International Journal of Engineering Trends and Technology
ISSN: 2231 – 5381 / https://doi.org/10.14445/22315381/IJETT-V70I11P217
Volume 70 Issue 11, 163-169, November 2022

Original Article

Topic Modeling Techniques for Document Clustering and Analysis of Judicial Judgements

Amar Jeet Rawat[1], Sunil Ghildiyal[2], Anil Kumar Dixit[3]

[1,2]*Department of Computer Science and Engineering, Uttaranchal University, Dehradun, Uttarakhand, India*
[3]*Law College Dehradun, Uttaranchal University, Dehradun, Uttarakhand, India*

[1]*Corresponding Author : aj.amar.rawat@gmail.com*

Received: 30 August 2022 Revised: 02 November 2022 Accepted: 16 November 2022 Published: 26 November 2022

Abstract - *The digital world is growing rapidly in every dimension. Legal case information and judgements are now available online and are becoming a big problem because of their unstructured textual nature. The classification, analysis, and understanding of such unstructured textual data are complex. Various top-modelling techniques are used for the classification of such corpora. In this paper, two popular topic-modelling models, LDA and LSA, are implemented, and their performances are compared on a dataset of 1000 legal judgement documents. Coherence scores are used to evaluate the performance of both models. Tests show that LDA and LSA have different areas of strength. LDA is good at learning about descriptive topics, while LSA is good at making a short representation of the meaning of documents and words in a corpus.*

Keywords - *Latent Semantic Analysis, Topic Modeling, Natural Language Processing, Document Clustering, Latent Dirichlet Allocation.*

1. Introduction

A judicial system's workflow depends on common law and legitimate information, which is available in an unstructured manner and is difficult to understand. Both ongoing legislative changes and the rapid expansion of precedent cases place increasing pressure on legal experts. Due to the rapid generation of legal information in large volumes, all legal data is digitized. It raises the question of whether field machines can help. Such a system would help lawyers, prosecutors, judges, and other professionals, as well as the public, by saving time, minimizing errors, and improving consistency. Artificial Intelligence (AI) - based automation system can convert complex unstructured information into more presentable information for law professionals, practitioners, and people seeking legal information [1], [2]. Natural Language Processing (NLP) has gained consideration for processing and analyzing complex legal text. NLP has been widely used in the social science domain [3]. It is a very tedious job to get relevant information from the enormous textual case data. Therefore, this work seeks techniques for adequately organising, analysing, retrieving information and discovering insights in legal text corpora [4]. Topic modelling is a brand-new, highly effective method for automatically classifying documents, doing unsupervised analyses of massive document groups, comprehending massive amounts of textual data in a vast group of unstructured documented data, and summarizing enormous amounts of textual data[5]. In [6], an LDA-based approach is used for generating topics. Topic modelling plays a vital role and is helpful in online digital libraries for the creation of supplementary metadata by offering a straightforward method to assess enormous amounts of unlabelled text data and show the unseen links between objects and subjects conveyed in headings [7].

Documents are considered composites in topic modelling, while words are viewed as individual components. A generalized probabilistic model called Latent Dirichlet Allocation (LDA) generates parts of a composition [8]. Topic modelling plays an essential role in processing and classifying the text of legal judgments. Classification of legal texts is beneficial for correctly understanding legal judgement [9]. It makes it easier to find the necessary information by grouping documents with related subjects together and making them searchable using various keywords. The classification of legal texts, on the other hand, enables libraries to classify new judgements within the relevant groupings with more accuracy and in less time, thanks to the development of a classification model that can handle any new judgement document and is trained on a vast number of classified documents. Numerous studies employed topic modelling methods to classify raw information on a pre-prepared corpus, in which the topic model was applied in numerous fields. Text mining includes topic modelling. It employs unmonitored and controlled computable AI techniques to identify patterns in a corpus or a substantial amount of unorganized content. It can take a massive collection of documents, arrange the words into groups of words, and identify topics using an affinity-based procedure.

Semantic Models for Topic Modeling: LSA, LDA, and Probabilistic Latent Semantic Analysis (PLSA) are the popular semantic models. Table 1 depicts a brief comparison between the models based on their merits, demerits, and methodologies. LSA is primarily suitable for summarising documents, solving ranking problems, and processing information retrieval (IR) tasks. Location prediction and IR problems can be solved using PLSA. To recommend significant multimedia tags, LDA can be used.

Table 1. Comparison of topic modeling models

Parameter	LDA	LSA	PLSA
Approach	Dirichlet Probability	SVD and Dimensional Reduction Technique	Probabilistic Method, Expectation Maximization
Merit	Unsupervised generative model	Works with synonyms	Polysemy & Synonymy
Demerit	Suitable for large data only	Polysemy	High compute time

2. Related Work

Topic-modelling techniques have been used in much research on text classification in the past few years.
The LDA Model was introduced by [10] Deerwester et al. This model is based on a three-step probabilistic hierarchical Bayesian process used for large datasets. It attempts to generate a simple definition for the collection of documents.

[11] J. Huang, Finding subject correlations between documents is the central focus of LDA. The judgement of theme relationships between documents is the foundation of LDA. Their system This system focused on how critical bloggers were about specific topics at various times. The universality of the postings is a key component of the time factor used to rank the blog topics. This approach does not take into account a wide range of record linkages. Different unused theme links in the archive are not managed by it.

[12] The author concluded that topic models are a superior tool for examining latent themes in archives. Topic Generic Words (TGW) must be differentiated from a corpus to obtain the correct subjects. It can be accomplished by consistently removing regular stopwords from the set of documents, considering only the Topic Generic Words (TGW). Co-occurrence of higher order is necessary for topic modelling. A word's generality is a metric used to assess how well it connects to different subjects. As a result, generic terms have high all-inclusive statement scores.

In [13], the authors proposed an NMF algorithm-based discriminative label model. By endeavouring to locate the orthogonal-base matrix in the kernel space, it aims to demonstrate discriminant localization. The suggested methodology employs a unique discriminant NMF method to categorize the data in the pattern classification task into small dimensional, structural, and class indicator vector categories, which is accomplished by the foundation matrix. The integral multiple of the number of classes is the projection dimension.

[34] They recommend the associated topic model, which employs a simple logistic model for LDA to measure topic occurrences. Blei suggests using a dynamic-topic technique to represent temporal data in a data sequence [15]. The supervised category of LDA models topics such as a corpus of documents and answers, which are appropriate for information like product reviews that include both product descriptions and related evaluation scores [16]. The "Relational Topic Model (RTM)" simulates the subject of a pair of documents that are connected by links like paper citations [17].

In the study in [18], 62 documents on health-related topics were analyzed using in-house built software using the LDA model to grab an impression of different types of fitness-related data that the documents among connected corpus are labelling, as well as to grab a collection of documents about the scope of mental related health.

The author used topic modelling LDA and SVM techniques in medical documents to analyze the categorization of CT imaging medical reports into binary classification, demonstrating the system's capability to represent them effectively and understandably. The model was also effective in reducing dimensionality. Compared to baseline methods, this study demonstrated efficiency for datasets with similar class distribution [19].

In this study, an actual news text classifies using a topic vector supported by Softmax Regression and the LDA model. The outcomes of textual data classification techniques were appropriate and worthwhile to moderate the measurement of the feature, but that model has some flaws, such as the topic model parameters used and the magnitude of news text. Additionally, the proposed model has some flaws, such as the topic model parameter selection and news text size[20]

Many scholars have employed LSA to tackle various Information Retrieval (IR) issues. Some researchers [35] employed LSA to sort every document in order to accomplish the goal of a document summarizing in order to solve the challenge of document summarization in numerous user-generated documents.

The work done in [22] has focused on SQL injection attacks. LDA topic modelling approach was used for extracting features from log files. The feature vectorization was done for malicious data and a normal dataset, with future LDA transformations applied to create a labelled dataset.

In [23], topic modelling was applied to generate a summary of the Bible dataset in English. Most essential topics were generated using the LDA topic modelling approach. The proposed work outperformed the LSA and TextRank algorithms with an F-score of 0.37, 0.36, and 0.28, respectively.

3. Methodology

In this section, figure 1 depicts the workflow of the topic modeling process.

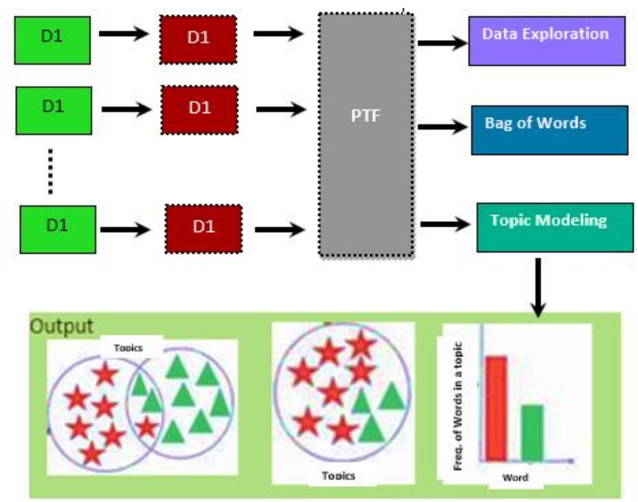

Fig. 1 Methodology

3.1. Data Collection

To create a case judgement document corpus, the judgement of cases have been extracted from the "LegalCrystal" website. Beautiful soap package of python is used to scrape case judgement and saved as a pdf file. The HTML page elements are parsed by BeautifulSoup using regular expressions, which makes for convenient navigation, searching for tags, and editing content through a parse tree. Each judgement file contains three sections.

1) Case Details, including the name of the court, the date of the decision, the ID of the case and the subject.

2) The case description includes the summary of the case judgement.

3) Judgement Section. For this work, 1000 case judgements are extracted into pdf files. The set of documents given by D = d1, d2, d3... dn. All pdf files are read and converted into a CSV file for further processing.

3.2. Preprocessing

Pre-processing is crucial for organizing unstructured material and preserving keywords that can be used to categorize text subjects [24]. The NLP pre-processing steps are given in figure 2.

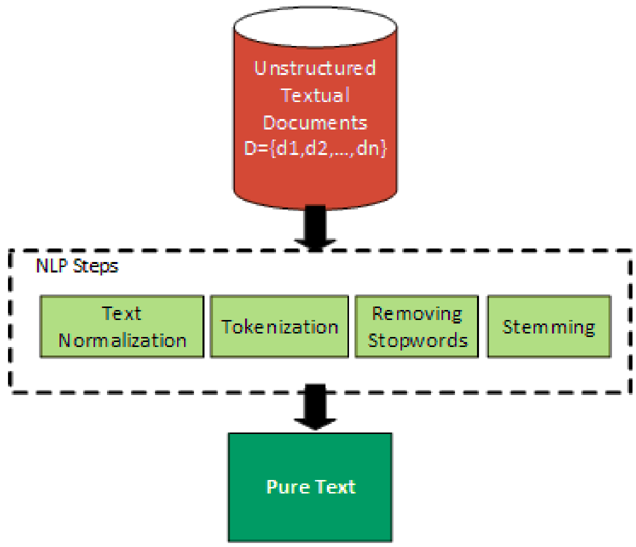

Fig. 2 NLP Preprocessing

3.2.1. Normalization

The raw text is first normalized by lowering the text case, removing the numeric parts, and removing special characters and punctuation. This transforms the text into a basic structure.

3.2.2. Token Generation

Tokenization The specified text is divided into tokens, which are then divided into smaller units called sentences. Line breaks or whitespace are used to separate tokens.

3.2.3. Remove irrelevant Words

Words having a minimal length (i.e., less than 3), such as and, ect, at, and hm, do not have an important impact in the sentence. Therefore, these words may be removed. Further stopwords such as the, has had, that, and thus do not have a significant role in the sentence, and removing them does not have any impact on the sentence.

3.2.4. Stemming and Creating Pure Text File (PTF)

Stemming is a process of reducing the size of a word up to its root, for example, "Limited => Limit". It helps to reduce the size of text data. After performing all NLP tasks, a Pure Text File (PTF) was created. Further vectorization and topic modeling models can be applied.

3.3 Topic Modeling

Topic modeling is a technique of statistical-based model that uses ML to catch latent semantic fashions in a group of documents [25]. It can take an extensive collection of archives and use an affinity-based algorithm to group words into groups and identify topics. A fundamental way to analyze significant amounts of unlabelled content is provided by topic models [26]. A topic model is employed to recognize the nonfigurative "subjects" that are present in a cluster of documents. Topic modelling is a standard text-mining method for tracing hidden semantic patterns in a text. [27].

In the topic modeling process, the document can be read as a variety of subjects, and each subject comprises a group of often recurring terms. Further, topics can be obtained by connecting words with similar meanings.

3.1.1. Latent Semantic Analysis (LSA)

It is a Single Value Decomposition (SVD) - based algebraic method. The idea of LSA is to take a matrix of terms and documents and convert them into separate topic-document matrixes and topic-term matrixes [28]. LSA assumes that words with similar meanings will appear in comparable texts if the goal is to reduce the dimension for classification or as a noise-reducing strategy [29].

3.1.2 Latent Dirichlet Allocation (LDA)

LDA is one of the unsupervised-based machine learning methods used in the fields of document modelling, classification, and information retrieval to identify the latent topic structure of text documents [30]. LDA assumes that every document in the corpora is a stochastic blend of latent topics, each specified by a distribution over the word pattern. [31]

3.1.3 Model Evaluation

Both LDA and LSA take input as a matrix format generated using bag-of-words. LSA handles matrix dimension reduction, and LDA resolves topic modelling issues. Topic models can be evaluated using a topic coherence matrix. It measures the grade of the semantic resemblance resultant values of the words occurring in a specific topic. Topic coherence measures can be taken by using UMass and UCI. UMass is used for intrinsic measures, and UCI is used for extrinsic measures[32].

4. Implementation and Results

After performing the NLP text cleaning process, a data set was explored using NLP data exploration methods. Figure 3 and figure 4 represent the word frequency of the judgement text and summary text, respectively.

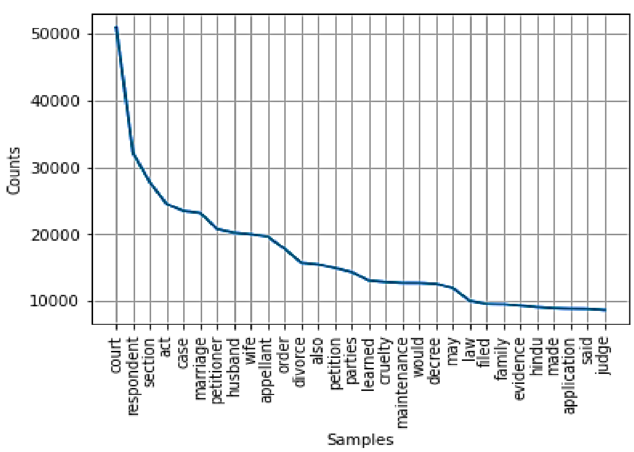

Fig. 3 Word Frequency of Judgements

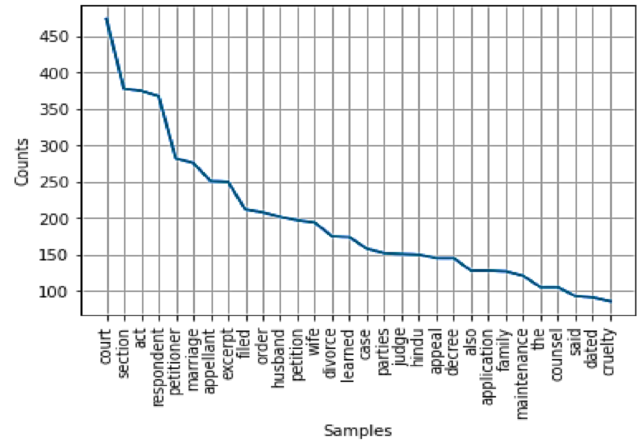

Fig. 4 Word Frequency of Summary

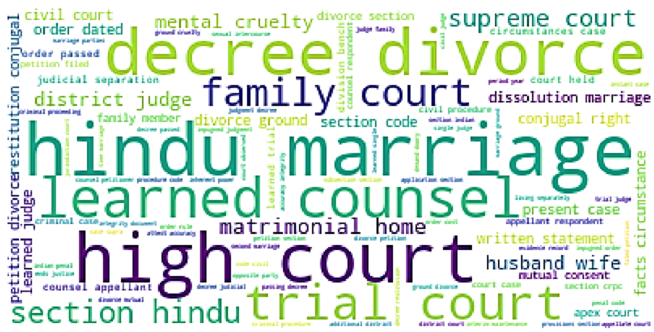

Fig. 5 Word Cloud of Judgement Pure Text

Fig. 6 Word Cloud of Summary Pure Text

In both cases, the word "court" has the highest frequency. The word cloud for judgement pure text and summary pure text is given in figure-5 and figure-6. A word cloud is generated to visualize the pure text and is used to display the words with high frequency. Prominent words appear in a bigger word cloud.

Table 2 and Table 3 demonstrate the words for the top 5 topic clusters using LSA and LDA, respectively. The most common words in both tables are court, respondent, section, marriage, appeal, wife, and husband. Using LSA, a few uncommon words are mental, cruelty, and decree. Words like petitioner, order, and offences, are generated using LDA.

Table 3 and Table 4 show the top three topics for calculation using LSA and LDA.

Based on the frequency of words in each document, each topic is linked to one or more of the group's documents in a specific mixing proportion. Topic models cannot guarantee that their output can be interpreted.

Hence topic coherence measurements were utilized to assess (LDA and LSA). By assigning each topic's interpretability a coherence score, the UCI and Umass topic coherences capture the ideal topics count.

Table 2. Top 5 topics and words using LSA

Topic	Words
0	Court, respond, section, case, marriage, appeal, husband, wife, party, divorce
1	Court, appeal, respond, section, cruelty, proceed, husband, civil, avid, order
2	Wife, appeal, respond, husband, decree, divorce, marriage, maintenance, section
3	Cruelty, marriage, court, decree, appeal, section, conduct, mental, maintenance
4	Marriage, wife, cruelty, husband, child, section, spouse, family, mental

Table 3. Top 5 topics and words using LDA

Topic	Words
0	Court, evidence, offences, act, husband, section, excerpt, respondent, well, wife
1	Court, respondent, section, act, marriage, petitioner, appellant, excerpt, divorce, order
2	Act, section, petitioner, court, order, filed, wife, marriage
3	Respondent, act, section, parties, maintenance, petitioner
4	Respondent, section, learned, case, appellant.

Table 4. Topic terms using LSA (training of a batch of ten topics run on 500 judgement documents)

Topic	Terms
1	-0.492*"court" + -0.251*"respond" + -0.227*"section" + -0.224*"case" + -0.218*"marriage" + -0.193*"appeal" + -0.170*"husband" + -0.165*"wife" + -0.143*"parties" + -0.140*"divorce"'
2	-0.434*"wife" + 0.422*"appeal" + 0.415*"respond" + -0.414*"husband" + 0.238*"court" + -0.173*"decree" + -0.140*"divorce" + -0.133*"marriage" + -0.125*" maintenance " + -0.112*"section"'
3	-0.281*"allegation" + -0.279*"petition" + 0.218*"decree" + -0.196*"case" + -0.157*"civil" + 0.152*" maintenance " + -0.151*"suit" + -0.145*"section" + 0.143*"parties" + -0.135*"avid"'

Table 5. Topic terms using LDA (training of a batch of ten topics on 500 judgement documents)

Topic	Terms
1	'0.008*"court" + 0.007*"evidence" + 0.007*"offences" + 0.006*"act" + "0.006*"husband" + 0.006*"section" + 0.006*"excerpt" + 0.005*"respondent" + '0.005*"well" + 0.005*"wife"'
2	'0.019*"act" + 0.018*"section" + 0.014*"petitioner" + 0.014*"court" + "0.010*"order" 0.009*"excerpt" + 0.009*"filed" + 0.009*"wife" + "0.008*"respondent" + 0.008*"marriage"'
3	'0.022*"respondent" + 0.016*"court" + 0.013*"appellant" + 0.012*"act" + "0.012*"petitioner" + 0.012*"decree" + 0.010*"husband" + 0.010*"marriage" + "0.010*"excerpt" + 0.009*"section"'

Table 6. Coherence Metrix score (UCI and Umass) for LSA and LDA

Topic Count	LDA - UCI Score	LDA-Umass Score	LSA-UCI Score	LSA-Umass Score
10	0.49	-0.52	0.47	-0.53
15	0.53	-0.46	0.46	-0.61
20	0.55	-0.41	0.44	-0.71

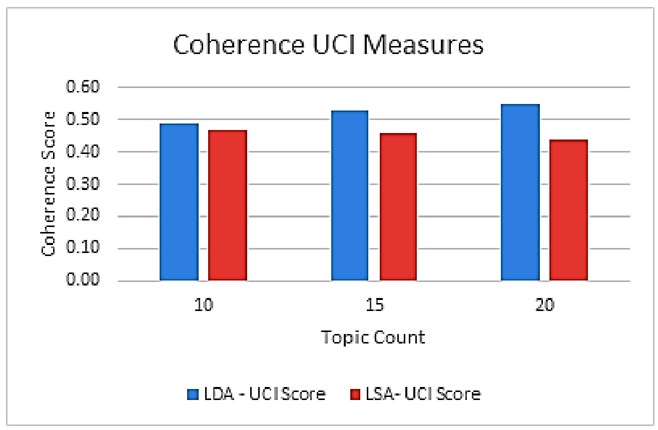

Fig. 7 Coherence UCI Score for LDA and LDA for 10, 15 and 20 topic count

5. Discussion

Here, the implementation of the proposed work is explained by performing LDA and LSA algorithms to extract topics from a case dataset. After performing pre-processing on the dataset, several latent topics are identified by the LDA and LSA. The dataset contained summaries of cases and complete judgement text. Both LDA and LSA modelling techniques are applied to summarise and detail case judgement text. In this study, a variety of (LSA and LDA) models was constructed using various values for the different k-number of topics, and the highest coherence values were chosen. Table 6 demonstrates the coherence matrix (UCI and

Umass) score of LSA and LDA modes for topic count 10, 15 and 20. Table 6 shows that LDA has an increasing UCI score for 10, 15 and 20 topic counts compared to LSA. LSA model's UCI score are 0.47, 0.46 and 0.44 for 10, 15 20 topic count, respectively. It was found that when the number of topics rose, the coherence value (UCL) of LDA climbed more than that of LSA, while the coherence UMass Score (LSA) decreased more quickly (LDA). The comparison of performances of LDA and LSA is given in figure 7.

6. Conclusion

This paper demonstrates and compares the workings of LDA and LSA topic modelling models. These two models have been used to classify the legal judgement documents based on dominant topics extracted from the documents. For working professionals in the legal field, classification is crucial for understanding unstructured language and classifying new documents. In the paper, LDA and LSA models are compared to classify documents. The performance is measured by the coherence scores (UCI and Umass) for LSA and LDA. These findings demonstrate that, depending on the scale, the LDA technique outperformed the LSA technique. The pre-processing stage is crucial for both methods since it effectively reduces dimensionality and eliminates superfluous words from the unstructured textual input. The topic coherence metric can be considered a practical approach to comparing various subject modelling techniques in terms of how easily comprehensible they are to humans, which helps to present a clear picture and, ultimately, helps them make sound judgments. The experiment's findings, however, indicated that the size of the dataset must be increased in the future for improved performance, as both LDA and LSA have performance constraints when using the current dataset.

References

[1] D. Ji, J. Gao, H. Fei, C. Teng, and Y. Ren, "A Deep Neural Network Model for Speakers Coreference Resolution in Legal Texts," *Information Processing and Management*, vol. 57, no. 6, pp. 102365, 2020. Crossref, https://doi.org/10.1016/J.IPM.2020.102365

[2] X. Li, P. Wu, and W. Wang, "Incorporating Stock Prices and News Sentiments for Stock Market Prediction: A Case of Hong Kong," *Information Processing and Management*, vol. 57, no. 5, pp. 102212, 2020. Crossref, https://doi.org/10.1016/J.IPM.2020.102212

[3] A. Y. Ikram and L. Chakir, "Arabic Text Classification in the Legal Domain," *2019 Third International Conference on Intelligent Computing in Data Sciences (ICDS)*, pp. 1-6, 2019. Crossref, https://doi.org/10.1109/ICDS47004.2019.8942343

[4] "Correlated Topic Models," *Proceedings of the 18th International Conference on Neural Information Processing Systems*, 2022. Crossref, https://dl.acm.org/doi/10.5555/2976248.2976267

[5] P. Anupriya and S. Karpagavalli, "LDA based Topic Modeling of Journal Abstracts," *ICACCS 2015 - Proceedings of the Second International Conference on Computer and Communication Technologies*, pp. 1-5 , 2015. Crossref, https://doi.org/10.1109/ICACCS.2015.7324058

[6] L. C. Chen, "An Effective LDA-Based Time Topic Model to Improve Blog Search Performance," *Information Processing and Management*, vol. 53, no. 6, pp. 1299–1319, 2017. Crossref, https://doi.org/10.1016/j.ipm.2017.08.001

[7] D. Newman, Y. Noh, E. Talley, S. Karimi, and T. Baldwin, "Evaluating Topic Models for Digital Libraries," *Proceedings of ACM International Conference on Digital Library*, pp. 215–224, 2010. Crossref, https://doi.org/10.1145/1816123.1816156

[8] "2014 International Forum on Materials Processing Technology, IFMPT 2014 and 2014 International Conference on Sensors, Instrument and Information Technology, ICSIIT 2014," *Applied Mechanics and Materials*, vol. 533, 2014. [Online]. Available: https://www-scopus-com-uttaranchaluniversity.knimbus.com/inward/record.uri?eid=2-s2.0-84896298158&partnerID=40&md5=2301bbf2de24db97215f8fe77a252bbc.

[9] S. H. Mohammed and S. Al-Augby, "LSA & LDA Topic Modeling Classification: Comparison Study on E-books," *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 19, no. 1, pp. 353–362, 2020. Crossref, https://doi.org/10.11591/ijeecs.v19.i1.pp353-362

[10] S. Deerwester, G. W. Furnas, T. K. Landauer, and R. Harshman, "Indexing by Latent Semantic Analysis Scott," *Kehidupan*, vol. 3, no. 12, p. 34, 2015.

[11] Y. Xu, J. Yin, J. Huang, and Y. Yin, "Hierarchical Topic Modeling with Automatic Knowledge Mining," *Expert Systems with Applications*, vol. 103, pp. 106–117, 2018. Crossref, https://doi.org/10.1016/J.ESWA.2018.03.008

[12] Y. Xu, Y. Yin, and J. Yin, "Tackling Topic General Words in Topic Modeling," *Engineering Applications of Artificial Intelligence*, vol. 62, pp. 124–133, 2017. Crossref, https://doi.org/10.1016/J.ENGAPPAI.2017.04.009

[13] W. Zhu and Y. 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. Crossref, https://doi.org/10.1016/J.JVCIR.2018.06.030

[14] Nilesh Yadav and Dr. Narendra Shekokar, "SQLI Detection Based on LDA Topic Model," *International Journal of Engineering Trends and Technology*, vol. 69, no. 11, pp. 47-52, 2021. Crossref, https://doi.org/10.14445/22315381/IJETT-V69I11P206

[15] I. T. Koponen and M. Nousiainen, "Lexical Networks and Lexicon Profiles in Didactical Texts for Science Education," *SStudies in Computational Intelligence*, vol. 882 SCI, pp. 15–27, 2020. Crossref, https://doi.org/10.1007/978-3-030-36683-4_2/COVER

[16] Kazuki Ashihara, Chenhui Chu, Benjamin Renoust, Noriko Okubo, Noriko Takemura, Yuta Nakashima and Hajime Nagahara, "Legal Information as a Complex Network: Improving Topic Modeling Through Homophily," in *Studies in Computational Intelligence*, vol. 882 SCI, M. J. F. M. E. R. L. M. Cherifi H. Gaito S., Ed. Springer, pp. 28–39, 2020. Crossref, https://doi.org/10.1007/978-3-030-36683-4_3

[17] M. Bockholt and K. A. Zweig, "Why We Need a Process-Driven Network Analysis," *Studies in Computational Intelligence*, vol. 882 SCI, pp. 81–93, 2020. Crossref, https://doi.org/10.1007/978-3-030-36683-4_7/COVER

[18] J. W. Uys, N. D. Du Preez, and E. W. Uys, "Leveraging Unstructured Information using Topic Modelling," *Portland International Center for Management of Engineering and Technology*, pp. 955–961, 2008. Crossref, https://doi.org/10.1109/PICMET.2008.4599703.

[19] E. S. Kayi, K. Yadav, and H.-A. Choi, "Topic Modeling Based Classification of Clinical Reports," pp. 67–73, 2013.

[20] Z. Li, W. Shang, and M. Yan, "News Text Classification Model Based on the Topic Model," *2016 IEEE/ACIS 15th International Conference on Computer and Information Science*, 2016. Crossref, https://doi.org/10.1109/ICIS.2016.7550929

[21] Vasantha Kumari Garbhapu and Prajna Bodapati, "Extractive Summarization of Bible Data using Topic Modeling," *International Journal of Engineering Trends and Technology*, vol. 70, no. 6, pp. 79-89, 2022. Crossref, https://doi.org/10.14445/22315381/IJETT-V70I6P210

[22] N. Yadav and N. Shekokar, "SQLI Detection Based on LDA Topic Model," *International Journal of Engineering Trends and Technology,* vol. 69, no. 11, pp. 47–52, 2021. Crossref, https://doi.org/10.14445/22315381/IJETT-V69I11P206

[23] V. K. Garbhapu and P. Bodapati, "Extractive Summarization of Bible Data using Topic Modeling," *International Journal of Engineering Trends and Technology*, vol. 70, no. 6, pp. 79–89, 2022. Crossref, https://doi.org/10.14445/22315381/IJETT-V70I6P210

[24] V. Srividhya and R. Anitha, "Evaluating Preprocessing Techniques in Text Categorization," *International Journal of Computers and Applications*, pp. 49–51, 2010.

[25] C. K. Yau, A. Porter, N. Newman, and A. Suominen, "Clustering Scientific Documents with Topic Modeling," *Science*, vol. 100, no. 3, pp. 767–786, 2014. Crossref, https://doi.org/10.1007/S11192-014-1321-8

[26] R. Priyadarshinia, K. Anurathab, N. Rajendranc, S. Jeyanthid and S. Sujeetha, "LeDoCl : A Semantic Model for Legal Documents Classification using Ensemble Methods," *Turkish Journal of Computer and Mathematics Education*, vol. 12, no. 9, pp. 1899–1908, 2021. Crossref, https://doi.org/10.17762/turcomat.v12i9.3619

[27] Z. Tong and H. Zhang, "A Text Mining Research Based on LDA Topic Modelling," pp. 201–210, 2016. Crossref, https://doi.org/10.5121/CSIT.2016.60616

[28] M. George, P. Soundarabai, and K. Krishnamurthi, "Impact of Topic Modelling Methods and Text Classification Techniques in Text Mining: A Survey, vol. 4, no " 2017.

[29] D. Korenčić, S. Ristov, and J. Šnajder, "Document-Based Topic Coherence Measures for News Media Text," *Expert Systems with Applications*, vol. 114, pp. 357–373, 2018. Crossref, https://doi.org/10.1016/J.ESWA.2018.07.063

[30] J. C. Campbell, A. Hindle, and E. Stroulia, "Latent Dirichlet Allocation: Extracting Topics from Software Engineering Data," *The Art and Science of Analyzing Software Data*, pp. 139–159, 2015. Crossref, https://doi.org/10.1016/B978-0-12-411519-4.00006-9.

[31] "In Search of Coherence and Consensus: Measuring the Interpretability of Statistical Topics," *The Journal of Machine Learning Research*, vol. 18, no. 1, pp 6177–6208, 2017. Crossref, https://dl.acm.org/doi/10.5555/3122009.3242026

[32] K. Stevens, P. Kegelmeyer, D. Andrzejewski, and D. Buttler, "Exploring Topic Coherence over Many Models and Many Topics." *Association for Computational Linguistics,* pp. 952–961, 2012.

[33] Sonali Rajguru, Ajay Chavan, Ashutosh Rane, Shwetali Shinde and Prof. Monika Dangore, "Interview Preparation by Chatbots," *SSRG International Journal of Computer Science and Engineering,* vol. 5, no. 12, pp. 14-15, 2018. Crossref, https://doi.org/10.14445/23488387/IJCSE-V5I12P103

[34] J. Devezas and S. Nunes, "Characterizing the Hypergraph-of-Entity and the Structural Impact of its Extensions," *Applied Network Science,* vol. 5, no. 1, pp. 1–42, 2020. Crossref, https://doi.org/10.1007/S41109-020-00320-Z/TABLES/9

[35] J. Y. Yeh, H. R. Ke, W. P. Yang, and I. H. Meng, "Text Summarization Using a Trainable Summarizer and Latent Semantic Analysis," *Information Processing and Management,* vol. 41, no. 1, pp. 75–95, 2005. Crossref, https://doi.org/10.1016/J.IPM.2004.04.003