Ethics Statement - Supreme Court Judgements Chunked

By - Vihaan Nama

Data Privacy and Anonymization

All judgment data will be sourced from publicly accessible platforms, such as the Supreme Court of India's website. No personal or sensitive information about individuals involved in the cases will be included that is not already present on the Supreme Court of India’s website, ensuring compliance with privacy laws and ethical standards.

Responsible Use

The dataset is intended solely for research and technological advancements in legal applications. Any misuse, such as for unethical profiling or unauthorized commercial purposes, will be explicitly prohibited in the terms of use.

Transparency and Reproducibility

The methods used for data collection, chunking, and embedding will be documented comprehensively to promote transparency. The dataset and code will be made publicly available through platforms like HuggingFace, enabling reproducibility and fostering open collaboration.

Bias and Fairness

Care will be taken to ensure the dataset does not reinforce or introduce biases inherent in the source material. Regular audits will be conducted to identify and mitigate any potential biases in the processed data whenever the data is reloaded.

Respect for Legal Frameworks

This project will strictly adhere to all applicable laws, including those governing intellectual property and access to government data. Efforts will align with the "eCourts Initiative," which promotes technology development for the Indian judiciary.

Minimizing Harm

The project will prioritize minimizing harm by preventing data misuse and avoiding

unintended consequences, such as misinterpretation of legal judgments due to incorrect data chunking or embeddings. No data has been added explicitly, all data in chunks have been mined from the original pdf documents.

Code Details and Methods

The project employs several strategies to preprocess and chunk textual data, ensuring optimal structure for LLM applications:

- `Recursive Character Chunking`: Uses the `RecursiveCharacterTextSplitter` with parameters (1000 characters per chunk, 200-character overlap) to create chunks while maintaining context.
- Token-Wise Chunking: Implements the `TokenTextSplitter` (100 tokens per chunk, 20-token overlap) for fine-grained segmentation based on token count.
- Semantic Chunking: Utilizes the `SemanticChunker` powered by OpenAI embeddings to split text into semantically coherent units.

The process also includes cleaning text to remove invisible and non-standard characters, enhancing the quality and utility of the dataset.

Automation and Transparency

The provided Python scripts automate the workflow, from extracting text from PDFs using `pdfplumber` to chunking with advanced text splitters. The source code will be made publicly available, ensuring transparency in data processing methods.

Ethical Data Processing

The project uses publicly available Supreme Court judgment PDFs, with no modifications to original legal content. Preprocessing steps strictly remove hidden or extraneous characters without altering the legal meaning or structure.

Data Integrity

By using semantic chunking and embedding methods, the dataset preserves the context and logical structure of legal judgments, ensuring that the processed data remains meaningful and accurate.

Responsible Use and Sharing
All datasets and associated code will be shared under appropriate licenses - MIT License that prohibit misuse, including unethical profiling or discriminatory applications. The emphasis will be on research and development to assist in reducing court backlogs.

Bias Mitigation and Fair Representation
The chunking algorithms are applied uniformly across all data, minimizing the risk of selective bias. Semantic processing aims to enhance data consistency and usability across diverse legal scenarios.

Ethics of the MIT License

1. Freedom to Use, Modify, and Distribute
    - The MIT License allows anyone to use, modify, and distribute the licensed software, whether for private, commercial, or academic purposes.
    - Ethical Implication: This aligns with the principle of knowledge sharing and the democratization of technology, fostering innovation and collaboration.
2. Attribution Requirement
    - The license requires users to include the original copyright notice and a copy of the license in distributed software.
    - Ethical Implication: This ensures proper credit is given to the original creators, recognizing their contributions and promoting transparency.
3. No Liability or Warranty
    - The license explicitly disclaims warranties and liability, meaning users take full responsibility for how they use the software.
    - Ethical Implication: While this protects developers from legal risks, it shifts the responsibility to users, who must ethically consider the impact of their use of the software.

4. Lack of Restrictions on Usage
    - The permissive nature of the MIT License allows the software to be incorporated into both open-source and proprietary projects.
    - Ethical Implication: This flexibility can lead to ethical dilemmas, such as the software being used for purposes the original developers might find objectionable (e.g., surveillance, weapons development). Developers using the MIT License should be aware of this possibility and decide whether they are comfortable with it.
5. Promotion of Open Collaboration
    - The license encourages a culture of openness by removing barriers to adoption and modification.
    - Ethical Implication: This supports the global sharing of technology and ideas, benefiting both the tech community and society at large.

Parts of this Document were generated using AI softwares like ChatGPT