International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 08 , August - 2026

AI-Based Summarization for Insurance Policies and website Terms & Conditions

Y.Sruthi, B.ManiHarika, Badana Karthikeya, G.Bhavana, Ch.Praneeth Reddy
Department Of Computer Science and Engineering(AI & ML)
Anil Neerukonda Institute of Technology and Sciences (ANITS),Visakhapatnam – 531162,India

Abstract - Policies like terms and conditions and insurance as well as legality which are on written matters tend to be long, rarely comprehended, most of the time by the common user. Because of this, people often go into such agreements without understanding fully the obligations, exclusions and the risks involved. In this paper, an AI-based system of document summarization is proposed to simplify such legal documents having the help of Natural Language Processing (NLP) models and transformer-based models. Documents of various formats such as PDFs, scanned images, web pages etc. can be processed using OCR and text extraction methods in the system. The Named Entity Recognition (NER) is utilized to distinguish major aspects of the law (obligations, coverage limits, penalties, and exclusions). An adapted transformer domain model is then used to produce summaries in the clause level but each at a concise form and keeping the original meaning of the content legal. The experimental assessment illustrates that the suggested system helps minimize the document complexity and increase user understanding and speed up legal agreement review processes, the legal information can be more readily available to under-experts.

Key Words - Natural Language Processing (NLP), Legal Document Summarization, T5 / BART models, Named Entity Recognition (NER), Optimal Character Recognition, Hugging Face Transformers, pdf plumber.

I. INTRODUCTION

The legal documents like insurances and terms and conditions on the web site are important to establishing the rights, responsibilities and obligations of service providers and users. Nevertheless, most of these documents are usually written in convoluted legal terms wherein long clauses that are not easily comprehensible by an average user can be found in documents. This has led to the situation where a good number of users end up accepting agreements without clearly understanding the consequences of having entered the agreement- its exclusions, penalties, and limitations to coverage. This will result in misunderstanding, financial risks and legal fights.Historically, the process of such document review is labor intensive as it is expensive, time consuming and requires legal experts to perform the task. Traditional methods of text summarization like extractive one fail to produce an accurate semantic meaning and contextual relations that are contained within a legal text. Moreover, the strategies are incompetent in distinguishing valuable clauses like obligations, risks, and exclusions that are important to users.In recent years, Artificial Intelligence (AI)

as well as Natural Language Processing (NLP) has experienced significant improvements, which allows creating automated systems to analyze and summarize complex text information. Implementing models based on transformers (e.g., BERT, T5, GPT) has demonstrated considerable advancements in the activities of document classification, named entity recognition, and abstractive text summarization. These models have the ability in making sense of a text in terms of contextual relationship and create meaningful summaries without distorting the initial intent of a document.In spite of such development, the majority of the current systems are tailored to general text or legal case documentation and do not focus on insurance policies and web site terms and conditions. These files have specific terminology and formatted clauses which involves using special processing methods.To overcome these issues, this paper presents an AI-based system to summarize insurance policies and terms and conditions of the website with domain designed transformer models and Named Entity Recognition (NER). The suggested system facilitates the extraction of important clauses, identification of obligations and risks, and the generation of succinct summaries, which turn out to be less complex to the users. The system is also compatible with several types of documents such as PDFs, scanned documents, and websites by using Optical Character Recognition (OCR) and text processing.

The key contributions of this work are: Creation of multi-format document processing pipeline of legal and insurance documents. Incorporation of Named Entity Recognition in the detection of significant legal objects like obligations,

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 08 , August - 2026

exclusions and penalties. Application of transformer-based document summarization model adapted to legal documents.Production of transparent and concise excerpts that enhance easy understanding of complex legal agreements by users.The rest of this paper will be structured as follows. Section II displays the literature review of the current studies of legal document summarization. Section III outlines the proposed system infrastructure and system architecture. The analysis and results of the experiment are discussed in the IV section. Lastly, Section V summarizes the paper and states the future research directions.

II. RELATED WORK

The sustained development of legal and policy documents has prompted researchers to use Natural Language Processing (NLP) to analyze and summarize documents using automated methods. Legal text summarization is predisposed to compress long legal texts, but leaves important arguments, decisions, and terms and conditions not to be lost, enabling the user to read only that information necessary. Initial studies on legal document summarization were predominantly based on extractive approaches like sentence ranking and graph-based and similarity measures. These techniques consider key sentences in a document and integrate them, to make up a summary. Though these methods save on the time taken to read through legal texts, in most cases, the methods do not recognize the contextual associations and semantic meaning that occur in legal texts.As the deep learning continued to emerge, scholars started to consider neural network and transformer based models to process legal text. BERT, Legal-BERT, and Longformer models have been used to address the following tasks: classifying legal documents, summarizing them, and extracting information. These models show better contextual discrimination of legal language and superiority in generation of meaningful summaries than in conventional methods of NLP.

Benchmark data sets have also been prepared to test NLP models in legal sphere. LexGLUE is one of the most popular datasets that are used to accomplish the combination of various law NLP datasets and offer a unified benchmark to accomplish different tasks, including classification, question answering, legal text analysis, etc. Research indicates that domain-adapted models that are trained on legal corpora perform much better than general-purpose language models applied in a legal task.Even more recent studies have investigated transformer-based abstractive summarization that includes T5, BART, and mT5 to produce summaries that are more coherent and human like of legal documents. These methods operate on the basis of pre-trained language models and fine-tuning so as to enhance the quality of the summarizations. Nevertheless, there are still issues of complexity, word length, and domain vocabulary of the legal documents, along with the lack of annotated datasets.

Even with these developments, the majority of the literature available is dedicated to case judgments and court rulings, and little effort has been put into the summarization of insurance policies and terms and conditions in websites. These documents are characterized by the peculiarities of the clause structures, requirements, and exclusion conditions which have to be processed in order to be excluded. Hence, the systems that are explicitly created to review and summarise such documents and preserve the traceability of clauses at the level of user intelligibility and maintainability are required.The limitations are addressed in the proposed work by creating a transformer-based summarization system specifically adapted to the needs of the insurance policies and terms and conditions of the webpage, which would combine named entity recognition and a clause-level analysis that would produce clear and user friendly summaries.

III. LITERATURE SURVEY

NLP based summarization and legal have been studied by several researchers.understanding of documents, disclosing the potential developments, the evident gaps towards.insurance and T&C domains.Kolambe and Kaur (2024) used the BERT-based extractive and abstractive models.to insurance policies, prove encouraging summary quality and clause. They were however limited by small datasets, weak mapping. absence of clauses and summaries, absence of clause traceability.

Chalkidis et al(2022) presented LexGLUE, a general benchmark of legal NLP across several corpora with Legal-BERT and RoBERTa. While influential, and they mostly concerned their attention on case law as opposed to insurance or Consumer T&C documents, restricting direct application to policyholders, and domain tasks of insurance. Ariai et al. (2024) have systematic surveyed 154 legal and policy datasets and pointed to problems with domain adaptation, model generalization and that there are no standard evaluation measures of legal summarization. Despite being aware of such gaps, they have not accomplished a full-end to end system. Swamy et al. (2025) proposed using T5 transformers, LSTM, as ValidEase. Post-processing networks, and rule based to large bodies of law. Their work presented competition on legal document summarization but coverage of insurance and consumer T&C content were still restricted and summaries were predisposed to be extractive, at the risk of exclusion of latent clause-level risks. Aggarwal et al (2023), examined the deep hybrid models in CNNs, RNNs, and attention mechanisms on news, legal, policy datasets but not fine-tuned on special texts in insurance or superset section traceability. In their particular case, Dikmen et al. (2025) used machine learning and NER on construction

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 08 , August - 2026

and insuring contracts, and solutions are strongly customized. Analytical analysis of contracts as opposed to overall policy summary. Their approach, while niche, was not generalized to wider insurance. Bhattacharya et al. (2024) constructed DEL Summ, an unsupervised extractive model. Indian and UK judgments and contracts, which have shown good performance on legal texts but that need adaptation to insurance specific clause always like extracting entities. Some of the gaps that exist after reviewing the literature include: Diverse datasets of care insurance. Public annotated datasets of care insurance it is not well trained and evaluated because of the lack of policies and T&C documents models. Lack of traceability Most current systems lack any means of connecting summaries to original clauses, which complicates the issue of verification and compliance audit. Lack of personalization: There are not so many systems that give an opportunity to personalize summaries differentiated by their area of interests or types of clauses. poor multi-format support: The majority of systems expect clean, digital text input and weaknesses in robust PDF and scanner document control. Weak evaluation metrics: There exists no benchmark of evaluation. Summarization of insurance policies or satisfaction of users. This project is aimed at filling these gaps by proposing a practical end-to-end system boasts domain-adapted transformers, clause level NER, multi-format input user-centric design, support, and design.

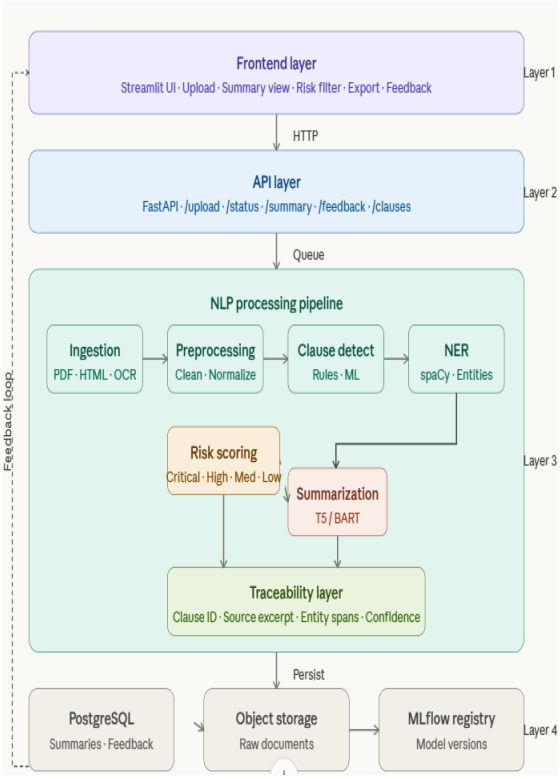
IV. METHODOLOGY

The offered system is aimed at exploring and summarizing insurance policies and websites terms and conditions with Natural Language Processing (NLP) and transformer-based models. The methodology can also be divided into several steps which are the document input, preprocessing, entity recognition and transformer based summarization. The design conceived of the architecture is such as to support different document types and produce simplified summaries without loss of the legal sense of the original text.

A. System Architecture

The system architecture is made up of a number of interrelated levels that work on documents and produce summaries in a structured format. The workflow starts with the input layer, i.e., users upload documents like PDFs, word files, pictures, or HTML based policy pages. Such documents are then given to the metadata extraction module that helps in detecting the significant information in the document e.g. the title of the document, the date and the organization of the document. The documents are then run through the preprocessing layer which performs the role of converting scanned images or PDFs into machine-readable formats with the help of Optical Character Recognition (OCR). The mined text is purged and divided into rational units like clauses, headings and subclauses. The processed text then gets processed in the NLP processing layer which identifies key legal elements like obligations, exclusions, penalties and terms of coverage using Named Entity Recognition (NER). Such entities identified assist the system to comprehend the situation and meaning of particular clauses. The summarization layer is based on a transformer-based model to provide summaries on the processed text in a concise form. The format in which the summary is generated is traceable as every point that is summarised gets connected to the corresponding clause in the original document. Lastly,the summarized result is saved in the database layer and the response is shown on the screen through a user interfaced and customizable results. The feedback mechanism serves to enhance the system since it makes the continuous refinement of the model possible.

B. Dataset and Preprocessing



The data employed in this research will include insurance policy documents, and terms and conditions of websites which were gathered in publicly accessible data. These records can be by use of various formats such as PDF files, scanned images, word documents, and web pages. The documents are first processed through a series of preprocessing in order to have quality and consistent data before the NLP technique is used. Primary, OCR Scanners included in the text extraction stage encompass Tesseract to scan paper documents and direct text parsing to read electronically a document. That is followed by cleaning the

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 08 , August - 2026

extracted text by eliminating superfluous things like headers, footers, page numbers and symbols of formatting. The text is normalized when it has been cleaned by converting it into a standard format including converting it to lowercase and any unnecessary characters are eliminated. The document is next divided into logical parts which include clauses, sentences, and paragraphs. This categorization assists in this process of distinguishing the key parts like details of cover, parts where the policy does not pay or applies penalties, and renewal details. As well, Named Entity Recognition (NER) is also used to label important legal entities including policy terms, financial amounts, periods and terms, as well as legal obligation. This formalized representation enables the model of summarization to prioritize on significant legal aspects in the process of summarizing.

C. Processing Algorithms

Algorithms employed in the Project.

[1]. Text Preprocessing

The project employs the use of text preprocessing methods to retrieve and purify information in a variety of form: Image OCR with Tesseract, Extracting text in PDF with pdf plumber, Beautiful Soup parsing in HTML. These procedures are in preparation to process the raw data.

[2]. Classification

The clauses are classified as: Rule-based classification, Pattern and pattern matching up of keywords. The system allows assigning categories according to the set rules and keywords rather than relying on a machine learning classifier.

[3]. Named Entity Recognition (NER).

The entity extraction is performed using spaCy in the project. This determines significant hostilities like: Monetary values Dates, Coverage details, Exclusions. spaCy internally operates with the use of statistical and neural NLP models.

[4]. Text Summarization

Transformer-based models are used to perform summarization:T5 (Text-to-Text Transfer Transformer), BART (Bidirectional Auto-Regressive Transformer). It is an abstractive method of summarization, i.e. it produces new sentences, but does not pick the old one.An extractive method is employed as a fallback, in the case where the model is not available.

[5]. Risk Scoring

The risk scoring is done with respect to:Rule-based scoring. Keywords and patterns identification. The system gives risk level like Critical, High, Medium, and Low depending on verifiable terms and patterns.

[6]. Evaluation

The project uses: ROUGE-L metric. This judges generated summaries quality by comparing with reference text.

[7]. Supporting Technologies

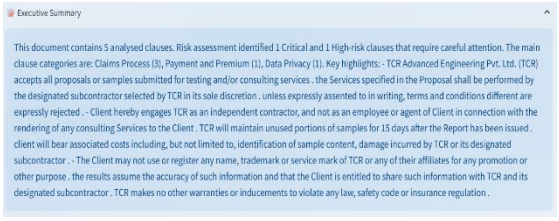
The system is built using: Transformers (Hugging Face), PyTorch , spaCy.

They allow deep learning and NLP processing in the project. The deep learning models that work with transformers with the help of NLP algorithms are applied in the processing stage and generate accurate legal document summaries. The general algorithm has three main parts, which include clause identification, entity recognition, and summarization.To begin with, the system carries out clause detection by subdividing the document into significant legal parts. They are then broken down into segments that can be used to determine how each of them would affect the policy like terms of coverage, obligations, and exclusions or penalties.The Named Entity Recognition (NER) then implements and recognizes crucial legal entities in each clause. These will be policy names, coverage, payment terms, due dates, and risk indicators. The summed out entities give contextual data to the summarization model.Lastly, the summarization model is transformer-based and handles the text that is segmented into pieces to produce summaries. The model makes use of contextual embeddings to learn the connection between clauses and gives summaries that maintain the legal sense of the document. Clauses that are associated with risks and vague words are also identified at the stage.The summaries generated are then presented in readable forms like bullet points or section by section summaries. All summarized points are connected to the original clause making them transparent and traceable. This will enable users to have a glimpse of the important parts of the document by not reading the entire policy.

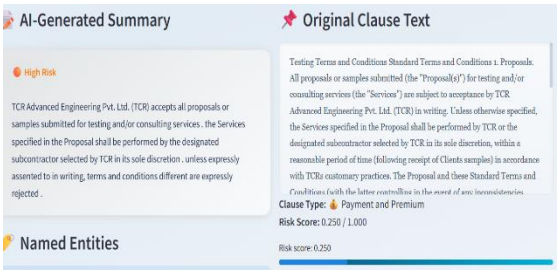
V. RESULTS & DISCUSSION

The evaluation of the proposed AI-based summarization system was done on a set of insurance policies and the terms and conditions on websites that were collected via publicly available sources. The assessment addressed three most important areas, as the summary quality, the processing efficiency, and the accuracy on a case-by-case basis.

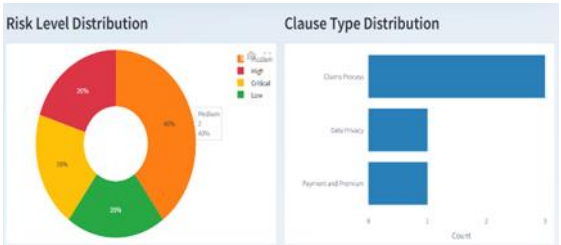
International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 08 , August - 2026



The summative model that uses transformer proved to be very effective at the task of creating concise, contextually accurate summaries. The proposed system provided more coherent and meaningful summaries compared to classical extractive algorithms like TextRank due to the ability to extract the semantic relationships of the legal text. The summaries still contained the necessary data like the coverage, the exclusions, penalties and obligations that are paramount in understanding the user. The clause level analysis was enhanced by the integration of Named Entity Recognition (NER), allowing the identification of important legal entities.



The system has been effective to identify important points that are important and are normally missed in traditional means like the risk-related statements and unclear clauses. Regarding efficiency, the system saved greatly in terms of time taken in the analysis of documents. Large legal documents were automated to reduce time wasted in manual processing as well as increasing the speed of decision



making. The system also accepted several input types i.e. PDFs, web pages and scanned documents with high performance. All in all, the outcomes of the conducted tests suggest that the suggested system does not only increase readability but also makes the process of simplifying legal documents that are hard to grasp considerably more transparent.

VI. CONCLUSION & FUTURE WORK

The present paper introduced an service summary of insurance policies and terms and conditions on websites as an AI-powered system with Natural Language Processing and the transformer-based model. The approach suggests loading several types of documents, using the Optical Character Recognition (OCR), Named Entity Recognition (NER), and domain-specifically trained transformer models to write a brief and significant summary-like short text of a complicated legal document.This is because the system handles the limitation of the conventional summary methods, which fail to accommodate the contextual information and have the ability to capture critical clauses including obligatory clauses, exclusion clauses and the presence of risk related clauses. The summaries generated are well structured, easy to read and traceable to the original document and hence enhances transparency and understanding of the user. As it has been experimentally found, the proposed system helps cut down on manual labor and increase the efficiency of law document review in the best way possible. Processing of various input formats also enhances its effectiveness in a real world environment. The subsequent research effort will be aimed at enhancing interpretability and inclusivity of models, building specific datasets, and adding multilingual capabilities that will allow serving a broader scope of legal documents and users.

VII. REFERENCES

[1] S. Kolambe and P. Kaur, "Insurance Policy Summarization Using BERT- Based Models and Named Entity Recognition," *Journal of Financial Technology*, vol. 15, no. 3, 2024.

[2] I. Chalkidis, M. Fergadiotis, P. Malakasiotis, I. Androutsopoulos, and O. Vida, "LEXGLUE: A Benchmark Dataset for Legal Language Un- derstanding in English," in *Proc. 60th Annual Meeting of the Association for Computational Linguistics (ACL)*, 2022.

[3] P. Ariai, R. Torres, A. Gomez, and M. Klein, "A Comprehensive Survey of 154 Legal and Policy Datasets: Challenges in Domain Adaptation and Evaluation Metrics," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 46, no. 2, 2024.

[4] M. Swamy, X. Chen, Y. Lee, and K. Wong, "ValidEase: Legal Document Summarization Using T5 Transformers and Rule-Based Post-Processing," in *Proc. Int. Conf. Natural Language Processing (ICONLP)*, 2025.

[5] A. Aggarwal, R. Patel, B. Sharma, and V. Kumar, "Deep Hybrid Ar- chitectures for Multi-Domain Document Summarization," in *Proc. 2023 Conf. Empirical Methods in Natural Language Processing (EMNLP)*, 2023.

[6] C. Dikmen, H. Ozer, A. Cakgin, and A. Unsal, "Machine Learning and Named Entity Recognition for Automated Contract Analysis in Insurance and Construction Domains," *Expert Syst. Appl.*, vol. 208, art. 118234, 2025.

[7] P. Bhattacharya, S. Pandey, S. Paul, K. Ghosh, S. Phadikar, and A. Das, "DELSumm: An Unsupervised Framework for Legal and Contract Summarization in Indian and UK Legal Systems," in *Proc. 2024 Joint Workshop on Legal NLP and Document Analysis (LNLP-DA)*, 2024.

[8] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," in *Proc. 2019 Conf. North American Chapter of the*

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 08 , August - 2026

Association for Computational Linguistics: Human Language Technologies (NAACL- HLT), 2019.

[9] C. Raffel *et al.*, "Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer," *J. Mach. Learn. Res.*, vol. 21, no. 140, 2020.

[10] Siino, M., Falco, M., Croce, D., & Rosso, P. (2025). Exploring LLMs Applications in Law: A Literature Review on Current Legal NLP Approaches. IEEE Access, 13, 18253–18276.

[11] Chalkidis, I., Fergadiotis, M., Malakasiotis, P., Aletras, N., & Androutsopoulos, I. (2020). LEGAL-BERT: The Muppets straight out of Law School. arXiv:2010.02559.

[12] Hendrycks, D., Burns, C., Chen, A., & Ball, S. (2021). CUAD: An Expert-Annotated NLP Dataset for Legal Contract Review. NeurIPS 2021.

[13] Vaswani, A., Shazeer, N., Parmar, N., et al. (2017). Attention Is All You Need. NeurIPS 2017, pp. 5998–6008.

[14] Lewis, M., Liu, Y., Goyal, N., et al. (2020). BART: Denoising Sequence-to-Sequence Pre-training. ACL 2020, pp. 7871–7880.