LEGALEASE: SIMPLIFYING LEGAL UNDERSTANDING WITH AI

20BCE2661, 20BCE2770, 20BCE2908 | Mylie E Mudaliyar, Diksha Adhikari, Muskan Sah | Dr. Akila Victor | SCOPE

Introduction

LegalEase, a comprehensive AI driven platform applies AI to simplify legal documents through summarization, translation, and text-to-speech, bridging gaps in prior work's limited AI/NLP use, lack of integrated accessibility features, inadequate legal text improvement suggestions, insufficient tailored evaluation metrics, and supervised models' limited effectiveness for legal language complexity. This comprehensive platform democratizes legal information access.

Motivation

Navigating the intricate legal landscape is a daunting task for non-experts due to the complex terminology and structure of legal documents. LegalEase is motivated by the need to democratize access to legal information by leveraging AI to simplify legal texts, enabling individuals to make informed decisions. It aims to promote transparency, fairness, and inclusivity by integrating translation and accessibility features.

Scope of the Project

LegalEase is an AI-powered platform that applies state-of-the-art natural language processing techniques, including the Pegasus transformer model, to generate concise and comprehensible summaries of legal documents. It encompasses document processing, summarization, language translation, text-to-speech conversion, and user-centric design to cater to diverse audiences and accessibility needs.

Methodology

1. Model Selection and Justification:

- The **Pegasus model** was chosen for legal document summarization due to its robust capabilities in handling complex language structures and its optimization for summarization tasks. Compared to alternative models considered, Pegasus demonstrated superior performance in capturing the nuances of legal language while generating concise summaries. It employs a transformer-based architecture, leveraging self-attention mechanisms to capture long-range dependencies in the input sequence which makes it well-suited for processing legal corpora.

2. Data Preprocessing:

- A diverse corpus of **Indian legal cases(ILC)** containing **3,073 court cases** is curated that are split into **2,058 training cases and 1,015 test cases**. The train and test cases consist of three features: Title, Summary and Case. The box plot and histogram diagrams provide substantial evidence of balanced partitioning in the dataset. However, the dataset portrays some outliers with very extensive text going over 6,400 words per Case and over 850 words per Summary. Hence, we dropped the cases and their corresponding summaries containing outliers.

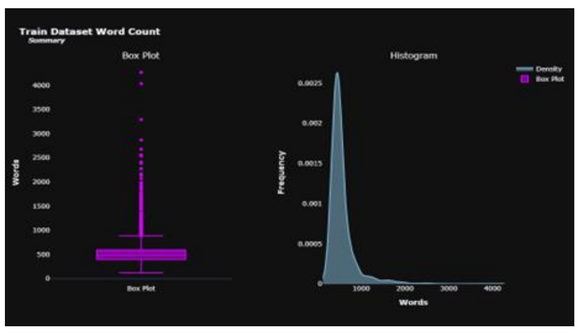


Fig. 1(a). Visual Representation of the training dataset

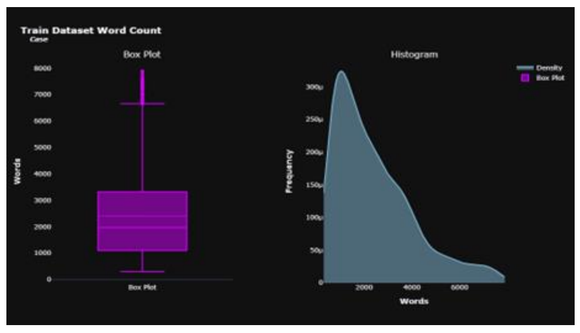
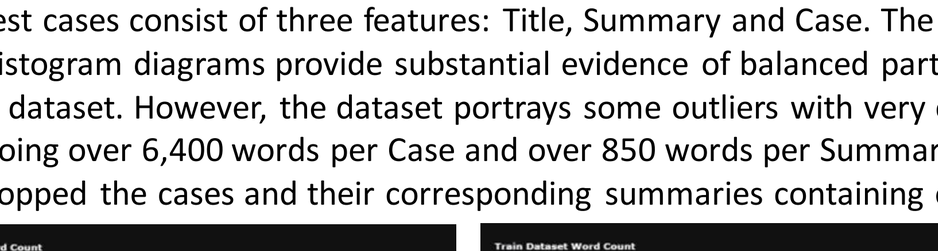


Fig. 1(b). Visual Representation of Summary column from Case column from the training dataset

3. Fine-tuning Strategy :

- A pretrained model name **Pegasus-billsum** is selected which undergoes fine-tuning on the dataset. Before fine-tuning, the data is converted into the format that the model understands i.e. tokenizing and cleaning the data.



4. Evaluation Metrics and Results :

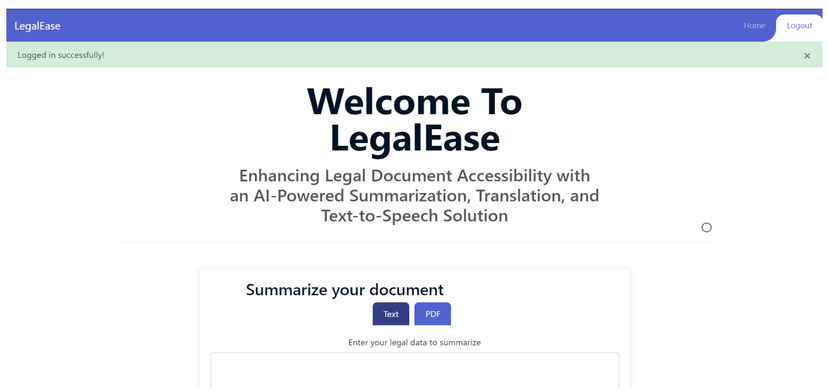
- The ROUGE scores assess the Pegasus model's performance on the ILC dataset using **ROUGE-1, ROUGE-2, and ROUGE-L** metrics. While the model effectively captures individual words (ROUGE-1), it faces challenges with longer sequences and common subsequences (ROUGE-2 and ROUGE-L). Scores decrease from training to test data, indicating generalization difficulties. These findings suggest areas for improvement, possibly due to dataset complexity or model limitations.

R1-recall $\longrightarrow \frac{number\_of\_overlapping\_words}{total\_words\_in\_reference\_summary}$

R1-precision $\longrightarrow \frac{number\_of\_overlapping\_words}{total\_words\_in\_system\_summary}$

5. Deploying the model :

- A user-friendly interface is developed to provide seamless access to the summarization service. The interface allows users to input legal query and receive a summarized output. Additionally, features such as language translation and audio output are integrated to enhance accessibility and usability.



Results

S.no.	ORIGINAL TEXT	SUMMARY GENERATED
1.	Non Reportable IN THE CRIMINAL APPELLATE JURISDICTION Criminal Appeal No. 319 Viram @ Virma .... Appellant The State of Madhya Pradesh â€¦. Respondent W I T H Criminal Appeal No.319 JUDGMENT L. NAGESWARA RAO J. 1. These Criminal Appeals are directed against the judgment of the High Court of Madhya Pradesh at Gwalior by which the conviction of the Appellants under Sections 147 302 149 324 149 323 149 and their sentences were upheld. ..... (truncated due to length)	The Madhya Pradesh High Court has dismissed the appeal filed by the state government against the conviction of 21 accused in the murder case of Babulal Lodha. The trial court had held the accused guilty of committing offences under Sections 147 302 149 149 324 149 323 149 and their sentences were upheld. The appellants contended that there is a contradiction between the ocular testimony of the witnesses and the medical evidence.

The Pegasus model's performance on the ILC dataset, evaluated using ROUGE metrics, shows promising results in capturing individual words from reference summaries. However, it struggles with longer sequences and maintaining context, as seen in lower ROUGE-2 and ROUGE-L scores, particularly on the test data. This suggests a need for improvement in coherence and capturing key points. Discrepancies between training and test scores hint at potential overfitting, likely due to the complexity of legal language and limited training data.

Conclusion

LegalEase employs AI to simplify legal documents, aiming to make them more understandable to the ordinary people. Although promising, it needs improvement in capturing longer text and improving translation accuracy. Future plans include refining techniques, expanding datasets, and adding language support to empower individuals by democratizing legal information access. Additionally, advanced NLP techniques, legal databases, and privacy measures will be integrated for enhanced user experience and security.

References

1. Jingqing Zhang, Yao Zhao, Mohammad Saleh, Peter J. Liu (2020 July). PEGASUS: Pre-training with Extracted Gap-sentences for Abstractive Summarization arXiv:1912.08777
2. Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, Illia Polosukhin (2017 Jun). Attention Is All You Need. arXiv:1706.03762
3. Mahesan, N., Hettiarachchi, S., Abeysinghe, E., & Sandaruwan, D. (2022). Application of Deep Learning Techniques for Legal Document Summarization. arXiv preprint arXiv:2211.12847.
4. Wu, J., Shang, Z., Li, H., Chen, X., & Dong, W. (2022). Legal Pre-trained Language Models for Legal Document Understanding. arXiv preprint arXiv:2211.02596.
5. Wolf T, Debut L, Sanh V, Chaumond J, Delangue C, Moi A, Cistac P, Rault T, Louf R, Funtowicz M, Brew J (2019) HuggingFace's Transformers: State-ofthe-art Natural Language Processing. CoRR abs/1910.03771: