.
Erstelle eine Pressemitteilung zum folgenden Gerichtsurteil.
Die Pressemitteilung sollte die wichtigsten Fakten und
Entscheidungsgründe enthalten und für die Öffentlichkeit
verständlich sein.

Aktenzeichen: 1 BvR 2456/18
Gericht: Bundesverfassungsgericht
Datum: 12.03.2022

[Urteilstext folgt hier]
```

4.3 Generation Pipeline

Our generation pipeline consists of several components:

(1) **Input preprocessing:** Court decisions are truncated to fit model context windows, with priority given to introductory sections, key legal reasoning, and the decision (tenor)

(2) **Prompt construction:** Synthetic prompts are combined with preprocessed decisions

(3) **Generation:** Models generate press releases with standardized parameters (temperature=0.7, top_p=0.9)
(4) **Output postprocessing:** Generated texts are cleaned and formatted consistently

The pipeline includes checkpoint functionality to handle interruptions in long processing runs and supports various output formats for further analysis.

5 Evaluation Framework and Results

We developed a comprehensive evaluation framework to assess the quality of generated press releases compared to the original ones.

5.1 Automatic Metrics

We implemented several automatic metrics to evaluate different aspects of the generated texts:

- **Lexical similarity:**
  - ROUGE (ROUGE-1, ROUGE-2, ROUGE-L)
  - BLEU (BLEU-1 through BLEU-4)
  - METEOR
- **Semantic similarity:**
  - BERTScore (using EuroBERT model)
  - Embedding similarity (using sentence transformers)
- **Factual accuracy:**
  - Named entity recognition accuracy
  - Legal reference accuracy

5.2 Human Evaluation

In addition to automatic metrics, we conducted a human evaluation with legal experts and laypeople. The evaluation assessed:

- **Legal accuracy:** Correctness of legal reasoning and facts
- **Completeness:** Coverage of key information from the decision
- **Readability:** Clarity and accessibility for non-legal audiences
- **Coherence:** Logical flow and organization
- **Overall quality:** General assessment of the press release

5.3 Baseline Results

The baseline results showed significant variations across models and evaluation metrics:

Human evaluation results showed similar trends but highlighted important nuances:

- Larger models generally produced more accurate and complete press releases
- German-specialized models (Teuken-7B) outperformed general models of similar size on legal terminology accuracy
- All models occasionally generated factually incorrect information, with smaller models doing so more frequently
- Human experts rated press releases from larger models as comparable to human-written ones in readability, but lower in legal accuracy

Table 2. Automatic Evaluation Results

Model	ROUGE-L	BLEU-4	BERTScore	Fact Acc.
GPT-4o	0.42	0.25	0.85	0.78
Llama-3-70B	0.38	0.22	0.83	0.74
Llama-3-8B	0.29	0.15	0.75	0.61
Teuken-7B	0.33	0.18	0.79	0.65
EuroLLM-9B	0.27	0.13	0.73	0.59

6 Conclusion and Future Work

The CourtPressGER dataset provides a valuable resource for researchers working on legal NLP applications in German. Our baseline experiments demonstrate that while current language models can generate reasonable press releases from court decisions, there remains significant room for improvement, particularly in factual accuracy and legal reasoning.

Future work could explore:

- Fine-tuning language models specifically for legal press release generation
- Developing specialized evaluation metrics for legal text generation
- Creating hybrid systems that combine extraction and abstraction approaches
- Extending the dataset with additional court levels and legal domains

The CourtPressGER dataset, cleaning pipeline, baseline models, and evaluation framework are all made available to the research community to facilitate further work in this important area.

Acknowledgments

We would like to thank the Technical University of Munich for supporting this research. We also acknowledge the courts that make their decisions and press releases publicly available, enabling research that can enhance access to justice through technology.

References

[1] Nikolaos Aletras, Dimitrios Tsarapatsanis, Daniel Preoţiuc-Pietro, and Vasileios Lampos. 2016. Predicting judicial decisions of the European Court of Human Rights: A natural language processing perspective. *PeerJ Computer Science* 2 (2016), e93.

[2] Michael J. Bommarito, Daniel Martin Katz, and Eric M. Detterman. 2021. LexNLP: Natural language processing and information extraction for legal and regulatory texts. *Research Handbook on Big Data Law* (2021).

[3] Ilias Chalkidis, Manos Fergadiotis, Prodromos Malakasiotis, Nikolaos Aletras, and Ion Androutsopoulos. 2020. LEGAL-BERT: The Muppets straight out of Law School. *Findings of the Association for Computational Linguistics: EMNLP 2020* (2020), 2898–2904. doi:10.18653/v1/2020.findings-emnlp.261

[4] Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. Measuring Massive Multitask Language Understanding. *Proceedings of the International Conference on Learning Representations (ICLR)* (2021).

[5] Ambedkar Kanapala, Sukomal Pal, and Rajendra Pamula. 2019. Text summarization from legal documents: a survey. *Artificial Intelligence Review* 51, 3 (2019), 371–402.

[6] Elena Leitner, Georg Rehm, and Julian Moreno-Schneider. 2022. A Dataset of German Legal Documents for Named Entity Recognition. In *Proceedings of the Language Resources and Evaluation Conference*. 4590–4598.

[7] Antoine Louis, Adrien Labbé, Mathilde Réty, and Adeline Nazarenko. 2022. JurisCorpusFR: A balanced dataset of French legal documents. *Legal Knowledge and Information Systems* (2022), 113–117.

[8] Chaojun Xiao, Haoxi Zhong, Zhipeng Guo, Cunchao Tu, Zhiyuan Liu, Maosong Sun, Yansong Feng, Xianpei Han, Zhen Hu, Heng Wang, et al. 2018. CAIL2018: A Large-Scale Legal Dataset for Judgment Prediction. In *The 17th China National Conference on Computational Linguistics*.