Legal Jargon Simplifier – Project Documentation

1. Overview

Legal documents such as Terms & Conditions are often written in complex and technical language, making them difficult for non-legal users to understand. This project aims to bridge that gap by converting legal jargon into clear, concise, and user-friendly English using a transformer-based Natural Language Processing (NLP) model.

The project demonstrates how transformer architecture and pretraining objectives influence real-world NLP applications, particularly in tasks that involve rewriting and simplifying text.

2. Key Concepts Understood

2.1 Transformer Architectures

Transformer models differ significantly based on their architecture. Encoder-only models such as BERT and RoBERTa are optimized for understanding tasks like classification, question answering, and masked language modeling. In contrast, encoder–decoder models such as BART are designed for sequence-to-sequence tasks where input text must be transformed into a different textual form.

This project reinforced the understanding that text rewriting and simplification require a decoding mechanism, which encoder-only models lack.

2.2 Abstractive vs Extractive Summarization

Summarization techniques can be broadly classified into extractive and abstractive methods. Extractive summarization selects and rearranges sentences directly from the original text, whereas abstractive summarization generates new sentences that convey the core meaning of the input.

This project uses abstractive summarization, enabling the system to rewrite complex legal language into simpler English rather than merely shortening the original text.

2.3 Model Selection Rationale

The model facebook/bart-large-cnn was selected because it aligns closely with the project requirements. It employs an encoder–decoder architecture and is fine-tuned on large-scale summarization datasets, allowing it to generate fluent, coherent, and human-readable summaries.

This selection highlights the importance of choosing models whose training objectives match the intended task.

2.4 Practical Limitations of Generic Models

While generic summarization models perform well on general text, legal documents contain domain-specific terminology and nuanced clauses. This can occasionally lead to loss of legal nuance or oversimplification.

The project highlights the importance of domain-adapted models for high-stakes applications such as legal and medical text processing.

3. Solution Implemented

3.1 System Input

The system accepts paragraphs extracted from legal documents such as Terms & Conditions and privacy policies. These inputs typically contain complex clauses, liability disclaimers, and formal legal language.

3.2 Processing Pipeline

The input legal text is passed through a summarization pipeline built using the BART Large CNN model. The encoder processes the text to understand its semantic and contextual meaning, while the decoder generates a simplified and concise version of the original content in plain English.

3.3 Output

The output is a rewritten version of the legal text that preserves the original meaning while significantly improving readability. The simplified text is easier for non-technical users to understand without requiring legal expertise.

3.4 Tools and Technologies Used

- Python
- Hugging Face Transformers library
- facebook/bart-large-cnn model
- Google Colab
- PyTorch backend

4. Outcome

The system successfully simplifies complex legal paragraphs into clear and accessible summaries. The project demonstrates that selecting the appropriate transformer architecture has a direct and significant impact on model performance and output quality.

5. Conclusion

This project strengthened the understanding of transformer architectures, particularly the differences between encoder-only and encoder–decoder models. It also provided practical experience in abstractive summarization and real-world NLP system design.

The Legal Jargon Simplifier serves as a practical example of how deep learning models can be applied to improve accessibility and comprehension of complex legal content.