Application Research on Semantic Analysis using Latent Dirichlet Allocation and Collapsed Gibbs Sampling for Topic Discovery

***Abstract*-Topic discovery is a process of identifying the main topics present in a collection of documents. It is a crucial step in text mining, digital humanities, and information retrieval, as it allows one to extract meaningful information from large volumes of unstructured text data. The most widely used algorithm for topic discovery is Latent Dirichlet Allocation (LDA). LDA assumes that the words in each document are generated by a small number of underlying topics, and the algorithm learns the topics from the text data automatically. One of the main problems of LDA is that the topics extracted are of poor quality if the document does notcoherently belong to a single topic. However, Gibbs sampling operates on a word-by-word basis, which allows it to be usedon documents with a variety of topics and modifies the topic assignment of a single word. The paper presents application research on Latent Dirichlet Allocation and Collapsed Gibbs Sampling Semantic Analysis for topic discovery.**

Keywords-Application, Semantic similarity, Topic modeling, LDA, Collapsed Gibbs sampling

I INTRODUCTION

"Topic discovery is a process of identifying the main topics present in a collection of documents. It is a crucial step in text mining, digital humanities, and information retrieval, as it allows one to extract meaningful information from large volumes of unstructured text data. Topic discovery can be achieved through a variety of techniques such as clustering, Latent Dirichlet Allocation (LDA), and Non-Negative Matrix Factorization (NMF), among others. The most widely used algorithm for topic discovery is Latent Dirichlet Allocation (LDA) introduced" by [1]. "LDA models each document as a mixture of latent topics, where a topic is defined as a distribution over words. In other words, LDA assumes that the words in each document are generated by a small number of underlying topics, and the algorithm learns the topics from the text data automatically" [2]. Topic discovery can help to uncover the main topic present in a large collection of documents, reduce the dimensionality of the data, and facilitate further analysis and exploration. For example, it can be used to analyze customer reviews, news articles, scientific publications, etc.

"Topic modeling is a type of statistical modeling technique used to identify latent topics present in a large collection of documents. It is widely used in fields such as digital humanities, topic discovery, information retrieval, and text mining. The most widely used algorithm for topic modeling is Latent Dirichlet Allocation (LDA) introduced" by [1].

"LDA is a popular statistical model for topic modeling, which is the process of discovering the abstract "topics" that occur in a collection of documents. In LDA, each document is modeled as a mixture of latent topics, where a topic is defined as a distribution over words. The basic idea is that the words in each document are generated by a small number of underlying topics, and the topics are learned automatically from the text data" [3].

"Collapsed Gibbs sampling is a popular algorithm for estimating the parameters of LDA. The algorithm is based on the Bayesian framework, which involves specifying a prior distribution over the parameters and using the observed data to compute a posterior distribution. The Gibbs sampler is used to generate a Markov Chain Monte Carlo (MCMC) algorithm that can be used to approximate the posterior distribution" [4]. In the Collapsed Gibbs sampling algorithm, the topic assignments for each word in the document are treated as latent variables and are inferred from the observed data. The algorithm alternates between updating the topic assignments for each word and updating the parameters of the topic distributions. After a sufficient number of iterations, the algorithm converges to a stable posterior distribution, from which the parameters of the topic model can be estimated [5].

Despite the development of a number of methodologies or models for assessing semantic analysis, the effectiveness of learnt subjects still needs to be much improved, according to the findings of their application. The research presents application research on LDA and collapsed gibbs sampling semantic analysis for topic discovery.

The goal of the study is to generate qualitylearned topics using LDA and Gibbs sampling method. The dataset consists of 11,000 newsgroup postings that were gathered from a database available online [6] that comprised 20 different categories, including ICT-related content from Nigerian newspapers' sports, entertainment, politics, and health sections. News sources considered were Punch Nigeria, the Sun daily, metro news, and other publications. The selection of the dataset was based on the ability to produce a range of topics using both LDA and Gibbs sampling. Generating topics for both algorithms is part of the research's overall conclusion. Other findings include the distribution of document word counts by predominate topic, the significance of topic keywords, and visualization of generated topics. Each of the images was edited at 300 dpi.

"Section 2 presents the review of relevant literature. The methodology, data source and text pre-processingare presented in Sect. 3 while Sect. 4 focuses on results and discussions. The conclusion drawn from the research is presented in Sect. 5" [18]

II RELATED WORKS

The study in [7] concentrated on building a semantic model of a textual document for efficient search and retrieval. A novel method for improving textual document search and retrieval was described in the study. The strategy attempted to make use of structured query languages for retrieval and search. The document's semantic model was developed with this objective in mind. The document's semantic model was an ontology-like structured semantic annotation that could be queried in a structured manner. The paper also covered how to generate a semantic model from a text source and how to efficiently search a group of semantic models using structured query language.

"Application research on latent semantic analysis for information retrieval was worked on" by [8]. The machine matching of the key word, or retrieval based on keywords, is the fundamental tenet of the classical conventional information retrieval paradigm. The study suggested a pre-clustering-based latent semantic analysis approach for document retrieval. The time-consuming computation of the similarity between each text vector and the query vector in the conventional latent semantic method for document retrieval might be solved by the algorithm. It first groups the documents using k-means clustering based on latent semantic analysis, identifies the central point of each cluster, and then computes the similarity between the query vector and each cluster's central points for retrieval. It developed a novel approach for computing the feature weights and adopted the pre-clustering method to pre-process document collection in light of the features of document retrieval. The experiment's findings demonstrated that the new method might speed up retrieval and cut down on search time.

Online reviews can reveal a reviewer's emotional tendencies, and one of the key techniques for analysing these tendencies is sentiment categorization.[9] proposed a method for classifying the sentiment of internet reviews based on topic analysis of text documents and semantic analysis of sentiment terms.The research team worked on sentiment classification of online reviews based on LDA and semantic analysis of sentimental words.First, the LDA approach was used to determine the topic of text documents and the topic of words. Then, Word2Vec was utilized for word vector training in order to obtain the vector representation of words in text documents and the words in the emotion dictionary. Calculated semantic similarity between words and documents as well as between text document words and words in sentiment dictionary. Using the two types of semantic similarity, the words with the greatest degree of similarity are chosen as the text document's primary categorization characteristics. The system had a strong performance and could enhance the sentiment classification impact of online reviews, according to experiments.

[10] worked on computer aided diagnosis semantic model for the report of medical image via LDA and LSA.The research team created a probabilistic model for medical text data that contains latent semantics as part of this study. They employed LDA and LSA, respectively, to handle hidden topics and build the semantic structure. The reports of medical images that were utilized had two parts: descriptions of the images, and diagnoses. Traditional tasks increased effort by requiring the expert to diagnose sickness based on the visual description and his or her own subjective assumptions, the accuracy of which depended on the expert's expertise. The team showed algorithms that automatically provided high-readability diagnoses without the need for operator intervention. The categorization and analysis of medical text data might both be done using the learning models.

Although various discriminant analysis approaches have been used in content-based image retrieval (CBIR) application, there have been relatively few concerns with kernel-based methods. Hence, [11] researched on a comparative study of PCA, LDA and Kernel LDA for Image Classification.Additionally, these CBIR applications continued to use face image discriminant analysis in the same manner as face recognition. The team's work focused on relating general semantic notions to visuals. To extract image visual information, they employed the presented symmetrical invariant LBP (SILBP) texture descriptor. The efficiency of the principal component analysis (PCA), fisher linear discriminant analysis (LDA), and kernel LDA algorithms in producing the best possible discriminating features was further investigated. In order to perform kernel discriminant analysis in the application, they proposed an LDA-based framework.

Semantic analysis and text data mining now frequently use Latent Dirichlet Allocation (LDA) and other modified topic models. A significant quantity of text-based data has been generated due to the quick evolution of medical knowledge, much of which is recorded in an illogical manner. [12] worked on a weighted LDA-based medical record semantic analysis. The study effort suggested a redesigned strategy based on LDA that clarified the concept of weighting. Additionally, it proved that the improved approach performed as desired, proving its efficacy.

[13] worked on "LDA-Based retrieval framework for semantic news video retrieval. A two-level LDA retrieval architecture with lexicon guidance was presented in the study. The parameter estimate for the first-level LDA model was guided by the HowNet, and the second-level LDA models were then built using the first-inference level's output. They assessed it using data from the TRECID 2005 ASR collection and contrasted it with results from the vector space model (VSM) and latent semantic indexing (LSI)". [14] worked on "LDA-based model for topic evolution mining on text. Through the technique of text latent semantic analysis on textual data, a text mining model for topical evolutionary analysis was put out. Analysing subject evolution by monitoring changing patterns in the topic over time. Using the LDA model for the corpus and text to obtain the topics, and then the Clarity method to assess the similarity of topics in order to identify topic mutation and unearth the topic concealed in the text. Experiments demonstrated that the suggested approach is capable of identifying significant topical evolution".

The limitations of the works presented in this section include scalability, that is, LDA can be slow and computationally expensive for very large datasets, as it requires a large number of computational resources to estimate the parameters of the model. Also, LDA does not guarantee that the topics discovered by the model will be semantically meaningful or coherent. The topics generated by LDA can sometimes be a mix of multiple concepts, making it difficult to interpret the results (Topic coherence). Other limitations include Hyperparameter tuning, that is, LDA requires the selection of several hyperparameters, such as the number of topics, the prior distributions over the topics, and the smoothing parameters. The performance of LDA can be sensitive to these hyperparameters, and finding the optimal values can be challenging. One of the limitations of LSA that LSA required a large number of computational resources to process large text corpora, and slow when dealing with very large datasets [15].

III METHODOLOGY

In the section below, we presented the algorithms for topic discovery using LDA and Gibbs.

A. Data Source

An online database [6] with 20 different categories, including ICT-related content from the sports, entertainment, politics, and health sections of Nigerian newspapers, was used to compile 11,000 newsgroup messages. Among the news outlets considered were Punch Nigeria, the Sun daily, metro news, and others.

B. Text Pre-processing

To increase accuracy, decrease data redundancy, and shorten model training times, text pre-processing was employed to clean and normalize the text data [16]. Each phrase was tokenized into a list of terms using regular expression patterns, removing all punctuation and superfluous letters. The bigram and trigrams were then created as the following phase. Trigrams are three words that regularly occur together, whereas bigrams are two words that frequently appear together in a text. To find words that commonly occur together and predict the conditional probability of the next words, bigrams and trigrams were used to the dataset. The LDA/Gibbs model was built utilizing 20 different topics, each of which is composed of a number of keywords, each of which gives the topic a certain weight. Table 1 outlines the step-by-step procedure for the first text cleaning.

TABLE 1: Text pre-processing process carried out on newsgroup dataset

Step	Description
Punctuation Removal	Remove punctuation
Word tokenization	Tokenize sentences to sets of words
Lowercase Conversion	Convert words to lowercase
Stopwords Removal	Remove stopword
Lemmatization	Lemmatize the words

Figure 1 shows the results of the pre-processed text.

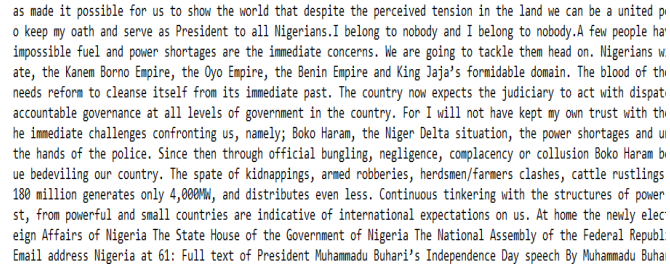
as made it possible for us to show the world that despite the perceived tension in the land we can be a united p
o keep my oath and serve as President to all Nigerians.I belong to nobody and I belong to nobody.A few people ha
impossible fuel and power shortages are the immediate concerns. We are going to tackle them head on. Nigerians w
ate, the Kanem Borno Empire, the Oyo Empire, the Benin Empire and King Jaja's formidable domain. The blood of th
needs reform to cleanse itself from its immediate past. The country now expects the judiciary to act with dispat
accountable governance at all levels of government in the country. For I will not have kept my own trust with th
he immediate challenges confronting us, namely; Boko Haram, the Niger Delta situation, the power shortages and u
the hands of the police. Since then through official bungling, negligence, complacency or collusion Boko Haram b
ue bedeviling our country. The spate of kidnappings, armed robberies, herdsmen/farmers clashes, cattle rustlings
180 million generates only 4,000MW, and distributes even less. Continuous tinkering with the structures of power
st, from powerful and small countries are indicative of international expectations on us. At home the newly elec
eign Affairs of Nigeria The State House of the Government of Nigeria The National Assembly of the Federal Republ
Email address Nigeria at 61: Full text of President Muhammadu Buhari's Independence Day speech By Muhammadu Buha

Figure 1: Snippet of the pre-processed text

C. LDA/Gibbs Algorithm

LDA is a generative probabilistic model for topic modeling that assumes that each document in a corpus is a mixture of latent topics, where each topic is characterized by a distribution over words. Gibbs sampling operates on a word-by-word basis, which allows it to be used on documents with a variety of topics and modifies the topic assignment of a single word. The algorithm for LDA/Gibbs is as follows:

i. Initialization: The algorithm starts by initializing the parameters of the model, including the topic distributions for each document, the word distributions for each topic, and the hyperparameters, such as the number of topics and the Dirichlet priors for the distributions.

ii. Gibbs Sampling: The core of the LDA algorithm is a Gibbs sampling process that iteratively updates the topic assignments for each word in the corpus. In each iteration, the algorithm samples a new topic assignment for a randomly selected word, based on the current topic assignments of all other words in the document and the current topic distributions.

iii. Convergence: The Gibbs sampling process continues until convergence, which can be determined using various convergence criteria, such as a fixed number of iterations, a minimum change in the log-likelihood of the model, or a minimum change in the topic assignments.

iv. Estimation of Parameters: Once the Gibbs sampling process has converged, the parameters of the model can be estimated, including the topic distributions for each document, the word distributions for each topic, and the hyperparameters.

v. Inference: The estimated parameters can then be used to perform inference, such as finding the most likely topics for a new document, estimating the likelihood of a document given the model, or finding the most likely words for a given topic.

A detailed analysis and mathematics behind the two algorithms were explained in [17].

IV RESULTS AND DISCUSSION

The experiment setup for topics discovery and semantic analysis was implemented on Anaconda Jupyter with GPU capability using the following packages: Numpy, Pandas, pyLDAvis, and the NLTK in python 3.6. LDA is a generative probabilistic model for topic modeling that assumes that each document in a corpus is a mixture of latent topics, where each topic is characterized by a distribution over words. Gibbs operates on the presumption that while the topic assignment of the supplied word is unknown, it is known for all other words in the text, and this information is utilized to deduce the topic that will be given to this word. The objective of the study is to present application research on LDA and Collapsed Gibbs Sampling Semantic Analysis for topic discovery. It is observed in Topic 1 that the country "Nigeria" suffered an "attack", people were "killed", "unknown" "people" killed "children" on "land" and the "government" failed to do anything. Topic 2 was about health and how "doctors" "treated" "patients" with different "diseases" (Table 2 and 3).

Table 2 – Output of Semantic analysis of Topic 1 score

Topic 1	
'people', 0.0231	'kill', 0.0116
'nigeria', 0.009	'attack', 0.009
'war', 0.00959	'government', 0.0092
'arm', 0.009	'unknown', 0.008347
'land', 0.00798	'child', 0.00792

Table 2 (A)– Output of Semantic analysis of Topic 2

Topic 2	
'drug', 0.01279	'study', 0.01138
'food', 0.009110	'doctor', 0.00906
'effect', 0.00825	'problem', 0.007498
'patient', 0.00718	'eat', 0.0067655
'disease', 0.00635	'show', 0.006228

Table 3 – Output of Semantic analysis of Topic 3

Topic 3	
'work', 0.01599	'year', 0.01382
'make', 0.01372	'money', 0.01311
'government', 0.01186	'pay', 0.01037
'job', 0.0101	'support', 0.00949
'people', 0.007417	'vote', 0.006959

Table 4 – Output of Semantic analysis of Topic 4

Topic 4	
'gun', 0.028036	'law', 0.02185
'crime', 0.0115	'fire', 0.01121
'people', 0.0110	'kill', 0.010598
'weapon', 0.01048	'case', 0.01004
'state', 0.01004	'police', 0.00987

Table 5 – Output of Semantic analysis of Topic 5

Topic 5	
'drive', 0.03188	'card', 0.02697
'problem', 0.0257	'system', 0.02026
'bit', 0.017416	'driver', 0.017268
'work', 0.01559	'run', 0.01348
'disk', 0.01348	online, 0.01285

Gibbs operates on the presumption that while the topic assignment of the supplied word is unknown, it is known for all other words in the text, and this information is utilized to deduce the topic that will be given to this word. The objective of the research is to present application research on LDA and Collapsed Gibbs sampling semantic analysis for topic discovery. The following obesrvations was derived:

i. In Topic 1 that the country “Nigeria” suffered an “attack”, people were “killed”, “unknown” “people” killed “children” on “land” and the “government” failed to do anything.

ii. Topic 2 was about health and how “doctors” “treated” “patients” with different “diseases” as seen in Tables 2A.

A. Word Count and Importance of Topic Keyword

The significance of the keywords in the topics as indicated by the weights is plotted in figure 4. Also, the frequency with which the words have occurred in the pre-processsed datasets is plotted. As can be seen, some words are shared among the topic [17]. The common words to stop words can be added to make sure they are not considered, but the reduction of several topics is a better solution as there is overlapping. Topics keywords is plotted as shown below in fig. 2.

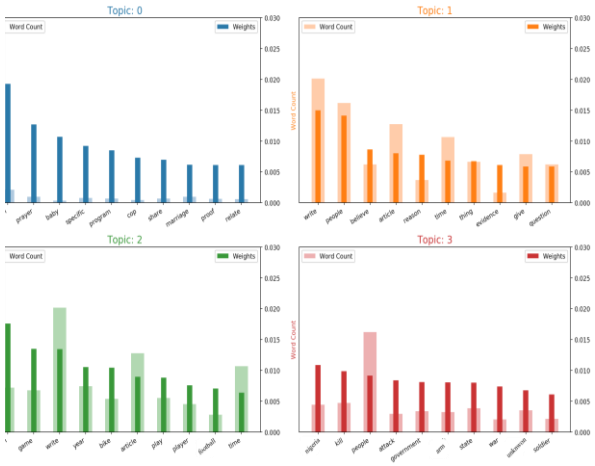

Fig. 2. Word Count and Importance of Topic Keyword

V CONCLUSION

This study provided with understanding of topic discoveryfrom a dataset. One of major issue of LDA is that the quality of topics extracted is poor if the document does not coherently discuss a single topic. However, using Gibbs sampling with LDA uses word by word basis which changes the topic assignment of one word and can be used on documents having different topics but cannot be used for very large dataset. The work sort to analysis semantically the weightage of each of the topics generated across the 20 topics considered over the dataset using LDA/Gibbs. There are few other techniques which can be used which can done in future works. Also, provision of document to word count visualization can also be done in future work.

REFERENCES

[1] D.M. Blei,Ng, A. Y., &M.I., Jordan. (2003). Latent Dirichlet allocation. Journal of machine learning research, 3(Jan), 993-1022.

[2] J. Chang, K.Q. Weinberger, &D.M. Blei, (2009). Reading tea leaves: How humans interpret topic models. In Advances in neural information processing systems (pp. 288-296).

[3] T.L. Griffiths, &M. Steyvers. (2004). Finding scientific topics. Proceedings of the National Academy of Sciences of the United States of America, 101(Suppl 1), 5228-5235.

[4] H.M. Wallach, M.N. Murty, &J. Featherston (2009). Evaluating topic models. In Proceedings of the 26th Annual International Conference on Machine Learning (pp. 1105-1112). ACM.

[5] P. Pinoli, D. Chicco and M. Masseroli, "Latent Dirichlet Allocation based on Gibbs Sampling for gene function prediction," *2014 IEEE Conference on Computational Intelligence in Bioinformatics and Computational Biology*, Honolulu, HI, USA, 2014, pp. 1-8, doi: 10.1109/CIBCB.2014.6845514.

[6] Newsgroup Master dataset. Retrieved from https://raw.githubusercontent.com/selva86/datasets/master/newsgroups.json

[7] E. Nyamsuren and Ho-Jin Choi, "Building a semantic model of a textual document for efficient search and retrieval," *2009 11th International Conference on Advanced Communication Technology*, Phoenix Park, 2009, pp. 298-302.

[8] C. Wenli, "Application Research on Latent Semantic Analysis for Information Retrieval," 2016 Eighth International Conference on Measuring Technology and Mechatronics Automation (ICMTMA), Macau, China, 2016, pp. 118-121, doi: 10.1109/ICMTMA.2016.37.

[9] Y. Jing, H. Gou, C. Fu and W. Sun, "Sentiment Classification of Online Reviews Based on LDA and Semantic Analysis of Sentimental Words," *2019 12th International Symposium on Computational Intelligence and Design (ISCID)*, Hangzhou, China, 2019, pp. 249-252, doi: 10.1109/ISCID.2019.00064.

[10] Bo Li and Ke Wang, "Computer aided diagnosis semantic model for the report of medical image via LDA and LSA," *2011 IEEE International Symposium on IT in Medicine and Education*, Cuangzhou, 2011, pp. 699-703, doi: 10.1109/ITiME.2011.6130756.

[11] F. Ye, Z. Shi and Z. Shi, "A Comparative Study of PCA, LDA and Kernel LDA for Image Classification," *2009 International Symposium on Ubiquitous Virtual Reality*, Guangju, Korea (South), 2009, pp. 51-54, doi: 10.1109/ISUVR.2009.26.

[12] X. Jin, Y. Lan and C. Ma, "Medical Record Semantic Analysis Based on Weighted LDA," *2016 9th International Symposium on Computational Intelligence and Design (ISCID)*, Hangzhou, China, 2016, pp. 158-161, doi: 10.1109/ISCID.2016.1044.

[13] J. Cao, J. Li, Y. Zhang and S. Tang, "LDA-Based Retrieval Framework for Semantic News Video Retrieval," *International Conference on Semantic Computing (ICSC 2007)*, Irvine, CA, USA, 2007, pp. 155-160, doi: 10.1109/ICSC.2007.26.

[14] Q. Wu, X. Deng, C. Zhang and C. Jiang, "LDA-based model for topic evolution mining on text," *2011 6th International Conference on Computer Science & Education (ICCSE)*, Singapore, 2011, pp. 946-949, doi: 10.1109/ICCSE.2011.6028792.

[15] T.K. Landauer, P.W. Foltz, &D. Laham. (1998). An introduction to latent semantic analysis. Discourse processes, 25(2-3), 259-284.

[16] Z.J. Hew, V.J. Olanrewaju, X.Y. Chew, K.W. Khaw: Text summarization for news articles by machine learning techniques. J. Appl.Math. Comput. Intell. (2022)

[17] M.O. Ajinaja, A.O.Adetunmbi, C.C. Ugwu&Olugbemiga S.P. Semantic similarity measure for topic modeling using latent Dirichlet allocation and collapsed Gibbs sampling. Iran Journal Computer Science (2022). https://doi.org/10.1007/s42044-022-00124-7

[18] Ajinaja MO, Adetunmbi AO, Ugwu CC, Popoola OS. Semantic similarity measure for topic modeling using latent Dirichlet allocation and collapsed Gibbs sampling. Iran Journal of Computer Science. 2023 Mar;6(1):81-94.