LeDoCl : A Semantic Model for Legal Documents Classification using Ensemble Methods

R. Priyadarshini[a], K. Anuratha[b], N. Rajendran[c], S. Jeyanthi[d], S. Sujeetha[e]

[a,c] Department of Information Technology, B.S. Abdur Rahman Crescent Institute of Science and Technology. rspdarshini@gmail.com, rajendran.n81@gmail.com

[b,e]Department of Information Technology, Sri Sai Ram Institute of Technology. anuvinaya2003@gmail.com, sujeetha.it@sairamit.edu.in

[d]Department of Computer Applications, New Prince Shri Bhavani College of Engineering and Technology. prabakaran. jeyanthi@gmail.com

Article History: Received: 10 January 2021; Revised: 12 February 2021; Accepted: 27 March 2021; Published online: 20 April 2021

Abstract: NLP is one of the components of Machine Learning. Topic Modeling is a sub component of information retrieval Information Retrieval is a broad domain research in Natural Language Processing (NLP). This downside has been broadly studied in the perspective of cluster algorithms like K–means and K-fold, that tends to converge to at least one of diverse native issue counting on the selection of format method. To overcome the instabilities and assumptions in existing systems such as Vector Space Model (VSM) and SVD, Semantic based topic modeling (SLDA) and ensemble model with generation and integration is proposed. In the case of topic modelling, instability is visible in two distinct aspects. First, when the topic descriptors are examined over multiple runs. During which there will be considerable change in the term rankings and few terms may appear or disappear completely as well. Next, there could be instability due to the extent to which topics have association with document, through several executions. In the proposed system, ensemble learning comprises of algorithms Kernel Support Vector Machine (KSVM) and Random Forest algorithm which overcomes the instability. The first issue of appearance and disappearance of words between multiple runs is overcome by Gibbs Sampling based Semantic LDA (GSLDA). The second issue of alignment of topics with document is aided by using ESLDA. This ensemble SLDA algorithm show increased accuracy in terms of retrieval and reduced time interval compared to conventional models. The accuracy increases up to 98% using ESLDA compared to SLDA (82%) and term frequency methods (78%).

Keywords: Topic modeling, LDA, SLDA, ESLDA, Random forest algorithm, NLP, Machine Learning, support vector machine

1. Introduction

Information retrieval system are utilized to recover authoritative records dependent on keyword search and Vector Space Model (VSM). Semantic information Retrieval method is beyond the standard information retrieval and to get related legal documents from the corpus. Vector Space Model based documents is not efficient in real time using with NLP. The proposed system facilitates the model by building stable topic modeling method. This will identify the legal documents and will increase the accuracy and performance of analyzing the legal document using ensemble model. Here Stable topic modeling algorithm will be used to analyse the document as it will help it in knowing the type of document with self-training process with help of ensemble algorithm. Semantic technique can used to retrieve the documents that considerably cut back the instability, whereas at the same time yielding additional correct topic models.

This proposed model in legal domain will help the lawyers and magistrates to refer the previous and related case documents. Semantic technique can used to retrieve the documents that considerably cut back the instability, whereas at the same time yielding additional correct topic models. In topic modeling, documents are considered as composites and the words are considered as parts. To gather composites of parts we have a probabilistic model which is a generative model, Latent Dirichlet Allocation (LDA).LDA's method of dealing with topic modeling is, it thinks each document as an accumulation of points in a specific extent and each topic as a collection of keywords, in a certain proportion.

1.2. Information retrieval (IR)

Information retrieval System objectives in designing and imposing systems to be able to capable of offer speedy and effective content primarily based get entry to to large amount of statistics along with text, pics and so on. The facts retrieval device should interpret the contents of the document in a set and rank them consistent with the relevance to the want of user.

Information retrieval is the study of scanning for data in a record, finding out archives themselves, and furthermore drilling down meta-data that portray information such as text, audio or images. In reponse,the IR system will retrieve the relevant deocuments about the necessary information as output about the required information and will return the documents related to the desired information.

Natural language processing (NLP) is an automatic and intelligent analysis of text data and meaningful insights can be derived from the analyzed data. Some of the popular NLP techniques are Text analysis, Sentiment

Analysis, Information Extraction and Retrieval (IR). The objective of NLP is perusing, interpreting, comprehending, and perceiving human languages in a significant manner.

The NLP mainly depend on Artificial Intelligence to extract context from human. The applications of Natural Language Processing are difficult since human in an exact, unambiguous, very organized programming language. Natural language processing allows computer systems talk with humans in their personal language and scales other language-related tasks. NLP makes it possible for computers to study text, listen speech, interpret it, degree sentiment and decide which components are important. NLP tasks breaks down language into shorter, elemental pieces, try to recognize relationships among the portions and discover how the portions work collectively to create meaning.

1.3 Topic Modeling

One kind of Text Mining is topic modeling. It utilizes unsupervised and regulated measurable AI approaches to recognize designs in a corpus or substantial measure of unstructured content. It can take your enormous gathering of archives and sort the words into bunches of words, recognize topics, by a utilizing procedure of affinity. Topic models give a basic method to examine substantial volumes of unlabeled content. A "topic" is considered as a group of words which frequently happen jointly. Utilizing relevant pieces of information, topic models can associate words with comparative implications. It can relate words with comparable implications utilizing logical intimations and recognize employments of words with different implications. Topic modeling is extraordinary for archive grouping, data recovery from unstructured content, and selecting features. It helps in investigating a lot of content information, discovering groups of words, likeness among reports, and finding unique topics.

1.4 Non-Negative Matrix Factorization (NMF)

NMF (Nonnegative Network Factorization) is a grid factorization technique where we oblige the frameworks to be non-negative.NMF is valuable when there are numerous traits and the properties are equivocal or have frail consistency. It can deliver significant patterns, topics, or themes by consolidating attributes. NMF is regularly valuable in text mining. NMF is appropriate for tasks where the basic elements can be deciphered as non-negative. NMF exploits the way that the vectors are non-negative. NMF powers the coefficients to abide nonnegative, by considering them into the small-dimensional structure.

1.5 Ensemble Classification

To get better performance in prediction, in Statistics and Machine Learning, ensemble methods adopt multiple learning algorithms than adopting only the constituent Machine Learning Algorithms. The statistical ensemble is typically infinite but a Machine Learning ensemble consists of fixed set of different models and allows for a better flexible structure to be present among the alternatives. To improve the performance of the model, Ensemble Modeling usually allows the application of ensemble learning over and above a variety of models that are been built.

2. Related Work

Lin-Chih Chen et.al., Blog web crawler focused on information and dodge the messy scraps of those web pages. This factor facilitates the bloggers to spot only on the information they are enthused about, than focusing on various kinds of pages. The inert semantic examination model used here to perceive the theme that exists between different records. Inactive etymology Analysis ILSA, PLSA, and LDA are semantic models to realize idle points from documents. LDA revolves around finding subject associations between documents.LDA bases on finding theme associations between documents. This system concentrated on the criticality of blogs on different subjects at various times. The time factor which ranks the blog subjects relies on the universality of the posts. This model does not consider a wide assortment of record relationships. It does not manage different idle theme connections in the archive [1].

Yueshen Xu et,al. concluded that to examine the latent topics in archives, topic models are superior tool. To get right topics for a corpus, Topic General Words(TGW) are to be distinguished from the corpus and can be done by a preprocessing step that regularly remove normal stop words from the corpus leaving only the Topic general words (TGW). TGWs can connect with expressions of various subjects, which drives theme models to create comparative points, which is unwanted. Topic modeling depends on higher-order co-occurrence. Generality is a metric measure grade of a word which will connect with various of topics. TGWs are therefore words with high all inclusive statement scores. GSLDA(Generality-touchy LDA) is a recently proposed theme model to join simplification scores. It utilizes the data to handle the TGW issue in the new space. GSLDA accomplishes the most elevated subject coherence It uses the information to unwind the TGW downside inside the new corpora. In this strategy extreme to spot point general words (TGW) in subject demonstrating [2].

Hanqi Wang et.al.,. In a document, particular words have either solid or feeble capability in transference actualities or expressing opinions. (The prior is sense of objective and later is sense of subjective) based on the topics they present in. The discriminative power (in transferring the right sense) of different words is different and dependent on the topics they are related with. A supervised topic model, ios LDA is proposed to search out

the sense of the words used in the topics. To employ this model each document should have two BoDW representations respectively [3].

Thiago Salles, Marcos Gonçalves,Victor Rodrigues,Leonardo Rocha et.al., The Random Forest (RF) classifier has been primarily effective throughout an outsized assortment of automatic distribution tasks. A lazy version of the conventional random forest (RF) classifier (called LazyNN RF- Lazy Nearest Neighbors Random Forest), supposed for classification tasks that are extraordinarily dimensional and noisy. The LazyNN RF "localized" projection is created out of models that higher seem like the take a look at precedent. This filter completely filters out unimportant information and decreases the effect of noisy factors within the classifier, maintaining a strategic distance from the genera-tion of needlessly difficult base tree learners within the RF classifier and excels in classification of texts automatically. n this method there is no run time efficiency [4].

Wenjie Zhu,Yunhui Yan et.al., proposed a discriminative label embedded NMF(LENMF) algorithm. It is devoted to indicate the discriminant localization by attempting to find the orthogonal basis matrix within the kernel space. In the pattern classification task the proposed LENMF uses a unique discriminant NMF technique to classify the data in to low dimensional, structures and label-indicator vector and is achieved by the basis matrix. The projection dimension is the integral multiple of the number of the classifications. LENMF has the limitations of tiny sampling problem [5].

Aytug Onan,Serdar Korukoğlu et.al., proved that, the factual catchphrase extraction techniques are feasible for base learning calculations and collecting strategies in archival classification. The collection strategies are thought as characterization precision, F - measure and zone below the curve values. To find the exact measure two-way ANOVA test is employed. The precise execution of classifiers is based on the quantity of catchphrase. The limitation is that the investigation on Semantics isn't done and separation of the factual strategies is done from everyone else who utilized the classifiers [6].

3. Problem Identification

3.1 Existing System

In existing System, examined the degree to which improved calculation instatement and procedures can create increasingly insecure models, which are conceivably not exact. Contrasting the execution of these methodologies and respects to various measurements that measure strength, exactness, and solidarity of topics. The outcomes show that K-Fold ensemble approach is not steady and erroneous arrangement of models, even if NMF is instated dependent on a Singular Value Decomposition estimation on matrix formed from the document-terms matrix can further lead to unsteadiness of the model dependent on topic modeling pertaining with concordance.

3.2 Issues in Existing System

This relates to the idea of "instability "which has recently been concentrated with regards to k - means clustering. This instability issue is not considered in numerous applications of topic modeling and theme models are treated as being conclusive, despite the fact that the outcomes may change impressively if the instatement procedure is modified. This is risky when looking to get an authoritative topic model for the given corpus and in these algorithms, issues on basic instability is addressed. Several trials of a similar algorithm on similar information will deliver diverse results. This complication is generally examined in the situation of combining algorithms like K - means, which has a tendency to join to one of various neighborhood issue contingent upon the decision of instatement process.

3.3 Proposed System

To solve the existing system, discrepancies, the proposed system takes on semantic based topic modeling and skip gram analysis for summarization of text documents. Data from legal repositories sites are pulled using a web scraper in un-structured data format. The extracted data will be then ingested into the database for pre-processing and storage. The pre-processed data which will be generating some sample data for the Integration phase. In integration phase the sample data will be passed to Ensemble based Semantic LDA (ESLDA). Semantic analysis will be done along through analysis and reasoning of machine learning based model. The Generation and Integration phase will be combined to give the summarized legal text documents that prevent from instability and performance issues. In Generation phase, the data is pre-processed and converted into Document Term Matrix (DTM) structured format. The Integration phase, the sample data which will be passed to the ESLDA for acquiring the model stability over multiple runs.

In proposed system, the two main criteria for evaluating topic modeling using Semantic Latent Dirichlet allocation algorithm with Gibbs sampler and word embedding separately. The parameters to improve model quality such as, by analyzing topic coherence, accuracy, and stability, which yields the support of examining topics and documents summary. Such a methodology may likewise be utilized to render increasingly stable subject models by means of DTM factorization and reduces the computational time. In this method, this can result in generation of significant topics over multiple documents of similar scenario over the legal corpus and predict the document class labels with the manifest accuracy using the learning strategies.

4. Methodology

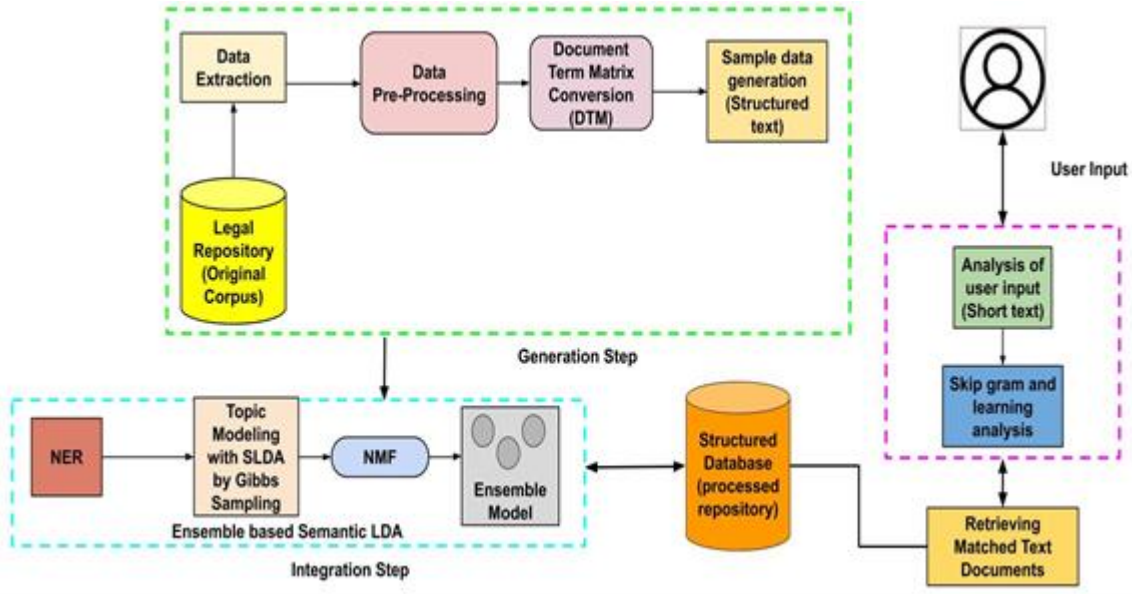

Fig.1. System Architecture

4.1 Data Extraction

The legal case text documents are extracted from the legal repository web pages using web Scraping. Web Scraping is a strategy utilized to draw out a lot of information from sites whereby the data is extracted and saved to a nearby document in the system. The extracted data is in the unstructured format.

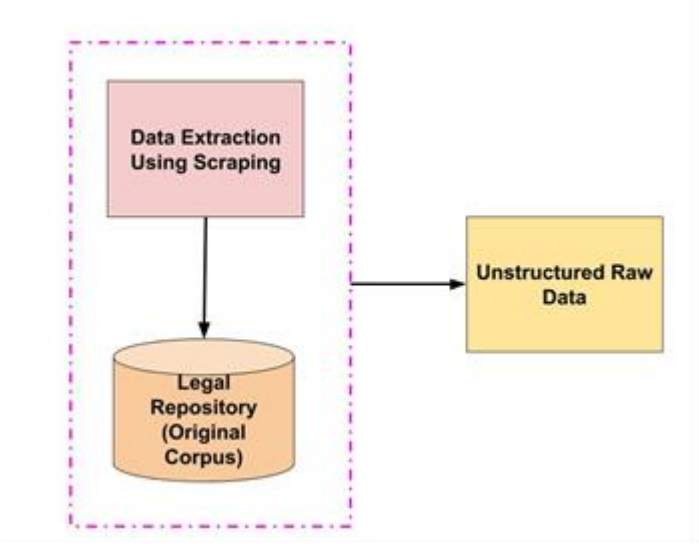

Fig.2. Data Extraction

The Input is the legal repository URL and the variables given is getHTMLLinks, URL, files pattern for (input url).

The URL is http://www.lexinexindia.co.in/en-in/products/lexis-india.page/~legal/ and the function grep is defined for *.txt; If (files == NULL) display <--"no files" otherwise parse individual text document into LFS end if return in .txt format.

4.2 Data Pre-Processing

The unstructured raw data (authoritative archives) are pre-processed using NLP methodologies, for example, stop words expulsion, stemming, lumping and tokenization. Stop words can't avoid being words which are filtered through beforehand or accordingly amid handling of regular language information (content). Stemming is the route toward diminishing adjusted words to their statement stem, base pull structure-generally of a created word structure. Phrase chunking is a natural language handling procedure that isolates and fragments a sentence into its sub constituents, for example, thing, verb and prepositional states. Tokenization is the strategy for separating and possibly gathering regions of a string of information characters. The generation of flow of data samples is done as follows. DTM, a lattice that lineup all occasions of words in the ideal substance, by record. In the DTM, the records are spoken to by lines and the terms (or words) by columns. Structured information is created

alongside weight appointed. The topic modelling is performed based on Latent Drichlet Allocation (LDA). The structured data will be topic modeled using Semantic LDA (SLDA). The named entities in a textual data are located and categorized the names of persons, organizations, locations and date using a process called Named Entity Recognition(NER). The topics of each document are weighted based on the term segmentation and entity recognition. The most weighted word is characterized in descending order in beta spread according to the topics. The Fig 3 shows the Data Pre-Processing.

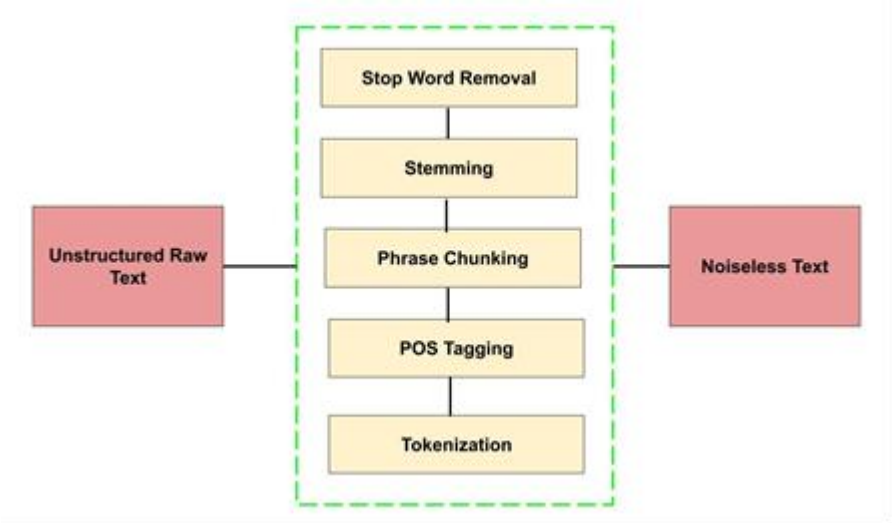

Fig.3. Data Pre-Processesing

4.3 Named Entity Relationship (NER) with OpenNLP : Algorithm

1) Input the data as text.
2) Covert the text document to “string format”.
3) Assign function to vector,
4) word_ann = maxent_word_token_annotator()
5) send_ann = maxent_sent_token_annotator()
6) pos_ann = maxent_pos_tag_annotator()
7) pos_annotation= annotator(reviews,list(sent_annen,word_annen,pos_annen))
8) wor_annotation = annotator(reviews,list(senten_annen,word_annen))
9) annotated plain text document (reviews,text_ann)
10) word(reviews) %>% head(10)
11) maximum_obj_annotator (en=”person”)
12) maximum_obj_annotator (en=”location”)
13) maximum_obj_annotator (en=”date”)
14) Function part for the entities
15) Ent=function(doc,kind){
16) S=doc$con
17) A=annotation(documen)[[1]]
18) (has Arg(kinds_obm)){

i. k=sapply (a $ features,’[[‘,”kind”)
ii. S[a[k==kinds_obm]]
iii. }

iv. else {

v. S[a_annen[a_annen 4 type_res==”entity”]]

vi. } }

19) Semantic LDA (SLDA) with Gibbs Sampler

20) Input the data as text in the function.

21) Covert the text document to “string format”.

22) Create the Corpus for the documents.

23) Convert the documents to DTM as rows and columns.

24) Apply LDA to the documents to extract the topics. function(input_text,plot=T,no.of.topics=4) { Corpus<-Corpus(vector source(input text))

i. DTM<-Document Term matrix(Corpus) Unique<-DTM[unique_indexes,] lda<-lda(DTM,K=no.of.topics,control=list(seed=1234)) topic<-tidy(lda,matrix=”beta”)

}

ii. Apply NER with OpenNLP to group the semantic terms in the documents.

iii. Plot the Graph for the documents using ggplot library else return the topics. The topics modelling is shown in the Fig. 4.

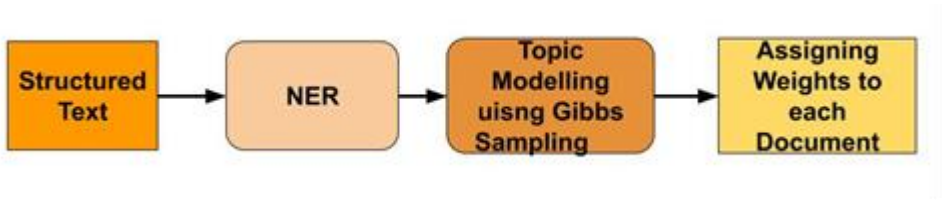

Fig 4. Flow of SLDA topic modeling

4.4 Generation of Ensemble Model

After, the topics modeled to each document, the NMF algorithm is applied to find the coherence between each document in the preprocessed database. NMF decrease the measure of the lattice by taking out the rows and columns. The ensemble model is used to frame the tree structure for the coherent document based on weightage. The step by step procedure for the Ensemble Based Semantic LDA (ESLDA) is as follows:

1) Input: structured legal case document
2) Key-Parameter:
3) n (d,k): No. of times document d use topics k, v (k,w): No. of times topic k uses the given words, αk: Dirichlet parameter for document to topic distribution, λw: Dirichlet parameter for topic to word distribution.
4) Process: Process the terms frequency and no. of documents within the corpus
5) Output: Extraction of top topics from the corpus
6) Extract the legal document from the repository
7) Pre-process the unstructured legal case documents
8) Convert the unstructured legal document to Document Term Matrix
9) Generate the sample data with corpus
10) Apply Name entity recognition
11) Read the document as String
12) Apply POS tagging for each term
13) The result as Person, Organization, Date, and Location
14) Return character terms with semantic terms
15) Apply semantic LDA with Gibbs sampler
16) Create a function of input, topics and plot
17) Input the no. of topics to be generated for the corpus
18) Group topics semantic terms
19) Write the probabilities of the topics and terms in the corpus
20) Plot the result using ggplot package
21) Split the corpus into training and testing sets
22) Initialize 0.7 for training and 0.3 testing datasets

23) Check the hypothesis summary for the dataset
24) Convert the corpus as data frame
25) Apply Matrix Factorization
26) Cluster the legal documents dependent on class labels or names
27) Correlate the corpus based on topics and terms coherence
28) Apply Random Forest Algorithm
29) Preprocess the dataset based in weightage
30) Summarize the model accuracy
31) If the p-value is < 0.5 and adjusted R square value < 0.70 then
32) The model eliminates null values
33) Predicts the accurate topics with stability
34) Else model fails
35) END

Table 1: Generated Keywords from the documents (Key terms)

S.No	Topic 1	Topic 2	Topic 3	Topic 4	Topic 5
1	case	high	shanthnu	account	premproti
2	civil	miscalling	pradesh	council	applic
3	final	learn	allow	advocate	curcumas
4	affidavit	applic	seek	suffice	civil
5	court may	day	consent	noncomplex	vinita

Table 2. Terms and Threshold value for the Identified Keywords

Key Terms (Topic 1)	Topic 1	Topic 2	Topic 3	Topic 4	Topic 5	Log - Ratio
Case	0.04563	0.00369	0.10299	0.00398	0.00497	-3.6286
Civil	0.17012	0.00369	0.00332	0.00398	0.00497	-5.5268
Final	0.00414	0.04059	0.00332	0.04381	0.00497	3.2901
Court May	0.04564	0.07749	0.00332	0.04321	0.00497	4.2231
Adjourn	0.15672	0.00368	0.00332	0.03984	0.00497	4.2230
High Court	0.00413	0.07749	0.00332	0.03984	0.00497	4.2230

The above table 1 shows that the keywords are generated from the document using ensemble algorithm and the matched keywords are prioritized and identified as priority based keywords. The table 2 shows the assigning of weight to the keywords for the five different topics which are selected form the content in the document. The log ratio is also generated using the proposed alogithm.

Table 3: Generated Keywords from the documents (Key terms)

S.No	Topic 1	Topic 2	Topic 3	Topic 4	Topic 5
1	madhya	state	case	restore	order
2	phay	number	file	dismiss	account
3	sri	account	mandoli	kemkemar	affidavit
4	support	advocate	omparkash	may	allow
5	mcc	review	date	cost	kamalesh

Table 4. Terms and Threshold value for the Identified Keywords

Related Terms (Topic 1)	Topic 1	Topic 2	Topic 3	Topic 4	Topic 5	Log - Ratio
madhya	0.12863	0.00369	0.00332	0.00398	0.004971	-5.5123
phay	0.04564	0.00369	0.00336	0.00467	0.00498	-3.6281
shri	0.00457	0.00478	0.00347	0.00457	0.00478	-3.2245
support	0.00478	0.00358	0.00389	0.00497	0.00398	3.2394
mcc	0.00412	0.00456	0.00678	0.00489	0.004789	-3.2216

The above table 3 shows that the related terms are generated from the document using ensemble algorithm and the matched keywords are prioritized and identified as relatied and matched keywords. The table 4 shows the

assigning of weight to the keywords for the five different topics which are selected form the content in the document. The log ratio is also generated using the proposed alogithm.

5. Experimental Classification Results and Analysis

The metrics used for analyzing and evaluating Semantic Topic modeling are recall, precision, weightage analysis of topic, probabilities of topic and graphical topic frequency. These metrics measure the quality of the system and also rate the accuracy of the retrieval. In this project, abstractive summarization of the legal document are processed and analyzed semantically. The analysis and retrieval process is measured by the weightage & probability of topics which are clustered and classified under different categories like most true positive, true negative, false positive and false negative in confusion matrix. The accuracy level using ESLDA increases up to 98% compared to SLDA and term frequency methods.

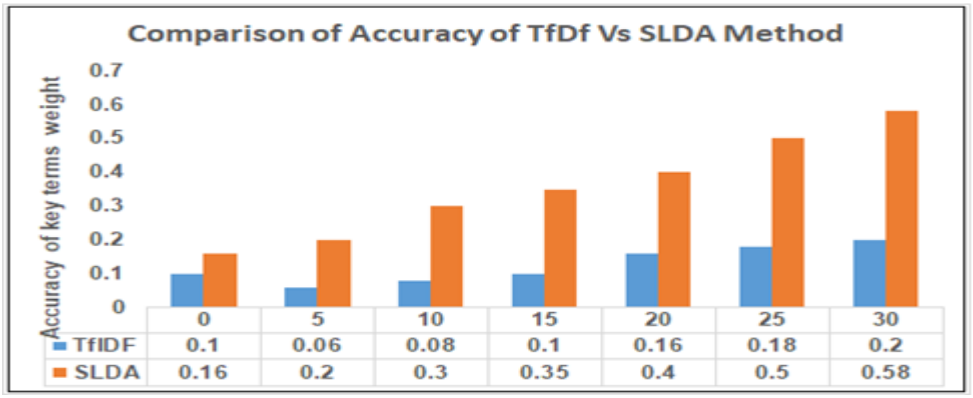

Fig 5.1 Comparison of TFIDF and SLDA

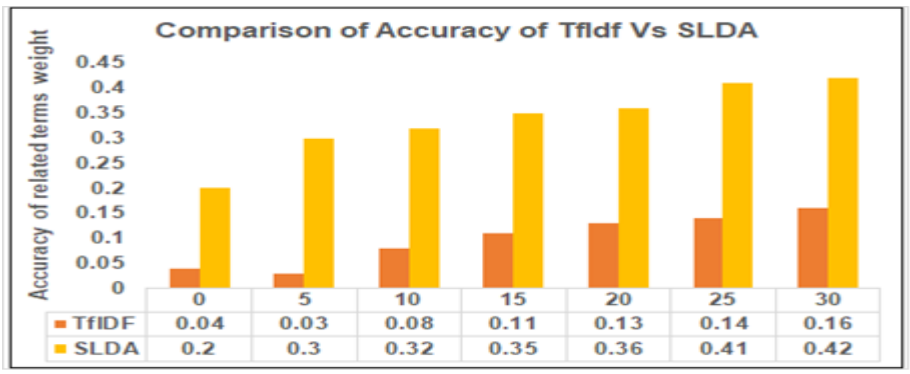

Fig 5.2 Comparison of TFIDF and SLDA

In Fig 5.1 the weightage graph of SLDA, represent the topic terms weightage of each topics of 30 terms in legal documents with percentile of value in between 0 and 1.

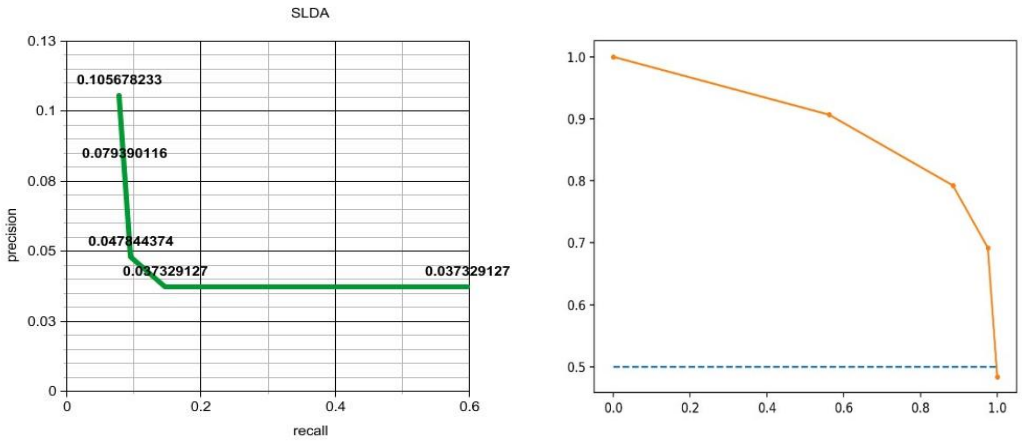

Fig 5.3 Precision and Recall

In Fig 5.2 the comparison graph of TFIDF and SLDA, shows that the topics extracted using SLDA is more stable than the TFIDF for the topics of 30 terms in legal documents.

In Fig 5.3 Precision and recall for the SLDA topics is generated using the topics probabilities values for the 5 topics terms in legal document. It denotes when the recall decreases the precision value increases.

6. Conclusion

In this project we first extracted the data from legal repository sites and then loaded it into Windows File System (WFS). Then preprocessed using Natural Language Processing technique such as removal of stop words, stemming, phrasing and loading of content to Document term matrix which converts the unstructured data into term matrix further the DTM is changed to rows and columns as matrix to return the terms frequency. Then we apply name entity recognition using part of speech tagging which result in tagging of semantic terms based on parameters of the rule based concern. The semantic LDA is used to sort out the topics from the corpus using the terms and documents. The clustering of document is done with naming the class labels for each document with the help of NMF which will correlate the document based on topic coherence. Finally, the ensemble strategy will the predict the documents using the topics weightage and terms document with Random forest Algorithm to render the model stability based metrics. The improved Ensemble LDA gives accuracy of 98% compared to conventional methods.

In future we will plan to widen this summarization methodology to xml, pdf, json and html documents type extracted from legal repository and process data upon cMapping and advanced text analytics involved in processing the legal documents. Also dynamic and live streaming of data from the websites can be included which will be an obvious added advantage to relate the case incident for better retrieval of the legal case documents.

References

1. Lin-Chih Chen, “An effective LDA-based time topic model to improve blog search performance”, Information Processing and Management vol.53, issue.6, pp. 1299–1319,2017.
2. Yueshen Xu,Yuyu Yin and Jianwei Yin, “Tackling topic general words in topic modeling.Engineering Applications of Artificial Intelligence”,vol.62,pp.124133,2017.
3. Hanqi Wang , Fei Wu,Weiming Lu,Yi Yang,Xi Li,Xuelong Li and Yueting Zhuang, “Identifying Objective and Subjective Words via Topic Modeling”, IEEE Transactions on Neural Networks and Learning Systems vol.29, issue.3,pp.718 – 730, 2018.
4. Thiago Salles, Marcos Gonçalves, Victor Rodrigues and Leonardo Rocha,“Improving random forests by neighborhood projection for effective text classification”,Information Systems 77,1-21,2018.
5. Wenjie Zhu,Yunhui Yan,”Non-negative matrix factorization via discriminative label embedding for pattern classification”, Journal of Visual Communication and Image Representation vol.55,pp.477-488,2018.

6. Aytug Onan,Serdar Korukoğlu,Hasan Bulut,"Ensemble of keyword extraction methods and classifiers in text classification", Expert Systems with Applications vol.57, pp. 232-247, 2016.
7. Suh, S., Choo, J., Lee, J., & Reddy, C. K. , "L-EnsNMF: Boosted local topic discovery via ensemble of nonnegative matrix factorization", In Proceedings of the IEEE 16th international conference on data mining,2016.
8. Xingwang Zhao,Jiye Liang, and Chuangyin Dang,"Clustering ensemble selection for categorical data based on internal validity indices", Pattern Recognition vol.69,pp.150-168,2017.
9. Tu Ding,Chen Ling,Lv Mingqi,Shi Hongyu, and Chen Gencai,"Hierarchical online NMF for detecting and tracking topic hierarchies in a text stream",Pattern Recognition vol.76,pp.203-214,2018.
10. Damir Korenčić,Strahil Ristov,Jan and Šnajder, "Document-based topic coherence measures for news media text",Expert Systems with Applications vol.114,pp.357-373,2018.
11. Yong Chen,Hui Zhang,Rui Liu,Zhiwen Ye,Jianying Lin,"Experimental explorations on short text topic mining between LDA and NMF based Schemes.Knowledge-Based Systems", vol.163, pp.1-13, 2019.
12. Yan Liang,Ying Liu,Chong Chen and Zhigang Jiang,"Extracting topicsensitive content from textual documents—A hybrid topic model approach",Engineering Applications of Artificial Intelligence vol.70,pp.81-91,2018.
13. Yueshen Xu,Jianwei Yin,Jianbin Huang and Yuyu Yin,"Hierarchical topic modeling with automatic knowledge mining",Expert Systems with Applications vol.103,pp.106-117,2018.
14. Y. Kim and S. R. Jeong, "Opinion-Mining Methodology for Social Media Analytics," KSII Transactions on Internet and Information Systems, vol. 9, no. 1, pp. 391-406, 2015. DOI: 10.3837/tiis.2015.01.024.