TermsBuster: An AI-Powered Smart Privacy Policy and Legal Document Analysis Assistant

S.Shobana[1], Dr.P.Anitha,M.C.A.,M.Phil.,B.Ed.,P.hd.,[2]

[1]Department of Computer Applications,Vellalar College for Women,TamilNadu, India.

[2]Assistant Professor, Department of Computer Applications, Vellalar College for Women, Erode, TamilNadu,India.

ABSTRACT

Legal, privacy, and policy documents contain important information about obligations, permissions, restrictions, conditions, and user rights. However, these documents are often lengthy and written using complex terminology, making them difficult for ordinary users to understand. TermsBuster is an AI-powered smart document analysis assistant developed to provide a first-level analysis of privacy policies and other legal documents. The system allows users to upload supported documents and processes their textual content using document extraction and natural language processing techniques. The proposed system performs document classification, text preprocessing, summarization, important-term identification, and rule-based risk analysis. TF-IDF-based text analysis is used to identify informative terms, while predefined rules, keywords, sentence-level analysis, negation handling, and safe-phrase information are used to support risk indication. The system supports different categories of legal and privacy-related documents, including contracts and agreements, legal notices, privacy policies, financial documents, ownership documents, intellectual property documents, criminal offense documents, regulatory compliance documents, employment documents, and court judgments. The results are presented through an interactive Streamlit interface. TermsBuster is designed as an informational assistant and does not replace professional legal advice. The proposed system provides a convenient approach for helping users understand complex legal and privacy-related documents before detailed review.

Keywords: Legal Document Analysis, Privacy Policy, Natural Language Processing, Text Classification, TF-IDF, Risk Analysis

1. INTRODUCTION

The rapid growth of digital services has resulted in users interacting with different types of legal and privacy-related documents on a regular basis. Privacy policies, terms and conditions, contracts, financial documents, employment documents, legal notices, and other agreements contain important information that can affect users. Before accepting or signing such documents, users need to understand the conditions, responsibilities, restrictions, and permissions mentioned in them.

A major difficulty is that legal documents are usually lengthy and contain formal language. Important information may be distributed throughout several sections, making manual examination time-consuming. Users without legal knowledge may also find it difficult to identify statements that require additional attention. This creates a need for software tools that can provide an initial and understandable overview of complex documents.

TermsBuster is proposed as an AI-powered smart privacy policy and legal document analysis assistant. The system provides a simple interface through which users can upload supported documents for analysis. The uploaded content is extracted and processed using natural language processing techniques. The system identifies the general document category, generates a structured summary, identifies important terms, and provides a rule-based risk indication.

1.1 Problem Statement

Legal and privacy-related documents are often lengthy, complex, and difficult for ordinary users to understand. Manual reading requires significant time and users may overlook important statements. Therefore, a simple system is required to analyze uploaded documents, identify their category, summarize their contents, identify important terms, and provide an initial risk indication.

1.2 Objectives

The main objectives of the proposed TermsBuster system are:

1. To provide a simple interface for uploading legal and privacy-related documents.
2. To extract textual information from supported documents.
3. To preprocess extracted text for further analysis.
4. To identify the category of the uploaded document.
5. To generate a meaningful summary of the document.
6. To identify important terms using TF-IDF-based analysis.
7. To identify potentially important or risk-related statements using predefined rules.
8. To present the analysis through a user-friendly Streamlit interface.
9. To provide informational support without replacing professional legal advice.

2. METHODOLOGY

The methodology of TermsBuster follows a sequence of document-processing and natural language processing operations. The overall process begins when the user uploads a document and ends when the analyzed information is displayed through the application interface.The major stages include document input, text extraction, preprocessing, document classification, summarization, important-term identification, risk analysis, and result presentation.

2.1 Document Input

The first stage of the system is document input. The application provides an upload interface through which users can submit supported legal and privacy-related documents.The system is designed to work with documents containing legal, privacy, financial, employment, ownership, regulatory, and other related information.

The uploaded document becomes the input for all subsequent processing stages.The input interface is intentionally kept simple so that users do not require technical knowledge to operate the application. After selecting a document, the user can initiate the analysis process.

2.2 Text Extraction

After the document is uploaded, the system extracts its textual content. Document extraction is an important stage because the subsequent NLP operations require readable text rather than the original document layout.For PDF documents, the application uses document-processing libraries to obtain available textual content. Where OCR-based processing is configured, text from image-based document content can also be processed.

The extracted text is passed to the preprocessing stage. Proper extraction is important because missing or incorrectly extracted text can affect classification, summarization, and risk analysis.

2.3 Text Preprocessing

The extracted text may contain unnecessary spaces, formatting elements, repeated characters, or other noise. Text preprocessing converts the extracted content into a more suitable form for NLP processing.The preprocessing stage organizes the text and prepares it for subsequent analysis. Sentence-level and word-level information can be used during later processing.

This stage improves consistency and allows the system to work with the actual textual content of the uploaded document.Preprocessing is especially important for legal documents because their content may contain headings, numbered clauses, long sentences, special symbols, and different formatting structures.

2.4 Document Classification

TermsBuster classifies uploaded documents into meaningful categories. Classification provides an initial understanding of the nature of the document and helps organize the analysis.

The supported document categories include:

- Contracts and Agreements
- Legal Notices
- Terms and Conditions / Privacy Policies
- Financial Documents
- Ownership Documents
- Intellectual Property Documents
- Criminal Offense Documents
- Regulatory Compliance Documents
- Employment Documents
- Court Judgments / Legal Precedents

Document classification allows the user to quickly understand what type of document has been submitted.

2.5 Document Summarization

Long legal documents may contain a large amount of textual information. Reading the complete document before obtaining a basic understanding can be time-consuming. Therefore, TermsBuster provides a summarized representation of the document.The summarization stage focuses on presenting the important content in a more accessible form. It reduces the need for the user to manually inspect every section just to understand the general nature of the document.The generated summary is intended to provide an overview rather than a complete legal interpretation. Users should refer to the original document when detailed information is required.

2.6 TF-IDF-Based Important Term Analysis

TermsBuster uses TF-IDF-based text analysis to identify informative terms within the document.TF-IDF represents the importance of a term by considering its frequency in the document and its relative usefulness for distinguishing content. Terms that occur frequently in a document but are not common across the overall vocabulary can receive higher importance.The important terms identified through this process help the system focus on relevant portions of the document. They can also support the summarization and risk-analysis stages.

3. MODELING AND ANALYSIS

The TermsBuster system consists of several connected modules. Each module performs a specific function and passes its output to the next stage.

The overall logical flow of the proposed system can be represented as:

User Upload → Text Extraction → Text Preprocessing → Document Classification → Summarization → TF-IDF Analysis → Risk Analysis → Result Presentation

3.1 System Architecture

The system architecture contains the following major components:

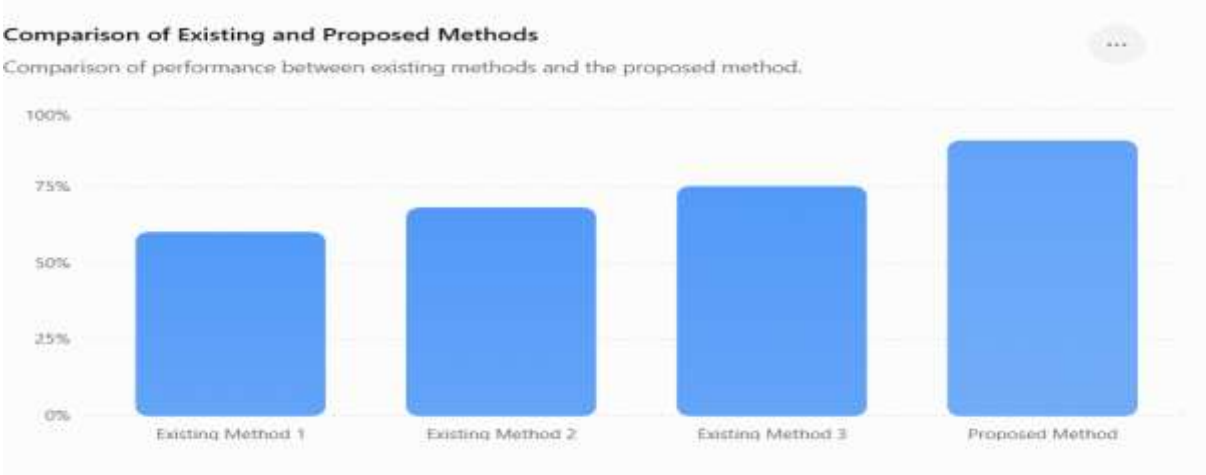
Component	Function
User Interface	Provides document upload and result display
Document Input	Accepts supported documents
Text Extraction	Extracts readable text
Preprocessing	Cleans and prepares text
Classification	Identifies document category
Summarization	Generates document overview
TF-IDF Analysis	Identifies important terms
Risk Analyzer	Identifies potentially important statements
Result Presentation	Displays analysis to the user

Table 1. Major components of TermsBuster

The user interacts with the system through the Streamlit interface. Once a document is uploaded, the document-processing components extract and prepare the content. The classification component determines the document category, while the summarization and TF-IDF components analyze the content. The risk analyzer then evaluates relevant terms and statements. Finally, the processed information is presented to the user.

3.2 Functional Modules

Fig. 1. Overall workflow of the TermsBuster system



4. RESULTS AND DISCUSSION

The developed TermsBuster application provides an integrated workflow for analyzing legal and privacy-related documents. The system allows users to upload a document and obtain a structured analysis through a single interface.The output provides information about the document category, summarized content, important textual information, and risk-related indications. This reduces the need for users to manually perform separate analysis tasks before obtaining an initial overview.

4.1 User Interface

The Streamlit-based interface provides a simple environment for interacting with the system. The user can upload a document and initiate its analysis. The results are then presented in a structured manner.

The interface is designed to be suitable for users who may not have technical or programming knowledge. The software-only implementation also makes the system easy to demonstrate and operate on a normal computer without requiring additional hardware.

4.2 Document Classification

Document classification provides an immediate indication of the nature of the uploaded file. Different legal documents contain different types of information. For example, privacy policies generally focus on information collection and usage, while contracts may contain obligations, responsibilities, and conditions.By classifying the document, TermsBuster provides additional context for the analysis.

4.3 Summarization

The summarization feature provides users with a concise overview of the document. This isparticularly useful when the uploaded document contains a large amount of text.Instead of reading the entire document immediately, users can first inspect the generated overview and then return to the original document for detailed examination.The summary should be considered an informational representation of the document and not a replacement for reading legally significant clauses.

4.4 Risk Indication

The risk-analysis module is designed to draw attention to potentially important statements. It uses configured rules, keywords, sentence-level information, negation handling, and safe-phrase information.The system does not claim that a document is legally valid, illegal, safe, or unsafe based only on automated analysis. Instead, the result helps users identify areas that may require further attention.This approach is important because legal meaning often depends on context, jurisdiction, definitions, and relationships between different clauses.

4.5 Advantages

The major advantages of TermsBuster include:

1. **Simple document upload:** Users can submit documents through a straightforward interface.
2. **Multiple document categories:** The system supports different types of legal and privacy-related documents.
3. **Automated text analysis:** Important processing stages are performed automatically.
4. **Structured summary:** The system provides an accessible overview of document content.
5. **Important-term identification:** TF-IDF analysis helps identify informative terms.
6. **Risk indication:** Rule-based analysis helps users identify content that may require additional attention.
7. **User-friendly interface:** Streamlit provides an interactive environment.
8. **Software-only implementation:** No special hardware is required.

4.6 Limitations

Although TermsBuster provides useful first-level document analysis, it has several limitations.

First, the quality of analysis depends on the quality of extracted text. Poorly scanned or incorrectly formatted documents may affect the subsequent NLP stages.

Second, legal documents contain context-dependent language. A keyword may have different meanings depending on the surrounding clause.

Third, the rule-based risk-analysis approach depends on predefined rules and configured information. It cannot identify every possible legal issue.

Finally, TermsBuster is not a replacement for professional legal advice. Users should consult qualified legal professionals when making important legal, financial, contractual, or regulatory decisions.

5. CONCLUSION

TermsBuster is an AI-powered smart privacy policy and legal document analysis assistant developed to make complex legal and privacy-related documents easier to understand at a first level. The system integrates document upload, text extraction, preprocessing, document classification, summarization, TF-IDF-based important-term analysis, and rule-based risk indication within a single Streamlit application.The proposed system reduces the effort required to obtain an initial overview of a lengthy document. Instead of manually searching through a document for its category, important terms, and potentially significant statements, the user can submit the document and receive structured information through the application.

The use of TF-IDF provides a method for identifying informative terms, while rule-based analysis provides an understandable approach for highlighting potentially important content.The classification module further improves the organization of the analysis by identifying the general type of document.

ACKNOWLEDGEMENTS

The author expresses sincere thanks to the Department of Computer Applications, Vellalar College for Women, Bharathiar University, Tamil Nadu, for providing the academic environment and support required to carry out this project.

6. REFERENCES

[1] M. Bokaie Hosseini, P. K. C, I. Reyes, and S. Egelman, “Identifying and Classifying Third-party Entities in Natural Language Privacy Policies,” *Proceedings of the Second Workshop on Privacy in NLP*, pp. 18–27, 2020.

[2] I. Chalkidis, M. Fergadiotis, P. Malakasiotis, N. Aletras, and I. Androutsopoulos, “LEGAL-BERT: The Muppets Straight Out of Law School,” *Findings of the Association for Computational Linguistics: EMNLP 2020*, pp. 2898–2904, 2020.

[3] D. Mamakas, P. Tsotsi, I. Androutsopoulos, and I. Chalkidis, “Processing Long Legal Documents with Pre-trained Transformers: Modding LegalBERT and Longformer,” *Proceedings of the Natural Legal Language Processing Workshop 2022*, 2022.

[4] N. Limsopatham, “Effectively Leveraging BERT for Legal Document Classification,” *Proceedings of the Natural Legal Language Processing Workshop 2021*, pp. 210–216, 2021.

[5] M. Masala, R. C. A. Iacob, A. S. Uban, M. Cidota, H. Velicu, T. Rebedea, and M. Popescu, “jurBERT: A Romanian BERT Model for Legal Judgement Prediction,” *Proceedings of the Natural Legal Language Processing Workshop 2021*, pp. 86–94, 2021.

[6] Y. Shvartzshanider, A. Balashankar, T. Wies, and L. Subramanian, “Beyond the Text: Analysis of Privacy Statements through Syntactic and Semantic Role Labeling,” *Proceedings of the Natural Legal Language Processing Workshop 2023*, pp. 85–98, 2023.

[7] A. Shukla, P. Bhattacharya, S. Poddar, R. Mukherjee, K. Ghosh, P. Goyal, and S. Ghosh, “Legal Case Document Summarization: Extractive and Abstractive Methods and their Evaluation,” *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing*, pp. 1048–1064, 2022.

[8] B. Sun and X. Zhou, “ignore at SemEval-2024 Task 5: A Legal Classification Model with Summary Generation and Contrastive Learning,” *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*, pp. 530–535, 2024.

[9] E. Lukose, S. De, and J. Johnson, "Privacy Pitfalls of Online Service Terms and Conditions: A Hybrid Approach for Classification and Summarization," *Proceedings of the Natural Legal Language Processing Workshop 2022*, pp. 65–75, 2022.

[10] This list has a problem: strict alphabetical order would require E. Lukose before I. Chalkidis, so let me give you the genuinely correct final sequence below.