

PROJECT PROPOSAL

LegalEase: Summarization of legal documents and query-answering chatbot

Kunal (2021330)

Ishan (2021465)

Aman Ranjan (2021376)

Megha (2021337)

10th March, 2024

Abstract

The project defines the issues consumers face when dealing with complex legal documents and getting trapped in unknown agreements. LegalEase aims to turn these documents into clear, comprehensible summaries and offer an interactive chatbot for customized inquiries. We have discussed various approaches to train our model and the features we plan to add.

Introduction

The project started with identifying the problem first. It proceeded with reading research papers and articles to get ideas of existing technologies used for summarizing and how to build upon them using fine-tuning.

Problem Statement

It takes a lot of time and confusion to sort through the complex legal documents;


The process is irritating. In particular, when it comes to contracts, loans, and real estate transactions, current solutions fall short of providing user-friendly experiences. This lack of simplicity and clarity is a significant obstacle to making confident, well-informed decisions in essential transactions.

- Time-consuming
- Confusing
- Contracts
- Unapproachable
- User-friendly
- Lack of simplicity
- Obstacle
- Intimidating
- Loans
- Real estate
- Informed decisions
- Lack of clarity

Importance of the Problem

Most of us have faced issues related to this problem. Let's see what Kunal faced

“A few years ago, my family bought an apartment. We were told that everyone would share the terrace. This was important to us. Later, the builders decided to add another floor and give the terrace to the top floor. When we complained, they showed us the agreement. It was 50 page



long and we needed to read it all carefully. It said that the builders could make these changes. We were disappointed but couldn't do anything.

This experience made us realize how hard it is to understand legal documents.”

A wide range of people are impacted by the challenge of understanding legal documents, including small businesses without dedicated legal departments and individuals engaged in personal transactions. Failing to comprehend the terms and conditions fully can have negative consequences, such as monetary losses and legal disputes. As a result, making it easier to understand these documents is vital for financial stability, legal safety, and convenience.

Related Work

Several services and tools, such as chatbots that provide legal advice and document summarization tools, aim to demystify legal documents. Nevertheless, these solutions frequently need to offer interactive, document-specific guidance and customized, thorough summaries.

Most currently available tools are generic and need to adjust to the particular complexities of individual documents. Moreover, they change the real meaning of the legal thing.

Some students have worked on documents related to the judgment of Indian courts and made a summarization tool. However, we would like to train better in another legal domain and integrate a chatbot.

Proposed Solution and Novelty

LegalEase is an innovative tool designed to simplify complex legal documents. It transforms contracts, loans, and real estate agreements into concise, easy-to-understand summaries. Users receive a clear breakdown of the key points and terms by uploading PDFs or scanning images of legal documents. We will also create an interactive chatbot tailored to each document, allowing users to ask specific questions and receive instant, understandable answers. This approach saves time and gives users confidence and clarity. In the future, we also plan to integrate it using a website extension.

Novelty: Existing products like AI, GPTs, and others are primarily trained in foreign databases. But in the Indian context there are less similar things. And also, there will be an option for multi-language support. People also sign on to various T&Cs online, a website extension will also be provided to take care of scams.

Technologies

Using information retrieval, we will build LegalEase. These techniques will include:

- Document summary algorithms to extract essential terms and points.

- Machine learning models that can comprehend and respond to user inquiries about the document in chatbot encounters. We are thinking of using torch,BERT,transformers and similar things
- Optical character recognition (OCR) technology creates machine-readable text from scanned documents and photos.
- A front-end app or a website.
- Creation of website extension.

Evaluation

Our evaluation approach includes

- User studies that assess time savings and comprehension improvements.
- There are some methods listed in the research paper listed below,we have used ROUGE scores as it is best for Natural languages.
- Usability testing is necessary to ensure the interface is accessible and easy to use.
- Many machine learning techniques can also be used to check the efficiency.

Contributions

- Each 4 of us will be involved in Model Training.
- Research papers were divided equally.
- Designing an app or a website and inking backend with front-end
- Website extension

Updated Literature Review

1.Indian Legal Text Summarization: A Text Normalization-based Approach: Satyajit Ghosh, Mousumi Dutta, and Tanaya Das' research work addresses the difficulty of legal text summary in India. Due to lack of summarized Indian legal documents dataset, they proposed a domain independent approach. The steps involve text extraction and cleaning,text normalization, then fragmentation and then these fragments are given as input to the model.Finally, all the outputs are merged.They are using the BART model as PEGASUS performs poorly.

Even without domain-specific training data, their experiments reveal that summarization efficacy improves significantly.

Conclusion : Not much useful for **abstractive summarization**, this method loses semantic method.It can also lose semantic meaning.

<https://arxiv.org/pdf/2206.06238.pdf>

2.The Right to remain plain : Summarization and simplification of legal : This stanford paper discusses about the existing work in the domain of text summarization and how to build upon these technologies using fine tuning within the legal domain.Also it further opens the possibilities of training the models on more specific domains in law. It uses BART as an example.

[Stanford Paper](#)

3.Legal Case Document Summarization: Extractive and Abstractive Methods and their Evaluation:

The paper written by Abhay Shukla and others addresses challenges in legal case summarization.This paper compares the extractive and abstractive models based on their performance on legal documents. The challenge with the abstractive method is the token limit. In Abstraction models ,the Chunking method is used to process large amounts of data into chunks and then pass through the model. Legal Pegasus gave the best result followed by BART. But for segment wise evaluation, Legal-LED performed better .In extractive methods, DSDR and SummaRuNNer are better.

Conclusion : Abstractive summarization method has limited tokens. Although chunking overcomes this, we still end up getting incomplete and unorganized output. Extractive methods represent final judgment and status better but misses important precedents and arguments.

<https://aclanthology.org/2022.aacl-main.77.pdf>

4. Summarization of Legal Documents: Where are We Now and the Way Forward :

Model:The model employs various machine-learning models, including gradient boosting, multilayer perceptrons (MLPs), and deep learning methods with long short-term memory (LSTM) units. These models were utilized to tackle the problem of generating gist statements from judgment documents as a sentence classification issue.

Evaluation:The performance of the models was assessed using precision, recall, and F1 measure. The best result achieved was an F1 measure of 0.9372, indicating high accuracy in selecting sentences that constitute the gist of legal documents.

Techniques:The techniques involved in the models' development include the use of legal linguistic statistical information, different word embedding methods for feature extraction, and the application of ensemble methods to combine predictions from multiple models. The research also explored the importance of contextual information surrounding sentence segments, employing LSTM units to capture such context effectively.

Conclusion:

Extractive summarization may miss the broader context of legal documents which can lead to wrong conclusions. The complex language and terminology is also an issue. Also some other evaluation metrics are needed, to capture quality. And it is country specific.

<https://www.sciencedirect.com/science/article/abs/pii/S1574013721000289>

5. Extracting the Gist of Chinese Judgments of the Supreme Court:

Model: It takes lengthy legal documents as input, preprocesses them by tasks like tokenization and sentence segmentation, then utilizes various summarization techniques, including extractive and abstractive methods, to generate concise summaries. Extractive summarization involves selecting the most important sentences or passages from the original document based on criteria like relevance and importance, while abstractive summarization may involve generating new sentences that capture the essence of the text.

Evaluation: The generated summaries are evaluated using metrics like ROUGE scores to assess their quality compared to reference summaries. It measures the overlap between system-generated summaries and human-generated reference summaries based on various units such as word pairs, n-grams, and word sequences.

Techniques: The paper discusses both domain independent and domain specific legal document summarization techniques. Despite the challenges of unique characteristics of legal documents and diverse structure and heavy usage of citations, the paper mentions the existence of tools like CaseSummarizer, which utilize word frequency and domain-specific knowledge to produce summaries tailored for legal professionals. Additionally, it presents two case studies focusing on automatic summarization of legal documents from different countries, providing a comparative analysis of various summarization techniques.

Conclusion: Deep learning methods face challenges such as reliance on extensive data, overfitting, and opaque decision-making processes. It suggested some areas of improvement and advance NLP and ML research that is continuously going.

<https://dl.acm.org/doi/10.1145/3322640.3326715>

Models Used

1. **Legal-Pegasus** : The existing model has the tokenization limit of 1024. So, we divided the complete dataset into chunks of 1000 tokens per dataset. Then generated the summary of each chunk. Finally we combined all the summaries and processed the new summary. This new summary is developed using all the outputs.

Final Summary:

It raises the question whether municipal property tax and urban immovable property tax payable under the relevant Bombay Acts are allowable deductions under section 9 (1) (iv) of the Indian Income tax Act. The assessee company is an investment company deriving its income from properties in the city of Bombay.

For the assessment year 1940-41, the net income of the assessee under the head "property" was computed by the Income tax Officer in the sum of Rs. 6,21,764 after deducting from gross rents certain payments.

The court held that the income from the property was not subject to the tax provisions of Section 9 (1) of the Income Tax Act and therefore the tax officer had no power to determine the amount of the tax payable.

The appeal was filed by K.M. Munshi (N.P. Nathvani) and M.C. Setalvad, Attorney General for India (H.J. Umrigar). A company had paid during the relevant year Rs. 1,22,675 as municipal property tax and Rs. 32,760 as urban property tax.

The Tribunal, however, agreed to refer two questions of law to the High Court of Judicature at Bombay, namely, (1) Whether the municipal taxes paid by the applicant company are an allowable deduction under section 55 of the provisions of section 9 (1) (iv) of the Indian Income tax Act; (2) Whether the urban immovable property taxes paid for the applicant company.

The question for our determination is whether the municipal property tax and urban immovable property tax can be deducted as an allowance under clause (iv) of sub-section (1) of section 9 of the Act. The decision of the point depends firstly on the construction of the language employed in sub-clause and secondly, on a finding as to the true nature and character of the liability of the owner under the relevant Bombay Acts for the payment of these taxes.

The court held that the tax shall be payable by an assessee under the head 'income from property' in respect of the bona fide annual value of property consisting of any buildings or lands appurtenant thereto of which

2. Facebook BART: It is also a summarisation tool but based on a transformer model and developed by facebook. It has a token limit of around 512. We divided For this long text ,the generated summary was:

```
Methods Methods Methods strategy strategy strategy renovations renovations
renovations797797797ConsoleConsoleConsole295295295JohnsonJohnsonJohnson masturb masturb masturbromisingromisingromisingGuardGuardGuard
Cellular Cellular CellularJohnsonJohnson touting touting toutingcicici hates hates hates412412412 hates hatesJohnsonJohnson reunit
reunit reunit TLS TLS TLS masturb masturb ENTER ENTER ENTER renovations renovations cultural cultural cultural categ categ categ
sizeable sizeable sizeableJohnsonJohnsonShopShopShopGuardGuard Hobbit Hobbit Hobbit EA EA EA hates hates Unleashed Unleashed Unleashed sim
sim simkgkgJohnsonJohnsonGuardGuardromisingromising144144144 climates climates climates144144 theaters theaters theaters Zero Zero Zero
masturb masturb employed employed employed masturb masturb AMA AMA AMA masturb masturbJohnsonJohnson masturb
masturbsendsendsends144144 hates
```

Time taken was: 2m 26 s

This Result was quite bad.

3. T5-transformer :

Using this the results were :

Generated Summary: The appellant was the tenant of a flat on the second floor of a building in Bombay. The tenancy was a monthly one, the rent being Rs. 215. The appellant wanted to go to the United Kingdom for treatment of his failing eye sight and he got into touch with the complainant through one Sayed for the purpose of making necessary arrangements about the flat occupied by him in view of his intended departure. The prosecution case was that the appellant demanded a sum of Rs. 30,000 which was later on reduced to Rs. 29,500 as consideration for putting the complainant in vacant possession of the flat and an additional amount of Rs. 2,000 for the furniture and

Out of the 3 ,we found that T5-Transformer and Legal Pegasus can be used .

Baseline Result :

```
ROUGE-1 Precision: 0.1313  
ROUGE-1 Recall: 0.7600  
ROUGE-1 F1 Score: 0.2240
```

```
ROUGE-2 Precision: 0.0568  
ROUGE-2 Recall: 0.4252  
ROUGE-2 F1 Score: 0.1002
```

```
ROUGE-L Precision: 0.1198  
ROUGE-L Recall: 0.6933  
ROUGE-L F1 Score: 0.2043
```

```
... Average ROUGE-1 Precision: 0.6228  
Average ROUGE-1 Recall: 0.1322  
Average ROUGE-1 F1 Score: 0.2083
```

```
Average ROUGE-2 Precision: 0.3264  
Average ROUGE-2 Recall: 0.0544  
Average ROUGE-2 F1 Score: 0.0888
```

```
Average ROUGE-L Precision: 0.5856  
Average ROUGE-L Recall: 0.1230  
Average ROUGE-L F1 Score: 0.1942
```

Legal Pegasus

T5 Transformer

We need to improve the model more and choose one of them .As it took a lot of time. And even with the chunks it is not giving very optimal output. For Legal Pegasus ,it was already fine-tuned on legal documents.We improved the input size. So that's why it performs better .But it is giving very long summary.Not that optimal.

We used ROUGE evaluation metrics.

For ROUGE 1, moderate F1 score (0.4 to 0.5)

ROUGE 2, F1 score (0.2 to 0.4)

ROUGE L, F1 score (0.4 to 0.5)