International Journal of Computer Applications Technology and Research
Volume 13–Issue 07, 30 – 34, 2024, ISSN:-2319–8656
DOI:10.7753/IJCATR1307.1004

Design and Implementation of Technology-Assisted Review of Legal Documents With Deep Learning

Nnaemeka .C Onyemelukwe
Department Computer Science
Chukwuemeka Odumegwu Ojukwu University
Anambra State, Nigeria

Ogochukwu C Okeke
Department Computer Science
Chukwuemeka Odumegwu Ojukwu University
Anambra State, Nigeria

Abstract: The research emphasizes the unexplored domain of utilizing artificial intelligence (AI) within the realm of Nigerian law. This is evident through the limited exploration and comprehensibility of AI's influence on legal procedures in the country. Consequently, there is a research gap regarding the effective integration of AI and machine learning (ML)-based systems, such as AILA, into Nigerian legal practices. This integration is crucial for fostering fairness, accountability, improved dynamic contract interpretation, and the appropriate proper integration of standards in digital modalities, which is lacking in the current system. The researcher designed two complementary approaches for legal document retrieval by introducing a Hybrid Model that incorporates semantic as well as implied word-order information

Keywords: Machine Learning, AILA, Hybrid Model and Legal Document

1. INTRODUCTION

In legal systems, discovery is a practice which administrates the right to attain and also has the responsibility to generate any non-relevant matter, relevant to the other party’s defences and claims. eDiscovery tools have helped in enhancing the data collection method and helps in reducing the effort, in terms of reviewing the data. Since the process related to reviewing the documents can be tedious, the legal analytics process is employed by practitioners as it helps in making decisions and assisting legal leaders. Legal analytics consists of legal strategy, financial operations, resource management, and eDiscovery efficiency [1].

Therefore, legal analytics tools assist lawyers in making data-driven decisions, which helps build several legal strategies. One of the legal analytical tools- eDiscovery helps review the data which has been collected and loaded into the storage platform. However, reviewing huge data can be a tedious, time-consuming process and requires lots of costs. Hence accuracy and speed, while reviewing the data can be increased by employing leveraging technology. Therefore EDRM suggested TAR (technology-assisted review) which is considered as a significant tool in eDiscovery. TAR is also known as predictive coding. TAR refers to the document review technique, which influences algorithms to detect and tag documents based on certain keywords and metadata [2].

Due to the requirement of huge labelled datasets and also skilled annotators, several domains are still untouched by deep learning. CUAD (Contract Understanding Atticus Dataset) is a dataset used for legal contract review. Many skilled experts from ‘The Atticus Project’ are involved in the creation of the CUAD dataset which also comprises 13,000 annotations [3].

The pre-processing stage consists of Word2Vec, which was developed by Google in 2013 and helps to process the text data. Word2vec algorithm is made up of 2 learning models such as skip grams and a common bag of word bag [4] Using CBOW, the word can be predicted based on its context and skip-gram is employed to predict the context word for the specific target word. In general, skip-gram is considered to be the reverse of the CBOW algorithm. since the target word is considered to be the input and the context word is the output [5].

Name Entity Recognition (NER) plays the most significant role. NER helps understand the structure of the text data and helps find the relationship between entities [6]. It has been revealed that only 241 documents in the Indonesia dataset have been performed NER, whereas the necessity to implement named entity recognition with the Indonesia dataset is still ongoing since NER provides various advantages such as enhancing the accuracy [7].

There are lots of traditional methods that lack handling large corpus of documents and are a bit time-consuming which is satisfied by the method which involves Word2vec and NER. The pre-processing method is carried out rapidly with enhanced quality in retrieving the information in the dataset.

The objective of the Technology-assisted review is to speed up the process of document reviewing. It can be used in legal documents medical articles etc. It can be accomplished by repeatedly integrating the ML (ML) algorithm and feedback from humans regarding the relevance of the document. However different types of algorithms are used in demonstrating higher performance when compared to the rest of the existing approaches, which helps in detecting and identifying the relevant documents. The suggested study employed a method along with the continuous Active Learning (CAL) algorithm as it is considered to be a non-iterative approach. CALemployed AI, which helped in retrieving the most relevant information present in the document. However, there are some of the challenges faced by the suggested study such as time to terminate the document which is presented to the reviewers, lack of transparency and also lack of efficiency additional costs have to be paid to assess the total number of relevant documents. From the experimental results, it has been identified that the approach, helps in retrieving the relevant documents effectively and also delivers precise obvious and effective stropping points [8].

Due to various advantages of Technology-assisted review, civil litigants in the US (United States) heavily depend on TAR since civil discovery is a process in which the lawsuit can attain evidence from another party. The main objective of civil discovery is to support specific party estimations, claims and defences, which helps litigants to decide, whether to settle a case or not based on the availability of the evidence. And supervised ML framework has been employed by technology-assisted review for the implementation of TAR. In the supervised ML framework, the algorithms understand how to differentiate between Non-responsive documents and responsive documents based on the two criteria, which is the presence or absence of a combination of aspects which includes punctuation, phrases, words, metadata and conceptual clusters.

However, the problems arise over time due to various factors, which include the requirement of the high cost to respond to the score of requests, timely and accurately. The traditional approach was a time-consuming and laborious process, which required a sustainable amount of money and time for investment. The cost of the review does not depend not only on the documents to be reviewed but also on the time taken by the attorney to review several categories of data. Hence the suggested study employs normalizing the data in the pre-processing stage which helps in reducing the time to review the document [9].

Even though some of the studies suggest using ML algorithms to differentiate the relevant documents from non-relevant documents based on training examples. It is coded as non-relevant and relevant by the experts, the suggested study employed systematic rules that help the experts in the decision-making process. Hence technology technology-assisted review process incorporates sampling techniques or even statistical models to conduct the process and also helps in measuring the overall effectiveness of the system [10].

Even though there are modern approaches to classifying documents based on input given by the expert reviewers, there were traditional approaches employed for document classification which include Boolean search, keyword search, manual review etc. On the other hand, modern approaches that employ ML for the classification of documents are denoted as predictive coding in the legal profession. However, it was demonstrated that modern approaches help in providing high recall and precision rate with the involvement of less labour and less time-consuming process in connection with the number of documents a human has to review[11]. Apple and Samsung gathered and handled around 3.59GB of data which is around 11,108,653 documents, which the processing cost was estimated at around 13$ million dollars for 20 month period since the clash for the market share is high due to the availability of highly enhanced techniques for classification of relevant document as quickly as possible.

Several documents to be reviewed in the HRR project can be minimized by employing the TAR process (Technology-assisted review). One of the commonly used workflows for review prioritization is pool-based active learning and iterative-based active learning. Some of the common approaches implemented in supervised learning methods for lexical features and metadata features are linear models which include LR (logistic regression) and SVM (Support Vector Machine). Even the suggested study demonstrated that linear models such as LR and SVM outperformed BERT in legal discovery topics (Jeb Bush email collection) [12].

Employing ML methods for document review has become one of the common practices to reduce time and cost in e-discovery, which is known as TAR. Even though the deep learning algorithm and other conventional ML algorithms have been employed for several tasks such as clustering the documents and text classifications, there is no particular application that deals with sounds, images and video files in documents for document reviewing. Hence, the suggested study employs image classification to identify the images in the legal document and review them. The process of classification images involved, the downloading of the images from the Google image in which the images are classified as positive samples and negative samples. The positive samples consisted of images from the text documents and the negative sample consisted of people, landscapes etc. 20000 images were employed and it was split between 50/50 and the accuracy rate is considered to be above 97.9%. The highest accuracy is obtained due to the ability of VGG16 to capture the critical features that differentiate the images present in the document from other types of images[13].

Classifying thousands of documents can be a tedious process, however suggested study revealed that employing Natural Language Processing (NLP) and ML Technologies (MLT) provided lots of scope for Technology Assisted Review (TAR). Even though human instruction is required to perform the technology-assisted review, including the creation of seed sets and conducting reviews it is still considered a vital part of e-discovery[14].

A semantic type of taxonomy has been suggested in the German civil law domain which consists of 9 diverse types of functional aspects which include permissions, prohibitions duties etc. A rule-based approach has been performed to classify the legal norms by employing a manually labelled dataset. The F1 score was improved constantly from 0.519 to 0.779, however ML approach for classification of documents was implemented and the performance of the F1 score obtained was 0.83. Even though the performance of the ML is higher than other methods, ML classifiers lack transparency in terms of the decision-making process. Hence to examine the behaviour of the classifiers, local linear approximation techniques were implemented [15].

It has been mentioned that, unlike the Western courts, public records of Indian courts are messy, unstructured, chaotic and disorganized. Therefore, big-scale annotated datasets of Indian legal documents do not exist publicly to date. Due to the unavailability of the datasets, room for legal analytical research was restricted. Hence the suggested study employed a dataset which consisted of 10,000 judgements which were delivered by the Supreme Court of India along with the handwritten summaries. The dataset employed was pre-processed by implementing the normalisation technique, which normalized legal abbreviations, and variations in spelling in named entities, handled bad punctuations and tokenization precise sentences. Several attributes such as the names of the defendants, plaintiffs and also names of the people representing them, the name of the judge who gave the judgement and several other attributes were mentioned in the annotated datasets. Apart from this, an automatic labelling approach was implemented in the study to find the sentences which consist of ‘summary-worthy’ information. Some of the applications of the suggested dataset, other than the summarization of legal documents were retrieval of the legal document, analysis of the citations and decisions can also be predicted by the judge who deliver judgement. From the experimental results, it was revealed that the suggested supervised technique outperformed the strong baseline methods [16].

In general, a law practitioner has to go through lots of lengthy documents of several categories, which include legal documents, corruption-related documents, civil-related documents etc., therefore it is vital to summarize the documents and summarized documents should comprise the phrases with intent equal the classification of the case. Hence the suggested study employed a summarization technique, i.e., an intent-based summarization technique called ‘intent metric’, which provided better results along with human valuation when compared to other existing metrics such as ROUGE-L and finally a dataset (Australia data) was curated and annotated the intent phrases in the legal documents [17].

2. PROBLEM STATEMENT

In general, the text utilizes a huge portion of the legal document. Legal documents consist of factors such as contracts, dates, legislative acts, treaties and many more. Legal academia and legal practice spent centuries identifying, analysing, reviewing, commenting reacting and explaining various legal documents.

However, it is practically not feasible to review thousands of documents and find similar document text or name based on category efficiently. Hence, to predict a similar document text or name, effective data pre-processing along with Named Entity Recognition (NER) has to be performed, which can remove the punctuation mark, normalize the word and convert the cases from lowercase to uppercase and vice-versa. Therefore an effective data pre-processing step should be implemented

3. AIM AND OBJECTIVES

The main purpose of the study is to identify the entity of the document using the Named Entity Recognition model. A word embedding algorithm called Word2Vec is used in the study during the data pre-processing stage since it is efficient and works much faster than other existing methods.

To predict a similar document text or name based on the category in the legal documents.

To identify the entity of the legal document using the Named Entity Recognition model.

To evaluate the performance of the model using the cos similarity.

4. RESEARCH GAP

The suggested study is not designed to deal with reviewing a large-scale collection, which is a collection containing millions of documents as it was employed for a small-scale collection of documents since the calculation of variance of R and calculation of mean is feasible. However, splitting the existing documents, running the suggested algorithm and finally concatenating the documents for final review is done. Yet the above-mentioned approach was not considered to be the best, hence more work such as sampling of several non-relevant documents has to be avoided and it can be refrained by providing training to the ranking model universally [8].

As a part of future work, the suggested study will employ various sampling strategies which are specially depicted for neural models which include DAL (Discriminative Active Learning). A document may contain more than 512 tokens, however, the present study lacks in handling those documents, hence in future, a widespread approach will be employed to handle documents which contain more than a certain number of tokens. Since many eDiscovery tasks work upon emails, a transformer model with huge email quantities will benefit, however, the carrying of biases by pre-training quantities into concluding retrieval results in technology-assisted review will be done as future research[12].

.

5. MODELS AND METHODS

Hybrid Model that incorporates semantic as well as implied word order information.

Load the **CUAD Dataset**.

Pre-processing and tokenize the data.

In feature extraction, a series of words or sentences are contained within the numeric vector.

Using Skip Gram model word embedding, we can represent a similar sentence in numbers in a variety of ways. Proposed a new approach that search engines might utilise to locate better-matched contents of documents being retrieved. The suggested method adds a new cosine similarity-based modification

.

5.1 Improved Latent semantic

Is a method of analysing a set of documents to discover statistical co-occurrences of words that appear together which then give insights into the topics of those words and documents. In Improved Latent semantic indexing Modified truncated singular value decomposition (SVD) to identify patterns in the relationships between the terms and concepts contained in an unstructured collection of text. To consider the most important values were taken into consideration, starting from the first singular values up to the desired value. Here we modified Truncated SVD by regularizing the parameters, it will overcome the unwanted character that intrinsically gives the Transformed data that may be difficult to understand and represent our original data set with a much smaller data set. This gives the cosine similarity matrix generating the enhanced feature vectors with optimal performances

To make it easier to extract generic entities (like agreement date, location, or organisation) from natural language texts of domains without generic named entities labelled domain data sets, we use the NER model.

In flow 2 a sample input query text will be given using the “correct match25” model the relevant text will be retrieved.

CUAD Dataset

In the Contract Understanding Atticus Dataset (CUAD), there exist 13,000+ labels included with the legal contracts of 510 that come under the commercial category. The labels and the contracts are being labelled with the supervision of skilled and experienced lawyers. The process of finding the labels involves figuring out the 41 clauses which comprise in the contact review that are found to be significant which consist of transaction corporate which has a connection, acquisitions and mergers. The Atticus project is the one that maintains and organizes the CUAD to enhance the research that deals with the NLP models and to improve the review of the legal contracts.

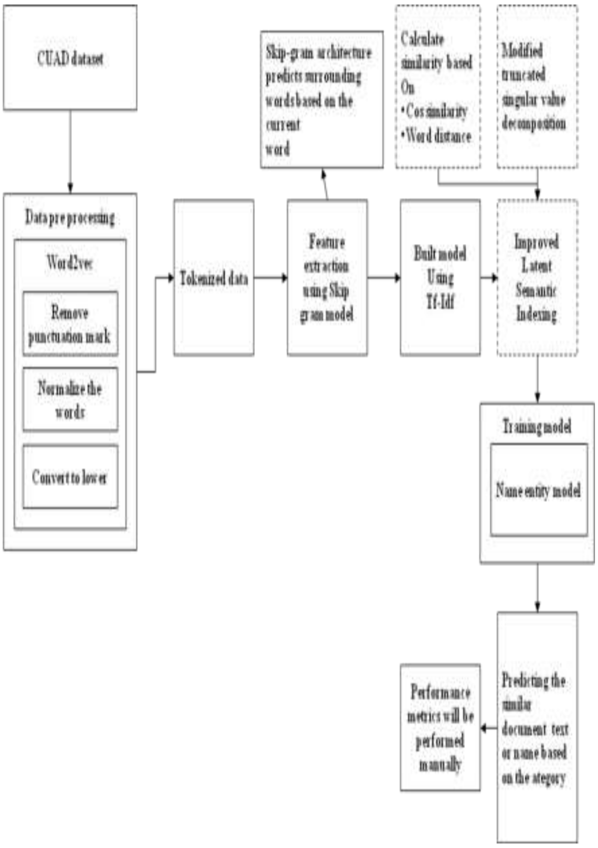
CUAD dataset
Data pre processing
Word2vec
Remove punctuation mark
Normalize the words
Convert to lower
Tokenized data
Skip-gram architecture predicts surrounding words based on the current word
Feature extraction using Skip gram model
Calculate similarity based On
•Cos similarity
•Word distance
Modified truncated singular value decomposition
Built model Using Tf-Idf
Improved Latent Semantic Indexing
Training model
Name entity model
Predicting the similar document text or name based on the ategory
Performance metrics will be performed manually

Figure. 1 Model flow diagram

Figure 1 depicts the overall performance of the method. The CUAD dataset is given as the input, the pre-processing is done by using the Word2vec algorithm. The punctuation marks are removed, the words are normalized and it is converted into the lower case. Feature extraction is performed by using the skip-gram model of the Word2vec. The model is built using the Tf-Idf method. The sentences are embedded based on the cosine similarity and the distance between the words. Semantic similarity is found using this method. For the training process, NER is used to identify the specific texts in the documents or to figure out the entities in the legal documents. The information is being retrieved based on the input. A similar text is being retrieved that depends upon the name in the document

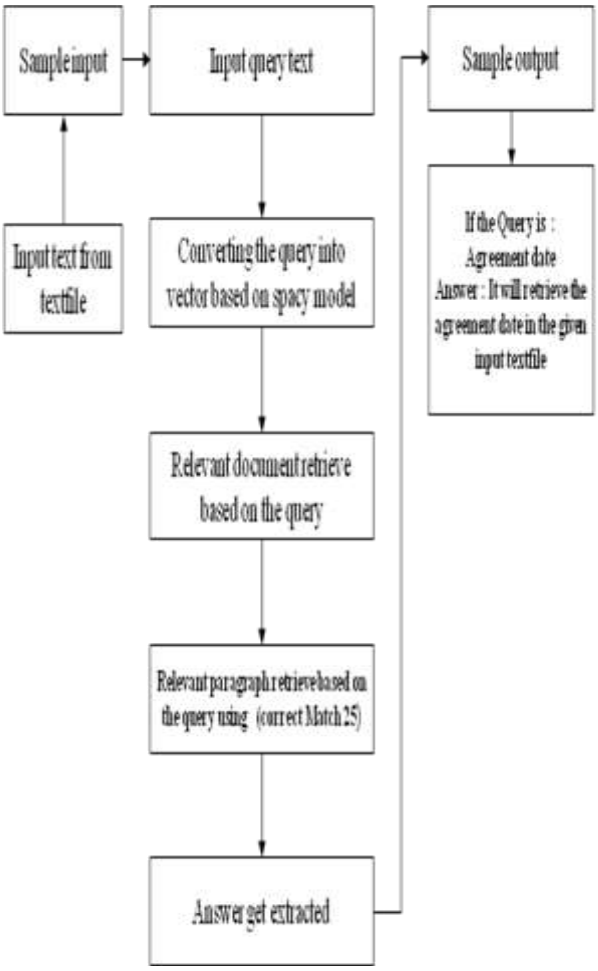
Sample input
Input query text
Sample output
Input text from textfile
Converting the query into vector based on spacy model
If the Query is :
Agreement date
Answer : It will retrieve the agreement date in the given input textfile
Relevant document retrieve based on the query
Relevant paragraph retrieve based on the query using (correct Match 25)
Answer get extracted

Figure 2 Flow of methodology (Agreement Date)

Figure 2 explains that the text file is given as the input. The text is converted into a vector based on the space model. The information is retrieved regarding the input that was given. The relevant paragraph is received by using the correct match 25. In this flow, information to retrieve the agreement date is given as the input. The Agreement is extracted as the output in the specific documents.

6. IMPLEMENTATION DESIGN

The environmental configuration of the system is tabulated in Table 1

Table .1 Environmental Configurations

Hardware Configuration	Software Configuration
CPU - Intel Core i7 – 7700 @ 2.80 GHz	Windows 10
GTX 1050	Python 3.7
16GB RAM	Anaconda Spyder

7. FINDINGS

Technology-Assisted Review (TAR) which indulges in reviewing legal documents is mainly processed for the retrieval of specific information that is very significant. The cost and the time are reduced by using the technologies for reviewing the legal documents. This increases the effectiveness of the large set of collections. The TAR is utilized in a wide range of applications to discover the information in legal documents, and literature review in the field of medicine, for organizing the collection of the evaluation.

The method of TAR outperforms the varying technologies that are used in detecting the information in legal documents that are ubiquitous. The CUAD dataset is taken and Word2vec a word-embedding algorithm is used for pre-processing the text in the legal documents as well as the skip-gram models support to figure out the similar sentences in the documents. The Tf-Idf model figures out the cosine similarity and the distance between the words in the documents. The NER, which is the form of the NLP model, is used to identify the entities in the text file that are given as input to it. A similar text is being found by these two pre-processing methods in this study.

8. CONCLUSION

The study explains employing the NER Model to detect similar document text or name based on the category. The entity of the legal document can be identified using the NER Model and the performance of the model using the cos similarity. The dataset implemented in the study is CUAD (contract understanding Atticus dataset) as it contains more than 12,999+ labels in 510 commercial legal contracts. Word2Vec and NER Model algorithms are used in the data pre-processing stage for an effective pre-processing process to detect similar text in the legal documents based on the names of the category and detecting the agreement date from the input text file which is expected to be the outcome.

9. ACKNOWLEDGMENTS

I wish to show my profound gratitude to Almighty God who gave me the insight and grace to carry out this thesis work notwithstanding the pressure at my place of work. To my Head of the Department Prof. Mgbeafulike and Post Graduate Coordinator Prof. Okeke O.C thank you for your guidance

I want to also thank my wife Mrs. Onyemelukwe Lilian Chigozie for her love and prayers my father for his financial support, and also my sibling and mother-in-law for their remarkable support during this research work.

.

10. REFERENCES

[1] A. Mullick, A. Nandy, M. N. Kapadnis, S. Patnaik, R. Raghav, and R. Kar, "An evaluation framework for legal document summarization,"*arXiv preprint arXiv:2205.08478,* 2022.

[2] B. Jang, I. Kim, and J. W. Kim, "Word2vec convolutional neural networks for classification of news articles and tweets, "*PloS one,* vol. 14, p. e0220976, 2019.

[3] B. Waltl, G. Bonczek, E. Scepankova, and F. Matthes, "Semantic types of legal norms in German laws: classification and analysis using local linear explanations, "*Artificial Intelligence and Law,* vol. 27, pp. 43-71, 2019.

[4] C. S. Patrons, "DISCOVERY PROPORTIONALITY MODEL A NEW FRAMEWORK," 2021

[5] D. Dayma, "Law Centre, Faculty of Law, University of Delhi, India, "*Cyber Crime, Regulation and Security: Contemporary Issues and Challenges,* p. 91, 2022

[6] D. Hendrycks, C. Burns, A. Chen, and S. Ball, "Cuad: An expert-annotated nlp dataset for legal contract review,"*arXiv preprint arXiv:2103.06268,* 2021.

[7] D. Li and E. Kanoulas, "When to stop reviewing in technology-assisted reviews: Sampling from an adaptive distribution to estimate residual relevant documents, "*ACM Transactions on Information Systems (TOIS),* vol. 38, pp. 1-36, 2020.

[8] E. Yang, S. MacAvaney, D. D. Lewis, and O. Frieder, "Goldilocks: Just-Right Tuning of BERT for Technology-Assisted Review,"*arXiv e-prints,* p. arXiv: 2105.01044, 2021.

[9] I. Budi and R. R. Suryono, "Application of named entity recognition method for Indonesian datasets: a review, "*Bulletin of Electrical Engineering and Informatics,* vol. 12, pp. 969-978, 2023.

[10] J. Cheng, J. Liu, X. Xu, D. Xia, L. Liu, and V. S. Sheng, "A review of Chinese named entity recognition, "*KSII Transactions on Internet & Information Systems,* vol. 15, 2021.

[11] M. Abdolahi and M. Zahedi, "A new method for sentence vector normalization using word2vec," *International Journal of Nonlinear Analysis and Applications,* vol. 10, pp. 87-96, 2019.

[12] N. Huber-Fliflet, F. Wei, H. Zhao, H. Qin, S. Ye, and A. Tsang, "Image Analytics for Legal Document Review: A Transfer Learning Approach,"*arXiv e-prints,* p. arXiv: 1912.12169, 2019.

[13] P. W. Grimm, M. R. Grossman, and G. V. Cormack, "Artificial intelligence as evidence, "*Nw. J. Tech. & Intell. Prop.,* vol. 19, p. 9, 2021.

[14] R. Dale, "Law and word order: NLP in legal tech, "*Natural Language Engineering,* vol. 25, pp. 211-217, 2019.

[15] R. Wang, "Legal technology in the contemporary USA and China," 2020.

[16] S. Krishnan, N. Shashidhar, C. Varol, and A. R. Islam, "Evidence Data Preprocessing for Forensic and Legal Analytics, "*Int. J. Comput. Linguist. (IJCL),* vol. 12, p. 24, 2021.

[17] V. Parikh, V. Mathur, P. Mehta, N. Mittal, and P. Majumder, "Lawsum: A weakly-supervised approach for Indian legal document summarization,"*arXiv preprint arXiv:2110.01188,* 2021.