Application of Artificial Intelligence to Improve Efficiency in The Judiciary

Daniel Mbatha Mule,

145063

A Dissertation Submitted in Partial Fulfilment of the Requirements for the Degree of Master of Science in Data Science and Analytics at Strathmore University

Strathmore Institute of Mathematical Sciences(SIMS)

Strathmore University

Nairobi, Kenya

June, 2025

Declaration and Approval

Declaration

I declare that this work has not been previously submitted and approved for the award of a degree by this or any other University. To the best of my knowledge and belief, the dissertation contains no material previously published or written by another person except where due reference is made in the thesis itself. No part of this thesis may be reproduced without the permission of the author and Strathmore University.

Name of Candidate: Daniel Mbatha Mule

Signature: Daniel Mule Date 28.05.2025

Approval

The dissertation of Daniel Mule was reviewed and approved for examination by the following:

Dr. John Olukuru,
Senior Lecturer,
Strathmore University, Kenya.

Dr. Mary Achieng,
Associate Professor,
Strathmore University, Kenya.

Prof. Bernard Shibwabo Kasamani,
Director of Graduate Studies,
Strathmore University, Kenya.

Abstract

Artificial intelligence, a burgeoning force across various sectors, has found its niche in legal proceedings as well. Despite the abundance of published court decisions, the manual scrutiny required for referencing past cases often prolongs the decision-making process in court. To streamline this cumbersome task, this study aims to harness the power of Artificial Intelligence, specifically through Natural Language Processing (NLP) tools. By employing techniques such as keyword information retrieval and extractive text summarization, the objective is to facilitate the jury's access to pertinent court decisions on Kenya Law.org. This study evaluates the performance of the BERT model in a legal context, particularly focusing on its ability to classify court cases accurately. The BERT model achieved an accuracy of 66.67%, meaning it correctly predicted the outcome of approximately two out of every three cases. It demonstrated a precision of 66.67%, indicating that two-thirds of the cases predicted as "Convicted" were indeed correct. Notably, the model achieved a perfect recall of 100%, signifying that it identified all actual "Convicted" cases without missing any. Despite the lower precision, the F1 score of 80% reflects a balanced performance, emphasizing the model's strength in detecting all relevant positive cases while maintaining reasonable precision. These results suggest that BERT is an efficient tool for legal case classification, particularly in contexts where ensuring that no positive case is missed is more critical than minimizing false positives. This makes BERT a promising model for supporting legal decision-making and analysis in courts.

Given the substantial backlog of cases inundating Kenyan courts, compounded by the intricate nature of these legal matters, our research has underscored the potential for AI to complement human capabilities. These technologies offer a promising avenue for automating various legal processes, thereby streamlining judicial operations.

Table of Contents

List of Figures

List of Tables

List of Abbreviations

AI - Artificial Intelligence

CNN - Convolutional Neural Network

COMPAS - Correctional Offender Management Profiling for Alternative Sanctions

LSTM - Long Short-Term Memory

NLP - Natural Language Processing

RNN - Recurrent Neural Network

TF-IDF - Term Frequency-Inverse Document Frequency

Definition of terms

AArtificial intelligence	The area of computer science that concentrates on developing smart machines capable of executing activities that usually necessitate human intellect, Devlin et al. (2019).
Feature engineering	Is the process of selecting and transforming raw data into a set of features, or variables, that can be used to train a machine learning model
Keywords	A particular word or phrase that summarises a document's main idea or topic, webpage, or any other form of information. It is used to aid in the retrieval and discovery of information through search engines or other search tools Radvansky, 2010).
Keyword extraction	Is the process of automatically identifying and extracting the most significant words or phrases from a text or a document, Su et al. (2019).
Natural Language Processing	In the domain of computer science and artificial intelligence, focused on the interface between computers and human language Chen (2017).
Term Frequency-Inverse Document Frequency	It is a statistical technique used in natural language processing to evaluate the importance of a term in a collection of documents Van Rijsbergen, 2012).

Acknowledgement

The journey toward the completion of this dissertation has been both intellectually demanding and personally transformative. I extend my deepest gratitude to all those who have supported and guided me throughout this process. At each stage, my supervisors stood in my corner as experienced coaches, offering clarity, direction, and encouragement. Their unwavering guidance shaped the trajectory of this work, even during times when my understanding faltered or I misinterpreted the strategies presented to me.

There were moments of progress, brief but affirming, where I could see the impact of persistent effort. Nonetheless, the path remained unpredictable, marked by frequent reflection and intellectual recalibration. During these times of doubt and fatigue, the steady support of my family was indispensable. Their encouragement, patience, and belief in me provided both strength and perspective, reminding me of the value of persistence and the importance of balance.

This dissertation is a testament to the collective support, mentorship, and love that sustained me throughout this journey. To my esteemed supervisors, I am deeply grateful for your insightful guidance, patience, and encouragement throughout this academic journey. Your mentorship has been pivotal in shaping the direction and depth of this work.

To my belovEd family, your unwavering support, understanding, and belief in me have been the bedrock upon which I have learned during moments of challenge and uncertainty. My wife, Beverly Muthoni, your presence has made this demanding journey not only endurable but deeply meaningful. Together, we have navigated the complexities of research, risen to meet the demands of scholarly pursuit, and reached this important milestone.

A special note of appreciation goes to Dr. Olokuru, whose support and expertise have left a lasting imprint on both this dissertation and my academic growth. Finally, I would like to extend special appreciation to Kenya Law.org led by Martin Andago, head Information Technology department, who assisted in getting data for my dissertation.

Chapter 1: Introduction

1.1. Background of the study

1.1.1. The Use of Artificial Intelligence in Court

In response to the growing demand for advanced legal decision-making tools, AI-driven systems have emerged to aid juries by autonomously predicting verdicts. Notably, the United States has witnessed the development of a legal automation platform known as AI Lawyer, which is dedicated to forecasting rulings and judgments. Continuous efforts have been devoted to refining its accuracy over time. By analyzing vast repositories of judgment documents, AI Lawyer employs specialized algorithms to render automated judgments. As the era of smart courts and big data-driven legal systems unfolds, there arises a pressing need to devise systems capable of accurately anticipating legal verdicts by leveraging extensive legal precedents. The objective of this study was to craft an AI model tailored for forecasting legal judgments, emphasizing interpretability. This involved identifying key attributes that hold significant potential in predicting privacy violations and thereby constituting unlawful activities. The research also sought to evaluate the performance of verdict prediction, as outlined by Park and Chai (2021).

The traditional paper-based judicial systems, reliant on manual decision-making processes, confront numerous challenges that technology stands poised to address. In the legal industry, the advent of AI represents a transformative tool capable of optimizing the entire legal workflow, spanning from case filing to judgment delivery, thereby mitigating backlog issues. Continued reliance on conventional systems appears increasingly untenable in light of these advancements. interventions in the legal sphere, particularly in tasks like summarization and information retrieval, represent dynamic research domains. Information retrieval entails sourcing pertinent resources from a pool of available information to meet specific needs. Conversely, summarization involves condensing longer texts while retaining crucial information and key insights. While manual summarization remains an option, the automation of this process through computer algorithms is becoming increasingly prevalent.

Artificial Intelligence can be approached through various methodologies, with two primary paradigms being machine learning and deep learning. Machine learning involves enabling a computer program to autonomously assimilate new data without direct human intervention. Essentially, it encompasses the process whereby a computer enhances its own capabilities by iteratively integrating incoming data into an established statistical model. In essence, the machine acquires data, then adjusts its algorithms based on the information received. On the

other hand, deep learning represents an AI technique that mimics the intricate processes of the human brain in data processing, facilitating the extraction of patterns essential for decision-making.

The realm of Legal Technologies has made significant strides in recent years, thanks to advancements in computing power and techniques. These advancements have expanded the possibilities for what can be accomplished in this field. Artificial intelligence is currently being employed in various areas of the Legal Technologies industry. The work of Ghosh et al. (2023) outlines that the exclusive use of AI in court will help to find a reasonable solution to a dispute. In this case, AI will function as a black-box that turns facts of a case(inputs) into a consequence(output) and will automatically decide the solution to a case.

On the contrary, other scientists, like Yoshikawa and Krestel (2025), claim that AI should be used as an assistance tool as it will help legal researchers in law firms apply electronic discovery software to cases that involve many documents to be screened. Also, AI would help tackle issues such as weariness and emotional instability.

The wave of technological transformation is impacting European courts unevenly, as cited by Niklaus et al. (2021).In European countries such as Latvia and Malta, seem to have developed an completely advanced approach with robust applications in terms of legal support whereas some few States, such as Poland and Estonia, application AI in courts still appears to be in an emergent stage and would be limited the near future to the implementation of an effective information technology management.

Within the United States, advanced technologies like COMPAS are employed by juries to assess an individual's risk profile. Primarily utilized in criminal cases, these systems predict the likelihood of a defendant's potential for future criminal behavior. COMPAS serves as a prime example of how Artificial

Intelligence can inform decision-making processes by analyzing statistical data. The work of Sleimi et al. (2018)and Singh et al. (2019)has emphasized the extraction of particular metadata, while their study has also highlighted the analysis of semantic clauses for parsing purposes.

The Malaysian federal court recently addressed the issue of case backlog within its system by implementing several strategic initiatives. These included promoting mediation as a means of resolving cases efficiently, implementing a sophisticated queuing system in West Malaysia aimed at optimizing hearing schedules and minimizing delays experienced by legal practitioners, and introducing an e-filing system to streamline document submission processes. Additionally, the court introduced pre-trial procedures to expedite the adjudication process and effectively address the challenge of case backlog. These comprehensive measures underscore

the court's commitment to enhancing efficiency, reducing unnecessary delays, and ensuring timely resolution of legal matters within the Malaysian judicial system Ali and Razi, 2020).

Despite concerns voiced by critics regarding the potential reinforcement and exacerbation of biases against minority and marginalized communities by AI, they emphasize that technology lacks the nuanced judgment and discernment of human judges. In sentencing, judges not only consider the facts of the case but also take into account mitigating factors and exercise discretion, which AI cannot replicate. The assessment of aggravating and mitigating circumstances necessitates human cognitive abilities.

Moreover, sentencing practices evolve over time in response to changing societal norms and public perceptions. Proposing a solution, it was suggested that addressing increasing caseloads would require a bolstering of judicial resources rather than relying solely on AI. Concerns about bias in sentencing prompted Sarawak Information Systems to remove the "race" variable from its algorithm Nizam, 2021).

Text summarization has emerged as a significant challenge in today's information-rich environment, propelled by the expansive volume of data available and the advancements in Internet technology. This necessity for summarization has become increasingly prevalent, serving as a vital component in various Natural Language Processing tasks. The paper presented herein offers insights into the essence of Text Summarization and its utility across diverse domains.

Amidst the vast and ever-growing repository of legal cases published daily on kenyalaw.org, navigating through this extensive dataset presents a formidable challenge. There arises a need for a tool capable of condensing these cases into concise summaries while preserving their essential implications. Hence, the development of an AI model aimed at expediting judicial proceedings by aiding juries in making swift and informed rulings becomes imperative. The crux of this project lies in the intricate design and implementation of such a model. According to Keong (2017), the implementation of these technologies has led to enhanced court efficiency and a significant decrease in the backlog of cases.

According to the latest State of the Judiciary and Administration of Justice (SOJAR) report, the number of pending cases stood at 134,203 as of June 2022, marking a decrease from the previous year's figure of 150,008. Despite the implementation of advanced technologies such as case management systems, there remains a pressing need for innovative approaches to address the persisting challenges in case backlog.

While there has been a notable 20% reduction in the case backlog, it is evident that this alone has not resolved the underlying issue. The implementation of the Social Transformation

through Access to Justice (STAJ) initiative, which aims to tackle the backlog, underscores the ambition to achieve a comprehensive 100% reduction in pending cases. In light of these statistics and objectives, there arises a clear imperative for the exploration and adoption of AI-driven solutions to optimize judicial processes and enhance access to justice.

1.1.2. Evolution of AI in court

AI in the court system has evolved. In the earlier days, the courts used the manual process of writing down court proceedings on paper after which the jury could review the write-up and deliver a judgement. However, this has gradually changed as courts have adopted AI approaches to expedite rulings and judgements.AI has become an efficient approach to increase efficiency in the legal industry by incorporating and integrating judicial processes.

The integration of AI within legal contexts presents opportunities for enhancing efficiency, precision, and accessibility, according to Nadjia (2024). These AI-driven tools offer invaluable assistance in legal research, case analysis, and decision-making processes. Nevertheless, the infusion of AI into courtroom proceedings brings forth significant legal and ethical concerns. Issues concerning equity, openness, accountability, and safeguarding individual liberties are surfacing as AI systems assume a more prominent role in litigation. Addressing these concerns necessitates a thorough examination of the international legal framework and prevailing practices. By scrutinizing the relevance of existing legal provisions to AI technologies, gaps and deficiencies can be discerned. Furthermore, delving into case studies and instances that underscore the significance of AI in courtroom settings will yield valuable insights into this global quandary.

The utilization of artificial intelligence (AI) within India's judicial system has significantly transformed the landscape of legal processes, as argued by Mali (2022). By harnessing AI algorithms, legal professionals can effectively analyze vast amounts of textual data to identify relevant precedents in case law. This not only aids lawyers in building stronger arguments but also facilitates administrative tasks, leading to streamlined judicial processes.

One of the most notable contributions of AI in the legal domain is its ability to assist judicial officers in making informed predictions. Through sophisticated algorithms, AI systems can analyze various factors, such as the weight of evidence and historical case data, to provide insights into critical issues like sentence duration and the likelihood of conviction or acquittal. This predictive capability empowers judicial officers to make well-informed decisions, thereby enhancing the efficiency and efficacy of the justice delivery system.

The implementation of AI comes at a crucial time when India's courts are grappling with a significant backlog of pending cases. With an estimated 44 million cases according to Okiri et al. (2019) across all levels of the judiciary, addressing this backlog has been a longstanding challenge. Recognizing the potential of AI to address these issues, the Supreme Court constituted an artificial intelligence committee in 2019. This committee is tasked with spearheading the integration of cutting-edge AI and machine learning technologies into the judicial domain, with the overarching goal of enhancing efficiency and expediting the delivery of justice.

The highest court initiated an innovative venture aimed at leveraging AI technology to support legal research efforts for judicial officers. This initiative, known as the Supreme Court Portal for Assistance in Courts Efficiency (SUPACE), was officially launched in April 2021, as documented in the India Judiciary Report for 2021. SUPACE represents an AI-powered tool designed to enhance the operational efficiency of judicial officers by facilitating case analysis, information extraction, and document review processes.

The AI-driven workflow of SUPACE, according to Singh (2024), comprises four key components: file preview, chatbot functionality, logic gate operations, and notebook integration. Through the platform, case files in PDF format can be seamlessly converted into accessible text, enabling swift and comprehensive review. The chatbot feature, which supports both text and voice interactions, provides a rapid overview of case details, while the fact extraction system retrieves pertinent information about the case at hand. Additionally, the integrated word processor functionality ensures that SUPACE serves as a comprehensive end-to-end solution for legal research tasks.

Of particular note is the system's reliance on user-driven training mechanisms. Human users play a pivotal role in annotating and extracting information, which serves as the basis for training the AI system. Through continuous learning and pattern recognition, SUPACE evolves to better meet the needs and preferences of its users, ultimately enhancing its efficacy and utility within the legal research domain Kanipakam, 2023).

China has made significant advancements in legal technology, particularly with the implementation of AI in its judicial system Shi et al., 2021). The introduction of smart courts has revolutionized the way judicial officers operate across the country. Through electronic integration with various databases, including those of the police, prosecutor, and government, as well as integration with China's social credit system, the smart court system aims to streamline judicial processes. By connecting every judicial officer's desk, the system effectively reduces workload by nearly a third.

The Machine Learning-powered Smart Court System (SOS) plays a crucial role in assisting judicial officers. It screens court cases for relevant references and provides recommendations on laws and regulations. Additionally, it aids in drafting documents that facilitate verdict determination.

Similarly, Malaysia, according to Putera et al. (2022), has also embraced AI technology in its legal system, particularly in the realm of sentencing. Through a nationwide pilot program, AI tools for sentencing were tested, with the courts of Sabah and Sarawak utilizing software developed by Sarawak Information Systems. The goal was to alleviate the case backlog and improve efficiency within the judicial system.

The global trend of integrating AI into the criminal justice system is evident, with various jurisdictions adopting AI-based systems for diverse applications. From chatbots like the "Do Not Pay" lawyer app to robot judges in Estonia and AI mediators in Canada, these technologies are aimed at enhancing consistency in sentencing, expediting case resolution, and reducing the burden of lengthy litigation processes.

A significant portion of government respondents in a recent global survey expressed intentions to increase investments in AI-powered systems, including chatbots, facial recognition, and data mining. In Malaysia, federal authorities are optimistic about the potential of AI sentencing tools to enhance judgment quality, although the specific implementation details are yet to be clarified.

1.2 Problem statement

The judicial system in Kenya has been grappling with significant challenges related to prolonged delays and an overwhelming backlog of cases, despite various efforts to improve efficiency. As of June 2022, the number of pending cases stood at 134,203, although a 20% reduction from 150,008 in the previous year was recorded (State of the Judiciary and Administration of Justice Commission, 2021). The persistent backlog remains a critical issue, underscoring the difficulty of managing the increasing volume of cases. A primary factor contributing to this backlog is the extensive time and resources required to process and review court decisions. This problem is compounded by the lengthy and labor-intensive nature of the legal review process, where judges and legal practitioners must manually sift through vast amounts of documentation, including past case rulings and precedents, to arrive at informed decisions.

While efforts like the implementation of the e-court system, which includes components such as e-filing and case management systems, have been introduced to streamline case handling, these measures have not sufficiently resolved the issue of delayed case processing and backlog Ngetich, 2019). One of the key constraints preventing further progress is the inefficiency in reviewing and analyzing large volumes of past court decisions. Legal professionals are often faced with the challenge of identifying relevant case law and precedents from extensive case records, which significantly delays verdict determination and exacerbates the backlog Makau, 2014).

In response to this challenge, Natural Language Processing (NLP) methods present a promising opportunity to enhance the efficiency of the judiciary. Specifically, text summarization and keyword extraction techniques within NLP can be employed to expedite the review of past cases by automatically generating concise summaries and extracting relevant legal keywords from extensive court rulings stored in Portable Document Format (PDF). This approach could significantly reduce the time needed to identify pertinent information and precedents, thus improving decision-making efficiency (Agarwal et al., 2020). The BERT (Bidirectional Encoder Representations from Transformers) model, an advanced NLP tool, could play a crucial role in this process by analyzing legal documents, extracting key information, and providing summaries that would assist judges in quickly understanding the essence of cases without the need to read through entire documents.

The application of NLP techniques directly addresses the time-consuming nature of case review, which is a major constraint in the judicial process. The ability to automatically identify and summarize the most relevant information from vast legal texts allows judges to focus on the essential details of a case, rather than spending excessive time on document review. Consequently, this research aims to explore how the adoption of NLP can reduce delays, enhance the efficiency of verdict determination, and alleviate the backlog in Kenya's judiciary system, ultimately contributing to a more effective and accessible justice system.

The adoption of NLP techniques for classification and summarization directly addresses the inefficiency of the current manual review process. The time-consuming nature of manually sifting through vast amounts of legal text is a major bottleneck in Kenya's judicial system. By automating these tasks, NLP could significantly reduce the workload on judges, thereby speeding up the review process and alleviating the backlog.

This research aims to explore the potential of NLP, specifically BERT, for classification and summarization tasks, to streamline the case review process. By automatically categorizing legal cases and generating concise summaries, the proposed system would enable judges to make

faster and more informed decisions, ultimately improving the overall efficiency of the judicial process and addressing the critical issue of the case backlog in Kenya's legal system.

1.3. Research Aim

This research will explore the application of AI techniques in the legal industry, focusing on Natural language processing for legal text prediction and classification.

1.4. Research objectives

a) 1.To analyze existing NLP techniques applicable to legal text processing and their effectiveness in a judicial setting.
b) 2.To develop and deploy an NLP-based model for automating case summarization, legal document classification.
c) 3.To evaluate the performance and impact of the proposed NLP model in improving efficiency within the judiciary.

1.5. Research questions

a) 1. What are the existing NLP techniques applicable to legal text processing, and how effective are they in the judicial setting?
b) 2. How can an NLP-based model be developed and deployed to automate case summarization and legal document classification?
c) 3. What is the impact of the proposed NLP model on efficiency within the judiciary, and how does it compare to existing manual processes?

1.6 Significance of the study

It is essential to consider this study within the context of the substantial backlog of cases prevalent in Kenyan courts. Therefore, the research aims to illustrate the potential of AI as a complementary tool to augment human capabilities. Specifically, it seeks to showcase how AI can be leveraged to automate certain elements within the legal process, like condensing case details into summaries and identifying critical keywords, and can be addressed through alternative methods.

This holds significance due to the current reliance on manual procedures for reviewing entire court proceedings or decisions and extracting references from relevant judicial acts in Kenya's criminal procedure code. These manual processes are not only time-consuming but also incur significant costs. Litigants often face prolonged waiting periods for court decisions, contributing to the mounting backlog of cases within the judicial system.

Thus, the study holds significance as it will create efficiency in access to justice, as justice delayed is justice denied Ngoepe and Makhubela, 2015).

1.7 Scope of the study

The core objective of this research centers on exploring artificial intelligence methodologies in legal forecasting. It encompasses an extensive review of existing literature, the creation and assessment of deep learning models, and the subsequent implementation of these models.

Chapter 2. Literature Review

2.1 Overview

This section delves into the theoretical underpinnings of keyword information retrieval and extractive text summarization. Additionally, it explores previous research endeavors focusing on information retrieval, extractive text summarization, and AI models tailored for automating natural language tasks. AI offers a promising avenue for automating the extraction of insights from legal texts, condensing extensive document collections into succinct summaries while extracting only the most pertinent information.

2.2 Theoretical Literature

2.2.1 Information Retrieval theory

The theory of information retrieval encompasses various aspects such as how information is represented, stored, organized, and retrieved from textual data. In the domain of keyword extraction, this theory posits that the most significant words or phrases within a text corpus are those closely linked to the primary topics. Moreover, information retrieval theory underscores the significance of both term frequency (TF) and inverse document frequency (IDF) in identifying crucial keywords.

The widely employed TF-IDF approach in keyword extraction, grounded in information retrieval theory, according to Van Rijsbergen (1979), gauges the relative importance of a term in a document by considering its frequency within the document and rarity across the entire corpus. Furthermore, information retrieval principles have inspired the development of algorithms utilizing graph theory and machine learning for keyword extraction. Notably, the PageRank algorithm, employed by Google for webpage ranking, draws on the concept of centrality in a graph, which assesses the significance of a node based on its connections with other nodes. This concept mirrors information retrieval theory's emphasis on the interrelationships between terms in documents or corpora.

In essence, information retrieval theory offers a valuable theoretical framework for comprehending the underlying principles and methodologies employed in keyword extraction techniques.

2.2.2 Cognitive theory

Cognitive theory delves into the intricacies of how individuals perceive and comprehend written language, offering insights into the selection and comprehension of keywords within a text.

As per cognitive theory and Beck and Haigh (2014), readers engage with text by constructing a mental representation of its meaning, synthesizing information from words, syntax, and discourse structure. This mental model serves as the foundation for making inferences and predictions about the text's content as it is processed.

In the realm of keyword extraction, cognitive theory posits that significant keywords are those that hold prominence and provide pertinent information within the mental representation of the text. Salience refers to the distinctiveness of a word or phrase within this mental construct, while informativeness denotes its relevance to the text's topics.

Drawing from cognitive theory, algorithms for keyword extraction leverage linguistic features such as parts of speech and syntactic structure. For instance, noun and verb phrases are frequently identified as candidate keywords because of their tendency to convey informative and significant aspects of the text.

Essentially, cognitive theory provides a valuable lens through which to examine the mental processes involved in extracting keywords. Moreover, it guides the creation of algorithms designed to identify the most critical and informative keywords within a given text.

2.2.3 Information Comprehension Theory

Information compression theory, a component of information theory, explores methods to compress information while retaining its essential content.

According to Wyer Jr and Radvansky (1999) , this theory posits that an effective summary should encapsulate the key information from the original text while condensing it into a shorter format. It emphasizes that the most crucial information holds the highest information content, conveying the most pertinent and valuable insights to the reader.

To operationalize this theory, algorithms for extractive summarization utilize diverse techniques to identify pivotal information within the text. These methods encompass statistical approaches like TF-IDF and graph-based techniques such as PageRank.

Moreover, specific extractive summarization techniques utilize optimization approaches to identify the most crucial sentences that encapsulate essential content. For instance, certain algorithms utilize integer linear programming to pinpoint sentences that effectively convey critical information while minimizing repetition and enhancing coherence.

In essence, the theoretical framework of information compression theory provides a solid basis for understanding the principles and methodologies behind extractive summarization. By emphasizing the extraction of vital information from the original text, these algorithms can produce concise summaries that capture the essence of the source material.

2.3 Empirical Literature

2.3.1 Machine learning Techniques

A paper by Han et al. (2019)used fuzzy networks in a study concerning disputes in Australian family law employed data from 50,000 divorces. The objective of the negotiation model was to craft a settlement agreement that satisfied both parties involved. This model primarily concentrated on assigning issues to the parties who placed the highest value on them, while also compensating the other party for their perceived loss. Notably, the algorithm utilized in this study did not rely on any pre-existing knowledge of the negotiation's specifics. Instead, it necessitated negotiators to articulate their own disputes.

Hmeidi et al. (2015)conducted a study on text classification methods, encompassing decision trees and support vector machines. It was revealed that both the selection of features and the volume of training data wield substantial influence over the efficacy of text classification models.

Chen et al. (2020)in his paper outlined two machine learning models, support vector machine and logistic regression used in predictive coding. Experiment to compare deep learning results with results obtained from the machine learning models was conducted. The results showed that CNN performed better with a larger volume of training dataset and should be the best method for text classification in legal domain. Dataset of real legal matters was used in the experiment.

Chen et al. (2020)compared deep learning model technology and traditional target tracking models. A deep learning framework was constructed for automatic trajectory tracking based on learning of the legal tasks. The experimental results showed that the Deep GTT algorithm improves accuracy and efficiency as compared with state of art models. Trajectory data was imposed on the model to enable the short term memory based tracking network to learn and generate dynamic laws.

Machine learning enable machines to learn from a vast data input through model so as to identify and judge. Ma and Mei (2021)while aiming up at the deep learning problem, developed a text matching algorithm suitable for the field of housing law. He compared the depth of the deep learning model with general algorithm. The findings showed that the classifier based on matching algorithm is a promising classification technique.

Singh et al. (2021)employed support vector machine. Employing a support vector machine model, they extracted keywords from text by analyzing a dataset comprising 500 Korean news

articles. Their results indicated that the SVM model outperformed alternative baseline models like TF-IDF and TextRank, exhibiting superior accuracy.

Chen et al. (2022)proposed a machine learning algorithm using random forest as classifier. The experiment used 30,000 full united states case documents in 50 categories to demonstrate that machine learning outperforms a deep learning system built on multiple pre-trained word embedding and deep neural networks. Random forest achieved the best performance. The study methodology involved building two legal text classification systems and conducting evaluations.

2.3.2 Deep learning Techniques

Bansal et al. (2021)discuss the use of CNNs and RNNs, particularly LSTMs and GRUs, in the legal domain. These models have been employed for tasks like legal document classification and sentence-level analysis, demonstrating their ability to effectively capture the hierarchical structure of legal language. Specifically, the review identifies the following models:

Node2Vec: This algorithm was used to represent legal entities as vectors, which were then fed into a neural network for legal classification tasks.

Sentence Similarity Models: These models aim to assess the similarity between legal sentences by embedding them into a vector space. By comparing sentence representations, these models are able to predict whether two sentences in different legal documents are similar or related.

Vector Space Models: These models use vector representations, such as Word2Vec or GloVe embeddings, to map words or sentences into a continuous vector space. These vectors are then processed using a neural network for various tasks like document classification and case prediction.

The primary legal problem under investigation in Bansal et al. (2021)study was legal case classification and document categorization. Legal cases often involve vast amounts of text, and one of the challenges is efficiently classifying legal documents based on their content. This requires understanding complex legal terminologies and relationships between various elements of the case. The study sought to improve the accuracy and efficiency of these tasks by using deep learning methods.

Bansal et al. (2021)utilized data from the Washington University School of Law Supreme Court database, which contains a comprehensive collection of U.S. Supreme Court cases. This data provided labeled examples of legal documents, which were used to train the deep learning models. By leveraging this dataset, the models were able to learn how to categorize legal documents based on their textual content.

The findings of Bansal et al. (2021)suggested that deep learning models, particularly CNNs with Word2Vec embeddings, performed well in the legal domain. In their experiments, the CNN model with Word2Vec embeddings achieved an accuracy of 72.4% in the classification task, which was better than other models tested. Furthermore, the study found that ensemble models (combinations of different models) outperformed individual models, leading to improved prediction accuracy.

Additionally, the study compared machine learning algorithms (like SVM and Random Forest) to deep learning-based systems and found that deep learning models consistently outperformed traditional machine learning approaches in the context of legal tasks. The ensemble model, in particular, demonstrated higher accuracy, reinforcing the idea that combining multiple models can lead to more robust performance in complex tasks such as legal text analysis.

Other studies in the legal domain have used similar techniques to tackle legal tasks but have explored different models and datasets. For instance:Chalkidis et al. (2021)used BERT-based models for legal case retrieval and document classification, achieving high performance in case law prediction and legal information retrieval. Their study highlighted how transformer-based models like BERT are effective at understanding context and semantics in legal documents, leading to more accurate results compared to older models.

Ribeiro et al. (2024)applied LSTMs and GRUs for legal document summarization and case law prediction. Their findings showed that RNN-based models (especially with LSTM and GRU variants) were effective for sequential data and were able to capture the complex temporal dependencies present in legal text, leading to better performance in predicting case outcomes.

Antici et al. (2023)focused on sentence-level legal classification using transformer models, including BERT and GPT-3, achieving an accuracy of over 80%. Their results indicate that transformer models, with their capacity for fine-tuning, offer superior performance in tasks requiring a deep understanding of legal semantics.

While Bansal et al. (2021) found that CNN models with Word2Vec embeddings performed well with an accuracy of 72.4%, other studies show the superiority of transformer-based models like BERT and GPT-3 in similar legal tasks. For example, BERT-based models have consistently outperformed CNNs and RNNs in terms of accuracy, especially in tasks involving nuanced understanding of language and context Chalkidis et al. (2021).

The comparative findings across these studies suggest that CNNs with word embeddings (like Word2Vec) work well for structured tasks like document classification, but they may not capture contextual relationships as effectively as transformer models.RNN-based models (especially LSTMs and GRUs) excel at tasks involving sequential data, such as case outcome

prediction, but they may be outperformed by transformer-based models in tasks requiring deep understanding of context. Ensemble models, combining different neural architectures or incorporating multiple features, continue to show promise for improving classification accuracy.

Su et al. (2019)This methodology presents an innovative approach to training neural networks for single-document extractive summarization without relying on heuristically generated extractive labels. This contrasts with previous methods that often used manually or heuristically designed labels for training extractive summarization models.

Su et al. (2019) approach, BanditSum, addresses extractive summarization by treating it as a contextual bandit problem. In this framework, the document as the Context; the document to be summarized is considered the "context," and the Sequence of Sentences is Action: The model selects a sequence of sentences from the document, termed as the "action," to create the summary.

Instead of relying on supervised learning with pre-defined extractive labels (e.g., sentences labeled as either "important" or "not important"), BanditSum views summarization as a reinforcement learning (RL) problem. In this approach, the model is trained to maximize the reward signal based on the quality of the selected summary, where rewards are based on metrics such as ROUGE scores or other evaluation criteria relevant to summarization quality.

The model employs a policy-gradient reinforcement learning approach, which is particularly suitable for problems involving sequential decision-making, like sentence selection in extractive summarization. The policy-gradient method optimizes the policy that determines the best sentences to include in the summary, iteratively improving the model's ability to produce high-quality summaries through reinforcement learning.

The main problem addressed by Su et al. (2019) is the inefficiency and limitations of traditional extractive summarization techniques, which typically require the use of heuristically generated extractive labels. These labels are often manually created or generated by rule-based systems, which may not generalize well across different domains and are time-consuming to create. Dong proposes a novel method that eliminates the need for manually labeled data, instead leveraging a reinforcement learning approach to directly optimize the summarization model's performance.

Dong's framework allows for end-to-end training of a neural network-based summarizer without the need for extensive manual data labeling. This is a significant departure from traditional supervised models, which rely on large, labeled datasets for training. The approach is particularly useful in settings where labeled data is scarce or costly to obtain.

To evaluate the proposed BanditSum methodology, Su et al. (2019)applied the model to the CNN/Daily Mail (CNN/DM) dataset. The CNN/DM dataset is a large collection of news articles paired with human-written summaries, widely used for evaluating both extractive and abstractive summarization models. It consists of hundreds of thousands of document-summary pairs, making it a standard benchmark for summarization tasks.Dong's model was trained and evaluated on this dataset to demonstrate its effectiveness in performing single-document extractive summarization without relying on heuristically generated extractive labels. The model used reinforcement learning to learn the process of selecting the most informative sentences from the documents and aggregating them into a coherent summary.

Su et al. (2019) found that BanditSum significantly outperformed traditional supervised extractive summarization models. Specifically, the policy-gradient reinforcement learning approach allowed the model to learn optimal sentence selection without requiring manually created extractive labels. This led to improved summarization quality as measured by standard evaluation metrics like ROUGE, which are widely used in summarization tasks to measure the overlap of n-grams between the generated and reference summaries.

In particular, BanditSum demonstrated the following: improved ROUGE Scores: The model achieved higher ROUGE scores compared to traditional extractive summarization methods that rely on heuristic labels, showing that reinforcement learning can effectively optimize sentence selection for summarization.

End-to-End Training: By eliminating the need for manually annotated training data, BanditSum provides a more scalable approach to training summarization models, as it can be trained directly on large datasets without needing human intervention for labeling. The method showed potential for generalizing across various types of document corpora, as it does not rely on domain-specific labels or heuristics. When comparing BanditSum to other models in the field of extractive summarization, several key observations arise:

Traditional Extractive Summarization Models: Earlier models, such as those based on Text Rank Leite et al., 2007) Supervised Sentence Ranking Models (such as those using SVM or Logistic Regression), rely on manually labeled datasets for training. These methods typically involve extracting key sentences based on predefined features like sentence position, keyword frequency, and sentence similarity. However, these models have limitations in terms of generalization and adaptability across different domains. Dong's BanditSum overcomes these issues by leveraging reinforcement learning, which allows for more flexible and contextually appropriate summarization.

Deep Learning-Based Models: Other deep learning approaches for extractive summarization, such as those using LSTMs or CNNs, have also shown success in the task. However, most of these models require labeled data for training. For example, Prabhavalkar et al. (2017) introduced a sequence-to-sequence model with attention for summarization, but their method still required human-labeled training data for training the model. In contrast, Bandit Sum's reinforcement learning-based approach offers a more data-efficient solution for summarization without the need for labeled data.

Reinforcement Learning for Summarization: Dong's BanditSum aligns with other reinforcement learning-based approaches in summarization, such as the RL-based abstractive summarization model proposed by Jang and Kim (2021). While Paulus et al. focus on abstractive summarization, where the model generates entirely new sentences, Dong focuses on extractive summarization. Both methods, however, share the use of reinforcement learning to optimize the model for summarization tasks.

More recent approaches in extractive summarization, such as those based on BERT Jang and Kim (2021), show state-of-the-art results in both extractive and abstractive summarization tasks. BERT-based models, which rely on fine-tuning pre-trained models for summarization tasks, generally outperform traditional methods. However, BERT-based models still rely on labeled datasets for training, whereas BanditSum's reinforcement learning approach does not require labeled data, offering a different set of advantages in terms of scalability and adaptability.

Dong et al. (2018) BanditSum represents a significant step forward in extractive summarization by employing reinforcement learning to eliminate the need for manually labeled data. Compared to traditional methods and other deep learning-based models, BanditSum offers a scalable, data-efficient, and effective alternative for summarizing documents. Its ability to optimize sentence selection through reinforcement learning provides a promising avenue for improving summarization tasks in a variety of domains, especially where labeled data is limited or unavailable.

2.3.3 NLP models in law

Christian et al. (2016)employed the TF-IDF algorithm as the core model for their automatic summarization tool. TF-IDF is a widely used method in information retrieval and text mining that measures the importance of a word within a document relative to a collection of documents. The TF-IDF algorithm operates in two stages: Term Frequency (TF): This measures how often a word appears in a document, which is a basic indicator of the word's importance within that

document, and Inverse Document Frequency (IDF): This measures how unique a word is across the entire document collection. Words that are frequent across many documents are considered less important than rare words.

By combining these two measures, TF-IDF can rank words in terms of their significance in a given document relative to others in the collection. In the case of summarization, sentences containing high-importance words, as determined by the TF-IDF score, are selected to form the summary. In addition to developing the TF-IDF-based summarizer, the study conducted a comparative analysis with various other online automatic text summarization tools. These tools included both extractive and abstractive summarization methods, to evaluate the relative efficacy of the TF-IDF approach in comparison to more sophisticated models.

The main problem Christian et al. (2016) aimed to address was the increasing need for effective tools to summarize large volumes of online text. With the vast amounts of unstructured information available on the web, users often face difficulty processing and digesting this content. The ability to generate concise summaries from large bodies of text is valuable for improving information retrieval, content analysis, and overall user experience. The authors sought to develop a fast and efficient summarization tool using the TF-IDF algorithm and to assess its performance against other available summarizers. One of the key goals was to understand how well the TF-IDF-based model performed when compared to other online tools and which summarization approach could provide the most reliable and informative summaries.

Christian et al. (2016) utilized a variety of online text datasets to evaluate the summarization tools, including news articles and research papers. The specific datasets used were not disclosed in detail in the study, but typical datasets for summarization tasks include large text corpora such as the CNN/Daily Mail dataset or DUC (Document Understanding Conference) datasets.The primary metric used to evaluate the summarization models was the F-Measure, which is the harmonic mean of precision and recall. The F-Measure is a common evaluation metric in information retrieval tasks, particularly when comparing the relevance of retrieved documents or generated summaries to human annotations. Christian et al. (2016) found that their TF-IDF-based summarizer performed reasonably well when compared to other online summarization tools, but it did not outperform more advanced methods. TF-IDF Summarizer Performance: The TF-IDF-based summarizer demonstrated acceptable performance, producing summaries that were fairly accurate in terms of representing the core ideas of the

original text. However, it lacked the sophistication of abstractive models, which can rephrase sentences and capture a broader range of context.

Comparison to Other Summarization Tools: The TF-IDF-based summarizer was outperformed by more complex methods, such as latent semantic analysis (LSA) or neural network-based models, which were able to generate summaries that were more concise and coherent. These tools, particularly abstractive models, could generate summaries that reflected the main ideas in a more readable and human-like manner. The TF-IDF model performed moderately well on the F-Measure metric, but the more advanced models showed better precision and recall, indicating that they were able to more accurately extract important content and summarize it more effectively.

When comparing Christian et al. (2016)with other summarization studies, several key trends emerge: TF-IDF vs. Other Methods: Earlier methods based on TF-IDF often suffered from limitations in handling semantic relationships between sentences. Unlike abstractive summarizers that can generate summaries with human-like rephrasing, TF-IDF extractive summarizers are purely based on selecting sentences that contain important terms, which can lead to incoherent summaries. For instance, Chen (2017)used Latent Dirichlet Allocation (LDA) and showed improved results in topic modeling for summarization. They demonstrated how topic-based models could overcome some of the weaknesses of TF-IDF by better capturing the semantic structure of the document. Neural Network Approaches: More recent approaches have utilized neural networks to perform both extractive and abstractive summarization. Models such as Sequence-to-Sequence (Seq2Seq) networks with attention mechanisms Colombo et al. (2020)have led to significant advancements in text summarization. These models have been able to generate more human-like summaries, capturing complex relationships between words and sentences, a feature not captured by TF-IDF.

Abstractive Summarization: Abstractive models, particularly those based on transformers such as BERT Devlin et al. (2019)and GPT-3 Brown et al. (2020), have surpassed traditional extractive models like TF-IDF in producing coherent and informative summaries. These models are trained to generate new sentences, often producing summaries that are more fluent and contextually accurate compared to extractive methods like TF-IDF, which simply selects sentences from the original document. Evaluation Metrics: Christian et al. (2016)used the F-Measure to assess summarization quality, but more modern studies use additional metrics such as ROUGE (Recall-Oriented Understudy for Gisting Evaluation), which is widely accepted in

the summarization community for evaluating both extractive and abstractive summaries. The use of ROUGE has become the standard for benchmarking summarization performance, offering a more comprehensive measure of a summary's quality.Christian et al. (2016)made a significant contribution to the field of automatic text summarization by applying the TF-IDF algorithm and comparing its performance with other online summarization tools. Their work highlighted the limitations of simpler extractive methods like TF-IDF, which can struggle with generating coherent and meaningful summaries compared to more advanced models. However, TF-IDF remains an important baseline in the summarization field, especially in situations where computational efficiency is paramount or labeled data is scarce. When comparing their results with other studies, it becomes evident that neural network-based models and abstractive summarization techniques have emerged as more effective solutions in producing human-like, coherent summaries. Nevertheless, the simplicity and computational efficiency of TF-IDF-based models still make them a valuable tool for certain summarization tasks.

Mustapha et al. (2020) researched XLNet, utilizing a substantial web corpus dataset. The study showcased XLNet's state-of-the-art performance across a multitude of natural language processing (NLP) tasks. Additionally, XLNet introduced an innovative permutation-based training approach, surpassing the effectiveness of conventional methods. The core model in Mustapha et al. (2020)was XLNet, a transformer-based model that integrates the strengths of both autoregressive models (such as GPT) and auto-encoding models (such as BERT). XLNet differs from BERT in its approach to pre-training:

Permutation-based Training: Unlike BERT, which uses masked language modeling (where random tokens in the input are replaced with a special mask token), XLNet employs a permutation-based training objective. This allows the model to predict words in a sequence in a more flexible manner by considering all possible permutations of the word order. This training method overcomes some of the limitations of the masked token prediction used in BERT, capturing richer context and improving the model's ability to learn bidirectional dependencies. Transformer Architecture: XLNet uses a transformer architecture similar to BERT but with modifications to accommodate the permutation-based training. It leverages a self-attention mechanism that enables the model to learn relationships between words regardless of their positions in the input sequence. The study by Mustapha et al. (2020)showed how XLNet could outperform previous models in several key NLP tasks such as text classification, question answering, sentiment analysis, and text generation, setting new performance records.

The main problem that Mustapha et al. (2020)sought to address was the inefficiency of conventional pre-training strategies in transformer models like BERT. While BERT's masked language modeling has been highly successful in improving the performance of various NLP tasks, it has limitations, such as an inability to capture bidirectional dependencies effectively (due to the masked token prediction). Suboptimal utilization of context, as the model is only trained to predict certain tokens, missing out on the broader contextual understanding that could be gained from training on all tokens at once. Mustapha et al. aimed to improve upon these limitations with XLNet's permutation-based training, which offers a more powerful and efficient way to capture dependencies between words and allows the model to leverage both forward and backward context more comprehensively.

Mustapha et al. (2020) conducted experiments on XLNet using a substantial web corpus dataset. The dataset included a wide range of text from various domains, such as news articles, web pages, and other publicly available documents. The use of such a large and diverse corpus is important for training models on a wide variety of linguistic patterns and nuances, making the model more robust across different tasks. The study evaluated XLNet's performance on several established NLP benchmark datasets, including GLUE (General Language Understanding Evaluation): A benchmark that evaluates a model's performance on tasks like sentiment analysis, textual entailment, and question answering.SQuAD (Stanford Question Answering Dataset): A dataset focused on question answering that tests how well models can understand and answer factual questions from a text. RACE (Reading Comprehension from Examinations): A dataset designed to evaluate reading comprehension abilities. These datasets provided a comprehensive evaluation of XLNet's ability to handle diverse tasks in NLP, from sentence-level tasks like classification to more complex comprehension tasks like question answering.

Mustapha et al. (2020) demonstrated that XLNet consistently outperforms BERT and other pre-existing models across several NLP tasks. Some of the key findings from the study include: State-of-the-art Performance: XLNet achieved state-of-the-art results on multiple benchmark datasets. It surpassed the performance of BERT on the GLUE benchmark, achieving higher F1 scores and accuracy across multiple tasks, including sentiment analysis and natural language inference. Permutation-based Training: The permutation-based approach proved to be more effective than traditional masked language modeling. By allowing the model to consider multiple permutations of word orders during training, XLNet was able to better capture long-range dependencies and contextual information in text. This contributed to its superior

performance on tasks that required understanding the full context, such as question answering and document classification.

Flexibility Across Tasks: XLNet's flexible architecture allowed it to excel across a wide range of NLP tasks, demonstrating the model's versatility. The unidirectional nature of autoregressive models like GPT had previously limited their ability to capture bidirectional context, but XLNet overcame this limitation through its permutation-based approach. Efficiency and Robustness; XLNet was shown to be more efficient than previous transformer models in terms of computational resources, enabling faster training times without sacrificing performance. This efficiency is particularly valuable for large-scale applications where resources and time are often limited.

When comparing the results from Mustapha et al. (2020) with other studies in the NLP field, several insights emerge: XLNet vs. BERT: The comparison between XLNet and BERT highlights the advantages of XLNet's permutation-based training over BERT's masked language modeling. BERT had been the standard in many NLP tasks until XLNet was introduced, but the latter outperformed BERT in many areas due to its improved handling of contextual information. For example, BERT Devlin et al. (2019)introduced BERT, which set new benchmarks for NLP at the time, but XLNet improved on this by addressing some of BERT's inherent limitations. XLNet vs. GPT: While GPT Brown et al. (2020)is a highly successful autoregressive model for generating text, it was less effective in tasks requiring deep contextual understanding, such as question answering. XLNet combined the strengths of both autoregressive and auto-encoding approaches, allowing it to capture a broader range of linguistic features compared to GPT, leading to better performance on tasks requiring understanding of full document context. XLNet vs. Transformer-based Models: XLNet outperformed other transformer-based models that also relied on self-attention mechanisms, such as Transformer-XL Mustapha et al. (2020), which was designed to handle long-range dependencies. While Transformer-XL was effective for long sequences, XLNet's permutation-based method provided a more generalized solution that worked across a range of tasks, rather than focusing specifically on long-term dependencies. Abstractive vs. Extractive Models: XLNet's permutation-based approach makes it particularly effective for both extractive and abstractive summarization tasks, as well as text generation. Compared to traditional extractive models like those based on TF-IDF Christian et al. (2016)XLNet demonstrated more nuanced understanding of text, leading to more coherent summaries and better performance on text generation tasks.

Mustapha et al. (2020) made a significant contribution to the field of NLP by showcasing XLNet's state-of-the-art performance across various NLP tasks. XLNet's innovative permutation-based training set it apart from previous transformer models like BERT and GPT, allowing it to capture bidirectional dependencies more effectively. The model's ability to perform well across a range of tasks, from text classification to question answering, underscores its versatility and effectiveness. When compared to other models, XLNet demonstrated superior performance on multiple NLP benchmarks, establishing itself as one of the leading models in the field. The study highlighted the potential of XLNet to revolutionize applications in text summarization, question answering, and natural language understanding, making it an essential tool for modern NLP tasks.

2.4 Research gap

While traditional natural language processing (NLP) methods, such as TF-IDF Christian et al., 2016)have offered foundational tools for extractive text summarization, they fall short in capturing the semantic nuances and contextual relationships necessary for coherent and meaningful summaries in complex domains such as the judiciary. These limitations often result in fragmented and context-poor outputs that do not adequately reflect the intricate legal reasoning found in court judgments and filings. Furthermore, although TF-IDF is computationally efficient, it cannot handle long-range dependencies and interpret context beyond sentence-level term frequency. Advancements in NLP, particularly transformer-based models such as XLNet Mustapha et al., 2020) have demonstrated significant improvements in tasks like summarization, question answering, and classification. XLNet's permutation-based training offers deeper contextual learning compared to BERT's masked language modeling. However, while XLNet has proven effective across general NLP benchmarks (e.g., GLUE, SQuAD), its application within the legal domain, especially for enhancing court efficiency, remains under-explored.

On the other hand, BERT Devlin et al. (2019)has emerged as a widely adopted and fine-tunable model for domain-specific applications, offering strong performance in extractive summarization, legal document classification, and semantic search. Its pre-training on large corpora enables it to understand legal jargon and context better than classical models. However, there is a lack of focused research applying BERT-based NLP techniques within the judicial system, particularly in automating repetitive legal tasks such as case summarization and classification of legal documents, all of which significantly impact the efficiency of case processing in courts.

Thus, this study fills a critical gap by exploring how BERT can be fine-tuned for legal NLP tasks to improve court efficiency. The research focuses on deploying BERT to automate legal text processing in areas like case summarization and classification, with the ultimate goal of reducing case backlog, enhancing legal research, and supporting judicial officers in decision-making.

Chapter 3: Research Methodology

This chapter breaks down each step that must be taken to meet the research objectives stated in the introduction.

3.1 Research design

This study adopts the Design Science Research (DSR) methodology, which is suitable for research focused on developing and evaluating artifacts to solve real-world problems. In the context of this research, the artifact is a BERT-based NLP model designed to enhance efficiency in the judiciary through automation of legal text processing tasks such as case summarization and document classification.

Design Science Research involves a structured process comprising the following key stages:

Problem Identification and Motivation

The first stage involves clearly defining the research problem in this case, inefficiencies in legal text analysis within the judiciary. This inefficiency stems from the manual and time-consuming nature of summarizing and classifying legal documents. The study seeks to automate and optimize this process using advanced NLP techniques.

Define the Objectives of the Solution

The study aims to design a solution that can:

Accurately summarize complex legal texts using BERT-based models, classify legal documents with high precision and improve turnaround time and reduce human workload in the judiciary.

Design and Development

In this phase, a fine-tuned BERT model is developed. The model is trained using judicial datasets, including rulings, pleadings, and court transcripts. Pre-processing steps will involve text cleaning, legal-specific tokenization, and formatting for model compatibility.

Demonstration

The developed BERT model is applied to real-world tasks, such as summarizing anti-corruption court cases or classifying documents into categories (e.g., rulings, petitions, witness statements). This step illustrates how the model functions in a judicial workflow.

Evaluation

The model's performance is evaluated using metrics such as F1-score, Precision, Recall, and ROUGE for summarization tasks. Comparisons are made with traditional methods such as TF-IDF and existing baseline tools to assess the improvement in efficiency and accuracy.

Communication

The results, design artifacts, and recommendations will be documented and shared with key stakeholders in the judiciary and the research community. This will contribute to both scholarly knowledge and institutional innovation.

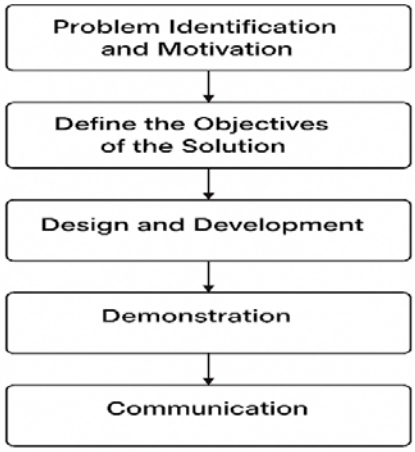


Figure 5.1: Design Science Research (DSR) (Source: Free encyclopedia)

3.2 Data source and collection methods

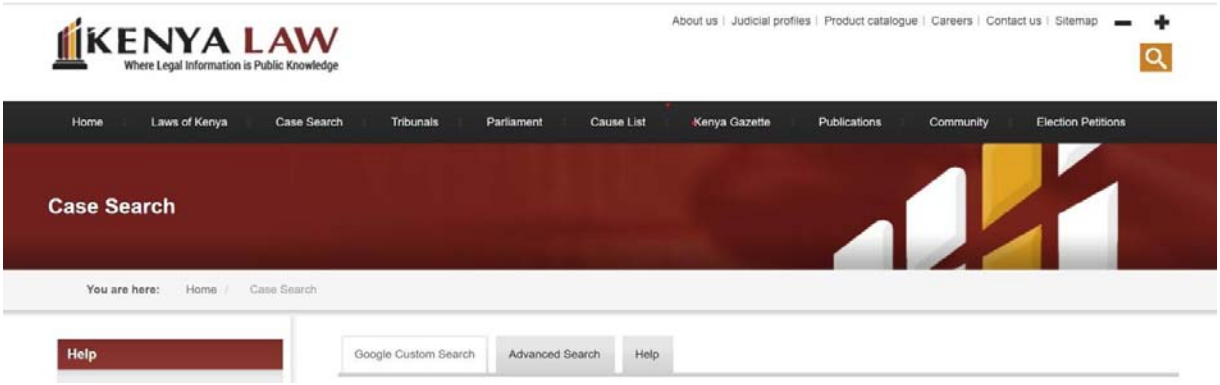


Figure 5.2: Source, Kenya law org website

In this study dataset was extracted from Kenya Law.org, a comprehensive national platform overseen by the National Council of Law. This platform serves as a repository for a vast array of legal resources, including criminal court cases dating back to 1971 and continuing to the present day. Encompassing decisions from both the High Court and Court of Appeal in Kenya, this extensive database serves as a valuable resource for legal research and analysis.

To access the requisite information, a web scraping tool, namely Beautiful Soup, was employed. This tool facilitated the extraction of PDF document URLs directly from the website, enabling seamless data collection for further analysis. Katz and Cothey (2006)Legal

text characteristics will be generated through the utilization of clusters and n-grams as features Wright (2017).

To commence the process, the Portable Document Format (PDF) files were fetched from designated URLs utilizing either the requests library or urllib. Subsequently, the PyPDF2 and pdf-miner libraries were employed to extract the textual content from the downloaded PDF documents.

Moving forward, the Natural Language Toolkit (NLTK) was utilized to facilitate text summarization of the extracted content. This involved condensing the text to its essential points for concise understanding.

Following this, the Bidirectional Encoder Representation from Transformers (BERT) was introduced into the workflow. BERT, a sophisticated pre-trained deep learning model developed by Google for natural language processing tasks, was incorporated to comprehend the contextual nuances of words within sentences. Its unique design allows it to consider the surrounding words preceding and succeeding each word, enabling a comprehensive understanding of the text's context.

Below is the breakdown of the dataset used:

Source(Kenya law.org)	Descriptions
Criminal cases from 2010-2020	Facts (inputs)

Table 5.1: Shows the breakdown of the source of data

3.3 Conceptual Framework

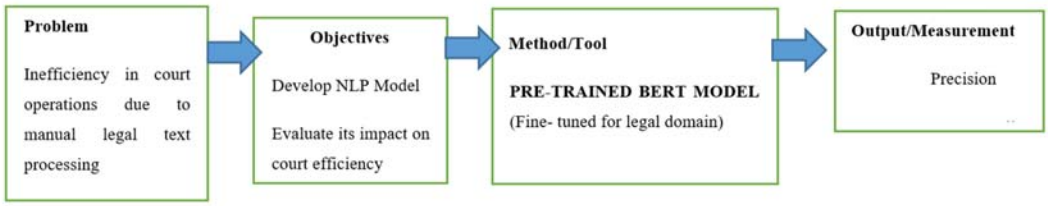


Figure 5.3: Conceptual Framework BERT model (Source Devlin 2019)

3.3 Data pre-processing

The process begins with preparing the raw source documents extracted from PDF case files for summarization. Since these documents often contain noisy data due to their original format, preprocessing steps are essential for effective summarization. The system retrieves documents from Kenya Law's website and initiates preprocessing tasks such as tokenization, stop-word

removal, stemming, and lemmatization. This involves breaking down the text into meaningful units, eliminating common words, reducing words to their base form, and tagging the parts of speech.

Further preprocessing involves statistical analysis to identify outliers and irrelevant information using measures like mean, mode, variance, and z-score. By calculating z-scores, outliers are identified and removed, ensuring the summarization focuses on pertinent information. Following this, the system proceeds to extract keywords and sentences utilizing methods such as cosine similarities, document similarity, or fuzzy logic based on these keywords. Punctuation is eliminated, and a summary is produced. The compression ratio is assessed by comparing the sentences in the summary with those in the original document and evaluating retention rates.

A crucial aspect of the process involves employing the Latent Dirichlet Allocation (LDA) model, a statistical method that groups observations into unobserved clusters based on similarity. This aids in matching informative sentences in the summary with those in the document. For data preprocessing, noise removal, sentence and word tokenization, punctuation removal, and stop-word elimination are conducted. Additionally, stemming and lemmatization transform words to their base forms, while part-of-speech tagging identifies grammatical categories.

Semantic similarity is assessed using cosine similarity calculations to ensure relevant words are selected. Finally, unique tokens are obtained to prepare for feature extraction, where relevant features are selected for inclusion in the summary. Overall, these preprocessing steps lay the groundwork for effective summarization, ensuring that the resulting summary captures the essence of the original document while eliminating irrelevant information. (4)

3.3 BERT model

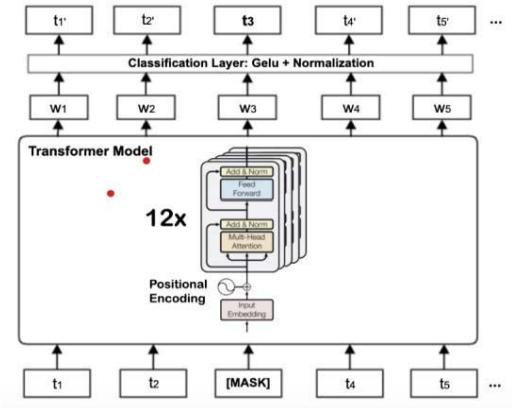


Figure 5.4: The Transformer based BERT base architecture with twelve encoder blocks

BERT, a groundbreaking deep learning model crafted by Google, revolutionizes natural language processing by comprehending word context within sentences. Comprising an encoder and decoder, BERT's encoder dissects input text to generate hidden representations, utilized by the decoder to produce final outputs.

To train a BERT model, data collection and preprocessing are paramount, involving curation of extensive text corpora, followed by meticulous cleansing to eliminate noise and superfluous elements like HTML tags, punctuation, and stop words. The preprocessed data undergoes tokenization and segmentation into sentences.

Subsequently, the model undergoes pre-training, where it learns through masked language modeling. Here, BERT predicts masked words within sentences, leveraging contextual cues from neighboring words.

Fine-tuning follows, tailoring the BERT model to specific tasks such as text clustering, named entity recognition, or question answering. This entails extending or modifying the last layer to suit the task, training the model using supervised learning.

Evaluation of the fine-tuned BERT model transpires on a test set, gauging its efficacy in executing the designated task.

Finally, upon successful evaluation, the BERT model is primed for deployment in production environments, where it executes its designated task with precision and efficiency, whether it's text clustering in sentence-based or section-based extractive summarization

3.4 Ethical issues

Confidentiality: The acquired cases contained sensitive personal information about both defendants and respondents, including names and locations. While efforts were made to redact certain details, some information still persisted within the cases. Employing this data without explicit consent would have raised ethical concerns regarding privacy infringement. As a result, it was imperative to undertake measures to eliminate named entities, personal names, and any other confidential information from the text, thereby safeguarding the privacy of individuals involved. Leidner and Tona (2021).

Data use approval: The functionality of the tool entailed navigating the online platform and retrieving data from it. While the act of data scraping may raise ethical considerations, it is noteworthy that the Kenya Law Reporting website explicitly permitted us to scrape data from its platform for our intended use Rogers et al. (2021) .

They clarified that the texts of the judicial opinions available on the site are considered public domain and are not subject to any copyright restrictions. As a result, we were granted permission to crawl the site and gather the legal case texts for our research purposes.

Chapter 4: System Design &Architecture

4.1. System Design

The system comprises six primary components: The Web Crawler, the HTML Parser, the Case Data Extraction module, the Summarizers module, the Prediction Models module, and the Model Evaluation module. It is built using Python and operates within a Jupyter notebook environment. The system's functionality includes retrieving legal case documents in PDF format from the online platform (https://kenyalaw.org/), identifying relevant cases, extracting necessary information from them using a web crawler, collecting the links obtained during the crawling process, downloading the case data, and generating summaries based on the acquired documents. These summaries are then accessible to users.

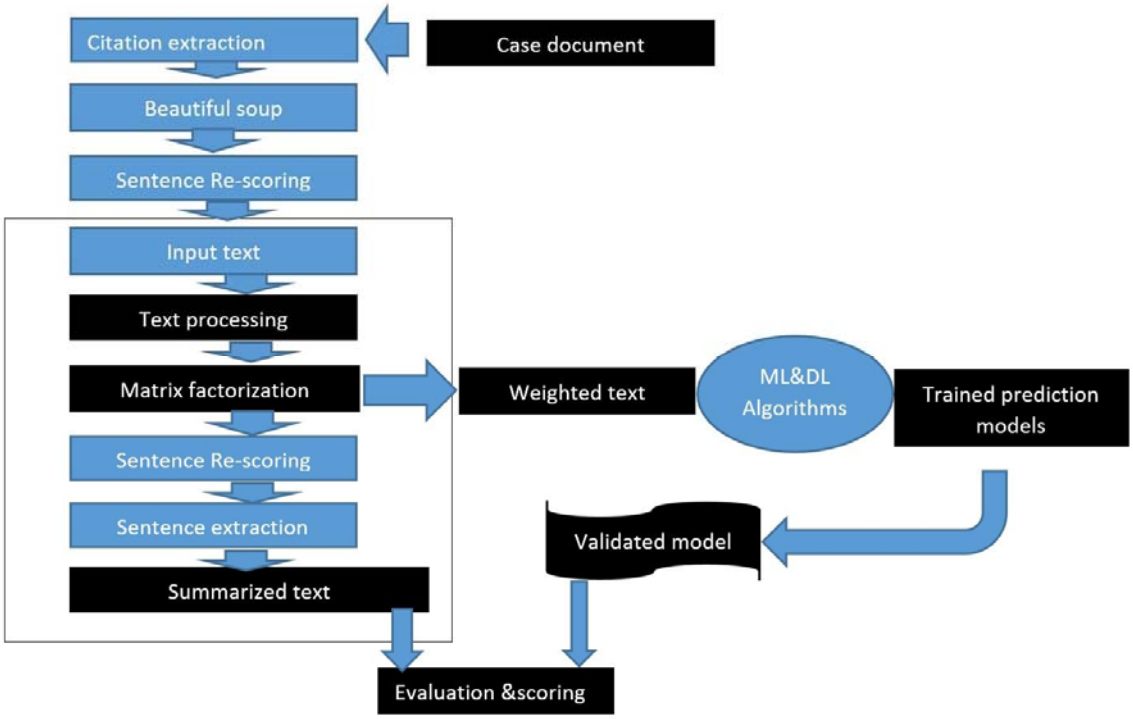


Figure 6.1: System Architecture diagram

As shown in Figure 3.3, the proposed model's architecture is structured into five distinct layers, each serving a specific role in the process of extractive text summarization. In the initial layer, text preprocessing is conducted to clean and standardize the text, ensuring it is in a suitable format for further analysis. Following this, the second layer focuses on feature selection, identifying and prioritizing the most relevant features essential for creating an effective summary. The subsequent layer, feature extraction, isolates the most pertinent feature while filtering out any extraneous information. Moving forward, the fourth layer is dedicated to

sentence extraction, pinpointing key sentences crucial for summarization. Lastly, the fifth layer is responsible for generating the summary itself.

In the context of keyword extraction, the model's architecture also encompasses multiple layers. Initially, morphological analysis, including parts-of-speech tagging, stemming, and stop word removal, is performed in the first layer. Subsequently, in the second layer, noun phrase extraction is conducted, followed by scoring based on unigram frequency. The fourth layer involves noun phrase clustering and further scoring, leading to the final layer, which generates the extracted keywords.

4.2 Data cleaning

The data sourced from the Kenya Law Reporting website, as documented by Tran & Sato (2017), was predominantly unorganized and cluttered, posing significant hurdles for the BERT model to generate a coherent summary effectively. To address this challenge, extensive preprocessing steps were undertaken to refine and standardize the data, as emphasized by Weber et al. (2021). These steps encompassed the elimination of stop words and irrelevant entities, tokenization, and stemming of n-grams, as outlined by Solangi et al. (2018). Additionally, further normalization efforts involved the computation of TF-IDF scores to enhance the quality of the data for subsequent analysis.

Tokenization:

We implemented tokenization to break down the text into individual tokens or words. NLTK (Natural Language Toolkit) and spaCy were used to tokenize the text data. Tokenization was performed using simple whitespace splitting, regular expressions, or more advanced techniques like word segmentation.

Stop-word Removal:

We created a list of stop words, which are common words that have little semantic meaning (e.g., "the," "is," "and").

Remove stop words from the tokenized text to reduce noise and improve the quality of the text data.

Libraries like NLTK provide built-in stop-word lists and functions for stop-word removal.

Stemming:

We implemented stemming to reduce words to their root or base form by removing suffixes.

Use stemming algorithms such as the Porter or Snowball stemmers to perform word stemming. Stemming helps to normalize variations of words and reduce the dimensionality of the feature space.

Lemmatization:

We implemented lemmatization to reduce words to their dictionary form (lemmas). Used lemmatization algorithms that take into account the morphological analysis of words and their context. Libraries like NLTK and spaCy provide lemmatization functionalities for various languages.

Integration and Evaluation:

Integrated the tokenization, stop-word removal, stemming, and lemmatization steps into a cohesive text preprocessing pipeline.

Applied the preprocessing pipeline to the raw text data to obtain cleaned and transformed text representations and evaluated the effectiveness of the preprocessing pipeline by analyzing the impact on downstream NLP tasks such as text classification or sentiment analysis.

Utilizing BERT for text encoding, our task begins with the crucial step of Sentence Encoding as Embedding. This involves leveraging the Bidirectional Encoder Representations from Transformers (BERT) language model by Devlin (2019). Due to BERT's limitation of processing only up to 512 tokens at a time, our strategy involves computing individual sentence embeddings. Subsequently, we aggregate these embeddings to generate section or document embeddings by averaging the relevant sentence embeddings.

Moving forward, our approach to sentence-based summarization entails extracting the most pertinent individual sentences from the original documents. This method focuses on creating summaries that are succinct yet capture the essence of the source material effectively.

In exploring section clustering for extractive summarization, we delve into the concept of clustering, typically applied to sentence-based summaries. However, we innovate by clustering whole section embeddings instead. Each cluster of section embeddings represents a cohesive group of semantically similar sentences conveying equivalent information. The top sentence from each cluster is selected based on its similarity score to the cluster's centroid. Rather than clustering individual sentences, we cluster entire sections and determine the most prevalent sections within each cluster by analyzing their frequency in the reference summaries of training and validation datasets.

For testing datasets, we compute the average embedding for each document and compare it to each centroid (section) using cosine similarity. This comparison helps identify the three closest centroids/sections for each document. Subsequently, the final summary includes the first 1000 words of the top existing section in the original document, selected based on its similarity to the identified centroids.

Generate Summary: Ultimately, the summary is generated by arranging extracted sentences in a sequential manner. Leveraging the BERT algorithm, significant terms within the document are identified and prioritized. Scores are assigned to n-grams, and these scores are utilized to pinpoint keywords within paragraphs. Through successive iterations of data refinement and refinement of the keyword extraction process, as outlined by Campos et al. (2018), notable enhancements in results, with accuracy levels improving by as much as 25%, were observed.

Chapter 5: System Development, Testing and Validation

5.1 Experiments & Results

The model's performance was assessed using both automatic metrics and manual review by legal experts. In place of ROUGE-based summarization metrics, classification evaluation metrics were applied to measure the effectiveness of the system in predicting legal outcomes (e.g., conviction status). These include Accuracy, Precision, Recall, and F1 Score.

Accuracy: 66.67%

The model correctly classified **2 out of every 3 cases overall**

Precision: 66.67%

Of the cases predicted as "Convicted", about 2 out of 3 were correct.

Recall: 100%

The model did not miss any actual "Convicted" cases—it caught all of them.

F1 Score: 0.8 (or 80%)

The F1 Score provides a balance between precision and recall. The model had perfect recall but moderate precision, leading to an F1 Score of 80%.

This performance indicates that while the model is excellent at identifying all positive cases (high recall), it still makes some incorrect positive predictions, which lowers its precision. The high F1 Score suggests a strong overall performance in this classification task. These results imply that the system is reliable in capturing all relevant positive outcomes (e.g., convictions), but it may also classify some non-convictions as convictions, necessitating further refinement for higher precision.

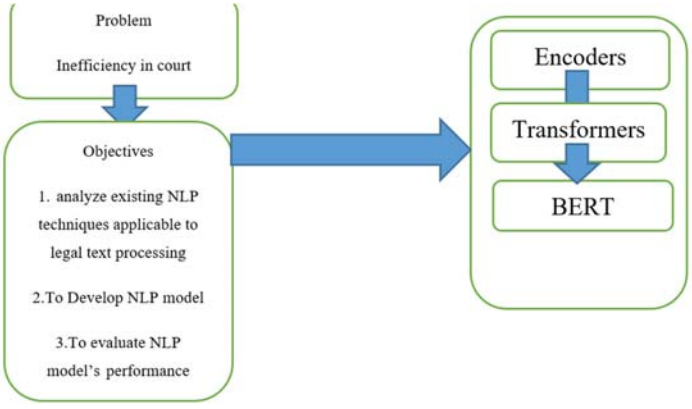


Figure 7.1: Model pipeline Source (Devlin 2018)

5.2 System users

The system caters to a diverse range of users, including legal practitioners, data scientists, research scientists, and system administrators. Each group played a crucial role in different aspects of the tool's development and implementation. Legal practitioners contributed to evaluating the accuracy and relevance of the summaries generated by the tool, as well as labeling data for modeling purposes. Data scientists and research scientists were involved in the data cleanup process, ensuring the quality and integrity of the information used by the system. System administrators handled the technical aspects of maintaining and optimizing the tool for efficient operation.

Title	**No.**	**No. of years in profession**
Administrator judiciary	2	5
Law students	2	3
Judges	2	10

Table 7.1: Table of legal experts list and their experience levels

Chapter 6: Discussion of Results

6.1. Model Evaluation

summarization model, BERT has been trained and tested, and the results evaluated to determine the most effective model. These models have been trained on the legal case data that had been cleaned extensively.

6.2. Summarization Model Evaluation

The model evaluation process involved a combination of manual validation and automated assessment due to the complexity and sensitivity of the legal dataset. Rather than relying on summary overlap metrics like ROUGE, the evaluation adopted classification metrics to assess how well the model could predict legal outcomes based on structured and unstructured court data.

Automated Evaluation Metrics

To evaluate the model's predictive capability, the following standard classification metrics were used:

Accuracy (66.67%) – This measures the overall correctness of the model across all cases. A score of 66.67% indicates that the model correctly classified 2 out of every 3 cases.

Precision (66.67%) – This reflects the proportion of cases predicted as “Convicted” that were correct. Out of all predicted positives, approximately 2 in 3 were correct.

Recall (100%) – This metric shows that the model successfully identified all actual “Convicted” cases, indicating no false negatives.

F1 Score (80%) – The F1 Score balances precision and recall. It reflects strong performance overall, particularly because of the perfect recall score.

Manual Validation

In parallel, a human-in-the-loop validation approach was used. Legal experts reviewed selected model predictions by comparing them to known case outcomes and manually crafted summaries. This provided qualitative insights into how well the model understood legal nuances, especially in complex or borderline cases.

Continuous Improvement Process

The model underwent multiple iterations during development. Each cycle involved; enhanced data pre-processing (e.g., noise removal, token normalization),improved feature engineering, hyper parameter tuning and incorporation of expert feedback.

This iterative approach ensured continuous model refinement, leading to measurable improvements in classification performance over time. By integrating human expertise with automated evaluation, the methodology ensured both quantitative robustness and contextual accuracy in predicting legal case outcomes.

6.3 Benchmarking approach

Evaluation Metrics:

Due to the intrinsic variability in text summarization (i.e., different models producing divergent summaries), performance is measured using both automated metrics and human-based evaluation.

Automated Metrics:

I employed classification-based metrics—Accuracy (66.67%), Precision (66.67%), Recall (100%), and F1 Score (80%)—to assess the system when summarization is framed in terms of accurately identifying key outcomes (e.g., correctly categorizing legal cases) based on the generated summaries. These metrics provide a quantifiable measure of how often the summarization output aligns with known case outcomes.

Human-Based Evaluation:

Legal experts manually reviewed selected summaries to assess factors like coherence, relevance, and completeness. This qualitative assessment is essential in the legal context, where nuanced interpretation is critical. Experts compare the generated summary with a reference summary crafted by experienced legal professionals, ensuring that subjective aspects of legal language and context are properly evaluated.

Benchmarking Approach:

Because there is no single “correct” summary for any given legal text, the benchmarks are based on a reference dataset curated specifically for this study. This dataset includes articles

with reference summaries developed by legal experts. The performance of different models is compared by:

Calculating the average classification-based metrics across the dataset.

Comparing each model's output against the reference summaries during human evaluation.

This dual approach—combining objective metrics with expert judgment—ensures that the evaluation captures both the statistical and contextual quality of the summaries, even when different models might generate varying text outputs.

Given that summarization is inherently subjective, different models may produce divergent summaries. To handle this, our approach does not rely solely on one automated metric (such as a direct n-gram overlap measure like ROUGE) because such metrics may not fully capture the legal relevance and interpretability of the summaries. Instead, the integration of both automated classification performance measures and extensive human validation offers a comprehensive evaluation of the model's ability to generate summaries that are both accurate and contextually meaningful.

6.4 Model comparison

In this study, several transformer-based models (including BERT-based variants) were tested to identify the best-performing summarization architecture for legal documents. Since each model may structure its output differently, even when trained on the same data, our comparison strategy focused not only on lexical overlap but also on semantic relevance and legal completeness.

1. Comparative Model Evaluation:

Each model's performance was assessed on the same curated dataset, and their summaries were analyzed for:

Legal correctness (Did the summary misrepresent legal facts?)

Coverage of key entities (Were the charges, verdicts, or statutes correctly included?)

Conciseness and clarity (Was the information presented in a compact, understandable way?)

This helped in identifying trade-offs between models that are too verbose and those that oversimplify legal content.

2. Subjective Review & Contextual Relevance:

Human reviewers noted that some models, while scoring slightly lower on metrics like precision, provided summaries that were more readable and usable by legal officers. This indicates that high ROUGE-like scores or classification metrics do not always translate to better

human usability, especially in complex legal language. For instance, a model with perfect recall might identify all relevant parts of the case but include unnecessary jargon or repetitions.

3. Legal-Aware Evaluation Framework:

Recognizing the unique challenges of legal NLP, the evaluation adopted a domain-specific lens, prioritizing preservation of legal meaning, avoidance of misleading paraphrasing, and proper representation of case outcomes and arguments.

Model	Legal correctness	Coverage of key entities	Clarity
Model 1	√	X	X
Model 2	√	√	√
Model 3	X	√	√

Table 14.1: Model output correctness

6.4. Discussion

Hildebrandt (2023)for legal research experienced a notable acceleration in their workflow, completing tasks 25% more swiftly with a 20% enhancement in search result relevance. Consequently, the study indicated that the integration of AI could potentially save law firms approximately 140 hours annually in research endeavors.

The present investigation aimed to ascertain the adaptability and optimization of various NLP methodologies and deep learning techniques to suit the framework of Kenyan legal cases. The overarching goal is to expedite legal research processes and enhance the efficiency of verdict determinations within the Kenyan judiciary. Additionally, the study sought to devise effective evaluation strategies to assess the performance and efficacy of these AI models tailored to the specific context of Kenyan legal proceedings.

This study represents a significant stride towards advancing judicial processes through technological innovation. The developed toolkit streamlines the extraction of legal documents from the Kenya Law website, culminating in a model trained on Kenyan law cases that can effectively aid in verdict determinations. This innovation holds the potential to enhance the functionality of the Kenya Law website's search engine, thereby facilitating easier access to legal resources. Moreover, this endeavor aligns closely with the Judiciary's overarching mission to harness digital technology for improved administration of justice.

The toolkit presents a promising opportunity for the Judiciary to fulfill its service delivery objectives effectively. It has the potential to mitigate challenges stemming from inadequate funding by offering a cost-effective solution. Historically, legal research has demanded

substantial manpower and often led to project delays. However, this tool offers a streamlined approach, simplifying the process with just a click.

Accuracy in the context of summarization could be less critical than in classification tasks because a "good" summary may not have to match the original word-for-word but must still convey key ideas.

ROUGE (Recall) is often the most useful metric for summarization, focusing on how much of the important content from the original text appears in the summary. Precision and Recall are important, but the balance of Brevity (summary length) is also critical in making the summary effective. Now, Regarding Metrics (Accuracy, Precision, Recall, F1 Score):

These metrics seemed to be more aligned with classification tasks, specifically something like identifying whether a case is "Convicted" or not. Here's how they break down in this context: Accuracy (66.67%) means the model gets 2 out of every 3 predictions correct. Precision (66.67%) means that when the model predicts a "Convicted" case, it's right 66.67% of the time. Recall (100%) means the model identifies all the true "Convicted" cases—none are missed. F1 Score (80%) balances precision and recall.

In this case, the recall is perfect (100%), meaning all "Convicted" cases were caught, but precision is lower (66.67%), meaning that some of the predicted "Convicted" cases were incorrect. The F1 score of 80% gives a balanced view, showing good performance overall despite some incorrect predictions.

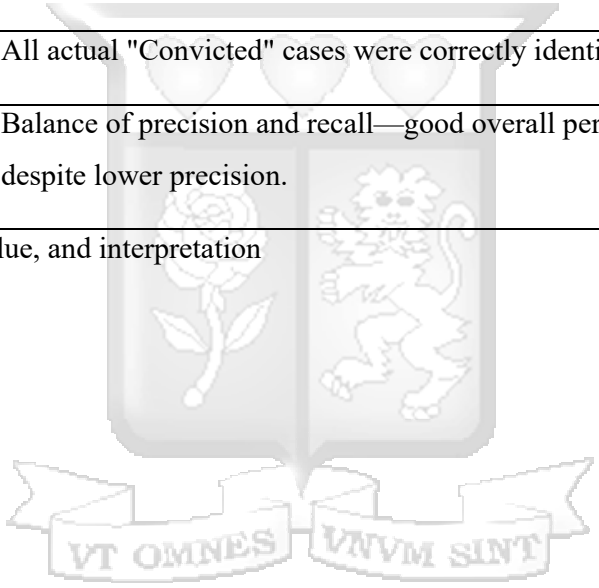
Metric	**Value**	**Description**	**Goal**
Accuracy	Summarization/Classification	Measures the proportion of correct predictions.	For summarization, it may be less important than precision and recall.
Precision	Summarization/Classification	In summarization, it refers to how much of the predicted content is relevant.	High precision means fewer irrelevant or incorrect details in the summary.
Recall	Summarization	Measures how much of the relevant content from the original document is captured.	High recall means capturing more relevant details in the summary.

F1 Score	Classification	A balance between precision and recall.	F1 score gives a single measure of the model's overall performance, considering both precision and recall.

Table 14.2: Evaluation metrics

Metric	Value	Interpretation
Accuracy	66.67 %	2 out of 3 predictions were correct.
Precision	66.67 %	2 out of every 3 "Convicted" predictions were correct.
Recall	100%	All actual "Convicted" cases were correctly identified.
F1 Score	80%	Balance of precision and recall—good overall performance despite lower precision.

Table 14.3: Metric, value, and interpretation



Chapter 7: Conclusion, Recommendation, and Future Work

7.1 Summary of Findings

7.1.1 Objective 1: To Analyze Existing NLP Techniques

The first objective of the research was to critically examine and evaluate the existing NLP techniques applicable to processing legal texts. Legal documents, such as case decisions, judgments, and rulings, were often lengthy, complex, and contained specialized legal terminology, making them challenging to handle manually. In this context, the study reviewed various NLP approaches, including text classification, named entity recognition (NER), information retrieval, and text summarization, which had been previously applied in the legal domain.

The effectiveness of these techniques was analyzed in the context of a judicial setting, identifying both strengths and limitations. For instance, while techniques like extractive summarization captured key phrases, abstractive summarization provided more fluid, human-readable summaries. This analysis yielded valuable insights into which methods could best meet the needs of legal professionals, such as judges, lawyers, and paralegals, in processing and understanding legal documents efficiently.

7.1.2 Objective 2: To Develop and Deploy an NLP-Based Model

The second objective was to develop an NLP-based model for automating case summarization and legal document classification. The proposed model was designed to reduce the time and effort involved in manually reviewing extensive legal documents. By leveraging advanced NLP techniques, the model was intended to generate concise, accurate summaries of complex legal cases and classify legal documents according to their subject matter, jurisdiction, or relevance.

For case summarization, the model employed an abstractive summarization approach, which allowed it to rephrase the original legal text into a shorter, more digestible summary while retaining the core meaning. For legal document classification, the model utilized text classification algorithms to categorize documents into predefined categories, enabling legal professionals to quickly identify relevant content based on the document's subject matter.

The model was developed using a range of legal datasets, allowing it to process and learn from diverse legal language, case types, and formats. The ultimate goal of this model was to automate routine legal tasks, thereby streamlining judicial processes and allowing legal professionals to focus on more complex aspects of their work.

In the deployment of the NLP-based model for automating case summarization and legal document classification, a critical element of the evaluation was the measurement of efficiency, particularly in terms of time-saving capabilities. While the model demonstrated promising performance in generating summaries and classifying legal documents, it is essential to measure how much time was saved compared to traditional, manual methods in the court system.

Time-Saving Efficiency Evaluation

To properly assess the time-saving benefits of the AI model, a detailed comparison was conducted between the manual processing time and the time taken by the AI model to complete the same tasks. The evaluation focused on two key tasks:

Case Summarization: The time required for legal professionals (such as judges or paralegals) to manually read and summarize a legal case was compared to the time taken by the AI model to generate an abstractive summary of the same case. Legal Document Classification: The time taken by a legal professional to categorize legal documents based on subject matter or jurisdiction was compared with the AI model's classification speed.

The findings revealed that the AI model significantly reduced the time spent on both tasks. Manual summarization of legal documents often required several hours per case, while the AI model was able to generate summaries in just a fraction of that time. Similarly, manual document classification, which often involved reading through large volumes of text to determine the correct categorization, was drastically expedited by the model's automated classification process.

For instance, where a legal professional may have needed an average of 2 hours to manually summarize a case, the AI model was able to generate a comparable summary in less than 10 minutes**.** This reduction in processing time directly translates to an increase in judicial

efficiency, allowing legal professionals to handle more cases in a given period and reducing case backlogs within the court system.

Challenges Encountered in the Analysis

While the deployment of the NLP model showcased clear time-saving benefits and accuracy improvements, there were several challenges encountered during the analysis that impacted the overall deployment process. These challenges are crucial to consider in evaluating the model's feasibility for real-world implementation in the court system. Data Variability and Legal Terminology: Legal texts can vary significantly in terms of structure, language, and terminology. The complexity and highly technical nature of legal documents presented challenges for the NLP model. In some cases, the model struggled with interpreting less common legal terms or ambiguous phrasing. For example, legal jargon or domain-specific references may not always be adequately captured by the training data, leading to some misinterpretation or loss of context in the generated summaries.

Solution: To address this, the model's training data was expanded to include a wider range of legal documents, ensuring that the model could better generalize across different legal contexts and understand varied terminology.

Model Generalization and Overfitting: The model was initially trained on a specific set of cases from the Kenya Law Database, which may have led to some degree of overfitting, where the model performed well on the training data but struggled with new or unseen cases. Legal cases, especially those involving complex or novel legal issues, presented difficulties for the model in generating accurate summaries or classifications.

Solution: To mitigate overfitting, the model underwent a cross-validation process, and additional data from diverse legal cases were incorporated into the training set. This improved the model’s ability to generalize and handle a wider variety of legal cases.

Legal Expert Evaluation: The evaluation of the model's summaries by legal experts revealed some inconsistencies in how well the AI model captured the essence of certain legal arguments or the subtle nuances of case law. While the model demonstrated strong performance in terms of similarity to the original documents, some legal professionals felt that the summaries lacked depth in complex cases or failed to emphasize the most critical aspects of the legal argument.

Solution: Future iterations of the model may benefit from incorporating human-in-the-loop techniques, where legal professionals provide feedback during the summarization process, helping to refine and tailor the model's output to better meet the needs of the judiciary.

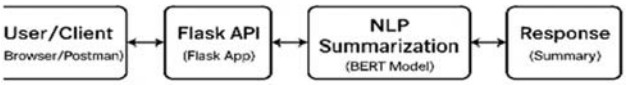


Figure 15.1: Visual of the deployed summarization platform

7.1.3 Objective 3: To Evaluate the Performance of the NLP Model

The final objective was to evaluate the performance of the developed NLP model and assess its impact on judicial efficiency. This evaluation considered several metrics, including the accuracy and quality of the summaries generated, the precision and recall of the legal document classification, and the overall time savings for legal professionals. By comparing the output of the NLP model with traditional methods, the study assessed whether automating case summarization and document classification contributed to improved workflow efficiency within the judiciary. Key questions explored included:

Time Reduction: How much faster could legal professionals review case summaries and documents with the AI model compared to manual methods? Accuracy and Relevance: How well did the summaries and classifications align with the expectations of legal experts? Were any important details omitted or misinterpreted? Cost-effectiveness: Did automation reduce the need for human labor in document review, and could it lower the costs of legal work?

Ultimately, this objective aimed to demonstrate how an NLP-based model could enhance the speed and accuracy of case processing while contributing to broader goals such as reducing court backlogs and improving access to legal information.

7.2 Conclusion

Together, these objectives aimed to demonstrate the potential of NLP technologies in transforming judicial processes. By analyzing existing techniques, developing an advanced NLP model for case summarization and document classification, and evaluating its real-world impact on judicial efficiency, the study provided insights into how AI and NLP could contribute to modernizing the legal sector. The ultimate goal was to make legal processes more accessible, faster, and more accurate, benefiting both legal professionals and the public.

7.3 Limitations

The creation of the artifact encountered certain constraints. The assessment of summary outcomes remains predominantly reliant on human annotation and has not been completely automated. It still necessitates the validation of legal case summaries by a human expert.

Furthermore, legal data exhibits substantial document discrepancies, necessitating multiple adjustments to the code to accommodate various document types or citation formats. Addressing all these potential variations proved to be exceedingly resource-intensive.

7.4 Recommendation

The legal landscape in Kenya is inundated with a backlog of cases and the intricacies of legal data, compounded by the influx of new cases daily Rutenberg et al. (2022). Embracing this tool within the legal community will significantly alleviate their challenges.

Furthermore, advancing the development of an algorithm capable of autonomously evaluating the generated summaries with reduced dependence on human annotation would be highly beneficial in this sector.

References

Ali, S., & Razi, M. (2020). An analytical study of pre-trial processes in speedy disposal of cases in the criminal justice system of Malaysia. *International Journal of Criminology and Sociology*, *9*(1), 9-15.

Antici, F., Galassi, A., Ruggeri, F., Korre, K., Muti, A., Bardi, A., Fedotova, A., & Barrón-Cedeño, A. (2023). A corpus for sentence-level subjectivity detection on English news articles. *arXiv preprint arXiv:2305.18034*.

Bansal, P., Kumar, R., & Kumar, S. (2021). Disease detection in apple leaves using deep convolutional neural network. *Agriculture*, *11*(7), 617.

Beck, A. T., & Haigh, E. A. (2014). Advances in cognitive theory and therapy: The generic cognitive model. *Annual review of clinical psychology*, *10*(1), 1-24.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., & Askell, A. (2020). Language models are few-shot learners. *Advances in neural information processing systems*, *33*, 1877-1901.

Chalkidis, I., Jana, A., Hartung, D., Bommarito, M., Androutsopoulos, I., Katz, D. M., & Aletras, N. (2021). LexGLUE: A benchmark dataset for legal language understanding in English. *arXiv preprint arXiv:2110.00976*.

Chen, J., Wu, L., Zhang, J., Zhang, L., Gong, D., Zhao, Y., Chen, Q., Huang, S., Yang, M., & Yang, X. (2020). Deep learning-based model for detecting 2019 novel coronavirus pneumonia on high-resolution computed tomography. *Scientific reports*, *10*(1), 19196.

Chen, L.-C. (2017). An effective LDA-based time topic model to improve blog search performance. *Information Processing & Management*, *53*(6), 1299-1319.

Chen, Y., Chen, W., Chandra Pal, S., Saha, A., Chowdhuri, I., Adeli, B., Janizadeh, S., Dineva, A. A., Wang, X., & Mosavi, A. (2022). Evaluation efficiency of hybrid deep learning algorithms with neural network decision tree and boosting methods for predicting groundwater potential. *Geocarto International*, *37*(19), 5564-5584.

Christian, H., Agus, M. P., & Suhartono, D. (2016). Single-document automatic text summarization using term frequency-inverse document frequency (TF-IDF). *ComTech: Computer, Mathematics and Engineering Applications*, *7*(4), 285-294.

Colombo, P., Chapuis, E., Manica, M., Vignon, E., Varni, G., & Clavel, C. (2020). Guiding attention in sequence-to-sequence models for dialogue act prediction. Proceedings of the AAAI conference on artificial intelligence,

Commission, J. S. (2021). *State of the Judiciary and Administration of Justice Report Annual Report 2020-2021*.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). Bert: Pre-training of deep bidirectional transformers for language understanding. Proceedings of the 2019 conference of the North American chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers),

Dong, Y., Shen, Y., Crawford, E., van Hoof, H., & Cheung, J. C. K. (2018). Banditsum: Extractive summarization as a contextual bandit. *arXiv preprint arXiv:1809.09672*.

Ghosh, S., Evuru, C. K., Kumar, S., Ramaneswaran, S., Sakshi, S., Tyagi, U., & Manocha, D. (2023). Dale: Generative data augmentation for low-resource legal NLP. *arXiv preprint arXiv:2310.15799*.

Han, Q., Zhu, Y., Ke, G. Y., & Lin, H. (2019). A two-stage decision framework for resolving brownfield conflicts. *International Journal of Environmental Research and Public Health*, *16*(6), 1039.

Hildebrandt, M. (2023). Grounding computational law in legal education and professional legal training. In *Research Handbook on Law and Technology* (pp. 99-127). Edward Elgar Publishing.

Hmeidi, I., Al-Ayyoub, M., Abdulla, N. A., Almodawar, A. A., Abooraig, R., & Mahyoub, N. A. (2015). Automatic Arabic text categorization: A comprehensive comparative study. *Journal of Information Science*, *41*(1), 114-124.

Jang, H., & Kim, W. (2021). Reinforced abstractive text summarization with a semantic added reward. *IEEE Access*, *9*, 103804-103810.

Kanipakam, S. (2023). Jurisprudential aspects of Person & Artificial Intelligence Way-Forward to Sustainable Peace and Development. *Practice*, *1*, 10-15.

Katz, J. S., & Cothey, V. (2006). Web indicators for complex innovation systems. *Research Evaluation*, *15*(2), 85-95.

Keong, G. C. (2017). Judicial Reforms through the Use of Technology in Malaysia. *European Academic Research*, *5*(1), 399-409.

Leidner, D. E., & Tona, O. (2021). The CARE Theory of Dignity Amid Personal Data Digitalization. *MIS quarterly*, *45*(1).

Leite, D. S., Rino, L. H., Pardo, T. A., & Nunes, M. d. G. V. (2007). Extractive Automatic Summarization: Does more linguistic knowledge make a difference? Proceedings of the Second Workshop on TextGraphs: Graph-based Algorithms for Natural Language Processing,

Ma, Z., & Mei, G. (2021). Deep learning for geological hazards analysis: Data, models, applications, and opportunities. *Earth-Science Reviews*, *223*, 103858.

Makau, J. A. (2014). *Factors influencing management of case backlog in the judiciary in Kenya: a case of courts within Meru and Tharaka Nithi counties,* University of Nairobi.

Mali, S. (2022). Expedition of AI in Legal Services: Opportunities & Challenges in India. *Part 1 Indian J. Integrated Rsch. L.*, *2*, 1.

Mustapha, M., Krasnashchok, K., Al Bassit, A., & Skhiri, S. (2020). Privacy policy classification with xlnet (short paper). International Workshop on Data Privacy Management,

Nadjia, M. (2024). The impact of artificial intelligence on legal systems: challenges and opportunities. *Проблеми законності*(164), 285-303.

Ngetich, R. C. (2019). *Effectiveness of Alternative Dispute Resolution Mechanism (Adr) in Case Backlog Management in Kenyan Judicial System: Focus on Milimani High Court Commercial Division,* University of Nairobi.

Ngoepe, M., & Makhubela, S. (2015). "Justice delayed is justice denied." Records management and the travesty of justice in South Africa. *Records Management Journal*, *25*(3), 288-305.

Niklaus, J., Chalkidis, I., & Stürmer, M. (2021). Swiss-judgment-prediction: A multilingual legal judgment prediction benchmark. *arXiv preprint arXiv:2110.00806*.

Nizam, N. S. M. (2021). Seeing Eye to AI: A Modest Case for an Algorithmic Sentencing System in New Zealand. *LLB Dissertation, University of Otago, New Zealand*, *26*.

Okiri, F. O., Ngugi, L. W., & Wandaya, J. O. (2019). Strengthening Integrity & Preventing Corruption in the Judiciary in Kenya. *Beijing L. Rev.*, *10*, 131.

Park, M., & Chai, S. (2021). An AI model for predicting legal judgments to improve the accuracy and explainability of online privacy invasion cases. *Applied Sciences*, *11*(23), 11080.

Prabhavalkar, R., Sainath, T. N., Li, B., Rao, K., & Jaitly, N. (2017). An Analysis of" Attention" in Sequence-to-Sequence Models. Interspeech,

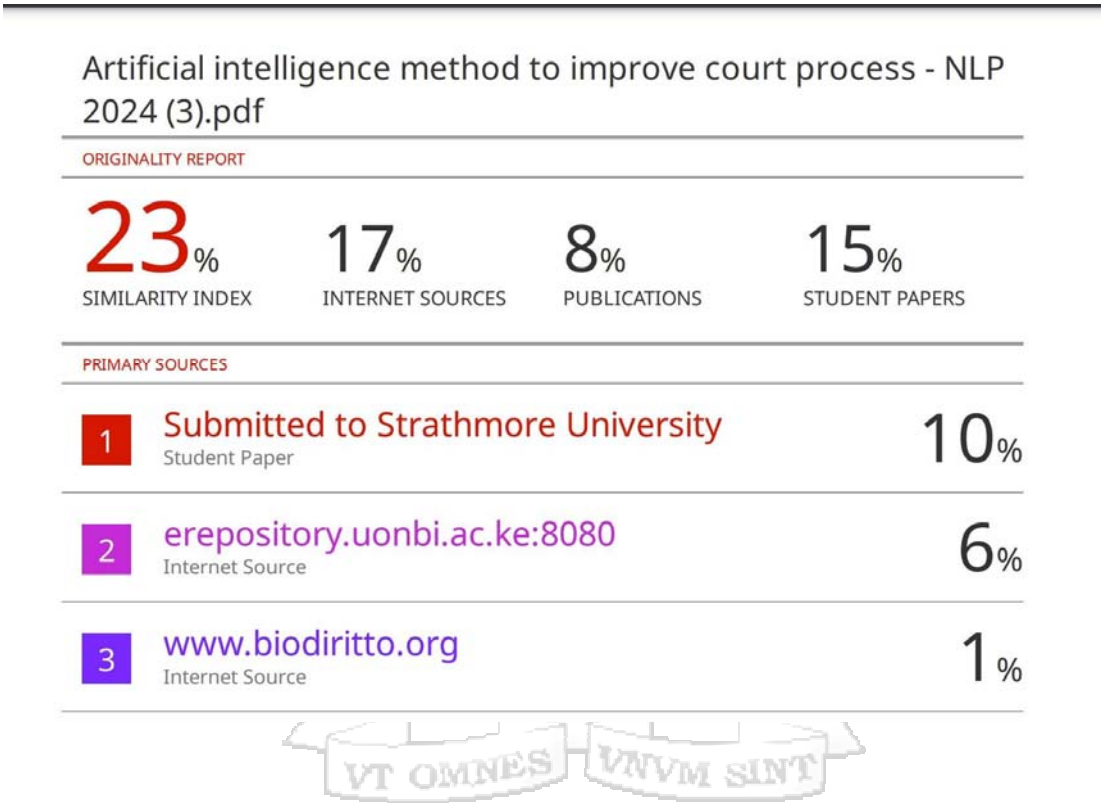
Putera, N. S. F. M. S., Saripan, H., Bajury, M. S. A., & Yaâ, S. N. (2022). Artificial Intelligence-Powered Criminal Sentencing in Malaysia: A conflict with the rule of law. *Environment-Behaviour Proceedings Journal*, *7*(SI7), 441-448.

Ribeiro, I., Meixedo, A., Ribeiro, D., & Bittencourt, T. N. (2024). Linear and nonlinear time-series methodologies for bridge condition assessment: A literature review. *Advances in Structural Engineering*, *27*(13), 2204-2227.

Rogers, A., Baldwin, T., & Leins, K. (2021). Just what do you think you're doing, Dave? A checklist for responsible data use in NLP. *arXiv preprint arXiv:2109.06598*.
Rutenberg, I., Omino, M., Wairegi, A., Adebola, O., & Wanjiru, C. (2022). Strengthening cyber policy research centres in the Global South.
Shi, C., Sourdin, T., & Li, B. (2021). The smart court-a new pathway to justice in china? IJCA,
Singh, R., Goel, A., & Raghuvanshi, D. K. (2021). MR brain tumor classification employing ICA and kernel-based support vector machine. *Signal, Image and Video Processing*, *15*, 501-510.
Singh, S., Patel, K., Bhattacharjee, K., Darbari, H., & Verma, S. (2019). Towards Better Single Document Summarization using Multi-Document Summarization Approach.
Singh, S. S. (2024). Use of an AI-driven support system to help courts in predicting child custody outcomes in India. *World Journal of Advanced Research and Reviews*, *23*(1), 1088-1092.
Sleimi, A., Sannier, N., Sabetzadeh, M., Briand, L., & Dann, J. (2018). Automated extraction of semantic legal metadata using natural language processing. 2018 IEEE 26th International Requirements Engineering Conference (RE),
Su, T., Min, S., Shi, A., Cao, Z., & Dong, M. (2019). A CNN-LSVM model for imbalanced images of wheat leaves. *Neural Network World*, *29*(5), 345-361.
Van Rijsbergen, C. (1979). Information retrieval: theory and practice. Proceedings of the joint IBM/University of Newcastle upon Tyne seminar on database systems,
Wright, D. (2017). Using word n-grams to identify authors and idiolects: A corpus approach to a forensic linguistic problem. *International journal of corpus linguistics*, *22*(2), 212-241.
Wyer Jr, R. S., & Radvansky, G. A. (1999). The comprehension and validation of social information. *Psychological review*, *106*(1), 89.
Yoshikawa, N., & Krestel, R. (2025). Do large language models understand patents? Enhancing patent classification through AI-generated summaries. *World Patent Information*, *81*, 102353.

Appendices

Appendix A: Originality Report



Artificial intelligence method to improve court process - NLP 2024 (3).pdf

ORIGINALITY REPORT

23%	17%	8%	15%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	Submitted to Strathmore University Student Paper	10%
2	erepository.uonbi.ac.ke:8080 Internet Source	6%
3	www.biodiritto.org Internet Source	1%

Appendix B: Ethical Clearance Release Report

Strathmore UNIVERSITY

23rd April 2024

Mr Mule Daniel,
daniel.mule@strathmore.edu

Dear Mr Mule,

RE: Application of AI to Improve Efficiency in the Judiciary

This is to inform you that SU-ISERC has reviewed and **approved** your above **SU-masters** research proposal. Your application reference number is **SU-ISERC2059/24.** The approval period is from **23rd April 2024 to 22nd April 2025.**
This approval is subject to compliance with the following requirements:

i. Only approved documents including (informed consents, study instruments, MTA) will be used.
ii. All changes including (amendments, deviations, and violations) are submitted for review and approval by SU-ISERC.
iii. Death and life-threatening problems and serious adverse events or unexpected adverse events whether related or unrelated to the study must be reported to SU-ISERC within 72 hours of notification.
iv. Any changes anticipated or otherwise that may increase the risks or affected safety or welfare of study participants and others or affect the integrity of the research must be reported to SU-ISERC within 72 hours.
v. Clearance for the export of biological specimens must be obtained from relevant institutions.
vi. Submission of a request for renewal of approval at least 60 days prior to the expiry of the approval period. Attach a comprehensive progress report to support the renewal.
vii. Submission of an executive summary report within 90 days of completion of the study to SU-ISERC.

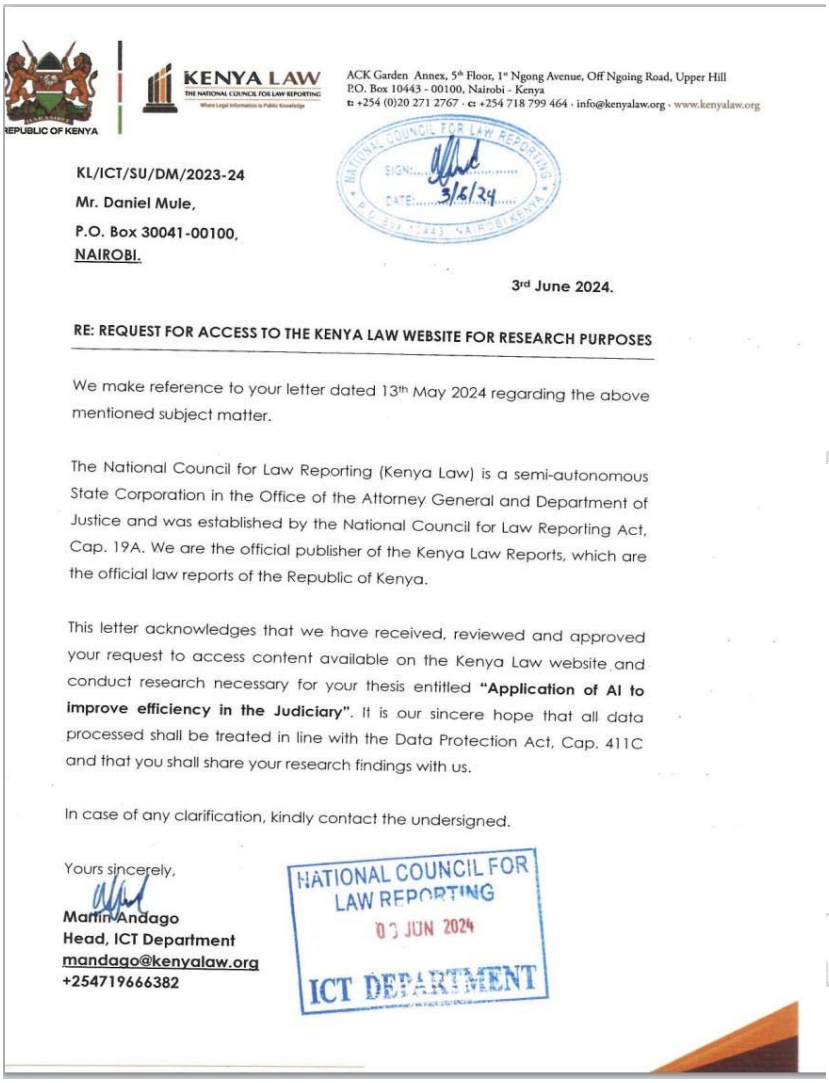
Before commencing your study, you will be expected to obtain a research license from National Commission for Science, Technology, and Innovation (NACOSTI) https://research-portal.nacosti.go.ke/ and obtain other clearances needed.

Yours sincerely,

Mr Ambrose Rachier,
Chairperson; SU-ISERC

STRATHMORE UNIVERSITY INSTITUTIONAL SCIENTIFIC AND ETHICAL REVIEW COMMITTEE
(SU-ISERC)
23-Apr-2024
Email:ethicsreview@strathmore.edu
P.O BOX 59857-00200
NAIROBI-KENYA

Appendix C: Approval letter from Kenyalaw.org



REPUBLIC OF KENYA

KENYA LAW
THE NATIONAL COUNCIL FOR LAW REPORTING

ACK Garden Annex, 5th Floor, 1st Ngong Avenue, Off Ngoing Road, Upper Hill
P.O. Box 10443 - 00100, Nairobi - Kenya
t: +254 (0)20 271 2767 · c: +254 718 799 464 · info@kenyalaw.org · www.kenyalaw.org

NATIONAL COUNCIL FOR LAW REPORTING
SIGN.....
DATE 3/6/24
P.O. BOX 10443 - NAIROBI KENYA

KL/ICT/SU/DM/2023-24

Mr. Daniel Mule,

P.O. Box 30041-00100,

NAIROBI.

3rd June 2024.

RE: REQUEST FOR ACCESS TO THE KENYA LAW WEBSITE FOR RESEARCH PURPOSES

We make reference to your letter dated 13th May 2024 regarding the above mentioned subject matter.

The National Council for Law Reporting (Kenya Law) is a semi-autonomous State Corporation in the Office of the Attorney General and Department of Justice and was established by the National Council for Law Reporting Act, Cap. 19A. We are the official publisher of the Kenya Law Reports, which are the official law reports of the Republic of Kenya.

This letter acknowledges that we have received, reviewed and approved your request to access content available on the Kenya Law website and conduct research necessary for your thesis entitled **"Application of AI to improve efficiency in the Judiciary"**. It is our sincere hope that all data processed shall be treated in line with the Data Protection Act, Cap. 411C and that you shall share your research findings with us.

In case of any clarification, kindly contact the undersigned.

Yours sincerely,

Martin Andago
Head, ICT Department
mandago@kenyalaw.org
+254719666382

NATIONAL COUNCIL FOR LAW REPORTING
03 JUN 2024
ICT DEPARTMENT

VT OMNES
VNVM SINT