Proceedings of the 4th International Conference on Signal Processing and Machine Learning
DOI: 10.54254/2755-2721/52/20241322

Performance evaluation of Latent Dirichlet Allocation on legal documents

A O Ogundare[1,5], A U Saleh[2,7], O A James[3,8], E E Ajayi[4,9] and S Gostojić[1,6]

[1]Department of Computing and Control Engineering, University of Novi Sad, Serbia
[2]Department of Computer Engineering, Istanbul Ticaret University, Turkey
[3]Department of Software Engineering, Federal University of Technology Owerri, Nigeria
[4]Department of Information Systems and Technology, University of Belgrade, Serbia

[5]adedotun@uns.ac.rs
[6]gostojic@uns.ac.rs
[7]abdullahi.saleh@aun.edu.ng
[8]orban.james@futo.edu.ng
[9]ea20213132@student.fon.bg.ac.rs

Abstract. Latent Dirichlet Allocation (LDA) is an algorithm with the capability of processing large amount of text data. In this study, the LDA is used to produce topic modelling of topic clusters from corpus of legal texts generated under 4 topics within Nigeria context– *Employment Contract, Election Petition, Deeds, and Articles of Incorporation*. Each topic has a substantial number of articles and the LDA method proves effective in extracting topics and generating index words that are in each topic cluster. At the end of experimentation, results are compared with manually pre-annotated dataset for validation purpose and the results show high accuracy.

The LDA output shows optimal performance in the word indexing processing for *Election Petition* as all the documents annotated under the topic were accurately classified.

Keywords: Latent Dirichlet Allocation, Probabilistic Latent Semantic Indexing, Latent Semantic Indexing.

1. Introduction

Topic Modeling allows us to analyze a large volume of texts by clustering them into topics. The texts are usually unlabeled making it impossible to apply conventional supervised learning approaches to create machine learning models on them [1].

As the field of legal informatics continue to evolve, there have been an exponential growth of digital contents within the legal domain and this has presented both opportunities and challenges. There are vast repositories of legal texts out there, ranging from court cases and statutes to contracts, legal opinions and so on. It has become paramount for researchers, practitioners and policymakers to harness the potentials of this textual data. The field of legal text mining emerged in response to this and to navigate the plethora of legal information out there in order to uncover valuable insights from them.

Topic modeling exists at the heart of legal text mining. It is a computational technique that discerns and categorizes underlying themes and patterns present in large collection of legal documents. It does this by identifying and extracting latent topics automatically from unstructured textual data. As such, it has been widely perceived in the research world as having the potential to revolutionize the way legal professionals digest, analyze and leverage the vast quantities of legal information they contend with every day [2].

In this paper, topic modeling is explored within the tenets of legal text mining to explore the broader context of legal informatics. The paper aims to evaluate the effectiveness of LDA in analyzing and classifying legal documents obtained from Nigerian jurisprudence. This leads to the research question: How effective is LDA in classifying legal documents from Nigeria? The relevance of this study can be envisioned from two perspectives. First, the application of Natural Language Processing, and in particular LDA, is still relatively new in the law domain. Secondly, the legal domain in Nigeria is currently transitioning into the digital space with the introduction of new frameworks for digitizing legal documents such as case files and court records, aimed at improving case management. Prior to 2019, Nigerian legal documents were rarely available in digital format [3]. Effective classification of legal documents is important for an effective workflow of the various government agencies in Nigeria. However, the various departments of government still face challenges as it becomes more difficult to manually classify the huge number of legal documents they deal with periodically. This study is a step towards solving this problem by evaluating the performance of LDA in classifying legal documents within Nigerian jurisdiction.

To the best of our knowledge, there has been no prior work of this scope that analyzes legal documents within Nigerian jurisprudence.

2. Literature Review

LDA is based on the Dirichlet Distribution, which is a popular probability distribution in literature. For its effective application in Topic Modeling, LDA thrives on the assumptions that documents with similar topics use similar group of words [4].

Topic modeling concept is often explained in context with key terms like “words”, “documents” and “corpora.” Word is the basic unit of discrete data in a document. As expressed by the authors in [5], each document in the corpus contains its own proportions of the topics discussed according to the words contained in them. Topic model is a set of algorithms that uncover the thematic structures hidden in a collection of documents [6]. This algorithm helps us develop new ways to search, and summarize large text archives. As such, topic modeling has been used extensively in the application of Natural Language Processing.

Latent Dirichlet Allocation (LDA) is currently the most popular topic modeling method. Since its successful application in analyzing very large documents, it has been used to summarize, cluster, connect or process very large data. It works by generating a list of topics that are then weighted for each document. The Dirichlet distribution is used to obtain the distribution per-document topics.

There have been a number of text classification studies arising from the huge number of electronic data that is being generated in today’s electronic age. However, a relatively few numbers of those studies have been tailored to the legal domain

The authors in [7] carried out a performance evaluation of LDA in text mining where they introduced three classic models of statistical topic models and compared their performance. The other two models that were compared with LDA in their work are Latent Semantic Indexing (LSI) and Probabilistic Latent Semantic Indexing (PLSI). They used perplexity as the evaluation metric by splitting the corpus into training set and test set and then measure perplexity on the test set. They found LDA to outperform the other two models in topic modeling.

The authors in [8] applied topic modelling on twitter data using Latent Dirichlet Allocation method and they did topic clusters on tweet data generated under 4 topics – economic, Military, Sports and Technology. The pre-labelling of the corpus enabled them to assess how well the model categorized the data. Their evaluation method was similar to the one employed in this research work because the legal data used for the analysis were pre-categorized into four flavors.

The authors in [9] applied LDA to classify news articles in Indonesia. The method employed in their work followed similar approach to this study but it has been applied in the context of media in Indonesia. They perform their analysis using several parameters which includes the number of topics. They presented an overall accuracy of about 70% for classifying the news articles into five classes—economic, tourism, criminal, sports and politics. Classification of news articles has been one of the domains that have been widely studied.

Beyond the realm of articles, LDA has also been applied to classify lyrics in songs with the goal of identifying and categorizing the storyline of songs. In their study, authors in [10] examined Indonesian Song Lyrics Topic Modeling using LDA and discovered that LDA provides an effective means of interpreting the topics within numerous songs by providing information about the top words in each topic and the topic probabilities for each document (song).

Similarly, authors in [11] conducted a comparable study in which they classified the mood of Hindi songs based on lyrical analysis, utilizing LDA. Their research was inspired by the growing popularity of Hindi music on the web due to the rapid increase in the digitization of Hindi content. They observed positive results in classifying the mood of different songs based on lyrical taxonomy.

LDA has also been applied in the topic modeling of Twitter data, as demonstrated in the work of [12]. The LDA method serves as an algorithm for producing topic modeling, determining topic similarity, and visualizing topic clusters from the tweet data, resulting in the identification of four topics (Economic, Military, Sports, Technology). All of these studies highlight the promise of LDA in classifying documents based on similarities in words. However, no study has been specifically tailored to the Nigerian jurisdiction.

3. Research Methodology

Several steps were carefully followed in this study in order to obtain an accurate analysis. The study uses the Latent Dirichlet Allocation (LDA) method to model topics and produce different clusters in each topic. Python programing language was used as the implementation language to process the data and implement LDA methods. In the reference section, a link to the repository containing the source codes and data is provided [13].

The data were collected from the registry department of the Office of the Accountant General of the Federation. Upon requesting access to a corpus of legal documents, the procedures of the office were followed in releasing the documents, which were provided in print format. Thereafter, the documents were converted to a digital format by scanning them. Subsequently Optical Character Recognition tool is used to extract words from each document in order to convert them into textual data. Thereafter, each document was prepared into a comma separated file for analysis.

Some preprocessing was done using count vectorizer to vectorize the data. Rather than tokenizing each article, stop-words were removed using Count Vectorizer. Generic legal terms such as law, constitution, government, Nigeria were removed from the data. Thereafter, the data was converted into Document Term Matrix which indicated a total of 42 articles by a total of 5,700 words. Each article represents a legal document which falls in one of the four categories – *Employment Contract, Election Petition, Deeds, and Articles of Incorporation*.

Following that, LDA was performed on the data. Maximum document frequency was set to 90% such that words that show up in 90% of the documents are not considered. This is an additional step taken to the common legal terms that were added to stop words in the preprocessing stage. This measure ensures that words that are really common across the documents are gotten rid of. In same vein, the minimum document frequency was set to 5% such that for a word to be counted in the vectorizer, it has to show up in at least 5% of the documents. This allows avoiding vectorizing words that are totally unique to a single article.

The number of components is set to 4 to represent the categories of legal texts that were originally collected for the study— *Employment Contract, Election Petition, Deeds, and Articles of Incorporation*. A random state is set to 77 in order to make it possible to repeat the analysis and get exactly the same outcome. The document term matrix created from the dataset is fitted into the instance of LDA.

Consequently, the top 15 words for each topic is queried and assessed. The probability of each document belonging to a topic is also queried and they are compared to the words associated to each topic for assessment.

4. Results, findings and evaluation

After performing the LDA, three steps were followed to interpret the results.

a. Get the vocabulary of words
b. Get the topic components (total components: 4)
c. Get the highest probability words per topic

The assessment of the results was done by assessing the top keywords for each topic; and it does truly appear that the keywords were peculiar to each topic that makes up the component. This is in line with the fundamental principle of LDA which thrives on the assumption that documents with similar topics use similar group of words. For instance, the top 15 words for Employment contract came out as employee, employer, terms and conditions, salary, probational period, termination, notice period, job description, full-time, part-time, severance package, contract duration and insurance. Table 1 below shows the top 15 keywords by each group

Table 1. Top keywords per topic.

Topics	**Top Words**
Topic 1 (**Employment contract**)	employee, employer, terms, salary, probation, termination, notice, job, full-time, part-time, severance, contract duration and insurance
Topic 2 (**Election Petition**)	Respondent, non-compliance, election, lawful, vote, computation, INEC, EC8A, transmitted, accreditation, petitioner, winner, declaration, number, registered
Topic 3 (**Deeds**)	Title, vendor, purchaser, property, transaction, amount, defaults, remedies, served, agreement, arbitration, procure, record, conditions, failures
Topic 4 (**Articles of incorporation**)	Guarantee, registered, indemnities, mortgage, financial, commercial, authorities, pension, funds, stocks, securities, bonds, liquidate, deposits, managers

Table 2. classification performance.

	No. of Docs manually annotated	**LDA accurate classification**	**Misclassified**
Topic 1	10	10	1
Topic 2	10	10	0
Topic 3	10	10	2
Topic 4	12	9	0

Evaluating the accuracy of topic modeling, particularly Latent Dirichlet Allocation (LDA), can be a bit challenging as it is an unsupervised learning technique. Unlike supervised learning, where there are labeled data to measure accuracy, topic modeling does not have predefined topics or ground truth to compare against. However, there are some evaluation metrics and techniques that have been used in past literatures to evaluate the quality of the topics generated by LDA. Among these standardized techniques are manual inspection, human evaluation and topic labeling. While quantitative metrics can provide some guidance, qualitative assessment by manually inspecting the topics is essential, especially in this case where a relatively small number of documents is involved. In this experiment, the generated topics are examined to see if they make sense, and if they represent the four original distinct themes upon which the legal texts were collected— *Employment Contract, Election Petition, Deeds, and Articles of Incorporation*. As indicated in Table 1 and Table 2, each of these topics were annotated as Topic 1, Topic 2, Topic 3 and Topic 4, respectively. This helped in determining how much the topic model captures meaningful patterns in the dataset. The quality of topic labels was further evaluated. After generating topics, meaningful labels were manually assigned to each topic based on the four categories of the legal documents obtained. Thereafter they were analyzed to see if the words within the topics align well with the assigned labels.

The outcome as shown in *Table 2* indicates that in few cases, the total article classified exceeds what is expected. For instance, the total number of articles expected in Topic 1 is 10, as manually inspected and annotated in the table. While all the topics in the category were correctly classified, one was misclassified, implying that the total number of documents returned in that category is 11 instead of 10. A further examination shows the accurate destination of the misclassified document is *Topic 4.* Similarly in Topic 3, two articles were misclassified.

5. Conclusion

The outcome of this research shows that Latent Dirichlet Allocation is effective in classifying legal documents. In the past, most of the topic classification works have been based on media, lyrics, and other sources. There have also been notable attempts in carrying out performance evaluations of LDA in text mining, such as the work of the authors in [12]. The strength of this research lies in tailoring the application to a specific domain of law and in specific jurisdiction, i.e Nigeria. Future research should expand on the performance metrics by considering other important parameters, such as an increased amount of data and computing speed. More specifically, future work should consider parallel or distributed computing to simulate a more realistic application for solving problems in the legal domain in Nigeria, where there is often a huge amount of data. Therefore, the experiment should be repeated on a huge corpus of data in varying computing environment.

References

[1] Aixin, S. and L. Ee-Peng. "Hierarchical text classification and evaluation ." *Proceedings IEEE International Conference* (2001): 521-528.

[2] Campbell, J.C, A. Hindle and E. Stroulia. "Latent Dirichlet Allocation: extracting topics from software engineering data." Morgan, Kaufmann. *The art and science of analyzing software data*. 2014. 139-1599.

[3] Desola, L.-O. (2023). Nigeria’s quest for digital transformation; Mirage or reality? Hertie School Centre for Digital Governance. Retrieved from https://www.hertie-school.org/en/digital-governance/research/blog/detail/content/nigerias-quest-for-digital-transformation-mirage-or-reality.

[4] Edi, S. N and A. Ria. "A Review on Overlapping and Non-Overlapping Community Detection Algorithms for Social Network Analytics ." *Far East Journal of Electronics and Communications* (2018): 1-27.

[5] Edi, Surya, Triadi Dendi and Andryani Ria. "Topic Modelling Twitter Data with Latent Dirichlet Allocation." *IEEE* (2019): 718-129.

[6] Fabrizio, S. *Machine Learning in automated text categorization*. ACM Computing Surveys, 2002.

[7] Liu, Z., & Ponraj, M. (2011). Performance Evaluation of Latent Dirichlet Allocation in text mining. International Conference on Fuzzy Systems and Knowledge Discovery (FSKD), (pp. 2695-2698). Shanghai. Retrieved from https://ezproxy.ticaret.edu.tr:2166/document/6020066

[8] Negara, E., Triadi, D., & Andryani, R. (2019). Topic Modelling Twitter Data with Latent Dirichlet Allocation Method. *International Conference on Electrical Engineering and Computer Science*, (pp. 286-290). Batam

[9] Kusumaningrum, R., Wiedjayanto, M., & Adhy, S. (2016). Classification of Indonesian news articles based on Latent Dirichlet Allocation. 2016 International Conference on Data and Software Engineering, (pp. 1-5). Denpasar.

[10] Laoh, E., Surjandari, I., & Febirautami, R. (2018). Indonesian Song Lyrics Topic Modelling Using Latent Dirichlet Allocation. 5th International Conference on Information Science and

Control Engineering, (pp.270-274). Zhengzhou. Retrieved from https://ezproxy.ticaret.edu.tr:2166/document/8612562

[11] Chauhan, S., & Chauhan, P. (2016). Music Mood classification based on lyrical analysis of hindi Songs using Latent Dirichlet Allocation. 2016 International Conference on Information Technology,(pp.72-76).Noida, India. Retrieved from https://ezproxy.ticaret.edu.tr:2166/document/7857593

[12] Negara, E., Triadi, D., & Andryani, R. (2019). Topic Modelling Twitter Data with Latent Dirichlet Allocation Method. *International Conference on Electrical Engineering and Computer Science*, (pp. 286-290). Batam

[13] Ogundare, D. (2023). Topic Modelling Legal Documents (v1.0.0). GitHub. https://github.com/dotun-ogundare/topic-modelling-legal-documents

Acknowledgment

We specially acknowledge Ambassador (Dr.) Eniola Ajayi, former Nigerian Ambassador to Hungary who is currently serving the same function as Nigerian Ambassador to the Kingdom of Netherlands, for relentlessly supporting research and educational initiatives of Nigerian scholars. We extend our gratitude to Barrister Adelani Ajibade at Wole Olanipekun & co, who facilitated the provision of data used for this study.