UNIVERSITY OF NAIROBI

FACULTY OF SCIENCE AND TECHNOLOGY

DEPARTMENT OF COMPUTING AND INFORMATICS

USING NATURAL LANGUAGE PROCESSING FOR CASE BRIEF GENERATION AND VERDICT SUPPORT IN THE KENYAN COURT SYSTEM

By

SHARON OBANDA

(P52/38830/2020)

SUPERVISOR

CHRISTOPHER A. MOTURI

A project report submitted in partial fulfillment of the requirements for the award of Master of Science in Computational Intelligence of the University of Nairobi.

June 2022

DECLARATION

This project is my original work and to the best of my knowledge this work has not been submitted for any other award in any University.

Date: ***Jun 6, 2022***

SHARON OBANDA

(P52/38830/2020)

This project report has been submitted in partial fulfillment of the requirements of the Master of Science in Computational Intelligence of the University of Nairobi with my approval as the University supervisor.

Date: ***June 6, 2022***

Christopher A. Moturi

Department of Computing and Informatics

ABSTRACT

The National Council for Law Reporting publishes judgements in an online searchable database, enabling large-scale machine learning and statistical analysis in the legal domain. This is the culmination of the transformation that had been going on in the Judiciary inspired by the new requirements of public service delivery under the 2010 Constitution and the increased awareness and demand for legal information by the citizens. The Kenyan Judiciary is now continually seeking to apply creative, innovative, appropriate and integrated technological solutions that enable efficient service delivery. This research focuses on how to integrate Natural Language Processing (NLP) systems and Artificial Intelligence (AI) in document review, legal writing and legal case predictions in Kenya leveraging on NLP's major purpose which is converting informal textual structures into formal representations for analysis. The aim is to demonstrate that NLP and Machine Learning (ML) algorithms can be exploited to provide a viable means of solving the problems bedeviling the Kenyan judicial system such as the increasing complexity of cases and huge backlog of cases in Kenyans courts.

Using selected lawyers and advocates to provide expert labeling of the downloaded cases and sample legal case briefs to fine tune and evaluate the outcome of the summary models, NLP and AI algorithms were used to automatically generate case briefs with relevant precedent cases and the likely outcome of the verdict associated with the case submitted to the Kenyan Judiciary. The toolkit developed is a trained NLP and AI model that can generate case briefs and predetermined verdicts of the specific case with 88% and 83.7% levels of accuracy respectively. Considering the huge backlog of cases in Kenyans courts, coupled with the complexity of the cases, this research has demonstrated that NLP and ML can augment human abilities and provide a viable means of automating some aspects of the legal process such as case brief generation and verdict prediction. The toolkit developed, when fully implemented, will result in improving service delivery through facilitating speedier trials and enhancing the efficiency and effectiveness of administrative processes.

Keywords: *Kenya Judiciary, Case Briefs, Verdict Prediction, Legal Case Prediction, Natural Language Processing, Machine Learning*

ABBREVIATIONS

AI – Artificial Intelligence

EKLR – Kenya Law Reporting

HTML – HyperText Markup Language

IE – Information Extraction

ML – Machine Learning

NLP – Natural Language Processing

SVM – Support Vector Matrix

TF-IDF – Term Frequency–Inverse Document Frequency

DEFINITION OF TERMS

Artificial Intelligence – The theory and development of computer systems being able to act, think and perform on tasks requiring humanlike intelligence.

Artificial Neural Networks – Computing systems inspired by the biological neural networks that constitute animal brains, based on a collection of connected units or nodes called artificial neurons, which loosely model the neurons in a biological brain.

Case Brief – A summary and analysis of a court's opinion.

Deep Learning – A subfield of machine learning concerned with algorithms inspired by the structure and function of the brain called artificial neural networks.

Machine Learning – A field of computer science that relates to the study of computer algorithms through pattern recognition, computational learning and use of data.

Natural Language Processing – A field of computer science that relates to the interactions between computers and human language with computers intelligently processing and understanding human language.

Term Frequency–Inverse Document Frequency – A statistical measure of how important a word is to a text document or a corpus of texts.

Verdict – An opinion, judgment or decision in a civil or criminal case or an inquest.

TABLE OF CONTENTS

CHAPTER 1
INTRODUCTION

1.1 BACKGROUND

In this era of big data, courts are increasingly publishing judgements enabling large-scale machine learning and statistical analysis in the legal domain. Recently, major advancements have been made in automatically summarizing and extracting information for statistical analysis using Natural Language Processing (NLP) tools to determine judicial verdicts automatically with up to 75% accuracy, highlighting the potential use of intelligent approaches in law (Medvedeva et al., 2020). Artificial Intelligence (AI) advancements in law have leveraged NLP and Deep Learning methods to identify patterns in legal information, draw conclusions, make policy recommendations, and predict legal outcomes (Alarie et al., 2018).

The law domain has proved to be both interesting and challenging for AI as it has a plethora of cases, norms, hierarchies of authority, rules, meta-rules, theories, and procedures. Cases include precedents that are previous cases that have been tried and whose decisions may have been appealed up through various court levels of the judicial system.. NLP technologies have enabled new approaches in the domain to improve the efficiency, consistency and comprehensibility of legal systems in allowing them to analyze, index and enrich big data from the web automatically (Kang et al., 2020).

The National Council for Law Reporting has in the last decade progressively made legal data publicly available through the Kenya Law Reporting website www.kenyalaw.org. This effectively opens up Kenyan legal data for analysis enabling the development of NLP and machine learning models specifically trained on Kenyan case law.

Legal practitioners need to effectively use sophisticated NLP technologies on large volumes of publicly accessible big data libraries for legal interpretation and reasoning in order to solve complex legal issues to the benefit of society (Robaldo et al., 2021). A number of tools that focus on using NLP and AI for legal cases and legislature have been developed. LexNLP, for example, is an open-source Python tool that allows users to segment documents, identify key text, extract structured information, transform the text, and build unsupervised and supervised machine learning models (Bommarito et al., 2021).

AI applications that utilize NLP models for legal writing are increasingly being exploited in the practice of law. Organization and quality of data define the intricacies of legal technology, rather than innovation itself (Haney et al., 2020). These AI applications are critical in advancing the practice of law. They include Automated Detection of potentially unfair clauses (Lippi et al., 2019), Case Corpuses (Solan et al., 2017), Deep Learning (Li et al., 2019) and Text Similarity Systems (Panagis et al., 2017). Today legal practitioners conduct a significant amount of their legal research online. Some of these practitioners are able to afford using software from proprietary data providers such as Westlaw, Lexis, or Bloomberg which have more powerful searching capabilities; some depend on resources available publicly such as Justia (United States), CanLII (Canada) or Kenya Law (Kenya); or general search engines like Google. Legal documents are currently available online including judicial opinions, regulations and legislation and can be accessed and analyzed by any legal researcher (Alarie et al., 2018).

Some of the current research focuses on mining textual arguments in legal cases and using court arguments and decisions to detect verdicts (Ruppert et al. 2018). These methods are used to sort published judgements or extract verdicts out of unstructured legal texts. Identification of arguments is critical for predicting court decisions or the automatic analysis of legal data. Machine learning techniques have also been used in case law analysis (Custers & Leeuw, 2017). In the US, for example, these techniques have been used to predict the voting behavior of judges or the verdicts of courts (Katz et al. 2012). Lately, Katz et al. (2017) developed a model that endeavors to predict the verdict of the US Supreme Court at Court and Judge level.

In other countries, some researchers have predicted court verdicts using ML models. Sulea et al. (2017b) made predictions on the area of law of a case and the court ruling on case using ML techniques on the French Court of Cassation case law achieving accuracy levels of more than 92%. Aletras et al. (2016) achieved an accuracy of 79% at the case outcome level by using ML to predict the court decisions using text extraction from relevant sections of the ECtHR judgements.

The Kenyan Judiciary, in its Strategic Plan 2019 - 2023, recognizes the impact of technology in increasing efficiency and improving service delivery through facilitating speedier trials and enhancing the efficiency and effectiveness of administrative processes (JSC, 2019). The Judiciary annual report stated that the Directorate of the ICT had developed specifications for the procurement of a speech to text software system for the judiciary (JSC Annual Report, 2019). This however, was halted due to lack of funds. Through the Ajira Digital Programme

of the Ministry of ICT, the judiciary has attempted to employ the youth to assist in digitizing the audio court proceedings by performing manual audio transcriptions to text. AI coupled with automatic speech recognition, a subfield of NLP, allows for proceedings to be recorded, processed, and transcribed faster than using traditional court or human transcriptionists.

Considering that NLP's major purpose is to convert informal textual structures into formal representations that computers can understand and analyze, there is a need for a strong case for integrating NLP systems and AI in document review, legal writing and legal case predictions in Kenya. Considering the huge backlog of cases in Kenyans courts which previous efforts by the judiciary to resolve the same have been unsuccessful (Ogonjo et al., 2021), coupled with the complexity of the cases, we seek to demonstrate that NLP and Machine Learning can augment human abilities and provide a viable means of tackling some of the problems bedeviling the Kenya judicial system.

1.2 PROBLEM STATEMENT

The Kenyan judiciary constantly faces an ever-growing backlog of cases of up to 100,000 cases per year (Ogonjo et al., 2021). There is a need to leverage technology to improve service delivery and enhance efficiency of the judiciary (JSC Annual Report, 2019). Therefore there is a need to convincingly demonstrate to the Kenyan Judiciary how NLP and AI can be used to automatically generate case briefs (legal summaries) with relevant precedent cases and the likely outcome of the verdict associated with a case submitted, taking into cognizance that Kenyan cases have a different format from that of other jurisdictions.

1.3 OBJECTIVES

1. To extract case information relevant to case brief generation and determination of judgment from legal cases using Natural Language Processing
2. To experiment on which Artificial Intelligence model is best suited in making judgment predictions on extracted legal data.
3. To design a prototype system that generates an intuitive case brief and likely verdict for a judge on a case given all the facts.

1.4 RESEARCH QUESTIONS

a) What case information is extracted in making court judgements?
b) How can Natural Language Processing be used to support extraction of case briefs?
c) How can Natural Language Processing be used to support court verdicts in Kenyan courts?
d) Which AI models are currently being used to analyse and make predictions on Legal case data and how do they perform?
e) How can we evaluate the generated legal case briefs and AI models used?
f) Which system architecture design best suits the prototype?

1.5 SCOPE

This research is limited to common law cases that contain precedent citations published on the Kenya Law Reporting website.

1.6 SIGNIFICANCE OF THE STUDY

The true significance of this project should be seen in the context of the huge backlog of cases in Kenyans courts, some of which are highly complex. It is therefore hoped that this research will convincingly demonstrate that NLP and Machine Learning can complement and enhance human abilities and provide a viable means of automating some aspects of the legal process such as case brief generation and verdict prediction.

1.7 ASSUMPTIONS

The research made the assumptions that the cases available on the Kenya Law Site are complete, clean and of good quality, and that the cases cited as precedent are available and accessible on the Kenya Law Site.

CHAPTER 2

LITERATURE REVIEW

2.1 LAW REPORTING IN KENYA

The main sources of laws in Kenya are the Constitution and Acts of Parliament while ancillary sources include by-laws, Quran, African traditional laws and Precedents. Precedents can be both authoritative which are supreme court cases and persuasive which are cases that if supported by strong legal arguments can be adopted. Lower courts are bound by the decisions made by superior courts. Judgements passed by courts can be cited later as part of law specifically on matters that lack a direct linkage to law (National Council for Law Reporting Act, 1994).

Legal Notice No. 29 of 2009 specifies that precedence greatly supports the administration of justice by ensuring certainty in the law as it provides a given level of sureness on predicting litigation outcome through referencing earlier court decisions. The doctrine guarantees transparency and independence of the courts because a judge is compelled to obey the law provided in preceding cases except when it can be overruled or distinguished. Precedence also enables evolution of jurisprudence and development of the law not afforded by Parliament. The judiciary therefore lays down new principles, or extends old principles, to meet novel circumstances faster.

The National Council for Law Reporting (www.kenyalaw.org) is the authorized publisher of the Kenya Law Reports and the Laws of Kenya mandated to observe and report through the Kenya law reports, the development of Kenya's jurisprudence; to perform revisions, consolidations and publishing of the Laws of Kenya; and any other future related functions conferred by law. This council as mandated by the National Council for Law Reporting Act and Legal Notice No. 29 of 2009 have published an online searchable database, the Kenya Law Reports eKLR site (https://kenyalaw.org/) which provides a current edition of the Laws of Kenya and the Kenya Gazette dating back from 2003. The site enables case searches of Kenya law reports from 1971 covering the High Court and Court of Appeal decisions in Kenya including recent unreported.

Case law is important in decision making and legal reasoning for countries governed by the Common Law system (e.g., Kenya, United Kingdom, United States, Canada, Australia, India).

Studies analyzed case law using machine learning techniques (Custers and Leeuw, 2017). In the US these techniques have been used to predict the voting behavior of judges or the verdicts of (Wongchaisuwat et al., 2017) Moreover, judges and lawyers make citations to precedents and law articles that they determine will provide support to their cases. Citations of Case law support legal argumentation and are referred to as a part of legal analysis.

2.2 NLP AND AI APPLICATIONS IN LAW

Dale (2019) defines legal research as the process of information discovery required to back up legal decision-making by examining both statute (as created by the legislature) and case law (as developed by the courts) to determine relevance of specific matters at hand. Artificial Intelligence and law research is mainly investigated for the formalization of arguments, rules and cases. Verheij (2017) developed the connections between the three and shows that cases can provide the logical basis for establishing which rules and arguments hold in a domain.

Search databases such as Westlaw and LexisNexis that contain legal data, have been present since the early 90s. Researchers today are endeavoring to perform the automatic summarization, information extraction, categorization and statistical analysis of legal information (Medvedeva et al., 2020). Information extraction (IE) stresses on unearthing valuable data from texts through named entity recognition (Weber et al., 2021), extraction of relations (Christopoulou et al., 2019), and event extraction (Liu et al. 2020). Extracting relevant information is fundamental in the legal field especially for countries following the common law that depend on precedent cases (Shao et al., 2020). Bhattacharya et al. (2019) developed an evaluation framework for the different retrieval approaches of acquiring relevant precedent cases and statutes given a current case. They discuss the use of Term-frequency inverse document frequency (TF-IDF) and textRank to perform keyword extraction and retrieval from queries using vector and language models. Recently, deep learning approaches for text mining of legal documents have become popular with Moreno and Redondo (2016) highlighting this increasing interest in deep learning and pointing to its potential for use with legal documents.

In legal advice, there are interactive systems that provide counsel suited to the situation and requirements needed based on questions given by the system (Dale, 2019). In many situations, the result is some type of legal document, thus legal counsel is essentially document automation. Electronic discovery involves finding and collecting electronically stored material to be used in a lawsuit or in making an inquiry (Sulea et al., 2017). When faced with numerous

files on a standard hard disk, one of the most difficult tasks is sorting through them to determine what is useful and what isn't. NLP goes a long way in making this process much easier to handle.

Classification techniques using AI have also been explored in Law as shown by developments such as the prediction model developed by Katz et al. (2017) that employed statistical ensemble methods achieving a score of 70.2% accuracy at the case outcome level and 71.9% at the justice vote level. Sulea et al. (2017) explore using linear Support Vector Machine (SVM) classifiers trained on lexical features, achieving an f1 score of 96% in predicting a case ruling. Craigle (2019) have explored and identified more diverse applications of AI in law that include: AI powered legal research platform such as Casetext; public legal data repository of judicial opinions, statutes and regulations; robot lawyers; and innovative AI tools such as Chatbots and virtual assistants that democratize access to basic legal services for the underserved.

2.3 NLP METHODS IN LAW

Natural language processing (NLP) enables the adaptation of machines while performing text evaluations. It allows users to identify relevant search materials whether or not they contain words or phrases expressly stated within the list of keywords unlike literal keyword searches that look for exact words or phrases. It applies to acquiring data (e.g., ascertaining document relevance) or information extraction (determining document key terms). NLP has several models that have been useful in document analysis. These NLP methods have been applied in law as shown below:

Named Entity Recognition (NER) - Extracting entities from texts is a central approach in Natural Language Processing as it emphasizes the most important ideas and references in the text. Named entity recognition (NER) extracts entities such as persons, places, dates, organizations, and so on from text. Supervised models and grammar rules are commonly used in NER. There are, however, NER systems like OpenNLP which are trained and have in-built NER models. NER makes use of tools which help in tokenization and sentence segmentation. Open libraries that can be used in NER include Stanford NER and NLTK. Stanford (NER), also known as CRF Classifier is a Java implementation program that arbitrarily implements linear chain Conditional Random Field (CRF) sequence models in a generic way. That is, you may use this code to create sequence models for NER or other activities by training your own

models using labeled data. NLTK includes tokenization, tagging, parsing, categorization, stemming, and semantic reasoning text processing packages.

Text Summarization - These are NLP approaches that aid in the summarization of huge amounts of text using extraction and abstraction techniques. Extraction techniques extract portions of the text to generate a summary. Abstraction methods generate new text that expresses the essence of the original material, resulting in a summary (Tran & Sato 2017). For text summarization, different methods such as Latent Semantic Analysis, TextRank and LexRank can be employed. The library often used for text summarization is genism. Gensim is a useful python library for performing natural language processing tasks. The TextRank Algorithm is used to summarize text using the gensim package. TextRank is a technique for extracting information from documents. It supposes that words that appear more frequently in text are more important. As a result, sentences with a high frequency of words are significant.

To evaluate summaries, we need to assess (a) the fluency of the output text itself (related to the language model aspect of a summarization model) and (b) the coherence of the summary and how it reflects the longer input text.

Aspect Mining - Aspect mining is a technique for identifying the many features of a text when applied together with sentiment analysis, it retrieves entire information from text. Part-of-speech tagging (POS) is one of the simplest ways of aspect mining. When aspect mining and sentiment analysis are applied to a sample text, the result reflects the text's whole purpose. Aspect mining makes use of tools such as spacy for tokenization and sentence boundary detection; and Neural Coref v2.0 to recognize and replace pronouns.

TF-IDF : TF-IDF is a statistical metric that assesses the word relevance of text to a document in a set of documents. It is an efficient method for the extraction of word features. The. It multiplies two metrics: the number of times a word occurrence rate in a document and its IDF over a corpus of documents (Tran & Sato 2017). TF-IDF is used for scoring words in machine learning models and automatic text analysis in Natural Language Processing (NLP).

FastText: FastText (Joulin et al., 2017) classifies text based on N-grams and Hierarchical SoftMax. It is a library for quick learning of word representations and categorizing sentences. This is effective since sentence classification and word representation are fundamental in NLP. It also allows you to train both supervised and unsupervised words and sentences.

Similar Case Matching (SCM) - This is an important issue in Legal AI to better anticipate the outcomes of judgments in the Common Law system. SCM focuses on discovering pairs of similar situations, with different definitions of similarity. SCM necessitates the modeling of case relationships from data at various levels of granularity, such as event, fact, and element. Essentially, SCM is a type of semantic matching that can aid in the retrieval of legal knowledge (Xiao et al., 2018).

Legal Element Prediction - Legal AI has its own unique symbols, in addition to the symbols used in general NLP, known as legal elements. The focus is on extracting critical items, such as a stolen item or the verdict. These features may be used to not only add intermediate supervisory material to the judgment prediction model, but also in making the model's prediction findings more understandable (Zhong et al., 2019).

2.4 AI METHODS FOR VERDICT PREDICTION IN LAW

Legal Judgment Prediction (LJP) is formalized by most existing studies under the text classification framework. The job LJP focuses on is predicting judgment outcomes based on both the facts of a case and the text of statute articles in the Civil Law system (Zhong et al., 2019). However, researchers have formalized LJP with machine learning methods. Machine learning algorithms have been employed in legal Judgment predictions and have shown impressive levels of up to 90% accuracy (Katz et al., 2016). These methods are as below:

2.4.1 K-Nearest Neighbor approach

This is one of the earliest prediction methods which calculates how similar or dissimilar the facts and pattern of a case are and determines which verdict should be assigned to the new case. The similarity measure is calculated using metrics such as Manhattan distance, hamming distance or Euclidean distance. These metrics simply sum the number of variables where the two cases differ values (Ashley et al., 2019)

2.4.2 Rule induction and decision trees

This method involves collecting data of cases judged, manually developing rules to explain the legal data, and evaluating and improving the rules on newer cases. A decision tree algorithm learns a tree-like set of questions and determines whether a new case will be classified as either positive or negative. It splits data on a single attribute and stops when each of the terminal nodes at the leaves all have instances with the same result (Ashley et al., 2019). Contradictory

data presents challenges for a clean split in decision trees which have a tendency to overfit data meaning they do not generalize well on unseen data. Katz et al., (2017) show the use of an extremely randomized forest of decision trees in case evaluation using case features to determine the verdict, based on all past verdicts made by the Judge, the Court, and all previous cases.

2.4.3 Case based reasoning models

Case law relies heavily on precedence relationships (Zhong et al., 2019). Case-based reasoning models have been applied to legal judgment prediction tasks, predicting outcomes based on the strength of competing arguments. Ashley et al (2019) discuss a case-based reasoning model that fits new cases into an existing database, mapping the cases to the different case bases using arguments. Quantitative weights are then propagated across a graphical model which represents the confidence level in a prediction and is dependent on the magnitude of promotion or demotion of the value in past contexts. It makes predictions based on the best fit realized from the scores of competing arguments.

2.4.4 Deep learning

Literature review on NLP carried out by Kang et al. (2020) shows how NLP can be harnessed as an analytical technique across multiple disciplines including law. Fusheng et al, (2019) discuss several neural network architectures that have been used for text classification. They refer to simple approaches of classification using linear models such as Linear Regression and Support Vector Machines (SVMs) and more advanced Deep Learning models such as Convolutional Neural Networks (CNN). They examine the capability of a CNN based deep learning model for binary classification which results in a higher performance compared to Support Vector Machines on the larger datasets and a more stable growth trend with the gradually increasing amount of training samples (Medvedeva et al., 2018). Deep learning methods perform better than a Support Vector Machine using bag-of-words techniques.

Convolutional Neural Network (CNN) has been proven efficient in text classification (Xiao et al., 2018). It makes use of tools such as gensim which is a topic modeling toolkit; and ConvNet which is able to learn useful features from data by itself. In CNN, each word is first represented using three-dimensional vectors. A 3x3 weight matrix is then dragged horizontally over the phrase, one step/stride at a time, collecting three words. This weight matrix is referred to as a filter, and each filter has an activation function typically found in feed-forward neural

networks. Each filter is used to detect different features of the text. The output is then calculated by summing the elements of each filter and multiplying by their corresponding weights in the filters (Jeong et al., 2020). When the size of the output is bigger than the input, we can use padding or pooling to align them. In padding, we can either pad the outer edges with zero matrices or ignore the part that does not fit the original text. Pooling is used to reduce dimensions. To do this, the output layer is divided into subsections and then the value that best represents the output is calculated. This is very effective as it helps our model to learn higher level interpretations of other texts (Jeong et al., 2020).

BiGRU with self-attention (BiGRU-Att): A Bidirectional Gated Recurrent Unit consists of two Gated Recurrent Units, one taking input forward, and the other taking input backward. It is a bidirectional recurrent neural network with only the input and forget gates and performs sequence processing. Using this model case facts are joined into a sequence of words and then mapped to embeddings which are fed into a BIGRU stack. The resulting embeddings are then summed up to compute a single case embedding which is then passed to the output layer using a sigmoid, a SoftMax, or no activation for case importance regression (Chalkidis et al., 2019).

Hierarchical Attention Network (**HAN)**: This is an advanced model for classification of text. It includes a BiGRU-Att layer, generating embeddings of facts and a second level self-attention layer that takes the embeddings as input and outputs one case embedding which then goes through another self-attention layer (Chalkidis et al., 2019).

The Label-Wise Attention Network (LWAN): This is robust in multi-label classification (Mullenbach et al., 2018). Rather than a single-attention mechanism like the HAN, LWAN uses L attentions each one representing a possible label. It generates L case embeddings for each case, with each embedding focused on predicting the corresponding label. It contains L separate linear layers each with a sigmoid, to decide if the corresponding label should be assigned through which each case embedding goes through. This multi-label deep learning model is only used on multi-label data (Chalkidis et al., 2019).

BERT: BERT is an NLP model pre-trained on large corpora based on Transformers ((Devlin et al.,2019, Vaswani et al., 2017). New layers are added at the top and trained concurrently on fine-tuned data, specific to tasks. BERT processes up to 512 words of text and therefore truncation of longer text must occur otherwise its performance is affected which is its limitation. This limitation is very common in case law text processing (Chalkidis et al., 2019).

HIER-BERT: Hierarchical -BERT that surpasses BERT's maximum length challenges. Here, the words a fact are read by a BERT-BASE constructing fact embeddings. Fact embeddings are then read by a self-attention mechanism, producing one case embedding that goes through a similar output self-attention layer (Chalkidis et al., 2019)

2.5. RESEARCH GAP

In legal research, NLP will aid in finding useful information to provide support for the judges when they are making decisions (Haney, 2020; Razzano, 2020) and lawyers while preparing their cases. In practice, this usually entails examining both statute and case law for relevant information on a given topic (Bafna & Saini, 2020). With the increased development of technology facilitating increased accessibility and navigation, it is interesting and value adding to pursue research in applying NLP and AI methods in Law. This research has however not been extended into the Kenyan judiciary which is currently dealing with the ever-growing backlog of cases of up to 100,000 cases per year. Since human efforts have proved insufficient in tackling the Kenya Judiciary challenges, AI tools will be useful in mitigating these challenges (Ogonjo et al., 2021).

2.6. CASE FOR NLP AND AI APPLICATION IN KENYA LAW

The Kenyan legal system, like many others throughout the world, is founded on precedent, with judges making decisions based on previous judgments on the same issue. As a result, judges must locate and retrieve important case material to aid their decision-making (Ogonjo et al., 2021). Because of the large quantity and complexity of the cases, this part of their job takes a long time and adds to the trial length. Artificial intelligence (AI) systems that assist with legal research might reduce a judge's job (Chalkidis, 2019).

NLP will aid in finding useful information to provide support for the judges when they are making decisions (Haney, 2020; Razzano, 2020). In practice, this usually entails examining both statute and case law for relevant information on a given topic (Bafna & Saini, 2020). Adopting NLP and AI into Kenya's justice system has several advantages. These advantages include contract review, legal research, electronic recovery, and legal advice and document automation (Dale, 2019).

2.7 CONCEPTUAL FRAMEWORK

The overall conceptual framework for this study is shown in Fig 1. This includes collecting data from the online database, processing it using NLP techniques, vectorizing and weighing it using advanced NLP techniques such as TF-IDF and passing this data into different machine learning models. The models will then be trained and evaluated and the best model selected

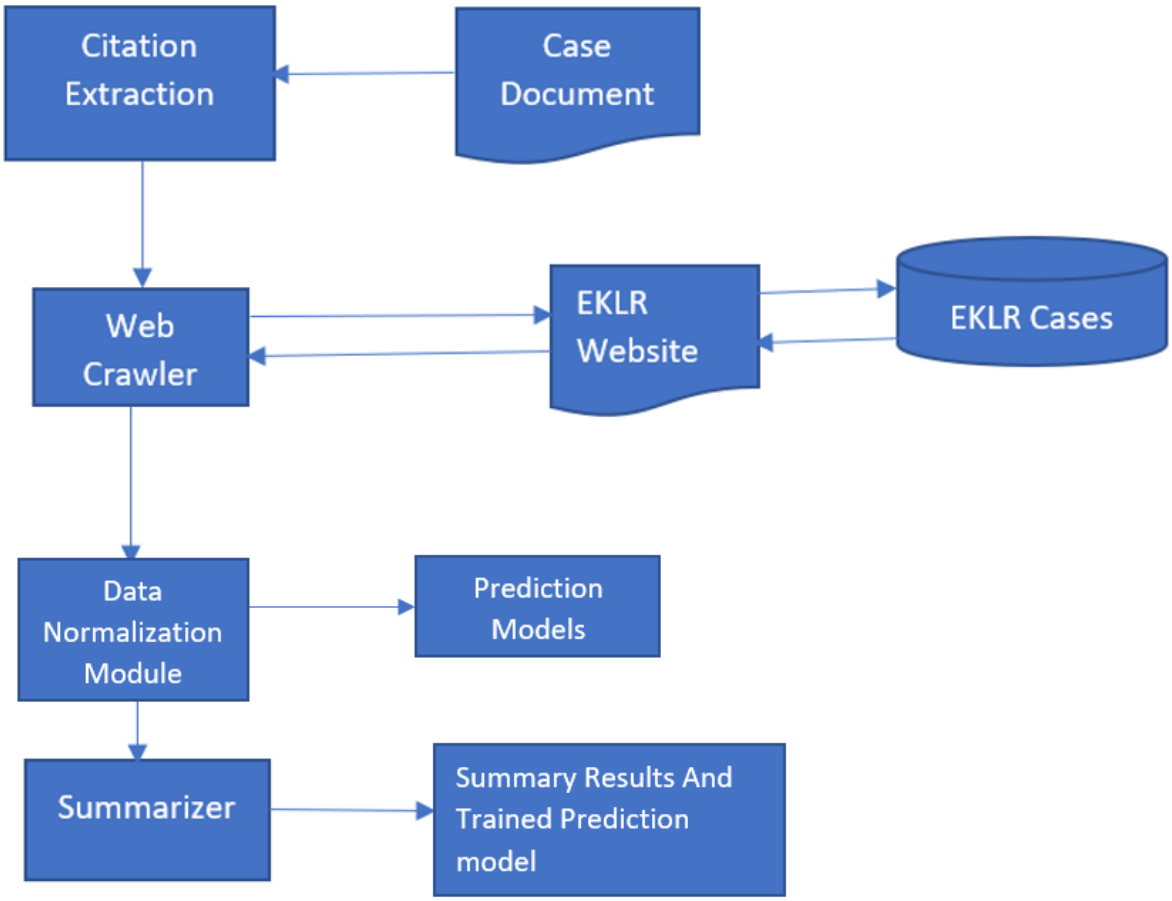


Fig 1: The Conceptual Model

CHAPTER 3

RESEARCH METHODOLOGY

3.1 RESEARCH DESIGN

This research used NLP to provide support for court case judgements in Kenya using quantitative research. This will enable the performing of statistical analysis on the data collected. The steps that were taken to meet the objectives of this study involved reviewing of existing literature on NLP and AI within the legal field, information extraction from legal cases within the Kenya Law reporting online database, summarization of legal information, categorization of legal resources, statistical analysis, prototyping, and finally evaluation of the model and generated results.

3.2 RESEARCH PHILOSOPHY

The research philosophy chosen was the pragmatism paradigm, which enables researchers to focus on the research problem, using all approaches available to understand the problem instead of focusing on specific methods (Creswell et al., 2017).

3.3 PROOF OF CONCEPT - THE PROTOTYPE

The proof of the integration of NLP and AI in document review, legal writing and legal case predictions by converting informal textual structures into formal representations for analysis was achieved by the development and testing of a prototype that automatically generates case briefs. Lawyers and advocates practicing in various firms in Kenya were used to provide expert labeling of the downloaded cases and sample legal case briefs that were used to fine tune and evaluate the outcome of the summary models. The labeled data sets were then used to train and test various machine learning models and the results recorded and analyzed. In developing the toolkit, this study used the Cross Industry Standard Process (CRISP) model with its six sections as shown in Fig. 2.

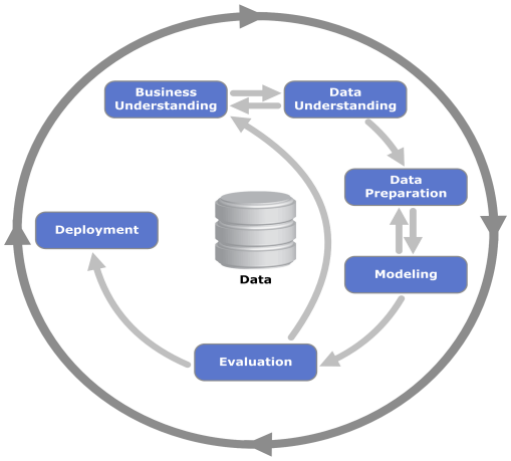


Fig 2: CRISP-DM Framework (Source: The Free Encyclopedia)

3.4 DATA SOURCE AND COLLECTION

The case data used was obtained online from the Kenya Law Reporting site www.kenyalaw.org. This is a national platform maintained by the National Council for Law Reporting with the mandate to observe and report through the Kenya law reports the development of Kenya's jurisprudence. The site enables case searches of Kenya law reports from 1971 covering the High Court and Court of Appeal decisions in Kenya including recent unreported decisions. Data was collected from the online database through web crawling techniques (Krotov & Tennyson, 2018) and legal case text features were derived using clusters and N-gram features (Aletras et al., 2016) . Legal Case briefs were also gotten from a local law firm and reviewed, the structure was then used to guide the model output structure.

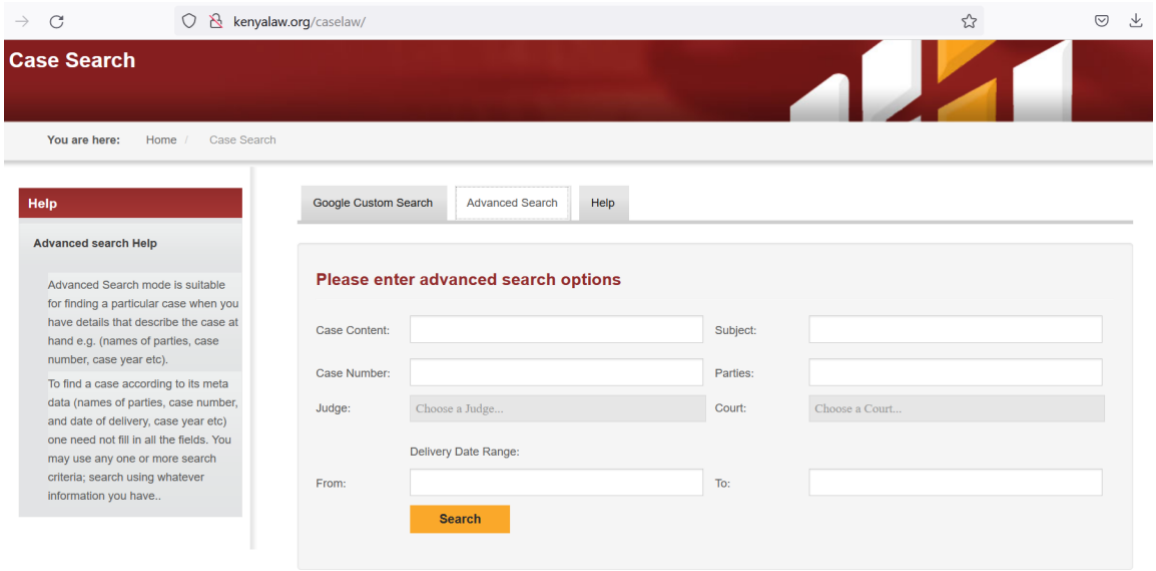


Fig 3: Kenya case law database query engine (Source: Kenya Law Reporting)

3.6 DATA ANALYSIS AND MODELING

Cases with similar characteristics were grouped together and analyzed to ensure they are similar or contain similar characteristics. A more thorough assessment was conducted using TF-IDF to perform feature selection and to further classify these cases into more compact groupings (Tran & Sato, 2017). TF IDF was used to evaluate word relevance in a document, scaling up to a large number of documents. To do this, data from the case filed was first cleaned and standardized. Each of these words were then tokenized according to their frequency. The TF for the words were calculated, followed by the IDF. Each of the keywords were then measured against their TF and IDF value and finally grouped together (Tran & Sato, 2017). Supervised Machine learning methods were then employed since the precedent cases already had the verdicts. These verdicts formed the various categories for classification of current cases. They were then run through CNN's input layer, convolution layer, and max pooling to determine the similar cases that exist. Since CNNs perform better at extracting local and position-invariant features, it was used to classify the related texts and cases (Yin et al., 2017). An assessment of the judges' verdicts was also analyzed and these cases were then grouped again based on their similarities. HAN and BERT were also employed to determine which model performed best in our use case (Chalkidis et al., 2019). These analyses assisted in bringing out similar cases and the verdicts that the judges gave and therefore provided support to the judges when ruling similar cases.

3.7 TESTING AND EVALUATION

The method of NLP case brief evaluation was determined to be a mixture of manual validation due to the nature of the work on the database and automated validation. Manual evaluation was done by comparing the summary against human generated summaries. Automated evaluation was carried out using Rouge and the results recorded. ROUGE-N, is n-gram co-occurrences between candidate summaries and reference summaries, where n is the number of words to match. In the case of multi-documents summarization, the average of all the rouge n-gram values was considered (Tran & Sato 2017).

In evaluating the model for the three algorithms, a stratified cross-validation with 10 folds was used; this ensured that the entire dataset is used for both training and evaluation of the model (Xiong et al., 2020). Precision of the algorithm and its recall values were used for performance evaluation of one algorithm against another, in terms of prediction accuracy (precision) and recalling how to predict (recall). Confusion matrix of each algorithm was used for both self-validation and cross validation against the other algorithms.

3.8 ETHICAL ISSUES

Confidentiality: The cases obtained contained defendant and respondent's personal information such as names and location. Some details were redacted however some details still remained in the cases. Using this data without their explicit consent would have been unethical. It was necessary therefore to ensure that we removed named entities, personal names and any other private information from the text so as to protect individual's privacy (Leidner et al., 2017).

Machine learning bias: The tool was fully automated and as such was able to acquire input and give out results in the form of the case brief and the likely verdict based on previous cases. These results could have been biased based on the cases seen, rather than the actual case being considered. These "unseen" steps were made transparent so as to ensure the model was able to account for the output and the inputs balanced out (Leidner et al., 2017).

Data use approval: The tool was expected to crawl the online website and acquire data from it. Scraping data might be considered unethical, however, the Kenya Law Reporting website expressly allowed us to scrape data and use it (Leins et al., 2020) They stated that the texts of the judicial opinions contained in the site are public and therefore free from any copyright restrictions. This thereby allowed us to crawl the site for the legal case texts for our research.

CHAPTER 4

RESULTS AND DISCUSSIONS

4.1 ARTIFACT DEVELOPMENT

Based on the research gap established, the end product is a trained NLP and AI model with the highest accuracy score from all the models used. This model is able to generate case briefs and predetermined verdicts of the specific case with confidence levels stated. The result of the various development phases is outlined below.

4.1.1 FEASIBILITY ANALYSIS

Feasibility analysis to determine the practicality of the tool was carried out as follows:

1) Operational Feasibility. This measures how the proposed solution will work within legal organizations. This was tested by providing a prototype of the developed solution for a private legal firm in Nairobi and the feedback on the system from the intended users captured
2) Technical Feasibility: This measures the availability of technical resources and expertise required to carry out the project. The software and data needed to undertake this project were open source and freely available making the project technically viable.
3) Schedule Feasibility: This measures how reasonable the project timeline is. The time needed to undertake the project sufficiently within the project timelines.
4) Economic Feasibility: This determines the cost-effectiveness of a project or solution. Since the software used in the development of this tool is open source and freely available, the project is therefore economically feasible. Minimal costs were incurred during the period of undertaking the project.

4.1.2 REQUIREMENTS ANALYSIS

The requirements of the system developed were analyzed and broken down into functional and non-functional requirements. The following were the identified system requirements:

A. FUNCTIONAL REQUIREMENTS

The system includes the following features:

1. Capturing data from a pdf document.
2. Converting pdf documents to docx documents.
3. Using regular expressions and document comprehension to extract case data from

legal case documents.

4. Crawling the Kenya Law Reporting website and acquiring links relevant to the cases required for extraction.
5. Following the weblinks, parsing HTML data from the page, and acquiring text from the page.
6. Summarizing legal text acquired from the legal documents.
7. Evaluating the summarization model.
8. Analyzing legal case data and making near accurate predictions of case outcomes.
9. Evaluating the prediction model accuracy.

B. NON-FUNCTIONAL REQUIREMENTS

The non-functional features of the system includes:

Performance Requirements: The system shall be able to perform optimally, but will depend on the speed of the internet and quantity of citation links.

Reliability: The system shall be able to provide services to users fast and in a secure manner.

Confidentiality: The system will ensure all personal information obtained from the cases is stripped off and not displayed to users.

Availability: the system will be available with an up time of 99%.

4.1.3 SYSTEM USERS

The users of the system include: Legal Practitioners, Data Scientists, Research Scientists and System Administrators. These users were involved in the prototyping and data cleanup section of the tool. The legal practitioners were instrumental in the evaluation of the appropriateness of the summaries generated by the tool and in the labeling of data used in modeling the machine learning models.

Table1: Table of legal practitioner list and their experience levels

Title	**No.**	**No. of years in profession**
Lawyer	3	5-9
Advocate of the High court	1	6
Systems Administrator judiciary	2	5
Law student	2	3-4

4.1.4 SYSTEM DESIGN

The system consists of six main parts: the Web Crawler, the HTML Parser, the Case Data Extraction, the Summarizers, the Prediction Models and the Model Evaluation. The open-sourced system is python based and runs on a Jupyter notebook environment. It captures legal pdf type case documents from the online website (https://kenyalaw.org/), locates the cases and downloads required features from them on the target website using a web crawler, captures the links generated from the crawl, downloads the case data, and generates summaries and predictions based on the documents acquired by following the links. The system summaries and prediction results are then made available to users. Fig 4 shows the overall system design:

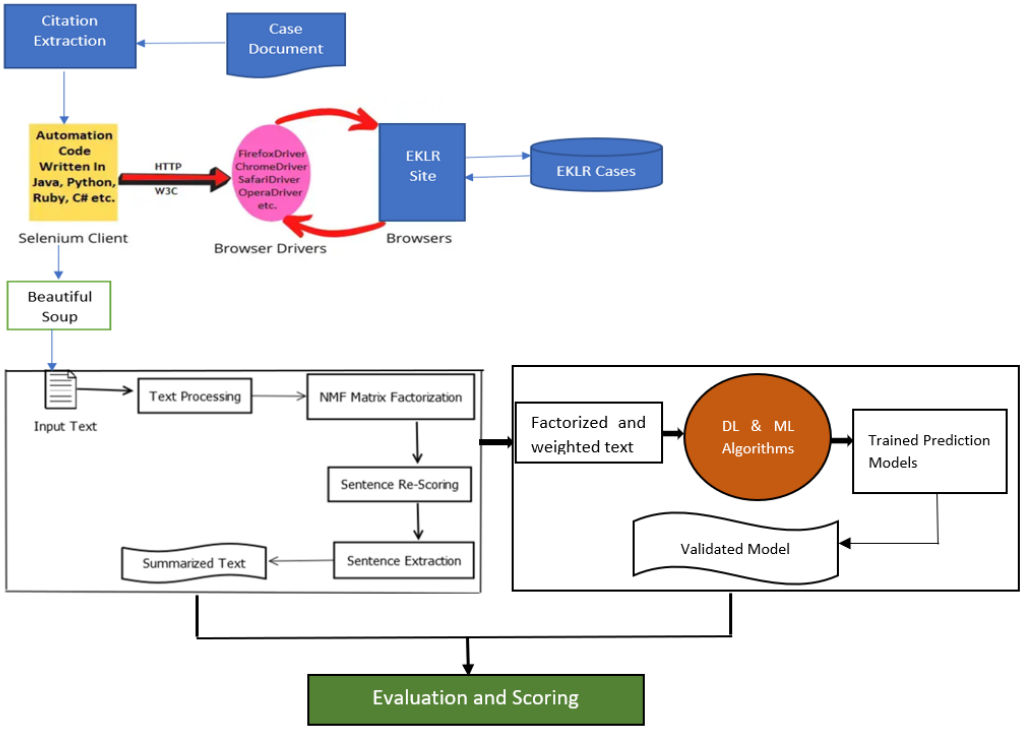


Fig 4: System Architecture diagram

4.1.5 DATA CLEANING

Data collected from the Kenya Law Reporting website (https://kenyalaw.org/) was majorly unstructured and ambiguous making it challenging for the machine learning models to effectively generate patterns from the data (Tran & Sato, 2017). This data had to be cleaned and normalized so as to achieve good results, through named entity recognition and removal (Weber et al., 2021), removal of stop words and noise entities, tokenization and stemming of n-grams (Solangi et al., 2018). Further normalization involved computing the TF-IDF

scores for the n-grams and using these scores to determine the keywords in the paragraphs, these were then structured into a bag-of-words that were then used in training our machine learning models. Our output labels were then one-hot encoded to be able to enable our models perform optimally. With each iterative cleaning of the data, and fine tuning of the Keyword extraction process (Campos et al., 2018), a positive improvement in results of up to 25% accuracy levels was observed.

4.1.6 MODEL EVALUATION

Several summarization and prediction models have 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. Below, we show the performance of these various models.

a) SUMMARIZATION MODEL EVALUATION

The method of model evaluation was determined to be a mixture of manual validation due to the nature of the work on the database and automated validation. Automated evaluation was carried out using Rouge and the results recorded. ROUGE-N, is n-gram co-occurrences between candidate summaries and reference summaries, where n is the number of words to match. In the 1-gram metrics, we collect the ratio of number of single words matching by number of words in reference summary. 2-gram metrics, is the ratio of two continuous words matching in both the summaries. In the case of multi-documents summarization, the average of all the ROUGE n-gram values is considered (Tran & Sato 2017). The system iterated from lower scores to improved scores with pre-processing of data. The summarization models evaluation results in comparison with the human annotated summaries are shown in figure 5.

```
#Lexrank summarizer
from rouge import Rouge
r = Rouge()

r.get_scores(str(summary_lr), text_data)
```

```
[{'rouge-1': {'r': 0.32753623188405795,
   'p': 0.6042780748663101,
   'f': 0.4248120255162106},
  'rouge-2': {'r': 0.18859649122807018,
   'p': 0.40186915887850466,
   'f': 0.2567164135627535},
  'rouge-l': {'r': 0.3072463768115942,
   'p': 0.5668449197860963,
   'f': 0.3984962360425265}}]
```

```
#Bert summarizer
from rouge import Rouge
r = Rouge()
r.get_scores(summary_bes, text_data)
```

```
[{'rouge-1': {'r': 1.0, 'p': 0.8194774346793349, 'f': 0.9007832848664522},
  'rouge-2': {'r': 1.0, 'p': 0.8231046931407943, 'f': 0.9029702920767768},
  'rouge-l': {'r': 1.0, 'p': 0.8194774346793349, 'f': 0.9007832848664522}}]
```

Fig 5: Rouge N-gram results for Lexrank and Bert Summarizers

b) MACHINE LEARNING MODEL EVALUATION

In the implementation of this research various NLP multilabel prediction models were evaluated. The evaluation of the predictive models sought to determine which model and what configurations for the model and data gave the best balance of speed and accuracy that the case prediction model needed (Medvedeva et al., 2018). Several machine learning models were selected, configured, implemented and evaluated. The models that were evaluated are those that could be run with the hardware resources and time constraints in play and were suitable to be structured on multilabel classification data. The clean labeled dataset was randomly split into a training and testing group (Xu et. al, 2018). The following are the models which were validated and tested:

Random Forest Classifier: This machine learning model was trained and validated on the preprocessed legal case data. This model is a classification algorithm that fits sub-samples of data onto several decision trees and averages the results thereby improving the accuracy and minimizing over-fitting (Shah et. al, 2020). It randomly builds each individual tree, to maximize unrelated forests and then aggregates each forest's predictions to make accurate decisions. This yielded an accuracy of 49.9%.

BinaryRelevanceClassifier: This method is very similar to the OneVsRestClassifier method mentioned above. In this category the GaussianNB (Gaussian Naive Bayes) model was

selected. If there are x labels, the binary relevance method creates x new datasets, one for each label, and trains single-label classifiers on each new data set. One classifier may answer yes/no, thus the "binary relevance." This is a simple approach but does not work well when there are dependencies between the labels and yielded an accuracy of 50.7%

ClassifierChain: This approach used the Logistic Regression Classifier to combine the computational efficiency of the Binary Relevance method while still being able to take the label dependencies into account for classification. This model achieved an accuracy of 50.6%

MultiOutputClassifier: This strategy uses the KNeighborsClassifier consisting of fitting one classifier per target. This is a simple strategy for extending classifiers that do not natively support multi-target classification. The accuracy level achieved with this model was at 52.6%

The above classification models are all machine learning models. Their accuracy levels were lower than the deep learning models considered next as shown in Figure 6.

```
In [80]: results.style.apply(highlight_min, color='red', axis=None).apply(highlight_max, color='lightgreen', axis=None)
```

Out[80]:

	OneVsAll	BinaryRelevance	ClassifierChain	MultipleOutput	DNN	CNN	LSTM
4	0.499492	0.507623	0.506209	0.526443	0.698241	0.837289	0.770470

Fig 6: Summaries of model accuracy levels

Tensorflow: Text classification has benefited from the deep learning architectures' trend due to their potential to reach high accuracy. There are different libraries available for deep learning such as Tensorflow and PyTorch which are the most popular libraries for the topic. Deep learning techniques normally give better results in NLP tasks, for instance, syntactic parsing and sentiment analysis. It is possible to either train the WordEmbedding layer or use a pre-trained one through transfer learning, such as word2vec and GloVe. For the following models, the vectorization used was texts_to_sequences, which transforms the words in numbers, and the pad_sequences ensures all the vectors have the same length. Class weights were calculated to address the imbalance problem in the categories.

DNN with WordEmbedding: We started with a simple model which only consists of an embedding layer, a dropout layer to reduce the size and prevent overfitting, a max-pooling layer, and one dense layer with a sigmoid activation to produce probabilities for each of the categories that we want to predict. The results of the individual layers are shown in figure 7. This model achieved an improved score of 69.8%

```
Epoch 22/30
10/10 [==============================] - 0s 27ms/step - loss: 0.0031 - auc: 0.9876 - val_loss: 0.4421 - val_auc: 0.7237 - lr:
0.0015
Epoch 23/30
10/10 [==============================] - 0s 37ms/step - loss: 0.0031 - auc: 0.9882 - val_loss: 0.4435 - val_auc: 0.7225 - lr:
0.0015
Epoch 24/30
10/10 [==============================] - 0s 19ms/step - loss: 0.0030 - auc: 0.9885 - val_loss: 0.4444 - val_auc: 0.7242 - lr:
0.0015
Epoch 25/30
10/10 [==============================] - 0s 16ms/step - loss: 0.0030 - auc: 0.9888 - val_loss: 0.4457 - val_auc: 0.7243 - lr:
0.0015
Epoch 26/30
10/10 [==============================] - 0s 28ms/step - loss: 0.0030 - auc: 0.9892 - val_loss: 0.4468 - val_auc: 0.7227 - lr:
0.0015
Epoch 27/30
10/10 [==============================] - 0s 29ms/step - loss: 0.0029 - auc: 0.9895 - val_loss: 0.4480 - val_auc: 0.7224 - lr:
0.0015
Epoch 28/30
10/10 [==============================] - 0s 23ms/step - loss: 0.0029 - auc: 0.9898 - val_loss: 0.4491 - val_auc: 0.7223 - lr:
0.0015
Epoch 29/30
10/10 [==============================] - 0s 17ms/step - loss: 0.0029 - auc: 0.9900 - val_loss: 0.4493 - val_auc: 0.7217 - lr:
1.5000e-04
Epoch 30/30
10/10 [==============================] - 0s 26ms/step - loss: 0.0029 - auc: 0.9901 - val_loss: 0.4494 - val_auc: 0.7212 - lr:
1.5000e-04
```

Fig 7:DNN Deep learning model layer performance

CNN with WordEmbedding: Convolutional Neural Networks recognize local patterns in a sequence by processing multiple words at the same time, and 1D convolutional networks are suitable for text processing tasks. In this case, the convolutional layer uses a window size of 3 and learns word sequences that can later be recognized in any position of a text. The results of the individual layers are shown in figure 8. This model achieved the highest score of 83.7%

```
Epoch 22/30
10/10 [==============================] - 1s 79ms/step - loss: 0.0074 - auc_1: 0.8905 - val_loss: 0.3434 - val_auc_1: 0.8300 - l
r: 0.0010
Epoch 23/30
10/10 [==============================] - 1s 66ms/step - loss: 0.0072 - auc_1: 0.8997 - val_loss: 0.3429 - val_auc_1: 0.8289 - l
r: 0.0010
Epoch 24/30
10/10 [==============================] - 1s 64ms/step - loss: 0.0070 - auc_1: 0.9093 - val_loss: 0.3408 - val_auc_1: 0.8338 - l
r: 0.0010
Epoch 25/30
10/10 [==============================] - 1s 64ms/step - loss: 0.0067 - auc_1: 0.9186 - val_loss: 0.3402 - val_auc_1: 0.8345 - l
r: 0.0010
Epoch 26/30
10/10 [==============================] - 1s 56ms/step - loss: 0.0065 - auc_1: 0.9291 - val_loss: 0.3422 - val_auc_1: 0.8325 - l
r: 0.0010
Epoch 27/30
10/10 [==============================] - 1s 62ms/step - loss: 0.0062 - auc_1: 0.9388 - val_loss: 0.3376 - val_auc_1: 0.8408 - l
r: 0.0010
Epoch 28/30
10/10 [==============================] - 1s 70ms/step - loss: 0.0059 - auc_1: 0.9476 - val_loss: 0.3402 - val_auc_1: 0.8345 - l
r: 0.0010
Epoch 29/30
10/10 [==============================] - 1s 81ms/step - loss: 0.0056 - auc_1: 0.9524 - val_loss: 0.3387 - val_auc_1: 0.8368 - l
r: 0.0010
Epoch 30/30
10/10 [==============================] - 1s 75ms/step - loss: 0.0054 - auc_1: 0.9632 - val_loss: 0.3425 - val_auc_1: 0.8313 - l
r: 0.0010
```

Fig 8:CNN Deep learning model layer performance

LSTM with GloVe WordEmbedding: We used GloVe word embedding to convert text inputs to their numeric counterparts, which is a different approach because this is a pre-trained layer. The model has one input layer, one embedding layer, one LSTM layer with

128 neurons, and one output layer with 21 neurons (the number of targets.) The results of the individual layers are shown in figure 7. This model achieved a score of 77.0% accuracy

```
Epoch 22/30
10/10 [==============================] - 4s 443ms/step - loss: 6.6587e-04 - auc_8: 1.0000 - val_loss: 0.5230 - val_auc_8: 0.784
7 - lr: 1.0000e-04
Epoch 23/30
10/10 [==============================] - 4s 426ms/step - loss: 6.1886e-04 - auc_8: 1.0000 - val_loss: 0.5253 - val_auc_8: 0.778
2 - lr: 1.0000e-04
Epoch 24/30
10/10 [==============================] - 4s 415ms/step - loss: 5.6501e-04 - auc_8: 1.0000 - val_loss: 0.5222 - val_auc_8: 0.786
7 - lr: 1.0000e-04
Epoch 25/30
10/10 [==============================] - 4s 444ms/step - loss: 5.4210e-04 - auc_8: 1.0000 - val_loss: 0.5242 - val_auc_8: 0.788
2 - lr: 1.0000e-04
Epoch 26/30
10/10 [==============================] - 5s 480ms/step - loss: 5.1377e-04 - auc_8: 1.0000 - val_loss: 0.5300 - val_auc_8: 0.782
5 - lr: 1.0000e-04
Epoch 27/30
10/10 [==============================] - 4s 393ms/step - loss: 4.9416e-04 - auc_8: 1.0000 - val_loss: 0.5344 - val_auc_8: 0.773
6 - lr: 1.0000e-04
Epoch 28/30
10/10 [==============================] - 4s 444ms/step - loss: 4.7473e-04 - auc_8: 1.0000 - val_loss: 0.5385 - val_auc_8: 0.771
1 - lr: 1.0000e-04
Epoch 29/30
10/10 [==============================] - 5s 450ms/step - loss: 4.5864e-04 - auc_8: 1.0000 - val_loss: 0.5408 - val_auc_8: 0.769
8 - lr: 1.0000e-04
Epoch 30/30
10/10 [==============================] - 4s 416ms/step - loss: 4.4226e-04 - auc_8: 1.0000 - val_loss: 0.5443 - val_auc_8: 0.768
2 - lr: 1.0000e-04
```

Fig 9:LSTM with Glove learning model layer performance

In conclusion, based on the benchmark, the Deep Neural Network showed the best accuracy scores, but the difference was minimal among the deep learning models, with CNN having the highest performance. On the other hand, the algorithms available in the scikit-learn package presented scores considerably lower, and they were therefore not suitable for this problem. Deep learning methods performed better as had been correctly deduced by other researchers (Medvedeva et al., 2018; Xiao et al., 2018). Medvedeva (2018) having examined the capability of a CNN based deep learning model for binary classification compared to Support Vector Machines noted that CNN resulted in a higher performance on the larger datasets and a more stable growth trend with the gradually increasing amount of training samples. In our case, with more training samples and an incremental cleaning and subsequent removal of legal stop words, **CNN** proved more efficient in text classification.

4.2 DISCUSSION

The National Legal Research Group (2018) found that lawyers who used AI tools to conduct legal research completed projects 24.5% faster and the search results were 21% more relevant. Their study concluded that use of AI would save legal practitioners 132 – 210 hours a year when conducting legal research. This research sought to determine which NLP methodologies accompanied by Deep Learning AI algorithms can be enhanced and adopted to fit the structure

of Kenyan cases, leading to swift provision of support in legal research and verdict determination of precedent cases in the Kenyan judiciary and how best to evaluate these models (Paulus et al., 2017).

This research has made a good attempt at the task of constructing a model for all of Kenyan law. The toolkit developed has automated the extraction of linkages from legal texts, allowing legal papers to be linked within a Kenyan Law Search Engine and has output a model trained on kenyan Law that is useful for making verdict predictions. This will enhance the usage of the Kenya Law search engine and its efficacy should it be implemented.

This research is in line with the digital strategy enshrined in the 2017-2021 Sustaining the Judiciary Transformation Blueprint that seeks to re-engineer its processes through ICT. Key in this strategy is the digitization of court records and proceedings, retiring archaic filing systems and modernizing document management (JSC Annual Report, 2019). According to the Ministry of ICT, 60 million records were digitized under the High Court Registry pilot digitization project (ICT Authority, 2019). Digitization of these records has made the use of AI to conduct legal research a viable strategy.

This toolkit offers a great opportunity for the judiciary to achieve its service delivery goals. The problems caused by insufficient funding and workforce could be mitigated by utilizing this tool. Legal research that has been a pain point requiring time and a lot of human resources can now be done automatically in just a few steps. Through deep learning techniques previously examined by other researchers (Medvedeva et al., 2018) we explored the capability of a CNN based deep learning model for multiclass classification compared to traditional machine learning models, LSTM and DNN. It was noted that CNN resulted in a higher performance on the larger datasets and a more stable growth trend with the gradually increasing amount of training samples. In our case, with more training samples and an incremental cleaning and subsequent removal of legal stop words, CNN proved more efficient in text classification..

In other countries, some researchers such as Sulea (2017b) have predicted court verdicts using ML models achieving accuracy levels of more than 92%, Aletras et al. (2016) achieved an accuracy of 79% at the case outcome level and Medvedeva et al. (2020) achieved 75% accuracy. These researchers, having highlighted the potential use of intelligent approaches in law, were beneficial in determining how best to carry out the research. Unfortunately, it was not possible to perform a systematic and thorough comparison of the results produced by the models since the corpus and in some cases the languages analyzed are very different. Despite this, our CNN model has achieved impressive results of 83.7%.

4.3 ETHICAL ISSUES

Scraping data might be considered unethical, however, the Kenya Law Reporting website expressly allowed us to scrape data and use it. The data was then preprocessed masking all the personal data of applicants and respondents thereby maintaining their anonymity.

CHAPTER 5
CONCLUSION AND RECOMMENDATIONS

5.1 SUMMARY OF FINDINGS

Objective 1: To extract case information relevant to case brief generation and determination of judgment from legal cases using Natural Language Processing

Legal case briefs were received from various law firms within the country and the structure used to model the output expected from the summarization models. These models were then tested on legal case data that had been extracted from the online case database eKLR. The output of the abstractive summary (Tran & Sato 2017) was then evaluated by legal practitioners to determine whether the summary was sufficient and appropriate. The summary was also evaluated against the original legal case briefs using Rouge-N (Tran & Sato 2017) and achieved an overall impressive score of 88% similarity.

Objective 2: To experiment on which Artificial Intelligence model is best suited in making judgment predictions on extracted legal data.

Several machine learning and deep learning models were implemented in this research. These models were then trained on Kenyan legal case data and their parameters fine tuned to determine the most suitable model parameters and the results were evaluated. The model with the highest accuracy levels achieved during the training and validation phase was then selected as the most suitable model. This was the CNN model (Xiao et al., 2018) which achieved an impressive score of 83.7% accuracy.

Objective 3: To design a prototype system that generates an intuitive case brief and likely verdict for a judge on a case given all the facts.

Following the research conducted (Katz et al., 2016), a prototype with summarization and case prediction features using the models uncovered during the research, was developed. This prototype has the capability to accept legal case data as input and output the summary and predicted outcome of the case. The prototype displays the probability of each class with the highest probability determining which class the case falls into.

5.2 CONCLUSION

In this research we sought to extract case information relevant to case brief generation and determination of judgment from the national online Kenya Law Reporting database,

experiment on which AI model is best suited in making judgment predictions on extracted legal data, and design a toolkit that generates an intuitive case brief and likely verdict for a judge. The results showed that our trained NLP and AI model can generate case briefs and predetermined verdicts of the specific case with 88% and 83.7% levels of accuracy respectively.

Considering the huge backlog of cases in Kenyans courts (Ogonjo et al., 2021), coupled with the complexity of the cases, this research has demonstrated that NLP and Machine Learning can augment human abilities and provide a viable means of automating some aspects of the legal process chain such as case brief generation and verdict prediction. This will aid legal practitioners in their legal research and case preparation efforts thereby improving their efficiency and effectiveness.

5.3 LIMITATIONS

The development of the artifact faced some limitations. The evaluation of summary results is still very much a human annotated task and has not yet been fully automated. A human expert still needed to confirm the adequacy of the legal case summaries. Legal data also has major document variations in meaning several updates to the code for analysis of different document types or citation formats. Coding all these possible variations proved to be a highly resource intensive task.

5.4 RECOMMENDATION

The Kenyan legal sector is overwhelmed with the backlog and complexity of cases and legal data, with new cases coming in every day (Ogonjo et al., 2021). Adoption of the tool by the legal fraternity will go a long way in easing their burdens.

Additional work on development of an algorithm that automatically scores the resulting summarization with less reliance on human annotation will also be impactful within this sector.

REFERENCES

Alarie, B., Niblett, A., & Yoon, A. H. (2018). How artificial intelligence will affect the practice of law. University of Toronto Law Journal, 68(supplement 1), 106-124.

Aletras, N., Tsarapatsanis, D., Preoţiuc-Pietro, D., & Lampos, V. (2016). Predicting judicial decisions of the European Court of Human Rights: A natural language processing perspective. PeerJ Computer Science, 2, e93.

Arunda, B. (2020). The Impacts of Emerging Technologies in the Future of Law and Legal Practice: A Case of Kenya. Statistica.

Ashley, K. D. (2019). A brief history of the changing roles of case prediction in AI and law. *Law Context: A Socio-Legal J.*, *36*, 93.

Bhattacharya, P., Ghosh, K., Ghosh, S., Pal, A., Mehta, P., Bhattacharya, A., & Majumder, P. (2019). Overview of the FIRE 2019 AILA Track: Artificial Intelligence for Legal Assistance. In FIRE (Working Notes) (pp. 1-12).

Bommarito II, M. J., Katz, D. M., & Detterman, E. M. (2021). LexNLP: Natural language processing and information extraction for legal and regulatory texts. In Research Handbook on Big Data Law. Edward Elgar Publishing.

Campos, R., Mangaravite, V., Pasquali, A., Jorge, A. M., Nunes, C., & Jatowt, A. (2018, March). A text feature based automatic keyword extraction method for single documents. In *European conference on information retrieval* (pp. 684-691). Springer, Cham.

Chalkidis, I., Androutsopoulos, I., & Aletras, N. (2019). Neural legal judgment prediction in English. *arXiv preprint arXiv:1906.02059*.

Christopoulou, F., Miwa, M., & Ananiadou, S. (2019). Connecting the dots: Document-level neural relation extraction with edge-oriented graphs. arXiv preprint arXiv:1909.00228.

Craigle, V. (2019). Law Libraries Embracing AI. Law Librarianship in the Age of AI,(Ellyssa Valenti, Ed.)

Creswell, J. W., & Creswell, J. D. (2017). *Research design: Qualitative, quantitative, and mixed methods approaches*. Sage publications.

Custers, B., & Leeuw, F. L. (2017). Legal Big Data (Legal Big Data). Custers BHM & Leeuw F.(2017), Legal big data: Toepassingen voor de rechtspraktijk en juridisch onderzoek, Nederlands Juristenblad, 34, 2449-2456.

Dale, R. (2019). Law and word order: NLP in legal tech. *Natural Language Engineering*, *25*(1), 211-217.

Haney, B. (2020). Applied natural language processing for law practice. Brian S. Haney, Applied Natural Language Processing for Law Practice.

Kang, Y., Cai, Z., Tan, C. W., Huang, Q., & Liu, H. (2020). Natural language processing (NLP) in management research: A literature review. Journal of Management Analytics, 7(2), 139-172.

Katz, D. M., Bommarito, M. J., & Blackman, J. (2017). A general approach for predicting the behavior of the Supreme Court of the United States. *PloS one*, *12*(4), e0174698. https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0174698

Kshetri, N. (2020). Artificial Intelligence in Developing Countries. IEEE Annals of the History of Computing, 22(04), 63-68.

Leidner, J. L., & Plachouras, V. (2017, April). Ethical by design: Ethics best practices for natural language processing. In *Proceedings of the First ACL Workshop on Ethics in Natural Language Processing* (pp. 30-40).

Leins, K., Lau, J. H., & Baldwin, T. (2020). Give Me Convenience and Give Her Death: Who Should Decide What Uses of NLP are Appropriate, and on What Basis?. *arXiv preprint arXiv:2005.13213*.

Li, P., Zhao, F., Li, Y., & Zhu, Z. (2018, June). Law text classification using semi-supervised convolutional neural networks. In *2018 Chinese Control and Decision Conference (CCDC)* (pp. 309-313). IEEE.

Li, S., Zhang, H., Ye, L., Guo, X., & Fang, B. (2019). Mann: A multichannel attentive neural network for legal judgment prediction. *IEEE Access*, *7*, 151144-151155.

Liu, J., Chen, Y., Liu, K., Bi, W., & Liu, X. (2020). Event extraction as machine reading comprehension. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP) (pp. 1641-1651).

Lippi, M., Pałka, P., Contissa, G., Lagioia, F., Micklitz, H. W., Sartor, G., & Torroni, P. (2019). CLAUDETTE: an automated detector of potentially unfair clauses in online terms of service. Artificial Intelligence and Law, 27(2), 117-139.

Mahony, C., Albrecht, E., & Sensoy, M. (2019). The relationship between influential actors' language and violence: A Kenyan case study using artificial intelligence. London: International Growth Centre.

McKamey, M. (2017). Legal technology: Artificial intelligence and the future of law practice. Appeal: Rev. Current L. & L. Reform, 22, 45.

Medvedeva, M., Vols, M., & Wieling, M. (2020). Using machine learning to predict decisions of the European Court of Human Rights. Artificial Intelligence and Law, 28(2), 237-266.

Moreno, A., & Redondo, T. (2016). Text analytics: the convergence of big data and artificial intelligence. IJIMAI, 3(6), 57-64.

Mullenbach, J., Wiegreffe, S., Duke, J., Sun, J., & Eisenstein, J. (2018). Explainable prediction of medical codes from clinical text. *arXiv preprint arXiv:1802.05695*.

National Legal Research Group, Inc. (2018). "The real impact of using artificial intelligence in legal research.," National Legal Research Group, Inc., Charlottesville, VA.

Ogonjo, F., Gitonga, J. T., Wairegi, A., & Rutenberg, I. (2021). Utilizing AI to Improve Efficiency of the Environment and Land Court in the Kenyan Judiciary.

Panagis, Y., Šadl, U., & Tarissan, F. (2017). Giving every case its (legal) due-the contribution of citation networks and text similarity techniques to legal studies of European Union law. In *Legal knowledge and information systems* (pp. 59-68). IOS Press.

Paulus, R., Xiong, C., & Socher, R. (2017). A deep reinforced model for abstractive summarization. *arXiv preprint arXiv:1705.04304*.

Procopiuck, M. (2018). Information technology and time of judgment in specialized courts: What is the impact of changing from physical to electronic processing?. Government Information Quarterly, 35(3), 491-501.

Robaldo, L., Antoniou, G., Baryannis, G., Batsakis, S., Governatori, G., Islam, M. B., ... & Tachmazidis, I. (2021). Large-scale legal reasoning with rules and databases. Journal of Applied Logics, 8(4), 911.

Ruppert, E., Hartung, D., Sittig, P., Gschwander, T., Rönneburg, L., Killing, T., & Biemann, C. (2018, August). LawStats–Large-Scale German Court Decision Evaluation Using Web Service Classifiers. In *International Cross-Domain Conference for Machine Learning and Knowledge Extraction* (pp. 212-222). Springer, Cham.

Shah, K., Patel, H., Sanghvi, D., & Shah, M. (2020). A comparative analysis of logistic regression, random forest and KNN models for the text classification. *Augmented Human Research*, *5*(1), 1-16.

Shao, Y., Mao, J., Liu, Y., Ma, W., Satoh, K., Zhang, M., & Ma, S. (2020). BERT-PLI: Modeling Paragraph-Level Interactions for Legal Case Retrieval. In *IJCAI* (pp. 3501-3507).

Solan, L. M., & Gales, T. (2017). Corpus linguistics as a tool in legal interpretation. *BYU L. Rev.*, 1311.

Solangi, Y. A., Solangi, Z. A., Aarain, S., Abro, A., Mallah, G. A., & Shah, A. (2018, November). Review on Natural Language Processing (NLP) and its toolkits for opinion mining and sentiment analysis. In *2018 IEEE 5th International Conference on Engineering Technologies and Applied Sciences (ICETAS)* (pp. 1-4). IEEE.

Sulea, O. M., Zampieri, M., Vela, M., & Van Genabith, J. (2017). Predicting the law area and decisions of french supreme court cases. *arXiv preprint arXiv:1708.01681*.

Sun, S., Sun, Q., Zhou, K., & Lv, T. (2019, November). Hierarchical attention prototypical networks for few-shot text classification. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)* (pp. 476-485).

Surden, H. (2019). Artificial intelligence and law: An overview. Georgia State University Law Review, 35, 19-22.

Tran, T. K., & Sato, H. (2017, November). NLP-based approaches for malware classification from API sequences. In 2017 21st Asia Pacific Symposium on Intelligent and Evolutionary Systems (IES) (pp. 101-105). IEEE.

The Judiciary (2020) Strategic Plan 2019 - 2023 (pp. 44-45)

Verheij, B. (2017, June). Formalizing arguments, rules and cases. In *Proceedings of the 16th edition of the International Conference on Artificial Intelligence and Law* (pp. 199-208).

Weber, L., Sänger, M., Münchmeyer, J., Habibi, M., Leser, U., & Akbik, A. (2021). HunFlair: an easy-to-use tool for state-of-the-art biomedical named entity recognition. Bioinformatics, 37(17), 2792-2794.

Wei, F., Qin, H., Ye, S., & Zhao, H. (2018, December). Empirical study of deep learning for text classification in legal document review. In *2018 IEEE International Conference on Big Data (Big Data)* (pp. 3317-3320). IEEE.

Wongchaisuwat, P., Klabjan, D., & McGinnis, J. O. (2017, June). Predicting litigation likelihood and time to litigation for patents. In *Proceedings of the 16th edition of the International Conference on Artificial Intelligence and Law* (pp. 257-260).

Xiao, C., Zhong, H., Guo, Z., Tu, C., Liu, Z., Sun, M., ... & Xu, J. (2018). Cail2018: A large-scale legal dataset for judgment prediction. *arXiv preprint arXiv:1807.02478*.

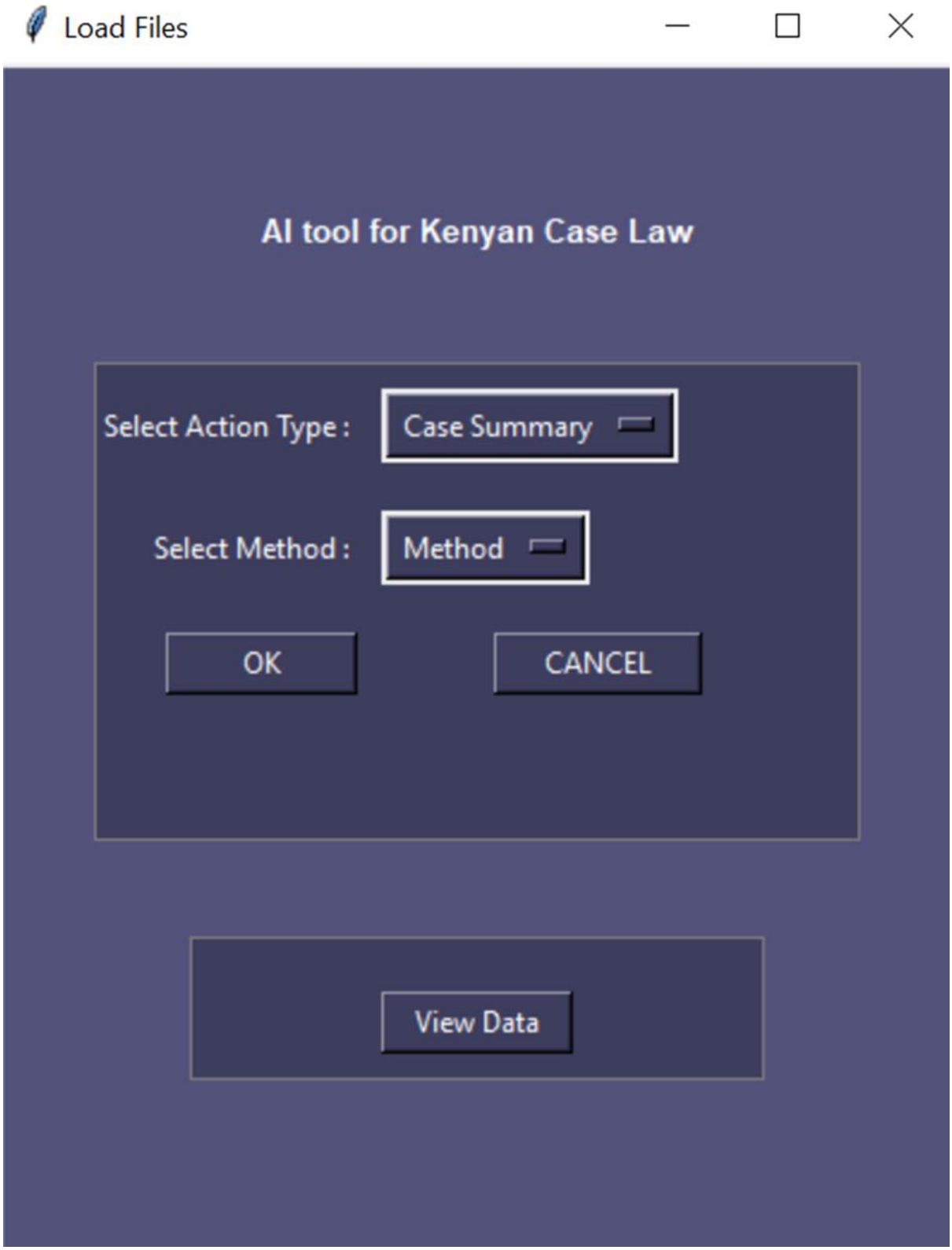
Xiong, Z., Cui, Y., Liu, Z., Zhao, Y., Hu, M., & Hu, J. (2020). Evaluating explorative prediction power of machine learning algorithms for materials discovery using k-fold forward cross-validation. *Computational Materials Science*, *171*, 109203.

Xu, Y., & Goodacre, R. (2018). On splitting training and validation set: a comparative study of cross-validation, bootstrap and systematic sampling for estimating the generalization performance of supervised learning. *Journal of analysis and testing*, *2*(3), 249-262.

Zhong, H., Guo, Z., Tu, C., Xiao, C., Liu, Z., & Sun, M. (2018). Legal judgment prediction via topological learning. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing* (pp. 3540-3549).

APPENDICES

Appendix 1: Toolkit UI



Appendix 2: Keras Model for NLP Category Prediction

```
from keras.models import load_model
from keras.models import Sequential
from keras.layers import Dense, Embedding, GlobalMaxPool1D, Dropout
from tensorflow.keras.optimizers import Adam
from keras.callbacks import ReduceLROnPlateau, EarlyStopping, ModelCheckpoint
import tensorflow as tf

model = Sequential()
model.add(Embedding(max_words, 20, input_length=maxlen))
#model.add(Dropout(0.2))
model.add(GlobalMaxPool1D())
model.add(Dense(num_classes, activation='sigmoid'))

model.compile(optimizer=Adam(0.015), loss='binary_crossentropy', metrics=[tf.keras.metrics.AUC()])
callbacks = [
    ReduceLROnPlateau(),
    #EarlyStopping(patience=10),
    ModelCheckpoint(filepath='model-simple.h5', save_best_only=True)
]


history = model.fit(X_train, y_train,
                    class_weight=class_weight,
                    epochs=30,
                    batch_size=32,
                    validation_split=0.3,
                    callbacks=callbacks)
```