e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:07/Issue:02/February-2025 Impact Factor- 8.187 www.irjmets.com

SUMMARIZATION OF LEGAL DOCUMENT: TERMS OF SERVICE

S.S. Jogdand[*1], Sahil Rane[*2], Jidnyesh Toke[*3], Harshwardhan Patil[*4], Prathamesh Bhoge[*5]

[*1]Lecturer, Department Of Computer Engineering, Pimpri Chinchwad Polytechnic, Pradhikaran, Nigdi, Pune, India.

[*2,3,4,5]Student, Department Of Computer Engineering, Pimpri Chinchwad Polytechnic, Pradhikaran, Nigdi, Pune, India.

DOI : https://www.doi.org/10.56726/IRJMETS67232

ABSTRACT

Legal documents often contain complex and dense information that can be time-consuming and difficult to understand. To tackle these challenges, this paper introduces an automatic summarization of legal texts can provide a valuable solution by extracting key information and presenting it in a more accessible format. This paper presents a approach to the summarization of legal documents. We propose a hybrid model that uses natural language processing (NLP) methods to identify and summarize critical legal concepts of third party websites. Our approach uses pre-trained language models that are adapted to a collection of legal documents mainly Terms and Conditions(T&C), helping the system better understand legal terms and structure. We tested the model using a dataset of labeled legal documents, and the results show that it produces more relevant and clearer summaries compared to traditional methods. The proposed system can make it easier to understand legal documents like T&C. By making T&C documents easier to understand, our approach can improve the user experience, especially on online platforms where users often deal with these agreements.

Keywords: Natural Language Processing, AI, Text Summarization, T5 Model, Machine Learning, Tokenization.

I. INTRODUCTION

Legal documents, and particularly those agreements in the form of Terms and Services (T&S), are usually extensive, brought in with complex legalese that makes it almost impossible for common users to understand. As more digital services keep popping up, users are equally being shoved with such agreements at times. Unfortunately, many times users would overlook or skim through those lengthy documents, which often create misunderstandings on their rights and privacy. The manual summary of such documents is a tiresome and error-prone process. Thus the rising interest in automation through machine-learning techniques, particularly natural language processing (NLP). NLP has provided an opaque light on the extraction of the main points and then the simplification of those detailed agreements into simple digestible summaries. Nevertheless, the legal field, both in terms of complexity and specialized diction, makes it more yielding than most typical NLP models are able to lend insight from, thus compounding the question of whether an accurate algorithm could solve the mystery for us. This paper talks about how machine learning could automatically summarize T&S documents and look at advanced models from detailed structure into proper context and then to valuable information.

The remainder of this paper is structured into the following parts: Literature Survey about related work appears in Section II; it provides the methodology in Section III; Proposed System is found in Section IV; this work concludes in Section V with possible directions for future research and Acknowledgement.

II. LITERATURE REVIEW

The field of text summarization has advanced considerably thanks to the development of Natural Language Processing (NLP) techniques. An extractive method using word frequency for identifying important sentences was proposed by Luhn [1], which may be viewed as the start of modern extractive settings. A later work by Radev et al. [2] introduces an unsupervised machine learning-based abstractive summarization by incorporating rankings from sentence centrality using graph-based algorithms toward sentence quality improvement. Using [3], machine learning algorithms to rank sentence importance fairly improved extraction, considering context.

Abstractive summarization-an undertaking to generate new sentences rather than the extractive approach-began with Rush et al. [4]. They used sequence-to-sequence models with embedded mechanism in order to have more coherence in their summaries. In [5], a hybrid strategy was introduced that combined extractive and

e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:07/Issue:02/February-2025 Impact Factor- 8.187 www.irjmets.com

abstractive qualities. The preliminary emphasis was to extract key sentences, and a deep learning model was employed to rephrase them ingenuously into a summary, amalgamating both techniques."

The introduction of BERT (Bidirectional Encoder Representations From Transformers) by Devlin et al. [6] has changed the entire direction for summarization. Models such as BART and T5, fine-tuned to summarize bodies of text, result in significantly better summaries. Criminisi et al. [7] used extractive methods to summarize T&C documents based upon domain-specific rules for legal text summarization. Legal contracts were summarized by using a framework based on deep learning, focusing on key clauses, making them more understandable, according to [8]. In [9], Transformer-based models such as BERT were fine-tuned for summarization on legal documents, therefore enabling the summarization of court rulings and legal briefs by understanding contextual entities.

III. METHODOLOGY

Requirement Analysis

The first step involves identifying the core requirements for the system, which are driven by the challenges in summarizing complex legal language. The primary requirements include the efficient extraction of relevant information from long legal documents without losing important details, and the presentation of this summarized content in an understandable format. The system must also have an easy-to-use interface, provided through a Chrome extension, that allows users to summarize the Terms of Service from third party websites they visit. An important feature is that the system must be scalable to handle a variety of website structures and legal document formats.

System Design

The system is designed with an architecture that balances usability and accuracy. The key components of the design include the Chrome extension, which interacts directly with web pages containing legal documents. The backend service is built in Python and handles processing of legal content, running summarization algorithms, and communication with the Chrome extension. The summarization process is primarily extractive, based on selection of key sentences in the document. For evaluation purposes, ROUGE (Recall-Oriented Understudy for Gisting Evaluation) is used to assess the quality of summaries.

Technology Section

The technology stack for this summarization system includes both frontend and backend components. The Chrome extension is developed using JavaScript, HTML, and CSS, making it work very well within web browsers. The backend is in Python and utilizes several libraries to perform the various tasks. The system uses BeautifulSoup and requests to scrape web pages, which enable it to retrieve legal content from the targeted web pages. For natural language processing and summarization, it uses Autotokenizer and pipeline libraries from transformers for tasks such as tokenization and extractive summarization. To measure the quality of the summarization, ROUGE is used, which provides a standardized way of measuring the quality of the summaries by comparing them to reference summaries.

Implementation

The process starts by creating a Chrome extension that grabs the current webpage's URL and sends it to the backend for processing. The backend service receives the URL, extracts the relevant legal content from the webpage. The extractive summarization algorithm is then applied to identify the most important sentences from the document, which are selected and presented as the summary. ROUGE evaluation is performed to compare the generated summary with reference summaries. The final summarized content is sent back to the Chrome extension, which then displays it in a user-friendly manner.

Monitoring and Maintenance

After deployment, the system continues to be observed and monitored to assure effectiveness. In tracking the effectiveness of the summarization system, relevance and accuracy in the generated summaries are continually followed. To understand areas on which the summary could be perfected, feedbacks from users continue to be sought. The summarization system must continue to be maintained due to changes made on the document structures over time. The backend service has an error logging mechanism in place that monitors any errors occurring to users or the system. This enables quick troubleshooting and improvements.

e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:07/Issue:02/February-2025 Impact Factor- 8.187 www.irjmets.com

Improvements

Continuous improvement is the hallmark of the summarization system. Periodically, the summarization algorithms are revised based on feedback from users and performance metrics. This might be achieved through parameter tuning in the extractive summarization model or incorporating more advanced techniques, like hybrid models, that combine the best of extractive and abstractive summarization approaches. Finally, the system is expanded to encompass a wider scope of legal documents beyond just Terms of Service to include privacy policies and end-user license agreements. Regular updates are also made to the user interface of the Chrome extension to enhance usability, ensuring that the system remains intuitive and effective.

IV. PROPOSED SYSTEM DESIGN

Input Text Data (Legal Documents in the form of Terms of Service)

The system starts by using legal documents, such as Terms of Service (ToS), which can be long and hard to understand. These documents explain the rules for using a service but are often full of complicated language. To make it easier, a Chrome extension is used to access the webpage where the ToS is located. It can automatically pull in the text, or users can manually paste the text into the extension for summarization.

Preprocessing

In this step, the text is cleaned up by removing things like special characters, unnecessary tags, and common words that don't add much meaning (like "the" or "and"). The text is also converted to lowercase for consistency. After that, it is broken down into smaller parts, called tokens, which helps the model better understand the content.

Tokenization

Tokenization is the process of splitting the text into smaller parts, such as words or subwords. This makes it easier for the model to understand and process the text. For rare words, the system can break them down into smaller parts so it can still understand them.

T5 Model (Core Model for Text Summarization)

The core of the system is the T5 model, a type of machine learning model that has been trained on lots of text. It looks at the input text and tries to pick out the most important parts to make a summary. T5 is good at summarizing because it can understand the structure of language and focus on the key points in a document.

ROUGE Evaluation (Evaluating Summary Quality)

Once the model generates the summary, it is checked for quality using a method called ROUGE. ROUGE compares the machine-generated summary to summaries written by humans to see how well the key points are captured and whether the summary is accurate.

Output Summary

The final output is a shorter version of the Terms of Service that focuses on the most important information, like what users can and can't do, and privacy policies. This summary is shown in the Chrome extension, making it easy for users to read. The extension also lets users input their own legal text to get a summary.

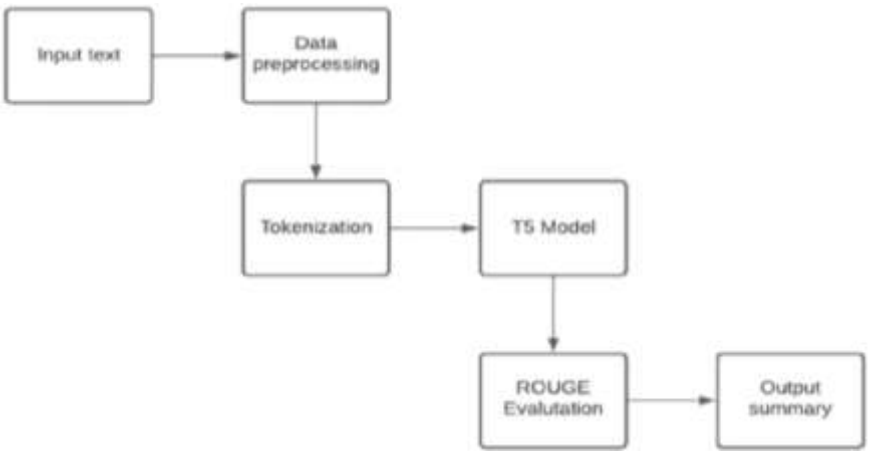

Figure 1: System Architecture diagram

e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:07/Issue:02/February-2025 Impact Factor- 8.187 www.irjmets.com

V. SCOPE

The Summarization of Legal Documents (Terms of Service) is a Chrome extension project that summarizes lengthy and complex legal text like Terms of Service agreements. The tool automatically summarizes the key points of such documents, making it easier to read and understand various aspects of the same in a relatively short period. It identifies key clauses and details from the text using NLP and then summarizes it into a simple statement, saving the users time and giving them a chance to make a more informed decision without reading the entire document. The tool is user-friendly and automatically activates when users visit third party websites with Terms of Service. It can also be extended to summarize other types of legal documents in the future.

Hardware and Software:

1) Chrome Extension

2) Python

3) JavaScript, CSS

4) NLP Libraries

VI. CONCLUSION

This paper introduces a Chrome extension that assists users in understanding legal documents, such as Terms of Service agreements, by summarizing them into easy-to-read key points. The extension uses natural language processing to make it faster and easier for users to find important information without reading long legal texts. This tool not only solves the problem of complicated legal language but can also be extended to summarize other types of legal documents in the future. In general, this project illustrates how technology can be used to improve user understanding and make legal information more accessible.

ACKNOWLEDGEMENTS

This project “Summarization of Legal Document: Terms of Service” would not have been possible without the kind support and help of many individuals and organizations. We would like to extend our sincere thanks to all of them. We are highly indebted to Mrs S.S. Jogdand for their guidance and constant supervision as well as for providing the necessary information regarding the project and for their support in completing the project. We would like to express our gratitude towards Prof. Malkar Mam, HOD of Computer Department, faculty and all the lab assistants of Pimpri Chinchwad Polytechnic for their kind co-operation and encouragement which helped us in the completion of this project.

VII. REFERENCES

[1] “Evaluation Measures for Text Summarization.” In: 2009. url:https://www.researchgate.net/publication/220106310_Evaluation_Measures_for_Text_Summarization.

[2] Ms. Anusha Pai. “Text Summarizer Using Abstractive and Extractive Method”. In: May 2014. url: https://www.ijert.org/research/text-summarizer-using-abstractiveand-extractive-method-IJERTV3IS050821.pdf.

[3] “Empirical Study of Deep Learning for Text Classification in Legal Document Review”. In: Dec. 2018. url: https://ieeexplore.ieee.org/abstract/document/8622157.

[4] “Legal Document Retrieval Using Document Vector Embeddings and Deep Learning”. In: Nov. 2018. url: https://link.springer.com/chapter/10.1007/978-3-030-01177-2_12.