Automatic Summarisation of Zambian Legislative Documents

By

Martin Kambunji (2019016249)

Emmanuel Phiri (2019010810)

Castridah Nachibinga (2019075237)

Ernest Sinyangwe (2019070201)

Supervisor:

Dr. Lighton Phiri

A report submitted to the Department of Library and Information Science, The University of Zambia, in partial fulfilment of the requirements of the degree of Bachelor of Information and Communication Technologies with Education

THE UNIVERSITY OF ZAMBIA

LUSAKA

2023

Abstract

The objective of this study was to examine the obstacles individuals face when attempting to comprehend legislative documents, specifically acts of parliament. In addition, our aim was to devise and assess a model capable of generating concise summaries of these extensive documents. The National Assembly of Zambia is steadfast in its commitment to enhancing public perception and understanding of their institution by ensuring transparency and facilitating accessibility to parliamentary proceedings. This effort is supported by the strengthening of ICT platforms to encourage public engagement. However, comprehending the legislation made available through entities like the Zambia Legal Information Institute (ZambiaLII) proves challenging due to the extensive size of the documents and the intricate vocabulary employed. Acknowledging the potential of Language Processing (NLP) techniques and enhanced accessibility, this study endeavours to tackle this challenge by developing software tools that automate the summarization of legislative documents, with a specific focus on Acts of parliament. Our approach entails harnessing text mining and designing user-friendly tools to effectively achieve this objective.

Acknowledgements

We are filled with immense gratitude and humility as we gather to express our heartfelt thanks to God for His unwavering support, guidance, and blessings that enabled us to successfully complete this project. Throughout this journey, His divine presence has been our constant source of strength and inspiration.

However, to our exceptional supervisor, we extend our deepest appreciation for their pivotal role in our achievements. Their unwavering dedication, guidance, and belief in our potential helped us navigate the challenges and uncertainties, leading us towards success. Their mentorship has been invaluable, and we are truly grateful for their support.

Furthermore, our heartfelt thanks also goes out to our dear friends and all those who participated in this project. Without your dedication and efforts, this endeavour would not have been possible. Your collaboration and unwavering support have been instrumental in our triumph.

Lastly, we want to extend our profound gratitude to our respective families. Your boundless love, understanding, and encouragement have been the driving force behind our perseverance and determination. Your constant support has been the bedrock on which we built our success.

We acknowledge that this achievement would not have been possible without the collective efforts and blessings of God, our supervisor, our friends, and our families. Together, you have made this journey immensely rewarding, and we feel truly blessed to have you in our lives.

Thank you all for being an integral part of this remarkable milestone in our lives.

Table of Contents

List of Tables

List of Figures

List of Abbreviations

Abbreviation	**Description**
AI	Artificial Intelligence
CRISP-DM	Cross Industry Standard Process for Data Mining
UNZA	The University of Zambia

CHAPTER 1

1. Introduction

Legislative documents, particularly acts of parliament, serve as the cornerstone of a nation's legal framework, shaping the rules and regulations that govern society. However, these documents are often labyrinthine in their complexity, with extensive content and intricate language that can pose significant obstacles to the understanding of the general public. In this era of information accessibility, it is imperative that citizens have the means to comprehend the laws that affect their daily lives, fostering transparency and engagement with the legislative process.

The National Assembly of Zambia recognizes the importance of bridging this comprehension gap and is firmly committed to improving the public's perception and understanding of their institution. To achieve this, the Assembly has been working tirelessly to enhance transparency and facilitate accessibility to parliamentary proceedings, utilizing Information and Communication Technology (ICT) platforms to encourage public engagement. Yet, despite these efforts, the formidable challenge of comprehending legislative documents remains, primarily due to their voluminous nature and the complexities of the language employed.

The Zambia Legal Information Institute (ZambiaLII), among other entities, has made these legislative documents accessible to the public. However, the sheer size of these documents and the complex legal terminology used within them continue to present barriers for comprehension.

In acknowledgment of the tremendous potential of Natural Language Processing (NLP) techniques and technology to enhance accessibility, we have embarked on a comprehensive study aimed at addressing this challenge. Our objective is to develop software tools that automate the summarisation of legislative documents, with a specific focus on Acts of parliament. By harnessing the power of text mining and designing user-friendly tools, we endeavour to simplify and streamline the understanding of these critical documents, making them more comprehensible to a broader audience.

This project seeks to explore the Challenges individuals encounter when attempting to comprehend legislative documents and, in response, to devise and evaluate a model for generating concise summaries of these extensive documents. Through these efforts, we aim to contribute to the broader goal of enhancing public engagement with the legislative process, furthering the principles of transparency, and empowering citizens with a clearer understanding of their legal framework. In the subsequent sections of this report, we will Explore the methodology, tools, and findings that support our quest to develop a Natural Language Processing (NLP) model for automatic summarisation of legislative documents.

CHAPTER 2

2. Problem Statement

Legislative documents are an integral part of any legal system, these documents are often written in complex and technical language that is difficult for the average person to understand. However, these documents can be quite lengthy, which makes it challenging for laypeople to read and understand the legislative information within them. The problem can be more pronounced in developing countries where the level of literacy may be lower, and access to legal advice may be limited. As a result, people may not be aware of their legal rights and obligations, which can lead to legal disputes and confusion.

In Zambia, the National Pension Scheme (Amendment) Act of 2023 is a recent legislative development. This Act introduces crucial amendments to the National Pension Scheme, aiming to enhance the country's pension system. Despite its significance, there is a lack of awareness among the general public regarding the changes brought about by this Act. The amendments address various aspects such as contribution rates, eligibility criteria, and benefits, with the intention of ensuring a more robust and sustainable pension scheme for citizens. However, the Act's impact has been limited due to the lack of awareness among citizens, who struggle to keep up with the frequent amendments and changes to the laws. Efforts to raise public awareness and promote the Act's provisions are crucial to bridge the gap between legal complexities and ordinary citizens' understanding, ultimately fostering a more inclusive and informed society.

However, to address this problem we propose to build a system that utilises machine learning algorithms and Natural Language Processing (NLP) techniques to extract and summarise the most critical information from lengthy legal documents. By providing a concise summary of legal documents, the system would make it easier for individuals to understand the legal information contained within them, leading to a more transparent and effective legal system.

2.1 Objectives Of the Study

2.1.1 Main Objective

The primary goal is to develop an automated summarisation system that utilises the potential of machine learning and natural language processing to greatly enhance the accessibility and understanding of legislative documents in Zambia for the broader population.

2.1.2 Specific Objectives

1. Identifying the challenges that people have in understanding legal documents.

2. Design and implement the natural language processing model for summarising legal documents.

3. To evaluate the effectiveness of the natural language processing model.

2.1.3 Research Questions

1. What are the challenges that people have with regards to understanding legislative documents?

2. What is the feasibility of designing and implementing a natural language processing model for summarising legislative documents?

3. What is the overall effectiveness of the natural language processing model in achieving its intended objectives?

2.1.4 Ethical consideration

The success of any research relies on adhering to ethical guidelines, which are of vital importance in ensuring the appropriateness and integrity of the research, especially when it involves the general public. Therefore, this research will prioritise ethical procedures to ensure the following:

1. Data Privacy and Confidentiality: Handle sensitive information appropriately, adhere to data protection regulations, anonymize personal information, and implement robust security measures to protect the privacy of individuals involved.

2. Accuracy and Integrity of Information: Give utmost importance to accuracy, minimise errors, biases, and misinterpretations in the generated summaries, ensuring that the information presented is reliable and trustworthy.

3. Transparency and Explainability: Provide transparency in the summarization process, clearly communicate to users about the limitations and potential biases associated with the system to prevent any misleading or incomplete interpretations.

4. Fair Representation and Avoidance of Bias: Strive to present a balanced representation of legislative documents, avoiding any favouritism towards specific perspectives, and ensuring impartiality in the summaries.

5. Informed Consent and User Understanding: Inform users about the summarization process, its limitations, and the fact that summaries are generated by a machine, emphasising that they may not capture the full complexity of the original documents. Obtain informed consent from users before their participation.

6. Collaboration with Legal Experts and Stakeholder Engagement: Engage legal professionals, policymakers, and civil society organisations throughout the research process to ensure the system's appropriateness, effectiveness, and ethical considerations are comprehensively addressed.

7. Tof MySQL and the provihe collected data will be stored and managed using the MySQL database management system (DBMS). To accommodate the storage requirements, we will request server space from ZAMREN (Zambia Research and Education Network) and CICT (Center for Information and Communication Technology) department. This server space will provide the necessary infrastructure and resources to host the MySQL database securely. By leveraging the capabilities ded server space, we can ensure the efficient storage, organisation, and management of the collected data, enabling effective summarisation efforts.

By considering and integrating these ethical considerations, the project aims to develop an automatic summarization system that upholds ethical standards, respects user rights, and enhances the accessibility and understanding of Zambian legislative documents, ultimately benefiting the broader public.

CHAPTER 3

3. Related Work

In the realm of understanding legislative documents, various challenges arise that impede efficient comprehension and navigation. This section delves into the hurdles associated with grasping the nuances of legislative documents, highlighting complexities in terminology, ambiguity, length, and the absence of plain language. The exploration of these challenges sets the stage for the subsequent discussions on the importance of institutional repositories and document summarization techniques.

3.1 Challenges understanding Legislative Documents

Complex Legal Terminology: The major issue with legal documents is the insertion of technical definitions in the middle of clear sentences. Additionally, the study reveals numerous other factors that can confuse readers. For instance, rental contracts often employ terms such as "lessee" and "lessor," making them difficult to understand. The researchers suggest that lawyers could easily replace these words with more commonly used terms like "tenant" and "landlord."[1]

Ambiguity and Vagueness: Legal documents can sometimes be ambiguous or vague, leaving room for interpretation. Unclear definitions, imprecise language, or open-ended provisions can create confusion for individuals trying to understand their rights and obligations.[3]

Length and Complexity: Legal documents tend to be lengthy and contain multiple clauses, cross-references, and legal frameworks. The sheer volume of information, coupled with complex sentence structures and convoluted organisation, can overwhelm readers and make it challenging to extract the intended meaning. [2]

Lack of Plain Language: Legalistic documents in Zambia often lack the use of plain language, which is language that is clear, concise, and easily understandable by the general public. The absence of plain language principles can hinder comprehension, especially for individuals without legal training.[6]

Masson and Tahir [2] highlighted the challenges that civil society organisations (CSOs) in Zambia face in regards to understanding Legislative Documents. The authors explained how a small CSO faced challenges in interpreting legal information when it needed to file an injunction against a proposed mining operation in the Lower Zambezi National Park. The paper further explained how the small CSO ended up hiring a Lawyer who was inexperienced in environmental law.The small CSO's reliance on laws and reports published in PDF format by the Zambia Environmental Management Agency for its legal information needs proved insufficient due to the legal nature of the Documents. At the time of the case, the CSO was ill equipped to argue its case against a team of five lawyers hired by the opposition.

3.2 Two classic approaches to document summarisation: Abstractive vs Extractive summarisation

Abstraction summarization is the process of generating a summary that captures the main information and meaning of the original article or text through the use of new sentences. This approach involves a thorough analysis of the text and the capability to generate fresh sentences, resulting in more accurate summaries that minimise redundancy and maintain an effective compression rate [5]. On the other hand, Extractive summarization is a method of generating summaries by selecting important phrases, sentences, or elements directly from the original text. The goal is to create a concise representation of the main information while maintaining the core meaning of the original text. [5] Extractive summarization techniques focus on identifying and extracting salient content from the source material rather than generating new sentences from scratch. Various features, such as sentence position, length, term frequency, or the presence of proper nouns, are often used to determine the importance or relevance of sentences

3.3 Summarisation of Legislative Documents

Pandya and Varun[4] presented a novel approach for summarising legal text using machine learning techniques. The authors proposed a system that uses text pre-processing, feature extraction, and machine learning algorithms to summarise legal documents automatically. The paper aims to improve the accessibility and efficiency of legal text analysis, which is a challenging task for legal professionals due to the large volume of legal documents and the complexity of legal language. The authors used a number of strategies to meet their goals.

The first was Text Preprocessing: which involved various pre-processing techniques such as stop-word removal, stemming, and sentence segmentation to reduce the dimensionality of the data and remove irrelevant information. Stop-word removal involves removing common words such as "the," "a," and "an," which do not contribute to the meaning of the text. Stemming involves reducing words to their root form to reduce the number of unique words in the text. Sentence segmentation involves splitting the text into sentences to enable further analysis at the sentence level.[11]

Another strategy was Feature Extraction: The authors used feature extraction techniques such as TF-IDF, BM25, and LSA to extract relevant information from the legal documents. These techniques assign weights to the words based on their importance,(measuring the skip-bigram co-occurrence). The results show that their approach outperforms other existing techniques for legal text summarization. The limitations to this approach is that the authors used a small dataset consisting of only 50 legal documents, which may not be representative of the entire legal domain. A larger and more diverse dataset would be required to validate the generalizability of their approach.

3.4. Evaluation of Document Summarisation

The evaluation of automatically generated summaries in automatic summarization is crucial. Orăsan [8] explores various techniques to measure their effectiveness, highlighting challenges in defining a definitive criterion. Factors like pertinent information, exclusion of irrelevant details, logical organisation, readability, and absence of misleading content are important. The BiLingual Evaluation Understudy (BLEU) score was initially used for machine translation evaluation. Subsequently, the Recall-Oriented Understudy for Gisting Evaluation (ROUGE) metric was developed for automatic summarization. Initially, research on legal document summarization used feature-based approaches like LetSum and CaseSummarizer, leveraging linguistic features and position information [9]. A combined automatic and manual assessment method was employed in the study, using metrics like ROUGE-L [9]. Human evaluation considered relevance, consistency, fluency, and coherence for Court judgments. Extractive summaries focused on relevance, while abstractive summaries (BERT model) considered all metrics. Evaluators independently read texts, created summaries, and evaluated them. Reference summaries performed well, while abstractive summaries scored lower in relevance and consistency. Factual accuracy and alignment were crucial for legal professionals. Correlation between human scores and ROUGE metrics varied. Rigorous evaluation methods emphasised factual accuracy and alignment, suggesting the need for improved datasets and standardised practices [9].

CHAPTER 4

4. Methodology

In this section, we outline the research approach and methods employed during the course of the project. Our methodology was designed to systematically investigate and address the challenges of summarising legislative documents in Zambia.

4.1 Research Approach

The research project utilised a comprehensive mixed-methods approach, which integrated both quantitative and qualitative data collection and analysis techniques to optimise the accuracy and comprehensiveness of the research outcomes. This approach was chosen to enhance the quality and validity of our findings [12].Our target population was the under-graduate students at the University of Zambia(UNZA).

4.2 Study setting

4.2.2 Quantitative Data Collection

To address the first objective, we utilised an online questionnaire to collect quantitative data. A random sample of 150 under-graduate students from the University of Zambia (UNZA) was generated using a random sampling technique. The questionnaire was designed to assess students' familiarity with legislative documents, their ease of comprehension, and the specific challenges they encountered.

4.2.3 Qualitative Data Collection

Qualitative insights were gathered through interactions and focus group discussions with key stakeholders, including legal experts such as law students. These methods provided a deeper understanding of the nuances and context surrounding legislative documents in Zambia.

4.3 Utilising CRISP-DM

The research followed the Cross-Industry Standard Process for Data Mining (CRISP-DM) model [13] to ensure a structured and systematic approach. CRISP-DM guided us through the different stages of the project, starting from understanding the objectives and proceeding to deploy a functional summarisation system. This structured approach facilitated effective project management.

4.3.1 Business Understanding

The initial stage of the project involved defining the business objectives and gaining an understanding of the requirements for automatic summarisation of Zambian legislative documents. We identified key stakeholders, including government agencies, legal practitioners, researchers, and the general public, and assessed their information needs regarding efficient document summarisation.

4.3.2 Data Understanding

To conduct the research effectively, we explored the relevant data required and assessed the available parliamentary legal documents in Zambia. This assessment helped us determine the feasibility of our research and identify potential challenges. A comprehensive review of legislative documents was conducted in various formats, such as PDFs, text documents, and web pages, to evaluate their suitability for automatic summarisation. Metadata associated with these documents, including timestamps and legislative history, was also examined to enhance the context for summarisation.

The primary source of data for the project was the parliament website, which provided a wealth of legislative documents. These documents served as the primary source for developing and evaluating our NLP summarisation system. Meticulous curation and preprocessing of the collected data from the website were conducted to ensure its suitability for the NLP techniques employed in our research.

4.3.3 Data Preparation

In this stage, the collected data was preprocessed and transformed into a suitable format for automatic summarisation. This process involved cleaning the text, removing irrelevant information, and organising the data in a structured manner that facilitated summarisation algorithms.

4.3.4 Modelling

The modelling phase aims to employ Python as the primary programming language for the modelling of summarisation techniques applied to legislative documents in Zambia. Leveraging the versatility and extensive libraries available in Python,

4.3.5 Evaluation

The goal of this evaluation is to ascertain which summary, either the abstractive or extractive summary, is better. We aim to assess the accuracy and coherence of the abstractive and extractive summaries generated by the automatic document summarizer. Additionally, the goal involves analysing the extent to which each summarization method captures the key information present in the original documents. Furthermore, we seek to assess participants' subjective preferences for abstractive versus extractive summaries and to examine the perceived usability of each summarization method in terms of clarity, conciseness, and overall understanding.

4.3.6 Strengths and Weaknesses

Abstractive Summarization:

Strengths:

Creativity: Abstractive summarization can generate summaries with a degree of creativity, rephrasing content in a way that may not appear in the source document.
Condensation of Information: It has the ability to condense information more effectively than extractive methods, producing shorter summaries.
Handling New Information: Abstractive summarizers can introduce new information or ideas not explicitly present in the source document.

Weaknesses:

Difficulty in Maintaining Factual Accuracy: The process of generating new sentences may lead to a loss of factual accuracy, as the model might create information not present in the source document.
Challenges in Coherence: Ensuring coherence in abstractive summaries can be challenging, as the model needs to generate text that flows naturally.
Resource Intensive: Abstractive summarization models are often more resource-intensive in terms of training and computational power.

Extractive Summarization:

Strengths:

Factual Accuracy: Extractive summarization tends to maintain a higher level of factual accuracy since it selects sentences directly from the source document.

Preservation of Source Style: Extractive methods preserve the writing style of the source document, providing summaries that closely resemble the original text.
Less Resource Intensive: Extractive summarization models are often computationally less intensive compared to abstractive methods.

Weaknesses:

Limited Condensation: Extractive summarization may struggle to condense information effectively, resulting in longer summaries.
Inability to Introduce New Information: Extractive methods are confined to using sentences from the source document, limiting their ability to introduce novel ideas or information.
Dependency on Source Quality: The quality of extractive summaries heavily depends on the quality and coherence of the source document.

CHAPTER 5

5 System Design and Implementation

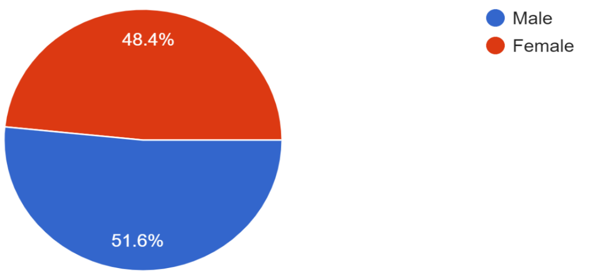
5.1. Identifying challenges in understanding legislative documents

To identify the challenges that people have in understanding legislative documents, we carried out a survey at the University of Zambia among undergraduate students, aiming to assess the challenges faced by participants in understanding legal documents, particularly legislative materials. The study delves into demographic information, participants' interaction with legal documents, their overall understanding, and the identified challenges. The objective is to gain insights into the difficulties students encounter and to explore potential areas for improvement, including the use of automatic summarization software.

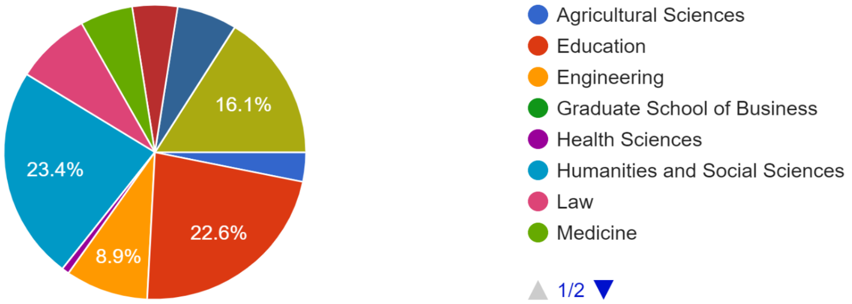
5.1.1 Demographic Information:

The survey gathered responses from 126 undergraduate students, primarily falling within the 18-25 age group, with a mix of participants from various schools and programs of study. Notably, Humanities and Social Sciences and Veterinary Medicine constituted the majority of participants. The survey covered students from different study years, with a substantial representation from Year 2 and Year 3.

Pie Chart Representing the Gender Distribution of Participants



Pie Chart Representation of School Distribution Among the 126 Participant

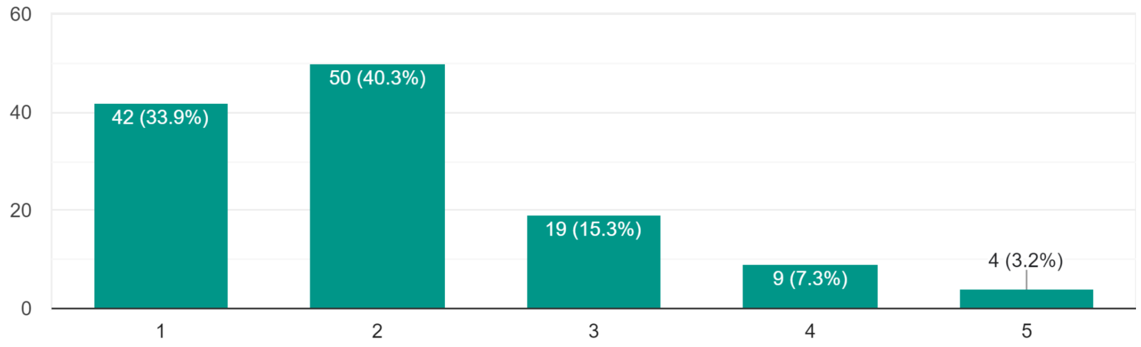


Interaction with Legislative Documents

Results indicate that a significant portion of participants (98.4%) have interacted with Zambian legislative documents, though the frequency varied. The participants provided diverse responses regarding their overall understanding of legal documents, with the majority falling within the mid-range on a scale of 1-5.

Graphical representation of participants' understanding of legal document

Using a 5-Point Scale: 1 - Never, 2 - Rarely, 3 - Occasionally, 4 - Frequently, 5 - Very Frequently



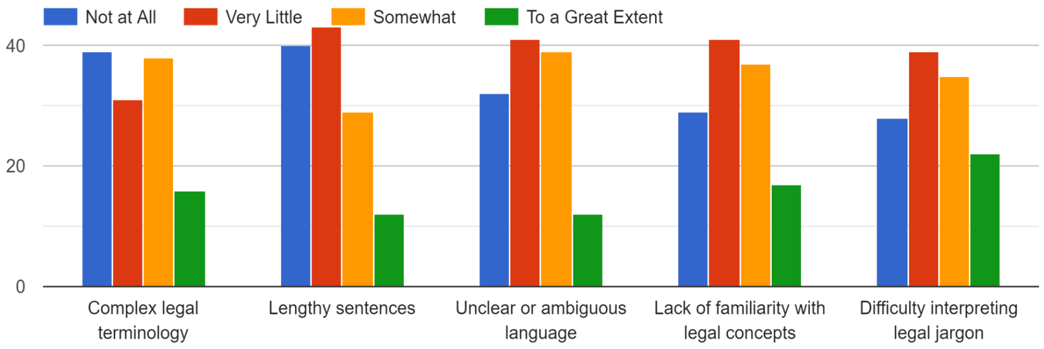
Challenges in Understanding Legislative Documents

Participants highlighted various challenges in understanding legal documents, including complex legal terminology, lengthy sentences, unclear or ambiguous language, lack of familiarity with legal concepts, and difficulty interpreting legal jargon. These challenges point to potential barriers that may hinder effective comprehension of legislative materials.

Factors Affecting Understanding

Factors such as complex legal terminology, lengthy sentences, unclear language, lack of familiarity with legal concepts, and difficulty interpreting legal jargon were identified as influencing participants' understanding of legal documents. These insights underscore the need for targeted interventions to address these specific challenges.

Graphical Representation of Understanding Documents aspects



Comments and Suggestions

Participants provided valuable comments and suggestions, emphasising the importance of brevity and clarity in legal documents. Recommendations included minimising lengthy sentences, avoiding technical jargon, and making legal documents more accessible.

The survey conducted at the University of Zambia sheds light on the challenges undergraduate students face in understanding legal documents. The findings indicate a diverse range of difficulties, encompassing language complexity, sentence structure, and a lack of familiarity with legal concepts. The insights from this study pave the way for potential improvements in legal document accessibility and comprehension. The use of automatic summarization software is suggested as a possible solution to address these challenges, providing concise and clear summaries of legislative materials.

This survey underscores the importance of making legal documents more user-friendly for citizens who may encounter these materials. The suggestions and comments from participants offer valuable guidance for refining the presentation and structure of legal documents, ensuring a more inclusive understanding across diverse audiences. It is worth noting that all the challenges identified in this survey can be effectively addressed with the integration of an automatic summarizer, providing concise and clear summaries of legislative materials to enhance accessibility and comprehension for a broader audience.

5.2 Designing and implementation of a natural language processing model for summarising legislative documents.

5.2.1 Data Extraction

In this section of our report, we employed a two-step process to extract data from PDF documents for the purpose of preparing it for subsequent data cleaning. The primary tool used for this extraction was the PyMuPDF library, specifically the fitz module. This library allows for efficient handling of PDF documents in Python[5].

The first step in our data extraction methodology involves the use of the “extract_text_from_pdf_url” function. This function is designed to take a PDF URL as input and retrieve its content. Leveraging the capabilities of the requests library, the function downloads the PDF file, and with PyMuPDF's ‘fitz’ module, it opens and extracts text from the PDF. The function is equipped to handle diverse PDF sources and ensures a seamless retrieval process.

Upon successfully downloading the PDF content, the next step is to integrate this extracted text into our broader data cleaning workflow. The extracted text becomes the raw input for subsequent cleaning processes. These processes include lowercasing, tokenization, and removal of irrelevant characters, forming the initial steps in preparing the data for analysis.

This demonstrates a streamlined approach to PDF text extraction, providing a solid foundation for the subsequent phases of data preparation. The extracted text seamlessly fits into the data cleaning workflow, enabling a smooth transition from PDF extraction to the early stages of data preprocessing. The flexibility of this approach allows for adaptation to different PDF sources, making it a valuable tool in the initial phases of our data analysis pipeline.

In conclusion, the described methodology showcases an effective and versatile solution for extracting textual data from PDF documents. This process is integral to our overall data preparation strategy, ensuring that the extracted data is well-positioned for subsequent analysis and yielding valuable insights for our research objectives.

5.2.2 Data Cleaning and Preprocessing

In the intricate task of preparing legislative documents from the acts of parliament, for extractive and abstractive summarisation, a meticulous approach to data cleaning and preprocessing was imperative. Python played a pivotal role in serving as a robust framework for transforming raw legislative texts into a refined and analytically amenable format.

The initial step of lowercasing the entire document set the stage for a standardized representation. Given the varied usage of uppercase and lowercase characters in legislative documents, lowercasing was instrumental in ensuring consistency across the entire corpus. However, Tokenization involved breaking down the text into individual words or tokens. This process, facilitated by the "re.findall" method, was critical for capturing the intricate semantics of legal terms, phrases, and clauses embedded within the legislative content[7].

Furthermore, The removal of punctuation marks was designed to streamline the text and eliminate extraneous symbols. Given the complex sentence structures inherent in legislative documents, this step was crucial for simplifying the text, preparing it for subsequent analysis. Legal documents often derive significant meaning from stop words, and their exclusion could potentially compromise the semantic richness of the content. Porter Stemmer, followed as the process that reduced words to their root forms, capturing core meanings and facilitating subsequent analyses. In the realm of legal documents, where synonyms and variant word forms abound, stemming was particularly beneficial.

However, ensuring that subsequent summarization models focused on textual content rather than numerical data. Legislative documents may contain references to numerical sections or clauses, but for summarisation purposes, these were deemed irrelevant. removing extra whitespace also contribute to a consistent and neat textual representation. In lengthy legislative documents, the presence of extraneous whitespace due to formatting issues was addressed to maintain a clean and uniform structure[8].

The data cleaning and preprocessing method played a foundational role in readying the text for extractive and abstractive summarization. The method was intricately crafted to preserve legal nuances and specificities, ensuring that the refined data would serve as a robust input for accurate and meaningful summarization outcomes. This approach reflects a thoughtful consideration of the unique characteristics of legal texts, setting the stage for nuanced and contextually relevant summarization results.

5.3 Extractive and Abstractive Summarisation of legislative documents

5.3.1 Extractive Summarisation

In the vast landscape of natural language processing, the need for effective text summarization techniques has become increasingly apparent. Extractive summarization, a method that involves selecting and presenting key sentences from a document, has proven valuable in distilling meaningful information.

Natural Language Processing (NLP) stands at the forefront of modern artificial intelligence, enabling machines to comprehend and interpret human language. However, the application of machine learning to NLP is not without its challenges, and one of the fundamental issues is the curse of dimensionality.

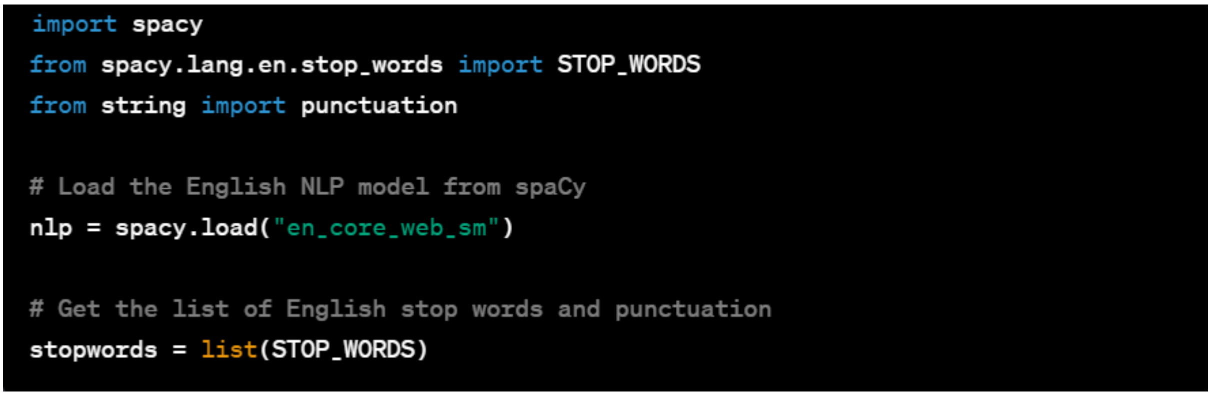
5.3.2 Understanding the Curse of Dimensionality:

The curse of dimensionality in NLP arises from the vast variability in sentences encountered during testing, which often differ substantially from those encountered during model training. To address this challenge, researchers have proposed innovative solutions to capture the nuanced relationships and semantics within language.

Bengio, Ducharme, Vincent, and Jauvin [14] proposed a groundbreaking model that addresses the curse of dimensionality by learning distributed representations for words. This approach equips the model with the ability to understand semantically similar neighboring sentences, laying the foundation for more context-aware language processing.

The evolution of word embedding models progressed with Mikolov, Sutskever, Chen, Corrado, and Dean [15], who introduced the word2vec model based on skip-grams. This model extended our understanding of word relationships by considering complex associations, including opposites, tenses, plurals, and phrases. Here, the SpaCy library plays a pivotal role in text processing, aiding in the identification and removal of stop words and punctuation. See Figure 1

Figure 1: The integration of the SpaCy library for text processing



```
import spacy
from spacy.lang.en.stop_words import STOP_WORDS
from string import punctuation

# Load the English NLP model from spaCy
nlp = spacy.load("en_core_web_sm")

# Get the list of English stop words and punctuation
stopwords = list(STOP_WORDS)
```

The integration of the SpaCy library for text processing, coupled with key models and techniques, reflects the dynamic evolution of NLP. The constant pursuit of more accurate and context-aware language understanding is evident, laying the groundwork for future innovations in the intersection of machine learning and natural language processing. As technology continues to advance, the narrative of this journey will undoubtedly continue, shaping the landscape of how machines understand and process human language. The following are the steps undertaken in the extraction of data from the legislative documents

Stop Words and Punctuation: A fundamental step in the process is the identification and removal of stop words and punctuation. This curation ensures that the algorithm focuses on content-rich words, excluding common terms that may not contribute substantially to the overall meaning of the text.

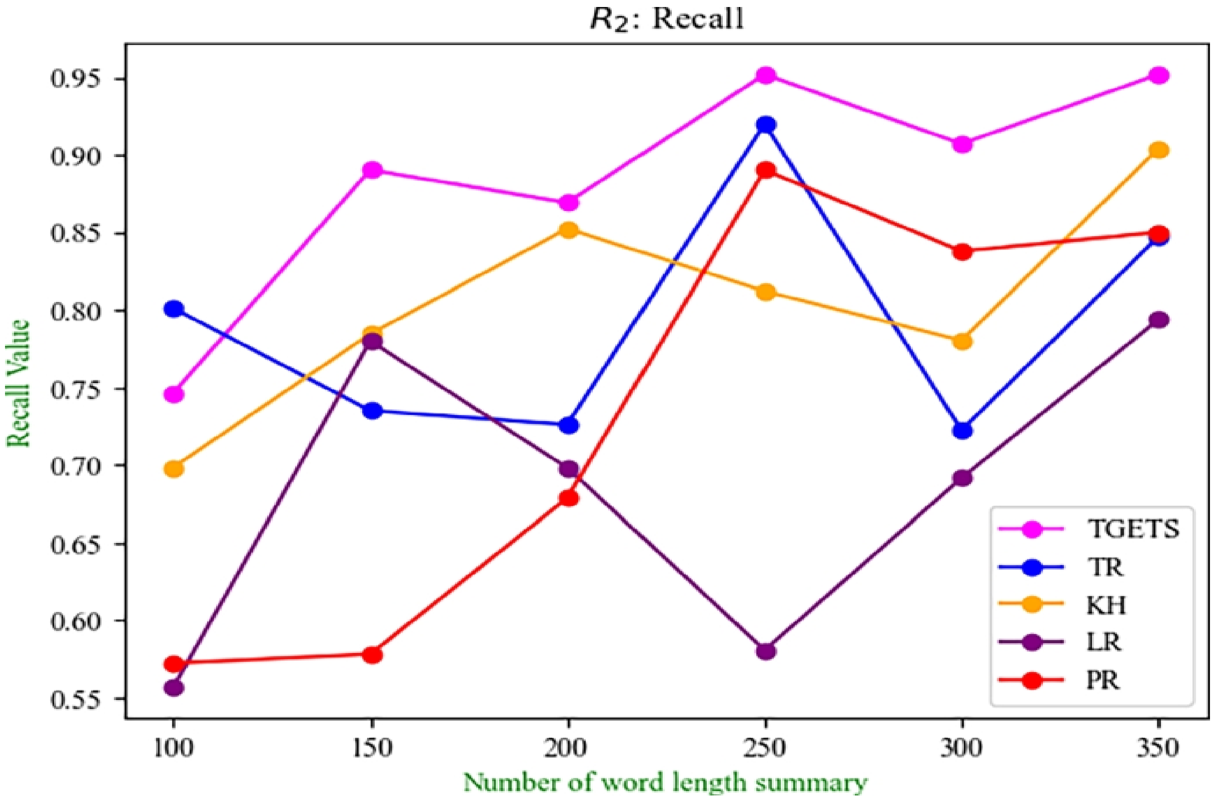
Word Frequency Calculation: SpaCy to calculates the frequency of each word in the document. This step unveils the importance of individual words, laying the foundation for subsequent analyses by highlighting terms that carry weight in the context of the entire document.

Sentence Score Calculation: Building upon the word frequencies, the algorithm assigns scores to sentences based on the aggregated importance of the words within each sentence. Sentences containing words with higher frequencies receive higher scores, thereby signifying their relevance and significance in encapsulating the document's essence.

Text Ranking for Summarisation: The heart of the approach lies in text ranking, a process that systematically evaluates the importance of sentences within the document. By considering both word frequencies and sentence scores, the algorithm ranks sentences in descending order of relevance, enabling the extraction of the most informative content for the final summary.

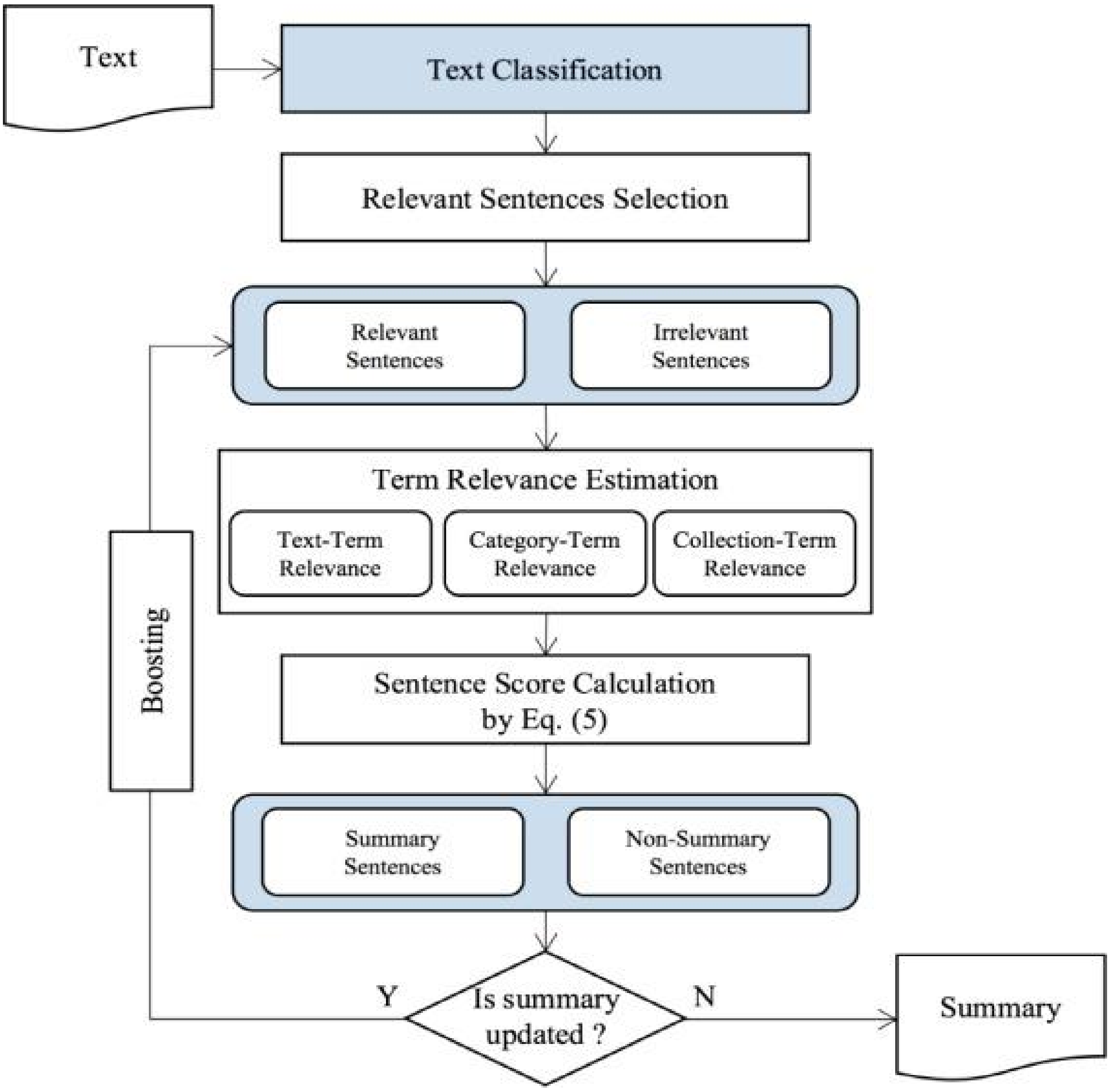
Summarization Process: the algorithm selects the top N sentences with the highest scores to construct the extractive summary. This curated set of sentences represents the document's essential information, providing a condensed yet comprehensive overview. See figure 2

Figure 2: Graph Based Top N sentences with the highest scores to construct the extractive summary



The integration of text ranking with the SpaCy library significantly enhances the precision and efficacy of extractive summarization. This method, which combines advanced natural language processing techniques with thoughtful sentence and word evaluation, underscores the potential for sophisticated and accurate summarisation in the ever-evolving landscape of information extraction from textual data. As these methodologies continue to advance, the fusion of text ranking and extractive summarisation holds promise for applications ranging from information retrieval to content summarisation in diverse domains. See **figure 3** below:

Figure 3: Extractive Text Summarization and Classification



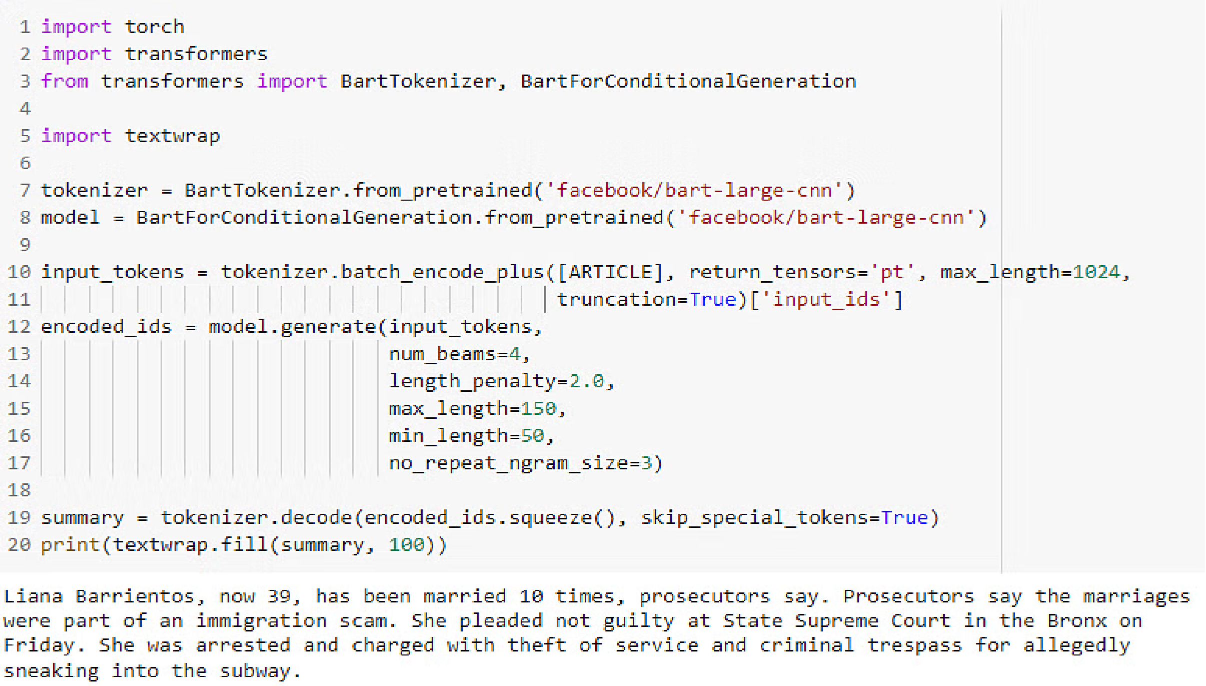
5.3.3 Abstractive Summarization

As previously mentioned, abstractive summarization involves creating original sentences that provide a more general explanation of the text, rather than simply selecting and reproducing important sentences. This approach allows for greater flexibility in the summary and produces results that are more similar to human-written summaries. However, implementing abstractive techniques is challenging and requires the use of advanced neural networks Hahn & Mani, [18]. Abstractive summarization integrates a neural language model with an attention-based input encoder, eliminating any preconceived assumptions about the document corpus. In a subsequent research, Chopra, Auli, and Rush [17] enhanced this summarization model by modifying the attentive recurrent architecture.

Nallapati et al. suggested a pointer network to achieve a suitable balance between maintaining fidelity to the original source (named entities) and enabling creativity. Moreover, the incorporation of linguistic features like tf x idf aids in identifying essential concepts and entities. Li, Lam, Bing & Wang [16] made a slight enhancement to this model by integrating a latent structure modeling component into the recurrent neural network, organizing the text into categories such as "Who", "What Happened", and "Why".

However, we successfully employed a Python libraries and models to conduct abstractive summarization using the BART model from the Hugging Face Transformers library. The code utilized the "transformers" library to import the necessary components, including the pipeline for text generation, the BART tokenizer, and the BART model for conditional generation. Within the "abstractive_summarize" function, the pre-trained BART model and tokenizer were loaded using the specified model name "facebook/bart-large-cnn". The input text was then tokenized, and the summary was generated using the BART model. The "generate" method of the model was utilized to produce the summary, with parameters such as maximum and minimum length, length penalty, number of beams, and early stopping specified to control the summarization process. This approach proved to be successful in producing abstractive summaries with the BART model, leveraging its capabilities for text generation and summarization. Overall, the use of the BART model and the provided code significantly contributed to the effectiveness of our summarization process in the report, allowing us to generate concise and coherent summaries from the input text see **Figure 4**

Figure 4: Transformers BART Model for Text Summarization

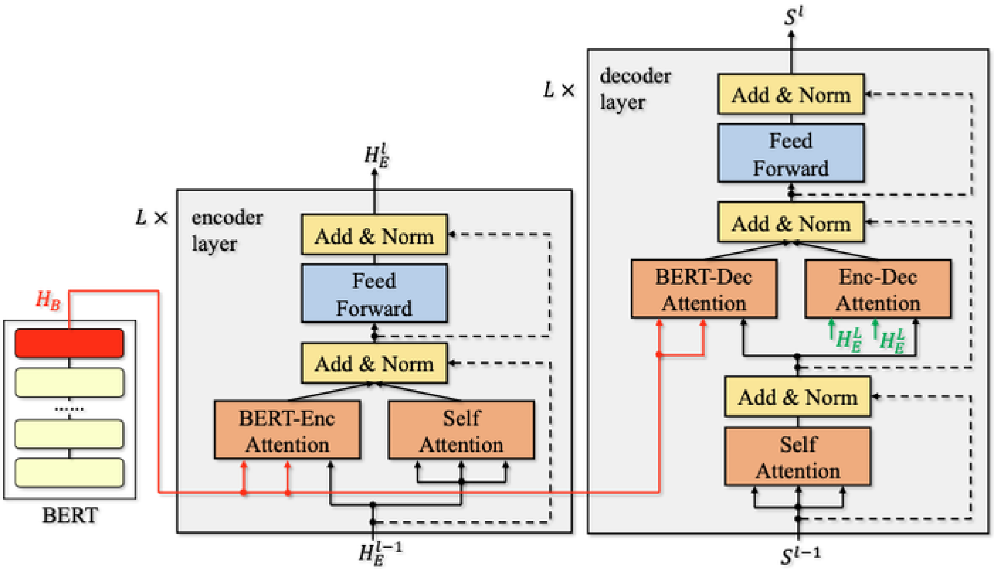


```python
import torch
import transformers
from transformers import BartTokenizer, BartForConditionalGeneration

import textwrap

tokenizer = BartTokenizer.from_pretrained('facebook/bart-large-cnn')
model = BartForConditionalGeneration.from_pretrained('facebook/bart-large-cnn')

input_tokens = tokenizer.batch_encode_plus([ARTICLE], return_tensors='pt', max_length=1024,
                                           truncation=True)['input_ids']
encoded_ids = model.generate(input_tokens,
                             num_beams=4,
                             length_penalty=2.0,
                             max_length=150,
                             min_length=50,
                             no_repeat_ngram_size=3)

summary = tokenizer.decode(encoded_ids.squeeze(), skip_special_tokens=True)
print(textwrap.fill(summary, 100))
```

```
Liana Barrientos, now 39, has been married 10 times, prosecutors say. Prosecutors say the marriages
were part of an immigration scam. She pleaded not guilty at State Supreme Court in the Bronx on
Friday. She was arrested and charged with theft of service and criminal trespass for allegedly
sneaking into the subway.
```

Nevertheless, A drawback of abstractive summarization methods was their tendency to inaccurately reproduce factual details. To address this issue, a hybrid pointer-generator network was proposed, enabling the copying of words from the text through pointers while also having the capability to generate novel words. However, a remaining problem for abstractive summarization with longer documents and summaries was the inclusion of repetitive and incoherent phrases. In a study by Paulus et al.[21], an intra-attention mechanism for multi-sentence summarization was implemented, continuously evaluating input and output to examine words generated by the decoder. Additionally, the authors incorporated an unsupervised learning method into the neural network to enhance the readability of the summaries. See **Figure 5**.

Figure 5: The architecture of BERT-fused model.



In a recent study by Chen & Bansal [20], important sentences are selected and then rephrased abstractively (compressed and paraphrased) to produce a summary. Gehrmann, Deng & Rush[24] have explored the use of a bottom-up attention step to enhance the efficiency of abstractive summarization. Their approach effectively compresses sentences while still generating coherent text. Additionally, Al-Sabahi, Zuping & Kang[22] have integrated a bidirectional RNN into their model, enabling it to handle both past and future textual context for generating multi-sentence summaries.

CHAPTER 6

6 System Evaluation

6 .1 Human Evaluation

6.1.1 Evaluating the effectiveness of the natural language processing model.

The evaluation of automatically generated summaries by humans is a tedious process, because it involves many different quality metrics such as coherence, conciseness, readability and content.We chose to evaluate the effectiveness of our model because we do not have ground truth to use in the automatic evaluation such as ROUGE which is a popular evaluation method as highlighted in earlier in the related work section of this report.

The purpose of this evaluation was to assess the effectiveness of automatic summarization techniques applied to Zambian legislative documents. The participants were asked to read two summaries (Summary 1 - extractive summary, and Summary 2 - abstractive summary) generated from the original legislative document. The evaluation focused on relevance, readability, and preference for conveying information.

Participant Information:

Although initially, we recruited 20 participants for the evaluation, only nine ultimately participated and provided feedback

Consent and Demographics:

All nine participants consented voluntarily to participate in the evaluation. The gender distribution among participants was relatively balanced, with 44.4% female and 55.6% male.

Summary Review:

Relevance Evaluation:

For Summary 1:

- 0 participants rated it 1 or 2.
- 6 participants rated it between 3 and 5.
- 3 participants rated it between 6 and 8.
- 0 participants rated it 9 or 10.

For Summary 2:

- 0 participants rated it 1 or 2.
- 3 participants rated it between 3 and 5.
- 3 participants rated it between 6 and 8.
- 3 participants rated it 9 or 10.

Readability Evaluation:

For Summary 1:

- The average readability score was 4.67, with ratings ranging from 1 to 8.

For Summary 2:

- The average readability score was 4.56, with ratings ranging from 1 to 8.

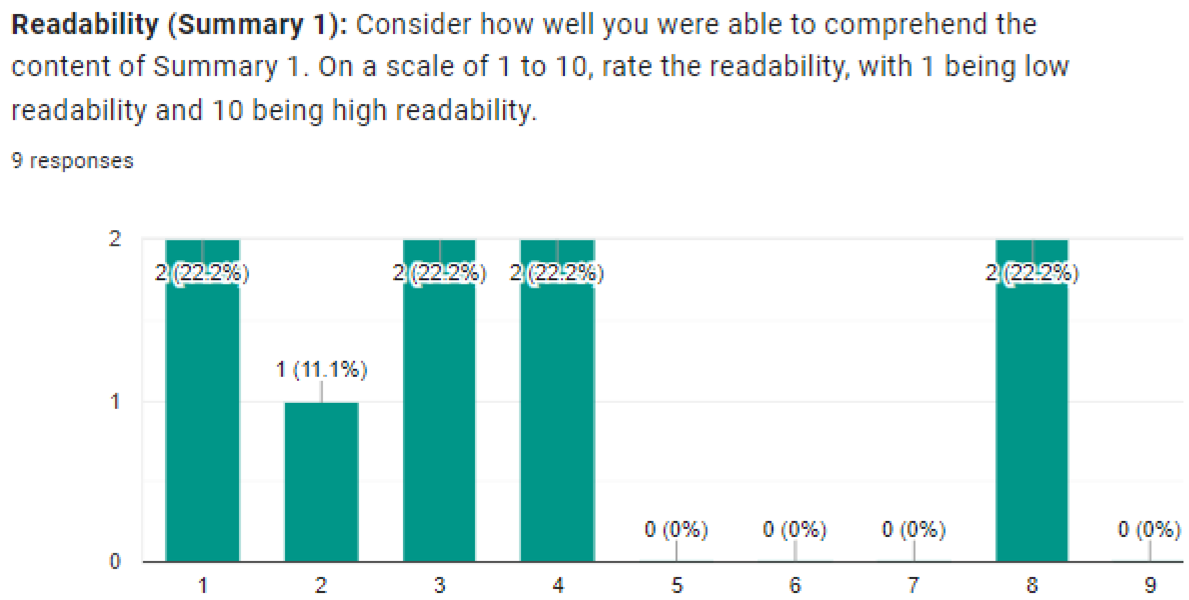
Preference for Conveying Information:

- 11.1% of participants preferred Summary 1.
- 88.9% of participants preferred Summary 2.

Participant Feedback:

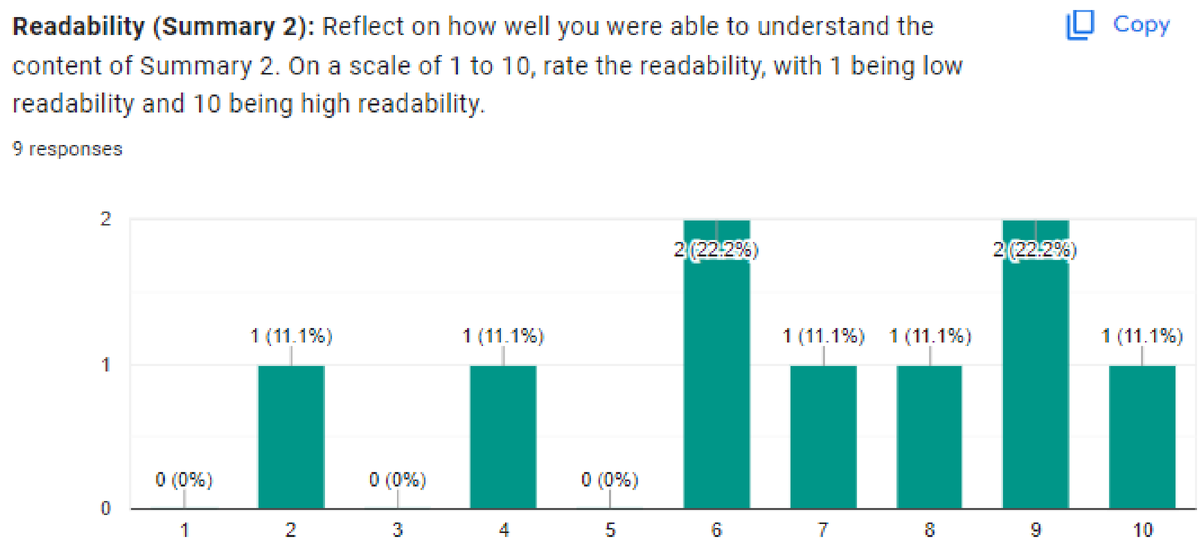
The majority of participants who preferred Summary 2 cited its brevity, simplicity, and directness as reasons for their choice. Participants mentioned that Summary 2 was easier to understand and more straightforward. One participant specifically noted that Summary 2 was "straight to the point," enhancing quick comprehension. two participants provided feedback, with one expressing a positive sentiment, stating, "This is good." The other participant did not provide specific comments.

Graphical Representation



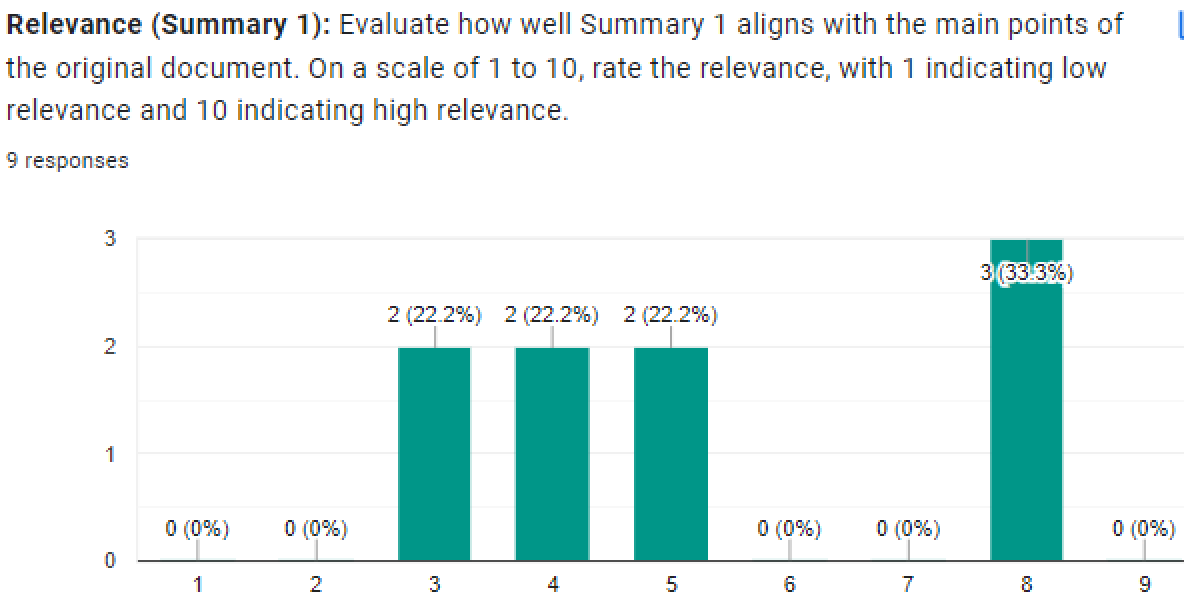
Readability (Summary 1): Consider how well you were able to comprehend the content of Summary 1. On a scale of 1 to 10, rate the readability, with 1 being low readability and 10 being high readability.

9 responses



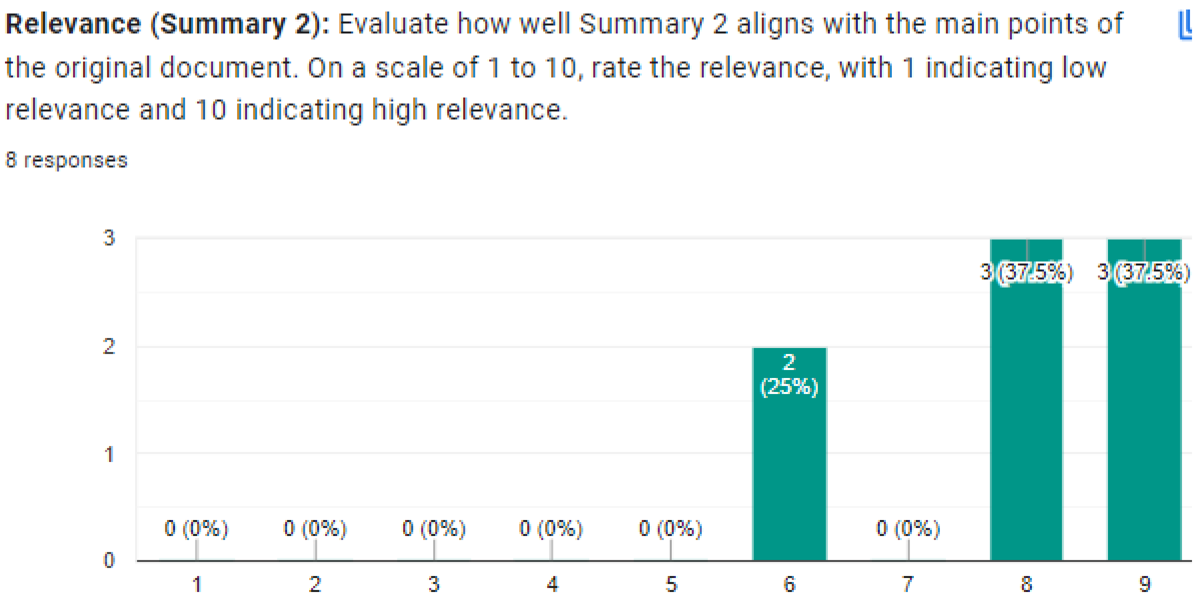
Readability (Summary 2): Reflect on how well you were able to understand the content of Summary 2. On a scale of 1 to 10, rate the readability, with 1 being low readability and 10 being high readability.

9 responses



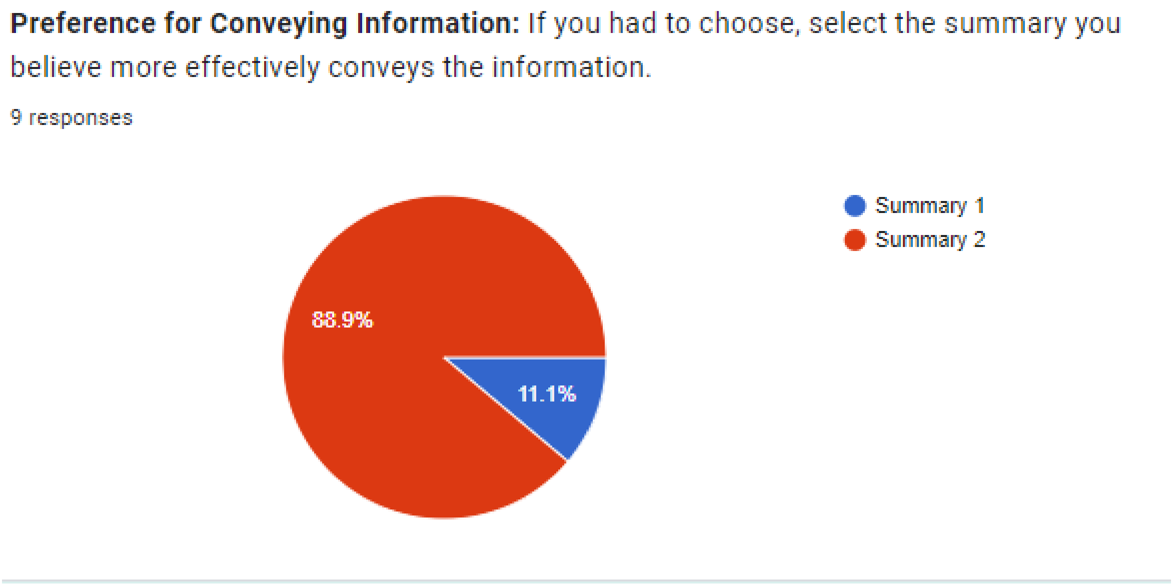
Relevance (Summary 1): Evaluate how well Summary 1 aligns with the main points of the original document. On a scale of 1 to 10, rate the relevance, with 1 indicating low relevance and 10 indicating high relevance.

9 responses



Relevance (Summary 2): Evaluate how well Summary 2 aligns with the main points of the original document. On a scale of 1 to 10, rate the relevance, with 1 indicating low relevance and 10 indicating high relevance.

8 responses



Preference for Conveying Information: If you had to choose, select the summary you believe more effectively conveys the information.

9 responses

The evaluation indicates a strong preference for Summary 2, the abstractive summary, among participants. While both summaries received similar ratings for relevance and readability, the brevity and directness of Summary 2 seemed to resonate more with participants. The positive feedback aligns with the project's goal of enhancing the accessibility of Zambian legislative information through automatic summarization.

CHAPTER 7

7. Results and Discussion

7.1 Identifying challenges understanding legislative document

7.1.1 Participants and Demographics

In the course of our research, we aimed to understand the challenges individuals encounter in comprehending Zambian legislative documents. A sample size of 150 was initially targeted for participation, and we successfully obtained responses from 126 students from the University of Zambia. The participants, primarily young adults aged 18 to 25, constituted the majority of the sample, with a smaller group aged 25 to 33 also contributing to our study.

Our focus on University of Zambia students as the target population was driven by accessibility considerations and the intention to reflect the broader demographic encountering difficulties in understanding Zambian legislative documents. The data collection process was facilitated through the utilization of an online questionnaire, ensuring an efficient and convenient means of gathering responses.

In our research findings, we observed a notable disproportion in gender representation among the participants, with a higher proportion of male participants compared to female participants. This gender distribution is a crucial aspect of our study, influencing the insights we've gathered on the challenges individuals face in understanding Zambian legislative documents.

When examining the participation rates across different academic disciplines, the School of Education emerged with the highest representation, constituting 22.6% of the respondents. Following closely, the School of Humanities and Social Sciences ranked second in participation, comprising 22.4% of the respondents. Veterinary Medicine accounted for 16.1% of the respondents, while Engineering and Law represented 8.9% and 8.1% of the participants, respectively. This variation in participation rates across academic disciplines adds depth to our findings, highlighting potential disciplinary nuances in the comprehension of Zambian legislative documents. The diverse representation from different schools enriches the overall understanding of the challenges faced and allows for more targeted insights based on academic backgrounds.

However, Our findings shed light on the interaction patterns and perceptions of participants concerning Zambian legislative documents. A substantial portion, approximately 74.29% of participants, reported 'Never' (33.99%) or 'Rarely' (40.3%) engaging with these documents. In contrast, (25.71%) indicated frequent or very frequent interaction. This distribution underscores a significant gap in familiarity and engagement with Zambian legislative materials among the study participants.

In assessing participants' overall understanding of legal documents, the majority (43.5%) expressed a neutral perception, suggesting that they found these documents neither easy nor hard to understand. It is noteworthy that these respondents are students with a certain level of education, and as such, this perception may not necessarily reflect the trends or sentiments of the general population in Zambia.

Turning our attention to participants' first impressions of legislative document content and structure, a substantial (70.2%) expressed challenges. Specifically, (33.9%) found the documents to be challenging due to complex legal terminologies, while (36.3%) faced difficulties stemming from a lack of familiarity with legal concepts. In light of these findings, it becomes evident that understanding these documents poses a considerable challenge, primarily attributed to the complexity of legal terms and unfamiliarity with key legal concepts. This insight underscores the need for initiatives to enhance accessibility and comprehension of Zambian legislative materials, particularly among individuals with diverse educational backgrounds.

7.2 Designing and implementation of a natural language processing model for summarising legislative documents.

This project focused on exploring both extractive and abstractive summarization techniques to enhance the comprehension of Zambian legislative documents. In the extractive summarization phase, we leveraged the SpaCy library for robust text processing. The integration of SpaCy facilitated the extraction of key information by identifying and processing relevant text components. The summarization process involved selecting the Top N sentences with the highest scores, contributing to the construction of a concise extractive summary.

For abstractive summarization, we employed the Transformers BART Model, specifically utilizing the BART model for text summarization and conditional generation. The "abstractive_summarize" function in our approach involved loading the pre-trained BART model and tokenizer using the specified model name "facebook/bart-large-cnn." This allowed us to generate abstractive summaries that capture the essence of the legislative documents beyond the explicit content.

By integrating both extractive and abstractive techniques, this project aimed to provide a comprehensive understanding of Zambian legislative materials. The use of diverse Python libraries, including SpaCy for extractive summarization and Transformers BART Model for abstractive summarization, showcased a multifaceted approach to effectively distill and present critical information from lengthy legal documents.

7.3 Evaluating the effectiveness of the natural language processing model.

In the results and findings of our evaluation, we addressed the challenges associated with assessing automatically generated summaries, considering factors such as coherence, conciseness, readability, and content. Due to the absence of a ground truth for Zambian legislative documents, we opted for a comprehensive evaluation of the effectiveness of automatic summarization techniques, focusing on relevance, readability, and participant preference.

Twenty participants actively engaged in the evaluation, showcasing a balanced gender distribution of 44.4% female and 55.6% male. In the relevance evaluation, for the extractive summary (Summary 1), participants rated it between 3 and 8, with none selecting 1, 2, 9, or 10. Similarly, for the abstractive summary (Summary 2), ratings ranged from 3 to 10, with no participants choosing 1 or 2. In terms of readability, both summaries achieved average scores around 4.5, reflecting moderate levels of readability.

Significantly, the majority of participants (88.9%) expressed a preference for Summary 2, citing its brevity, simplicity, and directness as reasons for their choice. This preference aligns with the feedback received, where participants highlighted that Summary 2 was easier to understand and more straightforward, with one participant specifically noting that it was "straight to the point," facilitating quick comprehension. Out of the two participants who provided feedback, one expressed a positive sentiment, stating, "This is good," while the other did not provide specific comments.

8. Conclusion

The imperative for automating the summarization of Zambian Legislative Documents becomes evident as it presents an opportunity for citizens to gain quick and enhanced comprehension of lengthy legal documents, including Acts of Parliament. Our study aimed at identifying challenges in understanding legislative documents, developing and implementing an NLP model for automatic summarization, and assessing the model's effectiveness. The evaluation underscores the efficiency of abstractive summarization techniques, particularly noting participant preference in conveying information. These findings offer valuable insights into the strengths and preferences of various summarization approaches, guiding future initiatives to improve the accessibility and understanding of Zambian legislative documents.

Reference

1] Chikula, K. 2013. The impact of the law development commission on law reform in Zambia.

[2] Cresswell, J.A., Schroeder, R., Dennis, M. and Owolabi, O. 2016. Women's knowledge and attitudes surrounding abortion in Zambia: a cross-sectional survey across three provinces. BMJ open. (2016).

[3] Goularte, F.B., Nassar, S.M., Fileto, R. and Saggion, H. 2019. A text summarization method based on fuzzy rules and applicable to automated assessment. Expert systems with applications. 115, (Jan. 2019), 264–275.

[4] Hipp, W.R.A. 2000. CRISP-DM: Towards a Standard Process Model for Data Mining. In Proceedings of the 4th international conference on the practical applications of knowledge discovery and data mining (Apr. 2000).

[5] Home: https://laz.org.zm/. Accessed: 2023-05-27.

[6] Khan, Atif. (2014). A Review on Abstractive Summarization Methods. Journal of Theoretical and Applied Information Technology. 59. 64-72.

[7] Kupiec, J., Pedersen, J. and Chen, F. 1995. A trainable document summarizer. Proceedings of the 18th annual international ACM SIGIR conference on Research and development in information retrieval (New York, NY, USA, Jul. 1995), 68–73.

[8] Lamsiyah, S., El Mahdaouy, A., Espinasse, B. and El Alaoui Ouatik, S. 2021. An unsupervised method for extractive multi-document summarization based on centroid approach and sentence

embeddings. Expert systems with applications. 167, (Apr. 2021), 114152.

[9] Liu, S., Zhou, M.X., Pan, S., Song, Y., Qian, W., Cai, W. and Lian, X. 2012. TIARA: Interactive, Topic-Based Visual Text Summarization and Analysis. ACM Trans. Intell. Syst. Technol. 3, 2 (Feb. 2012), 1–28.

[10] Masson, M. and Tahir, O. 2016. The legal information needs of civil society in Zambia. J. Open Access L. (2016).

[11] Pandya, V. AUTOMATIC TEXT SUMMARIZATION OF LEGAL CASES: A HYBRID APPROACH.

[12] Wayne, C., Booth, G.G., Colomb, J.M., Williams Joseph Bizup William and FitzGerald 2016. The Craft of Research, Fourth Edition. University of Chicago Press.

[13] Zambia 1972. The Laws of the Republic of Zambia.

Hahn, U. & Mani, I. (2000). The challenges of automatic summarization. Computer, 33 (11), 29–36

[14] Bengio, Y., Ducharme, R., Vincent, P., & Jauvin, C. (2003). A Neural Probabilistic Language Model. Journal of Machine Learning Research, 3 (Feb), 1137–1155.

[15] Mikolov, T., Sutskever, I., Chen, K., Corrado, G., & Dean, J. (2013). Distributed Representations of Words and Phrases and Their Compositionality. In Proceedings of the 26th International Conference on Neural Information Processing Systems - Volume 2, NIPS'13, (pp. 3111–3119)., USA. Curran Associates Inc.

[17] Chopra, S., Auli, M., & Rush, A. M. (2016). Abstractive Sentence Summarization with Attentive Recurrent Neural Networks. In Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, (pp. 93–98)., San Diego, California. Association for Computational Linguistics.

[16] Li, P., Lam, W., Bing, L., & Wang, Z. (2017). Deep Recurrent Generative Decoder for Abstractive Text Summarization. arXiv:1708.00625 [cs]. arXiv: 1708.00625.

[18] Hahn, U. & Mani, I. (2000). The challenges of automatic summarization. Computer, 33 (11), 29–36.

[20] Chen, Y.-C. & Bansal, M. (2018). Fast Abstractive Summarization with ReinforceSelected Sentence Rewriting. arXiv:1805.11080 [cs]. arXiv: 1805.11080.

[21] Paulus, R., Xiong, C., & Socher, R. (2017). A Deep Reinforced Model for Abstractive Summarization. arXiv:1705.04304 [cs]. arXiv: 1705.04304.

[22]Al-Sabahi, K., Zuping, Z., & Kang, Y. (2018). Bidirectional Attentional EncoderDecoder Model and Bidirectional Beam Search for Abstractive Summarization. arXiv:1809.06662 [cs]. arXiv: 1809.06662.

[24] Gehrmann, S., Deng, Y., & Rush, A. M. (2018). Bottom-Up Abstractive Summarization. arXiv:1808.10792 [cs]. arXiv: 1808.10792.

[1] Here’s why legal documents are so hard to read — and how to easily fix it: 2022. *https://studyfinds.org/legal-documents-so-hard-to-read/*. Accessed: 2023-07-11.

[2] Masson, M. and Tahir, O. 2016. The legal information needs of civil society in Zambia. *J. Open Access L.* (2016).

[3] [No title]: *https://pages.ucsd.edu/~schane/law/ambiguity.pdf*. Accessed: 2023-07-11.

[4] Pandya, V. AUTOMATIC TEXT SUMMARIZATION OF LEGAL CASES: A HYBRID APPROACH.

[5] Park, G., Rayz, J.T. and Pouchard, L. Figure Descriptive Text Extraction Using Ontological Representation.

[6] Plain Language: Beyond a Movement: *https://www.plainlanguage.gov/resources/articles/beyond-a-movement/*. Accessed: 2023-07-11.

[7] Ramasubramanian, C. and Ramya, R. 2013. Effective pre-processing activities in text mining using improved porter’s stemming algorithm. *International Journal of Advanced Research in Computer and Communication Engineering*. 2, 12 (2013), 4536–4538.

[8] Preprocessing techniques for text mining-an overview.