AUTOMATIC SUMMARIZATION OF LEGAL DOCUMENTS USING SUMY

Supported by tetfund

Shakirat Aderonke Salihu[1], Abdullahi Musa[1], Fatima Enehezei Usman-Hamza[1], Abimbola Ganiyat Akintola[1], Abdullateef Oluwagbemiga Balogun[2] , Hammed Adeleye Mojeed[1,3] and Ghaniyyat Bolanle Balogun[1]

[1]Department of Computer Science, University of Ilorin, Ilorin. Nigeria.

[2]Department of Computer and Information Sciences, Universiti Teknologi PETRONAS, Perak, Malaysia

[1,3]Department of Applied Computer Science, Institute of Ocean Engineering and Ship Technology, Gdansk University of Technology, Gdansk, Poland

corresponding author: salihu.sa@unilorin.edu.ng

Received: September 14, 2023 Accepted: November 28, 2023

Abstract : Automatic Text Summarization (ATS) is a natural language processing technique that attempts to extract or generate a shorter version of lengthy text while preserving the overall context of the text. The rise in digital organization has resulted in an influx of lengthy digital contracts, including Privacy Policy (PP) and Terms of Service (ToS). Consequently, this led to the aim of this paper which is to develop an ATS system. In this paper, Sumy, a Python-based library was utilized for the efficient summarization of these lengthy contracts both in plain text and URL. The Sumy library employs multiple extractive techniques such as Luhn, Edmundson, Latent Semantic, and Textrank to carry out text summarization. It also accommodates multiple languages as input. Following an in-depth assessment of these techniques, it can be concluded that the Latent Semantic Analysis (LSA) technique performs best on PP and ToS with an F1-score of 76.1% while Luhn has the lowest percentage of 46.3%. It is recommended that organizations adopt the use of this system to enhance contract readability and it also saves time.

Keywords: Contracts, Privacy Policy, SUMY, ATS, TOS

Introduction

At present, it has become common to create contracts for any agreement made between two parties; the number of documents managed by companies, businesses, and individuals is on a rapid increase. It has become challenging for corporate employees and stakeholders to review contracts that can be quite lengthy, often spanning hundreds of pages of complicated legal texts. A document is legal if it is crafted to be enforceable in a court of law (Jain, Borah, & Biswas, 2021). The presence of legality within contracts is a crucial component of enforceable agreements. A contract is a legally binding agreement that creates, defines, and governs the mutual obligations and entitlements among its contracting parties (Sancheti, Garimella, Srinivasan, & Rudinger, 2022). In contract law, the legality of purpose pertains to legally binding and enforceable terms in legal documents. These documents exhibit a quite long structure compared to a universal document, resulting in challenges when it comes to reading and comprehension. A potential solution to this is to summarize these lengthy documents into shorter versions manually by legal experts. However, this procedure is costly and demands substantial time. Some form of automation or simplifying the process could assist legal professionals in handling this workload better (Jain et al., 2021).

This study proposes the use of automated text summarization as a possible solution. Automatic text summarization is the creation of a summary that maintains both meaningful content and the overall context of the original documents (Tarun, Machiraju, Adarsh, Gorrela, & Suresh, 2022). A quality summary is a short form of a document that encompasses all the important information present in the original document (Jain et al., 2021). One area where automated text summarization can be particularly useful is in the legal domain, where contracts, bylaws, licenses, privacy policies, terms of service, and terms and conditions are critical documents that require careful attention. Automated summarization of legal documents has become more crucial in the legal field, as it helps legal professionals and stakeholders to efficiently understand the main points and arguments contained in lengthy legal documents.

With the various types of legal documents available, this study focuses on digital contracts with a particular emphasis on Privacy Policy (PP) and Terms of Service (ToS). In various situations, organizations involved in data management strive to abide by notice policies by providing individuals with consent materials, usually presented as privacy policies and terms of service policies. These policies are commonly found on websites, applications or distributed through mail, typically when an individual connects with the organization for the first time, and subsequently when policy changes (Obar & Oeldorf-Hirsch, 2016). Individuals utilize smartphones for collecting and exchanging digital information, connecting on social media, entertaining oneself, accessing online banking, and various other purposes. Virtually every application installed and website browsed has its own Privacy Policy (PP) and Terms of Service (ToS). These contracts bind individuals as soon as they power on their phones or browse a website, even though people might not be fully aware of the terms they have just accepted (Lippi et al., 2019). Through regression analysis, it was determined that information overload significantly impacts negative predictors of reading ToS during signup as well as when ToS and PP change (Obar & Oeldorf-Hirsch, 2016). Individuals usually ignore privacy policies and terms of service when engaging with digital media due to the length; however, Individuals would take more time to read policies if they were shorter and clearer (CSM, 2010).

Various algorithms and libraries have been developed to facilitate automatic summarization. One of the most widely used libraries is the Sumy library, which is an open-source library developed in Python. This library provides several algorithms for extractive summarization, which entails selecting the most significant sentences from the original text to construct a summary. Some of these algorithms include Latent Semantic Analysis (LSA), LexRank, Luhn, Edmundson, TextRank, Reduction, SumBasic and Kullback-Leibler (KL). The algorithms are designed to

quickly analyze the document and generate a summary that accurately reflects the original document, while also making the document easier to read and understand. The goal of this project is to explore the effectiveness of these algorithms for automatic summarization of legal contracts. Automatic Text Summarization poses several challenges, including preserving the meaning and context of the original document and maintaining coherence and fluency of the summary (Syed, Uddin, Faraaz, Faisal, & Abdul, 2020). The problem with automatically creating a summary is the identification of the main topics of the document and the subsequent extraction of sentences that best describe the identified topics (Belica, 2013). Summarizing an entire document may not be effective, as the resulting summaries may be too general and overlook important details given that each line or section of the document carries a different level of importance. Hence, the solution is to first identify the preferred topics or headings that hold significance for inclusion in the summary. This method ensures that the generated summary is more accurate and tailored to the specific requirements of individual users (Balachandar, Saatvik Reddy, Shahina, & Khan, 2021).

Automatic text summarization is one of the major difficulties in the field of Natural Language Processing (NLP) and has gained considerable interest in recent times (Syed, Uddin, Faraaz, Faisal, & Abdul, 2020). NLP, a subfield of Artificial Intelligence, is concerned with human-computer interaction using natural language. Its objective is to enable computers to comprehend, interpret and produce human language, and it has many applications, including text summarization. The two major approaches to automatic text summarization are Extractive and Abstractive (Hoorn, 2018). Extractive Text Summarization seeks to identify significant sentences or phrases from the original document and merge them to create a summary (Paheli, Soham, Koustav, Kripabandhu, & Saptarshi, 2017). NLP techniques like text parsing, part-of-speech tagging, and named entity recognition are used to identify and extract the most informative sentences. Abstractive Text Summarization techniques, on the other hand, involve generating a summary that captures the main concepts of the text in new sentences which is coherent with the context of the provided document while also incorporating words and phrases that might not exist in the original document (Tarun et al., 2022). It requires more advanced NLP techniques and often uses deep learning algorithms like neural network and transformers.

The focus of this paper is the use of Luhn, Edmundson, Latent Semantic Analysis (LSA), TextRank and LexRank algorithm of Python's Sumy library to extract key sentences from contracts, such as privacy policy and terms of service, and generate a concise summary that highlights the most significant information. This study will compare the summarization effectiveness, accuracy, and relevance of the proposed system against each algorithm. The aim is to develop a system that outperforms other techniques in terms of accuracy and relevance while providing a more efficient and effective method of summarizing legal documents.

Related Works

Farzindar and Lapalme (2004) developed a summarization system, called LetSum (Legal Text Summarizer), to produce short summaries for the legal decision of the proceedings of a court. Their approach focuses on analyzing the architecture and thematic structures of a document to create a table-style summary that enhances coherency and readability. They presented LetSum, a prototype system that identifies four themes—Introduction, Context, Juridical Analysis, and Conclusion—to determine the thematic structure of a legal judgment. Then it identifies the relevant sentences for each theme and presents them as a table-style summary. This methodology helps in organizing the summary effectively. The summary is built in four phases: thematic segmentation, filtering of less important units (such as citations of law articles), selection of relevant textual units, and production of the summary within the size limit of the abstract. LetSum is among the limited systems crafted specifically for summarizing legal documents.

Terms of Service; Didn't Read (ToS; DR) [ToS; DR 2012] is a free software project that originated in 2012 to tackle the issue of minimal user engagement with the terms of service for websites. This initiative involves an online community of volunteers who read, analyze, and rate privacy policies. ToS; DR focuses on topics concerning user data and privacy. This website clarifies the language within legal documents by providing summaries for specific sections of the original documents. Although privacy policies addressed in this project are read and rated by humans and discussed thoroughly. However, the project encounters a challenge due to comprehensive lengthy discussions potentially undermining the efficacy of ratings. One might read and selectively read the original privacy policy itself. Furthermore, it is worth noting that the coverage of ToS; DR is relatively limited compared to privacy seals and new formats. Currently, only 66 privacy policies have been evaluated and rated by the platform. [ToS; DR 2012].

Belica's (2013) research work deals with summarizing documents in HTML format. This study focuses on text summarization algorithms. This work briefly discusses general text mining and later focuses on summarization. The text of this thesis deals with creating a summary from documents in the HTML2 format, which is commonly used in web environments. The main benefit of the application is the automatic processing of documents in HTML format, which is the dominant format on the web today. This work describes three selected summarization methods: Luhn's method, Edmundson's method, and a method based on the analysis of latent semantics. The methods mentioned here are implemented as part of the Sumy module for the Python language. Findings from the evaluation of the summarization of individual methods show that the most suitable method for summarizing richly semantically marked HTML documents is the Edmundson method. For texts with or without a minimal amount of metadata regarding the document, the LSA-based method appears to be a more suitable option. The findings from the conducted research are used in the implementation of a freely available module that is capable of summarizing any HTML document in the Czech language only. The implemented module is distributed as an open-source library available at the URL https://github.com/miso-belica/Sumy

Previous studies have shown that automatic summarization can help reduce the time and effort required for document review while maintaining high levels of accuracy. Automatic summarization can be applied to legal documents to improve the efficiency of the review process. Several tools have been proposed for analyzing legal contracts, including LexNLP (Rosen-Zvi et al., 2015) and ContraxSuite (Katz et al., 2017). These tools use natural language processing techniques to extract information from legal contracts such as clauses, definitions, and obligations. LexNLP is an open-source Python library that provides tools to extract information from legal documents. The library includes modules for extracting information such as dates, amounts, and contract clauses. The ContraxSuite is another open-source tool that provides a suite of tools for contract analysis. ContraxSuite includes modules for identifying key clauses in

contracts, extracting obligation and rights information, and identifying potential contract issues.

Seth, Pooja, and Ruihong (2016) introduce CaseSummarizer, an automated text summarization tool specifically designed for legal documents. The tool utilizes standard summary techniques that rely on word frequency, complemented by domain-specific knowledge. CaseSummarizer is implemented as a Python-based solution that integrates the comprehensive Natural Language Toolkit (NLTK) module for preprocessing tasks, including sentence segmentation. Sentences are scored using a TF × IDF matrix built from thousands of legal case reports. The scores are summed across each sentence and normalized by sentence length. This normalization procedure ensures that the system does not bias lengthy sentences. This highlights the significance of sentence-level summarization in legal documents and validates the efficacy of sentence extraction methods in producing helpful summaries. The extracted sentences should ideally offer a representation of the various sections of the case file. Notably, CaseSummarizer demonstrates favourable performance compared to non-domain-specific summarizers. The generated summaries are capable of providing a fairly accurate idea of the case context, although some significant points are missed. CaseSummarizer uses a set of legal keywords to identify the importance of the sentences in the input document.

Paheli, Soham, Koustav, Kripabandhu, & Saptarshi (2017) discovered that many existing algorithms for legal case document summarization lack a systematic integration of domain knowledge, which is crucial for determining the essential information that should ideally be included in a summary. To address this gap, they introduce an unsupervised summarization algorithm called 'DELSumm,' meticulously designed to integrate legal expert guidance into an optimization framework. Their focus centres on extractive summarization due to its prominence in legal case documents. Recognizing the limitations of existing methodologies, they aimed to devise an algorithm that systematically incorporates the distinct rhetorical segments within a case document. The goal is to determine which parts from each segment should be incorporated into the summary, guided by law practitioner principles. Their proposal, DELSumm (Domain-adaptive Extractive Legal Summarizer), is an unsupervised extractive summarization algorithm tailored for legal case documents. They frame the task of summarizing legal case documents to maximize the inclusion of the most informative sentences while ensuring balanced representation from all thematic segments and minimizing redundancy. The implementation of DELSumm is publicly available at https://github.com/Law-AI/DELSumm.

A prior study indicates that only a small percentage of users read online privacy policies even though they implicitly agree to them while using a website. Prior research also suggests that users disregard privacy policies due to their length and, on average, require two years of college education to comprehend. The proposed technique addresses this issue by employing data mining models to automatically extract summaries from online privacy policies. The Chrome browser extension called 'PrivacyCheck' utilizes these models to summarize HTML pages containing privacy policies. PrivacyCheck is unique among existing solutions as it can be applied to any online privacy policy, offering a convenient graphical summary for users. Through a literature review and a survey of privacy experts, ten crucial questions have been identified that users should inquire about regarding businesses' utilization of their Personal Information. The PrivacyCheck browser extension has been developed to automatically address these ten questions for any provided privacy policy using data mining classification models, trained on 400 policies and operated through a server (Zaeem, German, & Barber, 2018).

Manor and Li (2019) researched unilateral contracts, specifically focusing on terms of service agreements that hold significant importance in the context of modern digital life. They propose the task of automatically summarizing legal documents in plain English for a non-legal audience. They hope that such technological advancement will facilitate a broader audience to engage in everyday contracts with a better comprehension of the agreements they're entering into. Automatic summarization is often used to reduce information overload, particularly in the news domain. Summarization application in the legal genre has been limited, except for instances like judicial judgments and case reports. They curated a dataset from websites dedicated to simplifying complicated legal documents into plain English. Instead of attempting to summarize the entire document, these sources summarize each document at the section level. This approach enables readers to access more detailed content when necessary. They assessed extractive summarization methods and compared their performance with human-written summaries.

The study conducted by Mojeed et al.(2020) focuses on the generation of micro-summaries for journal articles. The process involves several key steps, starting with preprocessing the articles and segmenting them into distinct sections. Each section is treated as an independent unit, and information retrieval techniques, particularly utilizing the Vector Space Model (VSM) with Cosine similarity, are employed to measure sentence similarity and assign numerical weights to sentences. The study utilizes TF-IDF (Term Frequency-Inverse Document Frequency) for computing term weights, followed by Cosine similarity to gauge the similarity between sentences based on their term weight vectors. The clustering aspect of the approach employs the k-means algorithm to group sentences with similar content and create micro-summaries. These micro-summaries are systematically combined to form a final summary, taking into account factors such as informativeness, redundancy, and coherence.

In the work of Jain, Borah, and Biswas (2021), an extensive comparative analysis of multiple classical extractive techniques like Luhn, Latent Semantic Analysis (LSA), Edmundson, Textrank, Reduction, Lexrank, Kullback-Leibler (KL), and Sumbasic is presented. This analysis is carried out using the BillSum dataset, which serves as a publicly available benchmark for legal document summarization. Through comprehensive experimental evaluation, the study finds that graph-based methods, specifically Textrank and Lexrank, exhibit superior performance compared to frequency-based techniques. Notably, graph-based methods leverage sentence similarity, incorporating more than just word frequencies. One important thing to note here is that this research exclusively considers classical extractive summarization methods within the legal domain. From experimental observations, the study's experimental observations conclude that graph-based summarization techniques generally excel in legal document summarization tasks.

In the paper by Balachandar, Saatvik Reddy, Shahina, & Khan (2021), they present a novel system aimed at generating summaries for commercial contracts, including Non-Disclosure Agreements (NDAs) and employment agreements. The objective of this system is to reduce the time spent on contract reviews and enhance comprehension by providing concise summaries of the contract content. Given the prevalent structure of commercial documents featuring paragraphs with headings/topics followed by content and context, the authors observe that extracting these topics and customizing their summarization to meet user requirements is a more efficient strategy. Rather than summarizing the entire document, the authors suggest that

focusing on specific paragraphs or topics aligned with specific needs yields a more effective outcome. This targeted summarization approach allows for greater relevance and customization, ensuring that the most relevant and essential information is captured for the intended purpose. They employ extractive summarization methods and assess their performance in comparison to summaries generated by humans. The study's findings indicate that the outcomes achieved through extractive techniques are deemed satisfactory.

Despite the advancements in summarization methods, it is important to note that these methods are not tailored specifically for legal contracts and may produce poor summaries when applied. Having reviewed all these articles, none of these studies have made use of the Sumy Library for summarizing Legal Contracts; hence, this research intends to find the best extractive algorithm that suits the summarization of Legal Contracts.

Methodology

This paper proposes Automatic Text Summarization (ATS) of legal documents such as Privacy Policy (PP) and Terms of Service (ToS), using various extractive SUMY approaches. This section describes Automatic Text Summarization (ATS) methods such as data collection, text pre-processing, summarization algorithms, and tools utilized throughout the development. In this paper, the system is enhanced to input data as plain text or URL links. The dataset used is PP and ToS, which is then preprocessed such that the data is in a usable format. Subsequently, the preprocessed data are subjected to Sumy algorithms for summarization. Figure 1 is the framework of the proposed method; this gives a detailed description of the methods used for the system

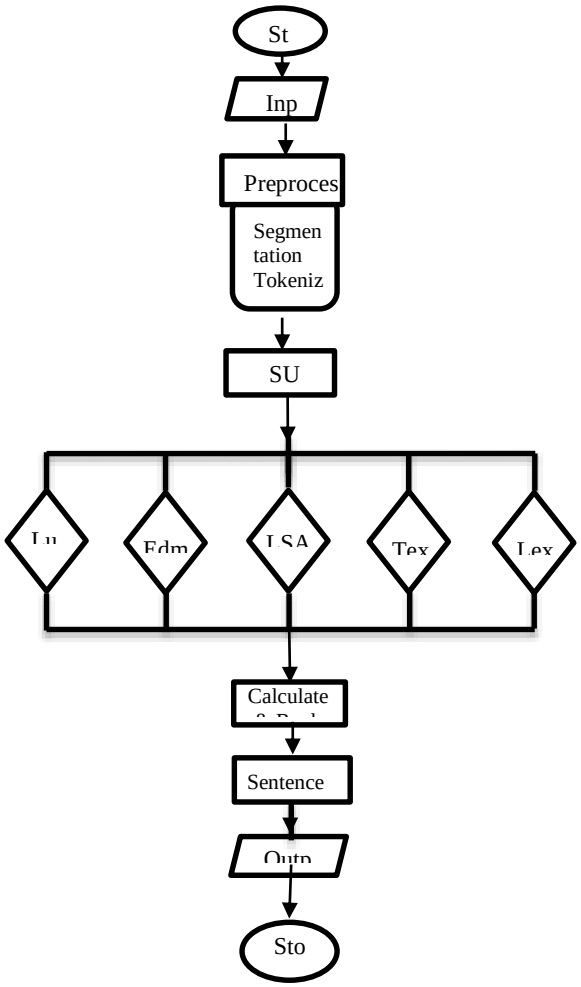

Figure 1. The framework of the Proposed Method

Description of the Dataset

The dataset collected consists of only PP and ToS both in plain text and URLs from diverse companies' websites. The plain text consists of 20 PP and 34 ToS, while the URL consists of 22 PP and 26 ToS extracted. Table 1 shows a brief description of the dataset.

Table 1. Description of the dataset

Dataset	**Privacy Policy (PP)**	**Terms of Service (ToS)**
Plain text	20	34
URL	22	26

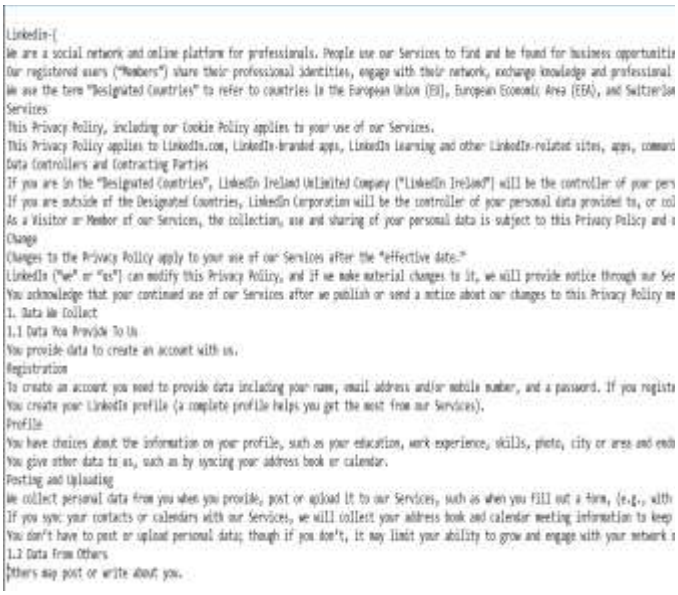

Figure 2: Samples of Privacy Policies collected

Text Preprocessing

Text preprocessing is an important step that involves taking raw data into a format that is comprehensible and analyzable through the identification and removal of nonfunctional contents from the data. Text preprocessing techniques were performed on the dataset to improve the overall performance. This includes sentence segmentation (automatically performed using period (.)), word tokenization (using the NLP model), stemming (using pre-built modules according to NLP), removing stop words (using pre-built modules) and converting uppercase letters to lowercase letters to have a balanced weight. Figure 3 depicts a sample of the dataset after preprocessing

Application of the SUMY Algorithms on the Preprocessed Dataset

This paper employs the use of only Luhn, Edmundson, LSA, Textrank and Lexrank algorithms from the Sumy library with their implementation shown in Figure 3.

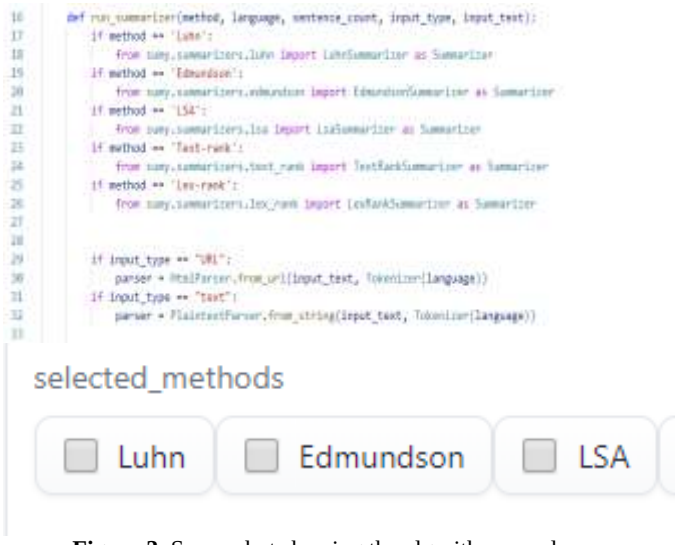

Figure 3. Screenshot showing the algorithms used

Results and Discussions

The results were presented through the creation of a user interface with the utilization of the Gradio library. The system allows multiple algorithms to be selected for summarization, which is accomplished by inputting either plain text or a URL. Furthermore, it provides the capability to select the language of the input text, thereby facilitating the selection of an appropriate preprocessing module for the system's operation. Figure 4 is a sample of the selection process.

Figure 4. Screenshot showing the User Interface on the web browser

Results for Luhn Summarizer

Figure 5 illustrates the application of the Luhn summarizer showing the input text and output summary and the results obtained can be seen in Table 2. It shows a comparison between the number of words generated by the Luhn summarizer and the number of words derived by human experts through manual summaries.

Figure 5. Screenshot showing input and output text using Luhn summarizer

Table 2. Number of words in summary generated by Luhn Summarizer

Contracts	No. of words in original text	No. of words in a generated summary	No. of words in a human summary
Behance PP	462	184	110
Snap ToS	5296	1051	620
Huawei ToS	1391	491	359
Proinvest PP	285	102	94
Proinvest ToS	517	180	142

The performance evaluation of the Luhn summarizer was done to show the effectiveness of the method using precision, recall, and F-measure with ROUGE metrics. These were calculated based on some selected PP and ToS documents. Table 3 depicts the results of the evaluation.

Table 3. Performance evaluation of the Luhn Summarizer

Contracts	Precision	Recall	F1-Score
Behance PP	0.634	0.547	0.419
Snap ToS	0.649	0.660	0.536
Huawei ToS	0.764	0.697	0.522
Proinvest PP	0.600	0.610	0.450
Proinvest Tos	0.516	0.659	0.385

Results for Edmundson Summarizer

The results are shown in Figure 6. The text was inputted and the output text was generated at the other side of the interface while Tables 4 and 5 depict the summary of the results and its performance evaluation respectively.

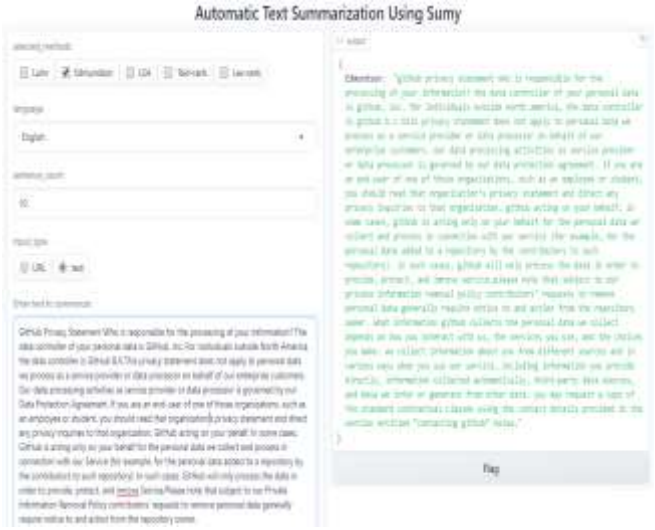

Figure 5. Screenshot showing input and output text using Edmundson summarizer

Table 4. Number of words in summary generated by Edmundson Summarizer

Contracts	No. of words in original text	No. of words in a generated summary	No. of words in a human summary
Behance PP	462	134	110
Snap ToS	5296	716	620
Huawei ToS	1391	566	359
Proinvest PP	285	111	94
Proinvest ToS	517	91	142

The accuracy of the system was evaluated using the ROGUE metrics as presented in Table 5

Table 5. Performance evaluation of the Edmundson Summarizer

Contracts	Precision	Recall	F1-Score
Behance PP	0.591	0.786	0.534
Snap ToS	0.667	0.777	0.618
Huawei ToS	0.728	0.794	0.615
Proinvest PP	0.669	0.709	0.537
Proinvest Tos	0.515	0.771	0.542

Results for LSA Summarizer

Figure 7 illustrates the application of the LSA summarizer showing the input text and output summary, while Tables 6 and 7 show the summary of the results and the performance evaluation.

Figure 7. Screenshot showing input and output text using LSA summarizer

Table 6. Number of words in summary generated by LSA Summarizer

Contracts	No. of words in original text	No. of words in a generated summary	No. of words in a human summary
Behance PP	462	144	110
Snap ToS	5296	1124	620
Huawei ToS	1391	594	359
Proinvest PP	285	97	94
Proinvest ToS	517	116	142

The accuracy of the system was evaluated using the ROGUE metrics as presented in Table 7

Table 7. Performance evaluation of the LSA Summarizer

Contracts	Precision	Recall	F1-Score
Behance PP	0.565	0.636	0.598
Snap ToS	0.801	0.762	0.781
Huawei ToS	0.766	0.699	0.734
Proinvest PP	0.849	0.792	0.819
Proinvest Tos	0.663	0.722	0.691

Results for Textrank Summarizer

The inputting of text and summary output is represented in Figure 8 for the Textrank summarizer. Tables 8 and 9 show the summary and the performance evaluation of the algorithm.

Figure 8. Screenshot showing input and output text using Textrank summarizer

Table 8. Number of words in summary generated by Textrank Summarizer

Contracts	No. of words in original text	No. of words in a generated summary	No. of words in a human summary
Behance PP	462	121	110
Snap ToS	5296	1468	620
Huawei ToS	1391	619	359
Proinvest PP	285	121	94
Proinvest ToS	517	178	142

The accuracy of the system was evaluated using the ROGUE metrics as presented in Table 9

Table 9. Performance evaluation of the Textrank Summarizer

Contracts	Precision	Recall	F1-Score
Behance PP	0.669	0.539	0.451
Snap ToS	0.771	0.456	0.384
Huawei ToS	0.712	0.514	0.436
Proinvest PP	0.617	0.484	0.437
Proinvest Tos	0.516	0.559	0.485

Results for Lexrank Summarizer

Text for Lexrank summarizer was also input and the output summary is depicted in figure 9, while tables 10 and 11 show the summary and the performance evaluation.

Figure 9. Screenshot showing input and output text using Lexrank summarizer

Table 10. Number of words in summary generated by Lexrank Summarizer

Contracts	No. of words in the original text	No. of words in a generated summary	No. of words in a human summary
Behance PP	462	156	110
Snap ToS	5296	801	620
Huawei ToS	1391	581	359
Proinvest PP	285	104	94
Proinvest ToS	517	111	142

The accuracy of the system was evaluated using the ROGUE metrics as presented in Table 11

Table 11. Performance evaluation of the Lexrank Summarizer

Contracts	Precision	Recall	F1-Score
Behance PP	0.522	0.579	0.440
Snap ToS	0.609	0.669	0.531
Huawei ToS	0.712	0.578	0.494
Proinvest PP	0.590	0.631	0.405
Proinvest Tos	0.526	0.684	0.494

The paper went further to do an empirical investigation into diverse text summarization techniques, as measured by their average ROUGE metrics—precision, recall, and F1 score. A careful analysis of these metrics yields significant insights into the efficacy of each of the summarizers. It can be deduced that the summarizers performed well but the LSA summarizer possesses the highest percentage across all the three metrics, this implies that LSA performs best in summarizing PP and ToS. The average summary for all the metrics and summarizers is depicted in Table 12.

Table 12. Average ROGUE metrics for the summarizers

Contracts	Precision	Recall	F1-Score
Luhn	0.633	0.635	0.463
Edmundson	0.634	0.767	0.569
LSA	0.729	0.723	0.761
Textrank	0.657	0.610	0.486
Lexrank	0.592	0.628	0.473

4. Conclusion

The automatic Text Summarization (ATS) technique involves the use of technology or software to produce a shorter version of documents that includes relevant information from the document. The main objective of this study is to utilize Sumy which offers a range of extractive summarization algorithms such as Luhn, Edmundson, LSA, TextRank, LexRank, and more to summarize PP and ToS. It supports multiple languages and can generate concise summaries from both web URLs and plain text. It can be deduced based on the outcomes of the summaries when subjected to ROUGE metrics that the Latent Semantic Analysis (LSA) algorithm presents a promising result as compared with other algorithms. It is imperative to emphasize that the adoption of Automated Text Summarization (ATS) holds the potential to significantly augment efficiency and productivity. However, this technology needs to align seamlessly with the proficiency of legal practitioners. Given that legal documents typically manifest as segmented compositions, an optimal ATS system should possess the capability to detect and summarize content on a section-by-section basis. This approach ensures a coherent and precise narrative flow within the generated summaries. The authors intend to break the documents down into sections by sections and use other languages in future.

Reference

Balachandar, K., Saatvik Reddy, A., Shahina, A., & Khan, N. (2021). *Summarization of Commercial Contracts*. Paper presented at the Machine Learning, IOT and Blockchain Technologies & Trends. http://10.5121/csit.2021.111202

Belica, M. (2013). *Methods of Document Summarization on the Web [online]*. (M.Sc), BRNO University of Technology, Brno, Czech. Retrieved from http://hdl.handle.net/11012/53529

Common Sense Media (2010) comment submitted to A Preliminary Staff Report on protecting Consumer Privacy in an ERA of Rapid Change, File No. PO95416, comment number 00457

Farzindar, A., & Lapalme, G. (2004). *LetSum, an Automatic Text Summarization system in Law field*. Paper presented at the Jurix 2004: the Sevententh Annual Conference, Berlin.

Hoorn, W. V. (2018). *Automatic Text Summarization As A Text Extraction Strategy For Effective Automated Highlighting.* (Bachelor Bachelor), RadBound University, RadBound. (s4018044)

Jain, D., Borah, M. D., & Biswas, A. (2021). *Automatic Summarization of Legal Bills: A Comparative Analysis of Classical Extractive Approaches*. Paper presented at the 2021 International Conference on Computing, Communication, and Intelligent Systems (ICCCIS).

Katz, D. M., Bommarito, M. J., Soellinger, T., & Wu, G. (2017). General purpose tools for modeling and reasoning about legal contracts. Artificial Intelligence and Law, 25(3), 319-349.

Lippi, M., Palka, P., Contissa, G., Lagioia, F., Micklitz, H.-W., Sartor, G., & Torroni, P. (2019). CLAUDETTE an Automated Detector of Potentially Unfair Clauses in Online Terms of service. *Artificial Intelligence and Law*, pp. 18. doi:10.1007/s10506-019-09243-2

Manor, L., & Li, J. J. (2019). *Plain English Summarization of Contracts.* Paper presented at the Proceedings of the Natural Legal Language Processing Workshop

Mojeed, H.A. et al. (2020). An Approach for Journal Summarization Using Clustering Based Micro-Summary Generation. In: Silhavy, R., Silhavy, P., Prokopova, Z. (eds) Software Engineering Perspectives in Intelligent Systems. CoMeSySo 2020. Advances in Intelligent Systems and Computing, vol 1295. Springer, Cham. https://doi.org/10.1007/978-3-030-63319-6_64

Obar, J. A., & Oeldorf-Hirsch, A. (2016). *The biggest lie on the internet: Ignoring the privacy policies and terms of service policies of social networking services*. Paper presented at the TPRC 44: The 44th Research Conference on Communication, Information and Internet Policy, https://dx.doi.org/10.2139/ssrn.2757465

Paheli, B., Soham, P., Koustav, R., Kripabandhu, G., & Saptarshi, G. (2017). *Incorporating domain knowledge for extractive summarization of legal case documents*. Paper presented at the Conference'17, Washington, DC, USA.

Rosen-Zvi, M., Marom, S., Finkelstein-Landau, A., Aharonovitz, A., Rappoport, A., & Shahaf, D. (2015). LexNLP: Natural Language Processing and Information Extraction For Legal and Regulatory Texts. In Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 225-234).

Sancheti, A., Garimella, A., Srinivasan, B. V., & Rudinger, R. (2022). What to Read in a Contract? Party-Specific Summarization of Important, Obligations,

Entitlements, and Prohibitions in Legal Documents*vol 1*, 15. Retrieved from https://doi.org/10.48550/arXiv:2212.09825

Seth, P., Pooja, J., & Ruihong, H. (2016). *CaseSummarizer A System for Automated Summarization of Legal Texts.* Paper presented at the Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: System Demonstrations, Osaka, Japan.

Syed, M., Uddin, H., Faraaz, M. K., Faisal, K., & Dr. Abdul, S. (2020). Text Summarization using Natural Language Processing. *Journal of Engineering Sciences, Vol 11*(Issue 4).

Tarun, S. K., Machiraju, S. S., Adarsh, V., Gorrela, R. D., & Suresh, C. (2022). Automatic Text Summarizer Application Using Extractive Text Summarization. *journal of Engineering Sciences, vol. 13*(issue 03), pp. 768-774.

TOS;DR, 2023,[Online]. Available: https://tosdr.org/

Zaeem, R. N., German, R. L., & Barber, K. S. (2018). PrivacyCheck- Automatic Summarization of Privacy Policies Using Data Mining. *CM Transactions on Internet Technology, Vol. 9*(Article 39), pp. 18. doi:10.1145/0000000.0000000