TnCSimplifier - Summarizes Terms and Conditions/ Privacy policies

Devansh Goswami
devansh21460@iiitd.ac.in

Ritisha Singh
Ritisha21089@iiitd.ac.in

Samanyu Kamra
samanyu21487@iiitd.ac.in

Shriya Verma
shriya21490@iiitd.ac.in

1 Problem Statement

Tnc-Simplifier is developing a web application that simplifies the process of reading and understanding privacy policies, which are often lengthy and filled with legal jargon. The main goal is to help users quickly grasp the key permissions they are granting to various applications, thereby reducing the risk of data exploitation amidst rising cyber crimes. The application will offer summarized versions of privacy policies, highlighting essential information while filtering out irrelevant parts using a keyword-based system. Users can input privacy policies in both text and image formats and can translate the summarized outputs into any language, enhancing accessibility and understanding. This service aims to create a safer digital experience for its users by making privacy policies more transparent and easier to understand.

2 Literature Survey

The literature survey section provides an overview of the existing research and applications related to simplifying terms and conditions and privacy policies. We review relevant studies, applications, and technologies that address similar challenges in the field of digital privacy and user understanding of legal agreements. There are two main types of text summarization, namely extractive and abstractive. Traditional extractive summarization techniques work by extracting phrases one at a time and modelling the relationships between them. A popular approach is the TextRank algorithm, which uses a graph-based model. As for abstractive techniques, Google's Pegasus [4] is one of the most popular models. For sequence-to-sequence learning, it employs a transformer-based encoder-decoder paradigm.

In order to help individuals and companies better understand the potential risks and consequences of the agreements they are entering into, AI and natural language processing are taking the fields of privacy policies and legal text summarization by storm. The evolution in summarizing legal documents through NLP and machine learning techniques is underscored by several key studies, each contributing to the field's development. Initially, the application of algorithms like Multinomial Naive Bayes [1] offered a pathway to distill complex legal texts into more accessible summaries, setting a precedent for innovation. Building on this, LexRank [2] utilized eigenvector centrality to enhance extractive summarization, demonstrating the effectiveness of graph-based methods over traditional approaches by more accurately determining sentence importance. Adding to the complexity, a study [5] introduced a CNN-based risk classifier to prioritize high-risk sections in privacy policies, marking a significant step towards tailored summarization focused on user concerns. This approach was further refined by integrating RoBERTa with sequence-to-sequence models, creating a hybrid system that extracts and paraphrases legal terms. Google's Pegasus has been fine tuned on legal corpus to create PEGASUSCourtOp/legalPEGASUS [3], this showcases the potential of legal domain specefic adaptations of existing systems.

Preprint. Under review.

Several existing platforms and services like Lexcheck and Lawgeex currently use AI and natural language processing to identify potentially risky clauses or language in contracts and policies, which is important for legal contracts and policies. These projects help those without legal expertise by highlighting key risks, preventing legal issues, and protecting against reputational harm. Our web application, Tnc-Simplifier, takes this up a notch by increasing accessibility with our multilingual features and utilization of optical character recognization to increase the amount and types of queries that can be addressed.

3 Motivation

The motivation behind Tnc-Simplifier's development is driven by the increased number of cyber crimes in today's world. Users tend to skip privacy policies primarily due to their length and the legal jargon within them that the average user cannot understand. The primary goal is to protect users and minimize the exploitation of data.

4 Novelty

1. **Multi-Lingual Summaries:** We used Google Translation API, to generate summaries in all possible languages for ease of users.

2. **Optical Character Recognition:** We used Tesseract, to generate summaries for images, like if user wishes to provide screenshot of privacy-policies, we plan to handle that too.

3. **Chat Bot:** We fine-tuned the Gemini LLM to generate outputs for user queries regarding doubts of privacy policies.

5 Methodology

1. **Data Collection:** We collected data from company/apps webpages and play store to accomodate a MySql database.

2. **Deciding Model:** We experimented with multiple models but later shifted to Legal Pegasus which is fine-tuned for legal documents and has great semantic similarity.

3. **Inferencing From Model:** We generated summaries for our dataset.

4. **Multi-Lingual Summaries:** We used Google Translation API, to generate summaries in all possible languages for ease of users.

5. **Optical Character Recognition:** We used Tesseract, to generate summaries for images, like if user wishes to provide screenshot of privacy-policies, we plan to handle that too.

6. **Chat Bot:** We fine-tuned the Gemini LLM to generate outputs for user queries regarding doubts of privacy policies.

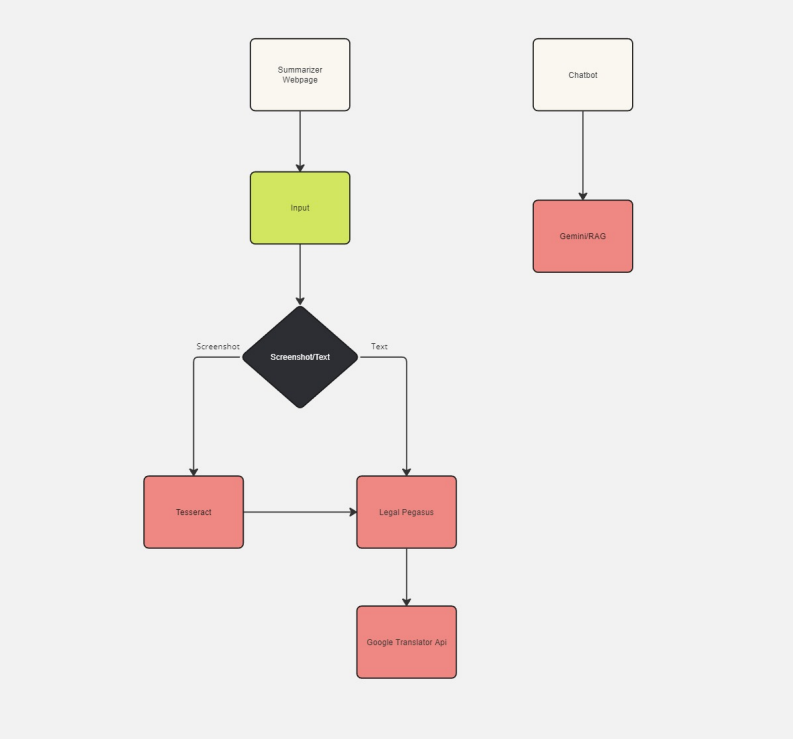

Figure 1: Pipeline

6 Updation from Midsem Report

The major updations from midsem have been:

1. Implementing all the Novelties.
2. Fine Tuning on Legal Pegasus instead of Pegasus.
3. Updating the database schema.
4. Implementing a fully-functional website.

7 Updated Database

Following the selection of the Legal-Pegasus model, we utilized it to condense the privacy policy texts contained within our database. These condensed summaries were then incorporated into a new table named 'generateddata'. This table serves as a valuable resource for future use, allowing users to access pre-generated summaries without the need for the model to regenerate the output each time.

We are also storing the short/medium/long length of summaries in our database.

Figure 2: Database Schema

8 Legal Pegasus Model

Legal Pegasus is a fine-tuned Pegasus model on legal documenents. Pegasus is a specific model that gives generated abstracts based on the article which means the model reads the article text and writes a suitable headline. This abstractive text summarization is one of the most challenging tasks,involving understanding long passages, information compression, and language generation.

The Pegasus model is impressive symbolic reasoning, as researchers concluded based on performance and evaluation results. Some variations show not just the human-level of natural language understanding, even beating the human performance in generating summarized output.

8.1 Architecture:

The current practice for this task would be to train a language model by predicting the masked out token at the end of the sequence. The task is known as self-supervised autoregressive language modeling. Next, fine-tune it on a dataset of labeled summaries and have the generalized output of an article or a document that hasn't seen on the training test, and contains the most information from the source documents. The aim of fine-tuning this kind of architecture is to gain the best results on downstream NLP tasks such as yelp-review, natural language inference, and question-answering.

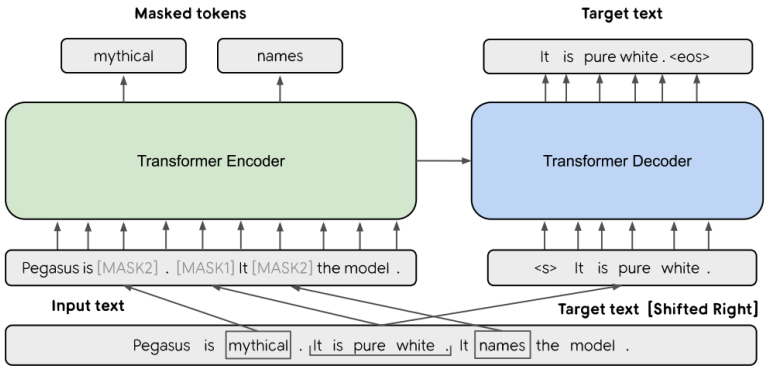

Figure 3: Pegasus Model Architecture

In our code we used the pre-processed tokens as inputs to generate the summary using the Pegasus pre-trained model.

9 Evaluation

1. High Difference in cosine similarities.
2. Rouge scores are double for legal pegasus.
3. Similar bert score but slightly better for legal pegasus.
4. Legal pegasus is better in all metrics.

Metric	Pegasus	Legal Pegasus
Cosine Similarity	0.37	0.75
Rouge 1	0.09	0.27
Rouge 2	0.04	0.17
Rouge L	0.09	0.27
Bert Score	0.8366	0.8620

Table 1: Comparison B/w Pegasus and Legal Pegasus

10 Functionality

1. **Chatbot**.

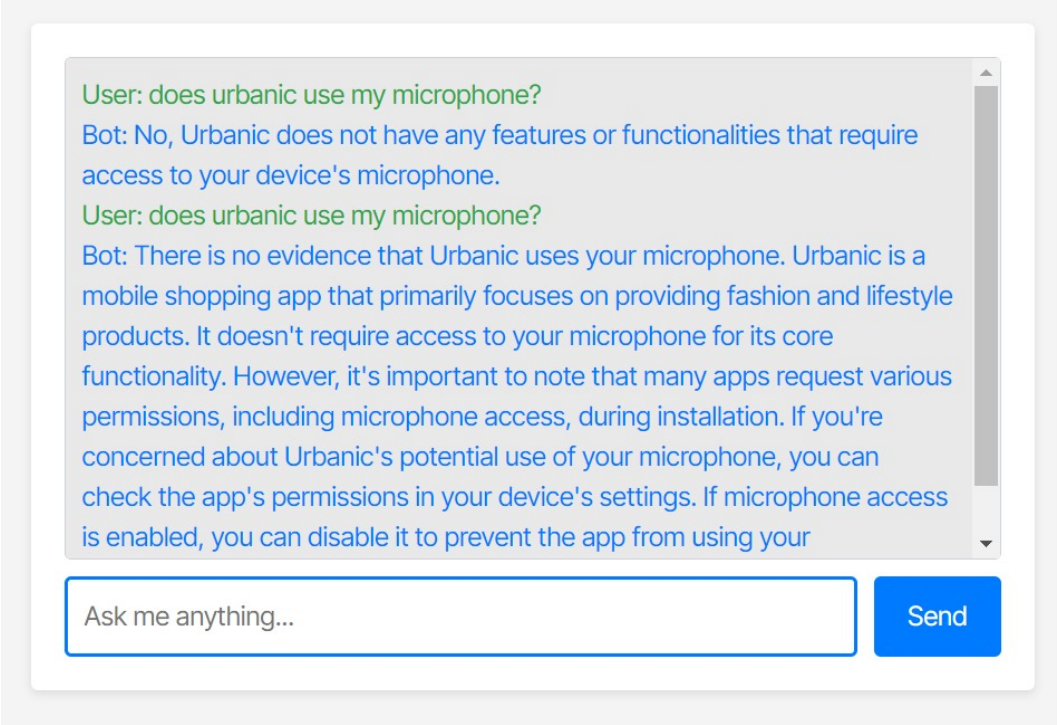

Figure 4: Chat Bot

2. OCR- Multi-linguality

Figure 5: Website OCR-Multilingual