Proceedings of the International Conference on Sustainable Communication Networks and Application (ICSCN-2025)
DVD Part Number: CFP25DW8-DVD; ISBN: 979-8-3315-9419-0

Analysis of Legal Text using Juxtaposed Pre-Processing for Effective Feature Extraction

1st Anjali V
Research Scholar
Department of Computing Sciences
Vels Institute of Science, Technology & Advanced Studies
Chennai, India
rushaugust21@gmail.com

2nd Dr. N. Shyamala Devi
Assistant Professor
Department of Information Technology
Vels Institute of Science, Technology & Advanced Studies
Chennai, India
shyamadevi@gmail.com

***Abstract*— Legal text processing is a domain that holds an unexplored path, due to the various confidentialities and constraints it may involve. Nonetheless, it is one of the domineering verticals that has methodically evaded technological integration primarily to ensure ethical and security concerns of the data involved. However, while there are still considerations, the need for automated legal text mining has accelerated informed litigative decisions, thereby augmenting the accuracy of rulings, verdicts and enabled refined results in multifarious ways. This paper pivots on accentuating the prominence of pre-processing through the juxtaposed results from two techniques such as the Bag-of-Words (BoW) and Term Frequency - Inverse Document Frequency (TF-IDF) effectuated prior and post pre-processing. Inorder to identify the accuracy of legal text mining, the assiduous steps involved in preprocessing delved in this study entail lowercasing, removal of numbers and punctuation, stopword removal, and stemming. Furthermore, it explores two key feature extraction methods: Bag of Words (BoW) and Term Frequency-Inverse Document Frequency (TF-IDF). Through visualization and analysis, this study demonstrates the impact of preprocessing on feature extraction, and further accelerate the efficacy of legal text mining. The simulations for this indagation are carried out in python, and the results are successfully obtained.**

Keywords—Natural Language Processing, Feature Extraction, Bag-of-Words (BoW), Term Frequency - Inverse Document Frequency (TF-IDF), Preprocessing, Lowercasing, Special Character Removal, Stemming, Legal NLP

I. INTRODUCTION

Legal institutions and advocate organizations in India hold a high number of caseloads and is constantly in need of automated services to increase their adeptness of decision-making, and swifter adjudication. The colossal impact of NLP in neoteric domains have promoted its application in crucial domains such as legal text mining, thereby advocating the profuse burgeon of documentation in various litigative realms [1]. NLP is instituted as an evolving tool to automate challenging litigative documentations, interpreting records and in better management of informed decisions. While the efficacy of text mining largely relies on punctilious effective processes relative to the domain, legal text mining requires domain- contingent processing pipeline that can augment the accuracy of processing. The automated processing using NLP especially with the involved complex legal tasks, domain-specific terminologies and statutory references has proved to redefine the nuances of litigative filing and decision-making. Nonetheless, meticulous and well-processed NLP pipelines can aid in establishing resilient automated systems especially with crucial realms like the legal systems. Natural Language processing techniques can aid in establishing efficacy and better accessibility of legal data, thereby emphasizing the indispensable role that they hold in legal dossier [2].

This paper delves into the methodologies and implications of preprocessing legal texts, underscoring its indispensable role in unlocking the full potential of NLP in legal analytics. The utilization of Natural Language Processing (NLP) in the legal domain through a meticulous pre-processing, and hierarchy-based application unsheathes the vital information from a gargantuan volume of data. This extraction of crucial information renders better clarity of decisions, and provides an unambiguous comprehension between the contextual relationship that might exists between data. The integration of NLP facilitates tasks such as legal document summarization, contract analysis, automated document generation, and legal research, thereby enhancing the efficiency and accuracy of legal processes [4]. The creation of hierarchical structures within documents, language translation, sentiment analysis, and compliance monitoring further contribute to the comprehensive utilization of NLP in the legal field. The impact of NLP in streamlining legal workflows, augmented information extraction models [3], to ensure compliance with evolving legal standards addresses the issues of time-consuming data retrieval, and aids in specifically pivoting on the need for pre-processing techniques to procure the zenith of accuracy in legal text mining. Preprocessing plays a pivotal role in normalizing the documents by converting raw, unstructured data into a structured format suitable for computational analysis. Common preprocessing techniques such as tokenization, stop-word removal, lowercasing, stemming, and punctuation elimination are adapted and refined to suit the nuances of legal language. The process involves pre-processing the text, conducting sentiment analysis to label sentences as positive, negative, or neutral, and then summarizing the document while incorporating the identified sentiments [6]. Special attention is paid to legal-specific considerations such as terminology and context.

This research explores the integration of preprocessing techniques as the foundation for hierarchy-based NLP applications in legal text mining. By cleaning and preparing legal corpora, it becomes feasible to extract vital information from vast volumes of legal documents, thereby facilitating downstream tasks such as entity recognition, document summarization, sentiment analysis, and legal compliance monitoring. Moreover, structured representations generated through preprocessing pave the way for constructing hierarchical knowledge models that reflect the logical and procedural flow within legal frameworks [5]. This paper is

structured with section II elaborating the literature review pertaining the study, section III delineates the methodologies used, and section IV and V elucidates the results procured and concludes the indagation respectively.

II. Literature Review

Olha Kovalchuk et al [1] delineated in the study about text mining in the legal domain. Pivoted on utilizing a Decision tree model using the Chi-square Automatic Interaction Detector (CHAID) technique for stratifying legal data. Real-time data for the study was garnered to for this study to effectuate automated legal text stratification. The study effectively showcased the decision tree model for criminal proceedings document, and further highlighted the best predictors using the X-Square criterion. Furthermore, text classification was elaborated through machine learning models. Another vital aspect of this study was the use of stratification through text documents through Positive Example Based Learning, while establishing topic modelling to determine the thematic belonging of text documents in document collections. Web crawling method was used as a part of model building. The research procured an overall accuracy of 95% for criminal proceeding document, and 85.5% for non-criminal proceeding documents.

Ashwini V. Zadgaonkar [3], et al in the paper titled “Legal Text Mining and Analysis Based on Artificial Intelligence”, delineates their objective to build a legal recommendation system based on the litigative documents procured as real-time data. The study is trifurcated into aspects of data collection, text expression representation and recommendation augmentation for legal argument mining. The simulative cognizance observed through this study delineates techniques from various researches ranging from LLMs to legal NER, BERT- based models, along with LDS and LTC to further improvise the usability of legal provisions. Challenges pertaining to this study entail the utilization of resilient pre-processing techniques, the ability to handle multilingual functionalities, and combining ontological models that can accelerate the predictive accuracy and overall processing performance.

III. Proposed Methodology

Legal text mining using NLP is a less traversed realm of research, but is swiftly catching pace inorder to regularize faster assessing and text pattern extraction in documents [7]. However, it entails several phases which require meticulous attention that need to be carried out in a punctilious manner. Some of the potential areas of research that NLP can be integrated for legal excellence is depicted through the figure (Fig 1) below:

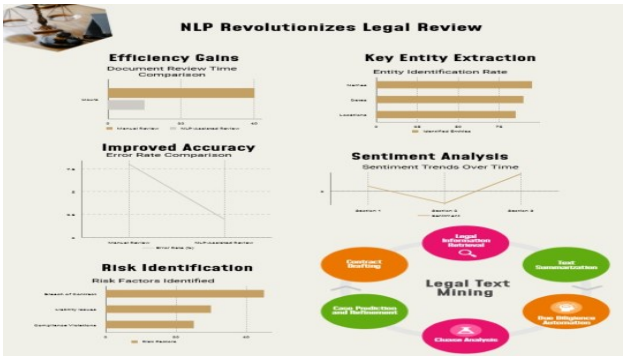

Fig. 1. NLP in Legal Domain Processing

The above figure illustrates the various uses of NLP in legal text processing and mining, and developed using pictochart.ai. Consecutively, a generic outline of the various steps implemented for text mining and extraction is illustrated in the below Fig 2.

Fig. 2. Generic Steps using NLP

This paper elaborates about the emphasis of pre-processing on legal datasets, comprising of a set of refining techniques that can aid in augmenting the accuracy of further phases, while showcasing the need for constant development through continuous refinement. The below figure (Fig 3) elaborates the various steps in pre-processing effectuated on the procured dataset. The dataset comprises of data obtained from Kaggle. The step-wise implementation of the below phases holds an impact on the processing accuracy, while further ensuring zenith of predictive performance.

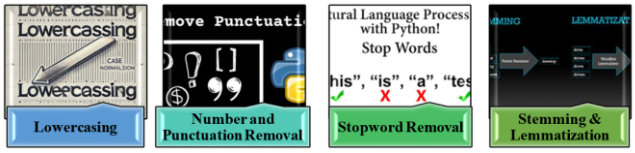

Fig. 3. Processes in Preprocessing

The process of lowercasing entails standardizing the procured dataset, inorder to evade mismatch of data with respect to lower and uppercase [12]. This process enables the unwarranted bias that might crop with the gathered data, thereby significantly mitigating complexity while incorporating semantic synonymity to the dataset. Lowercasing also by large augments the accuracy of further processing by distinctively applying redundant purging.

The subsequent phase post lowercasing entails removing numbers and punctuations, which effectuates data cleaning by removing the characters that may be irrelevant or not needed [8]. These characters may include all forms of punctuations that further helps in accelerating the accuracy of processing.

Furthermore, to augment predictive and classification accuracy, the inclusion of stopword purging is effectuated for the incorporated dataset. The rationale behind stopword purging is to remove all those words that may hold insignificance in the comprehensive entirety of the language [9], thereby expediting text processing to the pinnacle of processing accuracy [14]. However, the stopwords may differ from one dataset to another, and removing them in this research entails the use of Natural Language Toolkit in the incorporated dataset.

Stemming is implemented as the final phase of pre-processing, and is dedicated to represent the language in its most basic form [15]. The process entails condensing the word to its root, thereby to normalize and standardize words [10]. This reduction of the derivate words in a language database can enhance the process of information unsheathing to further equate them to structurally correlated words. This

equivalence pattern extraction can accelerate accuracy of language processing, while ensuring to reduce process complexity.

The next phase after pre-processing incorporates the process of feature extraction. In this research, feature extraction [11] is executed using two algorithms such as Bag of Words (BoW) [10] and Term Frequency-Inverse Document Frequency (TF-IDF). This paper pivots to demonstrate the effect of pre-processing on the data through the implementation of feature extraction prior and after the phases of pre-processing lineated above. BoW and TF-IDF are used inorder to cerate a corpus, with each document represented as a word vector. This technique is used with an objective to correspond every unique word as a feature. Thereby, establishing the feature count as the word frequency from the constructed document.

The TF-IDF on the other hand, overcomes the limitations of BoW interms of semantic context limitation and vulnerability to overfitting, and implements feature extraction by measuring the term weightage in correlation with the cruciality of words in the document relative to the entire corpus [13]. Two consecutive phases are designed with TF-IDF, with the frequency of words taken into account in the first phase, and the logarithmic inverse fractioning of the terms in the document implemented in the second. The TF-IDF technique refines the accuracy of language processing interms of their impact on the commonly occurring words, while instituting the priority of terms in the document, thereby accentuating the process of feature extraction than the former technique. The results of the delineated feature extraction techniques are established prior and after pre-processing, to comprehend the impact of pre- processing on the incorporated database. Furthermore, this juxtapose of pre-processing with feature extraction elucidates the accuracy analysis of language processing, while cognizing the vitality of pre-processing steps for any entailed dataset.

The post-analysis phase of this research focuses to showcase the results of implementation through a Principal Component Analysis (PCA) [16] method to emphasize on the language processing change and accuracy of processing. The results of this analysis are also delineated in the subsequent section.

IV. Results

This section elaborates the various results procured from the implementation of pre- processing and feature extraction. Legal text data is obtained from Kaggle, and when observed the data comprises of gargantuan quantity of unstandardized terms, thereby requiring a mandated normalization process to annihilate the redundancies and noise from the data. A screenshot of the raw data is shown in the figure below:The preprocessing and feature extraction was implemented in Python and applied to frames extracted from real and deepfake video datasets. The results of these implementations are presented and discussed below, with corresponding visualizations in Figures 4 to 7.

	case_id	case_outcome	case_title	case_text
0	Case1	cited	Alpine Hardwood (Aust) Pty Ltd v Hardys Pty Lt...	Ordinarily that discretion will be exercised s...
1	Case2	cited	Black v Lipovac [1998] FCA 699 ; (1998) 217 AL...	The general principles governing the exercise ...
2	Case3	cited	Colgate Palmolive Co v Cussons Pty Ltd (1993) ...	Ordinarily that discretion will be exercised s...
3	Case4	cited	Dais Studio Pty Ltd v Bullett Creative Pty Ltd...	The general principles governing the exercise ...
4	Case5	cited	Dr Martens Australia Pty Ltd v Figgins Holding...	The preceding general principles inform the ex...
5	Case6	cited	GEC Marconi Systems Pty Ltd v BHP Information ...	I accept that the making of a rolled up offer ...
6	Case7	cited	John S Hayes & Associates Pty Ltd v Kimber...	The preceding general principles inform the ex...
7	Case8	cited	Seven Network Limited v News Limited (2007) 24...	On the question of the level of unreasonablene...
8	Case9	applied	Australian Broadcasting Corporation v O'Neill ...	recent decision of the High Court in Australia...
9	Case10	followed	Hexal Australia Pty Ltd v Roche Therapeutics I...	Hexal Australia Pty Ltd v Roche Therapeutics I...
10	Case11	cited	Castlemaine Tooheys Ltd v South Australia [198...	Hexal Australia Pty Ltd v Roche Therapeutics I...
11	Case12	cited	R v McFarlane; Ex parte O'Flannagan and O'Kell...	quia timet proceedings, the court will have re...
12	Case13	followed	National Australia Bank v KDS Construction Ser...	It was not suggested in this proceeding that, ...
13	Case14	followed	George v Cluning (1979) 53 ALJR 767 (note)	Strictly speaking, a cheque, even a bank chequ...
14	Case15	followed	Australian Mid-Eastern Club Limited v Yassim (...	None of this is to suggest that the Deputy Com...

Fig. 4. Raw Data Prior to Pre-processing

The legal text data thus garnered is then effectuated into the feature extraction processes of BoW and TF-IDF prior to pre-processing, and the results thus obtained are illustrated in Fig 5 and Fig 6 respectively. The PCA analysis is also demonstrated prior to pre-processing inorder to demonstrate the language processing change that can be observed prior and post pre- processing phase. The result of the PCA implementation before pre-processing is demonstrated in Fig 7.

Bag-of-Words Feature Matrix:

	and	at	be	for	in	is	of	that	the	to
0	2	2	1	0	1	0	0	2	2	1
1	0	1	3	3	6	2	9	5	16	6
2	2	2	1	0	1	0	0	2	2	1
3	0	1	3	3	6	2	9	5	16	6
4	1	2	3	0	2	2	4	2	9	4
5	3	2	6	1	5	2	8	7	18	11
6	1	2	3	0	2	2	4	2	9	4
7	7	3	5	1	16	6	29	20	39	17
8	2	1	2	2	4	2	6	3	10	5
9	3	3	3	8	3	0	9	4	14	4
10	0	1	1	2	1	0	2	2	3	1
11	3	3	8	1	8	12	17	8	22	16
12	4	4	3	2	12	15	15	12	45	11
13	2	0	1	0	3	1	6	3	7	7
14	1	2	1	0	1	2	5	2	12	6
15	10	11	13	9	25	20	49	19	123	47
16	1	4	3	0	7	3	9	4	24	6
17	4	5	4	7	13	7	25	8	44	10
18	10	0	3	4	17	7	18	25	40	20
19	3	0	1	4	1	5	3	2	9	3
20	3	0	1	4	1	5	3	2	9	3
21	0	1	2	2	2	1	3	2	7	3
22	0	1	0	0	1	0	0	3	4	2
23	0	1	3	1	9	3	8	3	17	6
24	0	0	0	0	0	0	0	0	0	0
25	2	0	4	5	10	2	16	2	26	9
26	2	0	4	5	10	2	16	2	26	9
27	2	0	4	5	10	2	16	2	26	9
28	1	0	0	3	3	1	8	3	20	8
29	4	1	3	3	3	1	9	2	15	7

Fig. 5. Text Feature Extraction using BoW before Text Pre-Processing

TF-IDF Feature Matrix:

	and	at	be	for	in	is	of	that	the	to
0	0.452817	0.475711	0.263754	0.000000	0.224420	0.000000	0.000000	0.463877	0.429477	0.221569
1	0.000000	0.050616	0.168380	0.166164	0.286538	0.105843	0.415536	0.246782	0.731140	0.282899
2	0.452817	0.475711	0.263754	0.000000	0.224420	0.000000	0.000000	0.463877	0.429477	0.221569
3	0.000000	0.050616	0.168380	0.166164	0.286538	0.105843	0.415536	0.246782	0.731140	0.282899
4	0.086560	0.181873	0.302513	0.000000	0.171599	0.190159	0.331803	0.177348	0.738885	0.338839
5	0.121415	0.085036	0.282883	0.046527	0.200580	0.088910	0.310272	0.290220	0.690939	0.435671
6	0.086560	0.181873	0.302513	0.000000	0.171599	0.190159	0.331803	0.177348	0.738885	0.338839
7	0.123239	0.055487	0.102548	0.020240	0.279215	0.116030	0.489272	0.360711	0.651226	0.292897
8	0.144049	0.075666	0.167089	0.165601	0.285568	0.158227	0.414128	0.221351	0.683121	0.352425
9	0.148633	0.156148	0.173150	0.455658	0.147328	0.000000	0.427307	0.203018	0.657871	0.193942
10	0.000000	0.207413	0.229997	0.453941	0.195697	0.000000	0.378398	0.404507	0.561765	0.193211
11	0.081443	0.085561	0.253004	0.031209	0.215274	0.357835	0.442266	0.222486	0.566464	0.425078
12	0.070147	0.079997	0.066513	0.043770	0.226436	0.313657	0.273646	0.234022	0.812503	0.204929
13	0.163891	0.000000	0.095462	0.000000	0.243678	0.090011	0.471173	0.251841	0.544054	0.561359
14	0.070057	0.147197	0.081612	0.000000	0.069441	0.153903	0.335677	0.143535	0.797347	0.411354
15	0.070754	0.081764	0.107152	0.073206	0.175331	0.155435	0.332238	0.137716	0.825416	0.325435
16	0.036919	0.155141	0.129025	0.000000	0.256161	0.121657	0.318414	0.151282	0.840379	0.216777
17	0.075142	0.098677	0.087536	0.151173	0.242067	0.144441	0.450055	0.153955	0.783959	0.183840
18	0.175350	0.000000	0.061282	0.080634	0.295477	0.134826	0.302470	0.449083	0.665249	0.343205
19	0.239734	0.000000	0.093093	0.367471	0.079210	0.438883	0.229738	0.163726	0.682134	0.234610

Fig. 6. Text Feature Extraction using TF-IDF before Text Pre-Processing

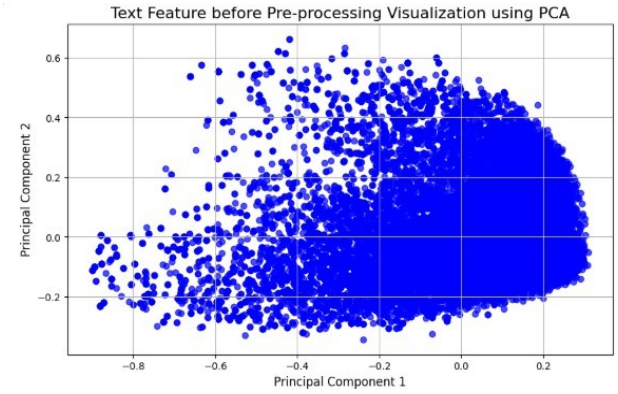

Fig. 7. Post-Analysis of Text Pre-Processing using PCA prior to Pre-processing

The following set of results illustrates the process of pre-processing through the different steps initiated with lowercasing to stemming. The following figure in Fig 8 implements all the processes and converts the raw data shown in Fig 4, to be standardized inorder to augment accuracy of processing.

	case_id	case_outcome	case_title	case_text
0	Case1	cited	alpin hardwood aust pti ltd v hardi pti ltd fc...	ordinarili discret exercis cost follow event a...
1	Case2	cited	black v lipovac fca alr	gener principl govern exercis discret award in...
2	Case3	cited	colgat palmol co v cusson pti ltd fcr	ordinarili discret exercis cost follow event a...
3	Case4	cited	dai studio pti ltd v bullett creativ pti ltd fca	gener principl govern exercis discret award in...
4	Case5	cited	dr marten australia pti ltd v figgin hold pti ...	preced gener principl inform exercis discret d...
5	Case6	cited	gec marconi system pti ltd v bhp inform techno...	accept make roll offer inclus cost interest ma...
6	Case7	cited	john hay amp associ pti ltd v kimberlyclark au...	preced gener principl inform exercis discret d...
7	Case8	cited	seven network limit v news limit alr	question level unreason necessari attract disc...
8	Case9	applied	australian broadcast corpor v oneil hca	recent decis high court australian broadcast c...
9	Case10	followed	hexal australia pti ltd v roch therapeut inc ipr	hexal australia pti ltd v roch therapeut inc i...
10	Case11	cited	castlemain toohey ltd v south australia hca clr	hexal australia pti ltd v roch therapeut inc i...
11	Case12	cited	r v mcfarlan ex part ollannagan okelli hca clr	quia timet proceed court regard degre probabl ...
12	Case13	followed	nation australia bank v ltd construct servic pt...	suggest proceed far commonwealth revenu debt c...
13	Case14	followed	georg v clune aljr note	strictli speak chequ even bank chequ form lega...
14	Case15	followed	australian mideastern club limit v yassim acsr	none suggest deputi commission oblig accept pr...
15	Case16	followed	deputi commission taxat v visidet pti ltd fca	none suggest deputi commission oblig accept pr...
16	Case17	followed	deputi commission taxat v guy hold pti ltd tas...	true posit applic statu creditor time applic m...
17	Case19	cited	motor term co pti ltd v liberti insur ltd hca clr	assum deputi commission ought present regard a...
18	Case21	referred to	appel v minist immigr multicultur affair appel...	satisfi find third tribun two appel homosexu m...
19	Case22	cited	minist immigr multicultur affair v wang clr	plain order set asid matter remit tribun recon...

Fig. 8. Data Pre-processing using NLKT

In comparison with the previous results, and to further understand the importance of pre-processing, the feature extraction techniques of BoW and TF-IDF are implemented on the pre-processed data, and the results of visualization implemented through PCA are shown in Fig 9, 10 and 11 respectively.

Bag-of-Words Feature Matrix:

	act	applic	case	claim	court	ltd	pti	reason	tribun	would
0	0	0	1	0	0	3	3	0	0	0
1	0	0	0	1	1	2	2	2	0	0
2	0	0	1	0	0	3	3	0	0	0
3	0	0	0	1	1	2	2	2	0	0
4	0	0	1	0	0	4	4	0	0	0
5	0	1	2	2	0	2	2	0	0	1
6	0	0	1	0	0	4	4	0	0	0
7	1	0	4	2	0	0	0	2	0	3
8	0	1	2	0	2	0	0	0	0	0
9	0	3	0	0	0	2	1	1	0	0
10	0	1	0	0	0	2	1	0	0	0
11	1	2	4	0	3	0	0	0	0	0
12	0	1	1	0	0	1	1	0	0	3
13	2	0	1	0	0	0	0	0	0	1
14	1	1	0	0	0	2	1	0	0	2
15	4	19	3	0	7	6	5	3	0	7
16	1	6	2	0	2	2	2	1	0	1
17	2	7	2	0	3	3	2	0	0	2
18	0	5	1	6	3	0	0	5	7	4
19	0	0	0	0	1	0	0	0	1	0
20	0	0	0	0	1	0	0	0	1	0
21	0	0	0	1	0	0	0	3	1	0
22	0	0	0	0	0	1	1	0	1	0
23	0	0	1	0	1	0	0	0	3	0
24	0	0	0	0	0	0	0	0	0	0
25	4	2	0	0	0	0	2	4	0	1
26	4	2	0	0	0	0	2	4	0	1
27	4	2	0	0	0	0	2	4	0	1
28	0	4	0	0	1	1	1	1	0	0
29	1	0	1	2	0	1	0	0	0	0

Fig. 9. Bag-of-Words (BoW) - Procured from Applying on the Pre-processed data

TF-IDF Feature Matrix:

	act	applic	case	claim	court	ltd	pti	reason	tribun	would
0	0.000000	0.000000	0.228352	0.000000	0.000000	0.653251	0.721885	0.000000	0.000000	0.000000
1	0.000000	0.000000	0.000000	0.348843	0.227008	0.481619	0.532221	0.558675	0.000000	0.000000
2	0.000000	0.000000	0.228352	0.000000	0.000000	0.653251	0.721885	0.000000	0.000000	0.000000
3	0.000000	0.000000	0.000000	0.348843	0.227008	0.481619	0.532221	0.558675	0.000000	0.000000
4	0.000000	0.000000	0.173251	0.000000	0.000000	0.660833	0.730263	0.000000	0.000000	0.000000
5	0.000000	0.218645	0.425757	0.586780	0.000000	0.405991	0.448647	0.000000	0.000000	0.000000
6	0.000000	0.000000	0.173251	0.000000	0.000000	0.660833	0.730263	0.000000	0.000000	0.000000
7	0.166969	0.000000	0.619658	0.427087	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
8	0.000000	0.356783	0.694748	0.000000	0.624524	0.000000	0.000000	0.000000	0.000000	0.000000
9	0.000000	0.783517	0.000000	0.000000	0.000000	0.484959	0.267956	0.000000	0.000000	0.000000
10	0.000000	0.426381	0.000000	0.000000	0.000000	0.791727	0.437455	0.000000	0.000000	0.000000
11	0.201350	0.383747	0.747255	0.000000	0.503792	0.000000	0.000000	0.000000	0.000000	0.000000
12	0.000000	0.256050	0.249298	0.000000	0.000000	0.237724	0.262700	0.000000	0.000000	0.000000
13	0.815812	0.000000	0.378457	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
14	0.307264	0.292802	0.000000	0.000000	0.000000	0.543691	0.300407	0.000000	0.000000	0.000000
15	0.178259	0.806880	0.124042	0.000000	0.268177	0.236567	0.217852	0.000000	0.000000	0.000000
16	0.142750	0.816188	0.264888	0.000000	0.238114	0.252598	0.279129	0.000000	0.000000	0.000000
17	0.233023	0.777196	0.216200	0.000000	0.291520	0.309244	0.227823	0.000000	0.000000	0.000000
18	0.000000	0.312582	0.060868	0.503329	0.164146	0.000000	0.000000	0.000000	0.000000	0.000000
19	0.000000	0.000000	0.000000	0.000000	0.506185	0.000000	0.000000	0.000000	0.000000	0.000000

Fig. 10. TF-IDF Procured from Applying on the Pre-processed Data.

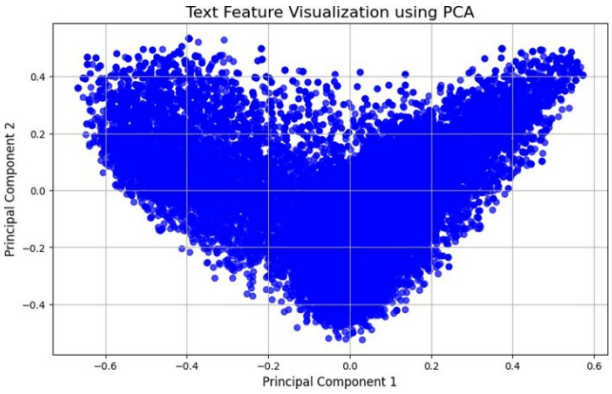

Fig. 11. Text Feature Post Analysis using PCA on the Pre-processed Data

The text representation after pre-processing can be observed to show improvements interms of uniform vocabulary initiated with lowercasing, purged stopword and unwarranted characters, and effective semantic clustering effectuated from stemming. It can also be further observed that TF-IDF vectors illustrates more significant patterns of legal expertise, with more feature emphasis observed in the language processed data applied after pre-processing.

V. Conclusion

Legal text mining is a crucial and most domain which deals with predictive data, and necessitates aid with respect to data pattern matching. This research pivots on the importance of preprocessing and feature extraction in legal document analysis. Through the application of techniques such as lowercasing, punctuation removal, stopword removal, and stemming, legal texts can be standardized and simplified. Feature extraction methods like BoW and TF-IDF transform the cleaned text into structured, computationally friendly formats, facilitating effective machine learning applications. Algorithmic and simulative accuracy can be further enhanced with the incorporation of deep learning models embedded with NLP techniques. Future work can entail advanced methods such as word embeddings and deep learning models to capture deeper semantic relationships in legal documents, thereby accelerating improvised legal data significance and higher establishment of predictive accuracy.

REFERENCES

[1] Olha Kovalchuk, Serhiy Banakh, Mariia Masonkova, Kateryna Berezka, Serhii Mokhun, Olha Fedchyshyn, "Text Mining for the Analysis of Legal Texts", Advanced Computer Information Technologies, 978-1-6654-1050-2/22, DOI: 10.1109/ACIT54803.2022.9913169, 2022

[2] Farid Ariai, Gianluca Demartini, "Natural Language Processing for the Legal Domain: A Survey of Tasks, Datasets, Models, and Challenges", arXiv:2410.21306v2, Mar 2025

[3] Subinay Adhikary, Dwaipayan Roy, Debasis Ganguly, Shouvik Kumar Guha, and Kripabandhu Ghosh, "Leda: a system for legal data annotation", Legal Knowledge and Information Systems. IOS Press, 367–370. 2023.

[4] Ashwini V. Zadgaonkar, Avinash J. Agrawal, "An overview of information extraction techniques for legal document analysis and processing", International Journal of Electrical and Computer Engineering (IJECE), Vol. 11, No. 6, pp. 5450~5457 ISSN: 2088-8708, DOI: 10.11591/ijece.v11i6.pp5450-5457, December 2021.

[5] A. Dyevre, "Text-mining for Lawyers: How Machine Learning Techniques Can Advance our Understanding of Legal Discourse," KU Leuven Centre for Empirical Jurisprudence, 2020.

[6] Ju. Xu, "Research on Judicial Big Data Text Mining and Sentencing Prediction Model," Journal of Physics: Conference Series 1883, 2021, pp. 1-6.

[7] Paheli Bhattacharya, Kaustubh Hiware, Subham Rajgaria, Nilay Pochhi, Kripabandhu Ghosh, and Saptarshi Ghosh, "A Comparative Study of Summarization Algorithms Applied to Legal Case Judgments", Advances in Information Retrieval, 413–428, 2019.

[8] Dang Hoang Anh, Dinh-Truong Do, Vu Tran, and Nguyen Le Minh, "The Impact of Large Language Modeling on Natural Language Processing in Legal Texts: A Comprehensive Survey", 15th International Conference on Knowledge and Systems Engineering (KSE), 2023

[9] Ting Wai Terence Au, Vasileios Lampos, and Ingemar Cox, "E-NER — An Annotated Named Entity Recognition Corpus of Legal Text", In Proceedings of the Natural Legal Language Processing Workshop, 246–255, 2022.

[10] Hiba J. Aleqabie, Mais Saad Sfoq, Rand Abdulwahid Albeer, Enaam Hadi Abd, "A Review of Text Mining Techniques: Trends, and Applications In Various Domains", Iraqi Journal for Computer Science and Mathematics, doi.org/10.52866/ijcsm.2024.05.01.009, 2024.

[11] D. A. Naik, S. Mythreyan, and S. Seema, "Relevance Feature Discovery in Text Mining Using NLP", 3rd International Conference for Emerging Technology (INCET), pp 1–6, 2022

[12] M. Işik and H. Dağ, "The impact of text preprocessing on the prediction of review ratings," Turkish J. Electr. Eng. Comput. Sci., vol. 28, no. 3, pp. 1405–1421, 2020.

[13] Nasa Zata Dina, Sri Devi Ravana, and Norisma Idris, "Legal Judgment Prediction using Natural Language Processing and Machine Learning Methods: A Systematic Literature Review", doi.org/10.1177/21582440251329663, SAGE Open, June 202

[14] Park M, & Chai S, "AI model for predicting legal judgments to improve accuracy and explainability of online privacy invasion cases", Applied Sciences, 11(23), 11080, doi.org/10.3390/app112311080, 2021

[15] Aachal Jakhotiya, Harshada Jain, Bhavik Jain, Charmi Chaniyara, "Text Pre-Processing Techniques in Natural Language Processing: A Review", International Research Journal of Engineering and Technology (IRJET), Vol 09 Issue: 02, 2022.

[16] Reza Drikvandi, Olamide Lawal, "Sparse Principal Component Analysis for Natural Language Processing", Annals of Data Science, DOI:10.1007/s40745-020-00277-x, 2020.