Classification of Insurance Legal Documents Using Natural Language Processing (NLP)

Abstract

This study explores the use of NLP to analyze and classify legal documents, aiming to build an intelligent system that automates legal text organization and enhances workflow efficiency. The novel approach outlines the motivation, methodology, experiments, and results of the proposed approach, emphasizing the importance of a thorough evaluation and the insights gained from the experimentation process.

Keywords: *Legal Document Classification (fraud, misrepresentation, non-disclosure, claims procedure/handling, indemnity), NLP, Machine Learning, Data Pre-processing, Performance Analysis, Compliance.*

1. Introduction

Legal document analysis is an essential aspect of legal practice but typically requires substantial time and human effort. This study introduces an innovative approach to automating this process through advanced Natural Language Processing (NLP) techniques. The motivation arises from the growing volume of legal documents and the need to improve efficiency and accuracy in their management.

Legal practitioners increasingly face large collections of documents such as contracts, case law, legal opinions, and statutes creating a pressing demand for automated tools to streamline the review process. Traditional methods often fall short, leading to delays and potential omissions. The proposed solution utilizes NLP to develop an intelligent system capable of understanding, classifying, and extracting meaningful insights from diverse legal texts.

2. Proposed Methodology

This section provides a detailed overview of the methodology, emphasizing the practical implementation of the proposed solution. It examines the selection of specific NLP techniques along with the rationale behind these choices. Additionally, it discusses the preprocessing strategies employed to ensure the model's adaptability to diverse legal document formats.

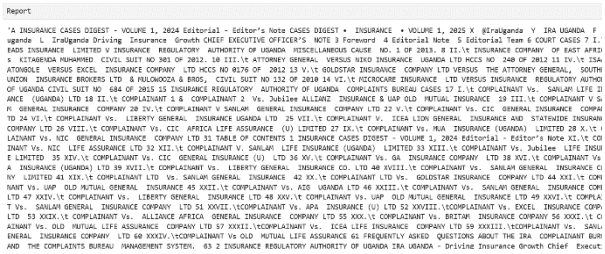
2.1. Data Collection

To train and evaluate the model, reference is made to the INSURANCE CASES DIGEST VOLUME 1, 2025 obtained from the Insurance Regulatory Authority of Uganda, a body instituted with the primary objective of promoting and facilitating the maintenance of a sound, effective, fair, transparent and stable insurance sector in Uganda.

The Insurance Cases Digest contains a broad collection of legal documents,

including court judgments, legal opinions, and statutes. The Digest informs readers on contemporary insurance issues, offering insights into the Authority's mandate, competition and consumer protection law and jurisprudence development. It serves as a quick reference of landmark cases from courts, the Insurance Appeals Tribunal and the Complaints Bureau that have shaped insurance law and practice.

Its selection was motivated by the richness and variety of its content, providing a representative sample of the linguistic and structural variations found in legal texts. An overview of the data is presented below.


Report

'A INSURANCE CASES DIGEST - VOLUME 1, 2024 Editorial - Editor's Note CASES DIGEST • INSURANCE • VOLUME 1, 2025 X @IraUganda Y IRA UGANDA F
ugandа L IraUganda Driving Insurance Growth CHIEF EXECUTIVE OFFICER'S NOTE 3 Foreword 4 Editorial Note 5 Editorial Team 6 COURT CASES 7 I.
BADS INSURANCE LIMITED V INSURANCE REGULATORY AUTHORITY OF UGANDA MISCELLANEOUS CAUSE NO. 1 OF 2013. 8 II.\t INSURANCE COMPANY OF EAST AFRI
s KITAGENDA MUHAMMED CIVIL SUIT NO 381 OF 2012. 10 III.\t ATTORNEY GENERAL VERSUS NIKO INSURANCE UGANDA LTD HCCS NO 240 OF 2012 11 IV.\t ISA
ATONGOLE VERSUS EXCEL INSURANCE COMPANY LTD HCCS NO 0176 OF 2012 13 V.\t GOLDSTAR INSURANCE COMPANY LTD VERSUS THE ATTORNEY GENERAL, SOUTH
UNION INSURANCE BROKERS LTD & MULONGOZA & BROS, CIVIL SUIT NO 132 OF 2010 14 VI.\t MICROCARE INSURANCE LTD VERSUS INSURANCE REGULATORY AUTHO
OF UGANDA CIVIL SUIT NO 684 OF 2015 15 INSURANCE REGULATORY AUTHORITY OF UGANDA COMPLAINTS BUREAU CASES 17 I.\t COMPLAINANT Vs. SANLAM LIFE I
ANCE (UGANDA) LTD 18 II.\t COMPLAINANT 1 & COMPLAINANT 2 Vs. Jubilee ALLIANZ INSURANCE & UAP OLD MUTUAL INSURANCE 19 III.\t COMPLAINANT V S
M GENERAL INSURANCE COMPANY 20 IV.\t COMPLAINANT V SANLAM GENERAL INSURANCE COMPANY LTD 22 V.\t COMPLAINANT Vs. CIC GENERAL INSURANCE COMPA
TD 24 VI.\t COMPLAINANT Vs. LIBERTY GENERAL INSURANCE UGANDA LTD 25 VII.\t COMPLAINANT V. ICEA LION GENERAL INSURANCE AND STATEWIDE INSURAN
COMPANY LTD 26 VIII.\t COMPLAINANT Vs. CIC AFRICA LIFE ASSURANCE (U) LIMITED 27 IX.\t COMPLAINANT Vs. MUA INSURANCE (UGANDA) LIMITED 28 X.\t
LAINANT Vs. NIC GENERAL INSURANCE COMPANY LTD 31 TABLE OF CONTENTS 1 INSURANCE CASES DIGEST - VOLUME 1, 2024 Editorial - Editor's Note XI.\t CO
INANT Vs. NIC LIFE ASSURANCE LTD 32 XII.\t COMPLAINANT V. SANLAM LIFE INSURANCE (UGANDA) LIMITED 33 XIII.\t COMPLAINANT Vs. Jubilee LIFE INSU
E LIMITED 35 XIV.\t COMPLAINANT Vs. CIC GENERAL INSURANCE (U) LTD 36 XV.\t COMPLAINANT Vs. GA INSURANCE COMPANY LTD 38 XVI.\t COMPLAINANT Vs
A INSURANCE (UGANDA) LTD 39 XVII.\t COMPLAINANT Vs. LIBERTY GENERAL INSURANCE CO. LTD 40 XVIII.\t COMPLAINANT Vs. SANLAM GENERAL INSURANCE C
NY LIMITED 41 XIX.\t COMPLAINANT LTD Vs. SANLAM GENERAL INSURANCE 42 XX.\t COMPLAINANT LTD Vs. GOLDSTAR INSURANCE COMPANY LTD 44 XXI.\t COM
NANT Vs. UAP OLD MUTUAL GENERAL INSURANCE 45 XXII.\t COMPLAINANT Vs. ADS UGANDA LTD 46 XXIII.\t COMPLAINANT Vs. SANLAM GENERAL INSURANCE COM
LTD 47 XXIV.\t COMPLAINANT Vs. LIBERTY GENERAL INSURANCE LTD 48 XXV.\t COMPLAINANT Vs. UAP OLD MUTUAL GENERAL INSURANCE LTD 49 XXVI.\t COMPLA
T Vs. SANLAM GENERAL INSURANCE COMPANY LTD 51 XXVII.\tCOMPLAINANT Vs. APA INSURANCE (U) LTD 52 XXVIII.\tCOMPLAINANT Vs. EXCEL INSURANCE COMP
LTD 53 XXIX.\t COMPLAINANT Vs. ALLIANCE AFRICA GENERAL INSURANCE COMPANY LTD 55 XXX.\t COMPLAINANT Vs. BRITAM INSURANCE COMPANY 56 XXXI.\t C
AINANT Vs. OLD MUTUAL LIFE ASSURANCE COMPANY LTD 57 XXXII.\tCOMPLAINANT Vs. ICEA LIFE INSURANCE COMPANY LTD 59 XXXIII.\tCOMPLAINANT Vs. SANL
ENERAL INSURANCE COMPANY LTD 60 XXXIV.\tCOMPLAINANT Vs OLD MUTUAL LIFE ASSURANCE 61 FREQUENTLY ASKED QUESTIONS ABOUT THE IRA COMPLAINANT BUR
AND THE COMPLAINTS BUREAU MANAGEMENT SYSTEM. 63 2 INSURANCE REGULATORY AUTHORITY OF UGANDA IRA UGANDA - Driving Insurance Growth Chief Executi

2.2 Data Cleaning

Prior to model development, the data underwent a thorough cleaning process to ensure its quality, consistency and reliability.

a. Unique Character Analysis

Analysis of legal texts identified special characters, symbols and formatting elements that lacked semantic value. By removing or encoding unique characters, the focus shifted to linguistic content. Examples include links, numbers and miscellaneous symbols.

	clean_text	text
0	insurance cases digest volume 1 2024 editorial...	A INSURANCE CASES DIGEST - VOLUME 1, 2024 Edit...
1	leads insurance limited v insurance regulatory...	LEADS INSURANCE LIMITED V INSURANCE REGULATO...
2	insurance company east africa vs kitagenda muh...	INSURANCE COMPANY OF EAST AFRICA vs KITAGEND...

2.3. Pre-processing for NLP

To enhance the model's adaptability to various legal document formats, a robust preprocessing pipeline was implemented.

a. Tokenization

The legal texts were tokenized into words (smaller units) to support NLP analysis, with the tokenisation strategy considering the specialized terms and phrases that carry significant legal meaning.

b. Stopword Removal

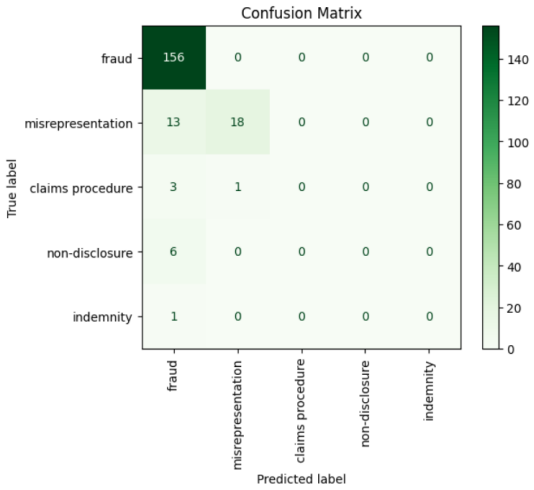
Common legal stopwords that add little meaning were identified and removed to reduce noise and help the model focus on substantive legal content.

c. Lemmatization

Lemmatization was applied to reduce legal terms to their base forms, ensuring uniform treatment of different inflections and improving the accuracy of legal language analysis.

3. Model Evaluation

The model adopts a systematic approach to model selection using a Logistic Regression classifier. A pipeline structure streamlines both preprocessing and modelling processes, yielding a performance score of 87.87%. Among the analysed categories, “claims procedure” demonstrated the strongest predictive performance. To enhance interpretability, the model leveraged Eli5, Lime, and SHAP, which consistently identified influential features such as claims, payment, payable, delays, premium and settlement. These insights indicate that the Digest captured recurring discrepancies in insurers’ claims procedures, particularly cases where premiums had been paid but claim settlement was delayed. The findings further showed that **fraud** was the leading contributing factor to these delays, followed by **misrepresentation**.



4. Discussion / Insights

	precision	recall	f1-score	support
Miscellaneuous	0.87	1.00	0.93	156
claims procedure	0.95	0.58	0.72	31
fraud	0.00	0.00	0.00	4
indemnity	0.00	0.00	0.00	6
misrepresentation	0.00	0.00	0.00	1
accuracy			0.88	198
macro avg	0.36	0.32	0.33	198
weighted avg	0.83	0.88	0.85	198

4.1. Precision, Recall, and F1-Score

Precision: Reflects the accuracy of positive predictions. For instance, the model achieves high precision for 'claims procedure' (0.95), indicating that when it predicts this class, it is usually correct.

However, some classes like ' misrepresentation, fraud & indemnity' (0.00) have a very low precision. Recall: Represents the model's ability to capture all positive instances. High recall values, such as for 'Miscellaneous ' (1.00), indicate effective identification of true positives. The implication is that the model is over-predicting certain classes while failing to discriminate accurately between closely related legal themes.

However, classes like 'misrepresentation', ‘indemnity’ and 'fraud' have a recall of 0.00, suggesting that the model struggles to identify instances of these classes.

F1-Score: The harmonic means of precision and recall. It provides a balanced measure of a model's overall performance. High F1 scores are observed for 'Miscellaneous' (0.93) and 'claims procedure' (0.72).

While the model demonstrates high precision and recall for some classes. The overall accuracy is 88%, indicating the proportion of correctly classified instances.

However, the macro and weighted averages for precision, recall, and F1-score suggest that the model's performance greatly varied, emphasizing the need for further investigation, especially in addressing class imbalances and improving classification for certain classes. The report serves as a valuable tool for understanding the model's strengths and weaknesses, guiding potential refinements for enhanced performance in legal document classification.

y=Miscellaneous top features		**y=claims procedure** top features		**y=fraud** top features		**y=indemnity** top features		**y=misrepresentation** top features		**y=non-disclosure** top features	
Weight[?]	**Feature**	**Weight[?]**	**Feature**	**Weight[?]**	**Feature**	**Weight[?]**	**Feature**	**Weight[?]**	**Feature**	**Weight[?]**	**Feature**
+3.125	<BIAS>	+5.359	claim	+3.053	fraud	+1.733	payment	+1.992	misrepresentation	+0.429	nondisclosure
+1.043	vs	+2.987	claims	+1.574	fraudulent	+0.860	entire	+1.645	ambiguous	+0.391	denied
+0.809	uganda	+1.359	payable	+1.109	employees	+0.848	premium	+1.090	vague	+0.391	grounds
... 1568 more positive ...		+1.343	claimed	+0.802	financial	+0.815	nonpayment	+1.064	wording	+0.337	liberty
... 814 more negative ...		+1.066	<BIAS>	+0.762	bond	+0.802	balance	+0.742	defendant	+0.330	liability
-0.711	vague	+0.869	settle	+0.744	dishonesty	+0.671	payments	+0.618	unambiguous	+0.319	breach
-0.760	insured	+0.805	delays	+0.711	performance	+0.617	loan	+0.615	provisions	+0.319	preexisting
-0.771	pay	+0.751	stage	+0.583	warehouse	+0.610	insured	+0.549	customers	+0.319	workmanship
-0.789	policy	+0.681	loss	+0.536	prove	+0.586	357742448	+0.527	case	+0.300	factors
-0.874	employees	+0.638	repudiated	+0.526	fraudulently	+0.546	lapsed	+0.502	action	+0.300	disclosed
-0.896	payments	+0.609	insurer	+0.514	defendant	+0.529	missed	+0.482	misselling	+0.285	collapse
-0.948	respondent	+0.572	bancassurance	+0.511	employee	+0.516	contract	+0.477	clear	+0.274	inception
-0.976	ambiguous	+0.572	accountable	+0.504	alleged	+0.502	agreement	+0.447	cause	+0.267	poor
-0.994	fraudulent	+0.552	cancellation	+0.503	gain	+0.491	directed	+0.437	plaintiff	+0.254	effect
-1.021	loss	+0.522	reported	+0.492	activities	+0.474	respondent	+0.406	road	+0.254	contract

Conclusion

The model's key insight is that classification is driven not only by positive indicators (i.e., 'fraud' for fraud cases, "claim" for claims procedure) but also by negative weights that capture the absence of terms associated with other categories, thus highlighting the distinctive language patterns of each class.

END