INDIAN JOURNAL OF SCIENCE AND TECHNOLOGY

RESEARCH ARTICLE

A comparative analysis of Latent Semantic analysis and Latent Dirichlet allocation topic modeling methods using Bible data

Vasantha Kumari Garbhapu[1]*, **Prajna Bodapati**[2]

1 Research Scholar, Department of Computer Science and Systems Engineering, Andhra University College of Engineering, Visakhapatnam, A.P, India
2 Professor, Department of Computer Science and Systems Engineering, Andhra University College of Engineering, Visakhapatnam, 530003, A.P, India

OPEN ACCESS

Received: 20.08.2020

Accepted: 03.12.2020

Published: 15.12.2020

Citation: Garbhapu VK, Bodapati P (2020) A comparative analysis of Latent Semantic analysis and Latent Dirichlet allocation topic modeling methods using Bible data. Indian Journal of Science and Technology 13(44): 4474-4482. https://doi.org/10.17485/IJST/v13i44.1479

* **Corresponding author**.

vasanthagarbhapu@gmail.com

Funding: None

Competing Interests: None

Published By Indian Society for Education and Environment (iSee)

ISSN
Print: 0974-6846
Electronic: 0974-5645

Abstract

Objective: To compare the topic modeling techniques, as no free lunch theorem states that under a uniform distribution over search problems, all machine learning algorithms perform equally. Hence, here, we compare Latent Semantic Analysis (LSA) or Latent Dirichlet Allocation (LDA) to identify better performer for English bible data set which has not been studied yet. **Methods:** This comparative study divided into three levels: In the first level, bible data was extracted from the sources and preprocessed to remove the words and characters which were not useful to obtain the semantic structures or necessary patterns to make the meaningful corpus. In the second level, the preprocessed data were converted into a bag of words and numerical statistic TF-IDF (Term Frequency – Inverse Document Frequency) is used to assess how relevant a word is to a document in a corpus. In the third level, Latent Semantic analysis and Latent Dirichlet Allocations methods were applied over the resultant corpus to study the feasibility of the techniques. **Findings:** Based on our evaluation, we observed that the LDA achieves 60 to 75% superior performance when compared to LSA using document similarity within-corpus, document similarity with the unseen document. Additionally, LDA showed better coherence score (0.58018) than LSA (0.50395). Moreover, when compared to any word within-corpus, the word association showed better results with LDA. Some words have homonyms based on the context; for example, in the bible; bear has a meaning of punishment and birth. In our study, LDA word association results are almost near to human word associations when compared to LSA. **Novelty:** LDA was found to be the computationally efficient and interpretable method in adopting the English Bible dataset of New International Version that was not yet created.

Keywords: Topic modeling; LSA; LDA; word association; document similarity; Bible data set

1 Introduction

There are many text mining methods to turn unstructured textual data into actionable information. While traditional methods to analyze texts are limited in processing large amounts of data, some researchers have applied text mining to qualitative research projects. Due to these research advancements, text mining is viewed as a viable qualitative research method in machine learning and natural language processing efficiently[(1–3)]. These computer applications closely follow the paradigm of a common technique, topic modeling in the field of text mining. The topic models allow in analyzing a set of documents based on statistics of words in each, to express what the topic might be and what each document's balance of topics. It also refers to a probabilistic topic model to use statistical algorithms for discovering hidden topics of the collection of documents[(4–6)]. The significant and crucial step in the accuracy and storage of the information is quality management and extraction according to the information that is present.

There are various methods of text mining to identify the underlying topics in the text. This study compares the results of applying Latent Semantic Analysis (LSA)[(7–9)], a natural language processing technique, and Latent Dirichlet Allocation (LDA)[(10–13)], a type of probabilistic topic modeling, to the text field. The outputs may help to determine and to demonstrate the feasibility of the technique if the use of these two models leads to additional insights when applied to the English language bible as a dataset. The dataset used in this comparative study is from the New International Version (NIV) available online at h ttp://www.biblegateway.com/passage/?search=Genesis+1&version=NIV.

The dataset includes text fields that describe each incident and the length of the text field different from a few words to paragraphs with more than a few sentences. These text data were mined to reveal additional knowledge about incidents in the bible. Data were collected from the Book of Genesis, the first book of the bible and the old testament, It is an account of the creation, life on earth, beginning of sin, the fallen state of the world, the need for a redeemer, and the promise of His coming. All these centre on the covenants that linking God to his chosen people and the people to the Promised Land.

2 The used methodologies

The details of the two well-known information retrieval methodologies, LSA, and LDA are presented in this section. This section demonstrates how these two text mining algorithms use different mechanisms to automatically generate the topics (A topic is a grouping of related words) in the text corpus.

2.1 Latent Semantic Analysis (LSA)

Latent Semantic Analysis is a method for representing and extracting the contextual meaning of words through statistical computations over a text corpus[(6)],[(14,15)]. It is formerly known as Latent Semantic Indexing (LSI)[(16)], before LSI, Information is fetched by accurately matching words in documents with the queries using lexical matching methods. These methods made the Information retrieval difficult because of two problems one is synonyms (missing documents regarding ”automobile” when querying on ”car”) and another polysemy (retrieving the documents about a financial bank when querying on the river bank)[(16)]. To work out these two problems and other similar issues, the documents are expressed as concealed concepts in preference to terms. The hidden structure is not a fixed mapping between terms and hidden concepts, but it depends on the underlying document and correlation between the words it contains.

Moreover, it has been recently established that it is possible to give a statistical interpretation of the traditional Latent Semantic Analysis (LSA) paradigm, which collects hidden concepts from the document of the corpus using a linear algebra technique known as ”Singular Value Decomposition” (SVD) technique[(17)]. The SVD represents the term-document matrix A_{nxm} as the product of three matrices $A=USV^{T.}$ Where $S = (\sigma_1, \sigma_2,.., \sigma_r\,)$ is an r x r matrix, $U=(u_1,\ldots,u_r\,)$ is an n x r matrix, $V=(v_1,\ldots v_r\,)$ is an m x r matrix; However, columns in both matrices are orthonormal and r is minimum (n,m). The algorithm, as shown in Supplementary Table 1, LSA works by keeping the K largest singular values in the above decomposition, for some appropriate k. Let $S_k =(\sigma_1,.., \sigma_k\,)$, $U=(u_1,\ldots,u_k\,)$ and $V=(v_1,\ldots v_k\,)$. Then $A=U_kS_kV_k^T$

A is a matrix of rank k, which is our approximation of A. The rows of V_kS_k above are then used to correspond to the documents. This new space (latent semantic space) is used to analyze semantic relatedness among the documents (within-corpus and outside of the corpus) and words. It is also useful for information retrieval and information filtering and performs well if the corpus is a collection of meaningfully correlated documents[(16)],[(18)].

2.2 Latent Dirichlet Allocation (LDA)

Topic modeling algorithms are statistical methods that analyze the words of unstructured original texts to discover the themes that run through them automatically. LDA is a generative probabilistic model for the collection of discrete data as a corpus. It was first introduced by Devid Blei et al.[10]. The basic idea is that documents are represented as a random mixture over latent topics, where a Dirichlet distribution over words characterizes each topic to find topics in documents, or LDA identifies a set of topics by associating a set of words to each topic[19,20]. The underlying assumption of LDA is that a text document will consist of multiple themes and has a three-level hierarchical Bayesian model where each item of a collection of text is modeled as a finite mixture over an underlying set of topics. Each topic is, in turn, modeled as an infinite mixture over an underlying set of word probabilities.

According to the algorithm shown in Supplementary Table 2, in words, this means that there are K topics. Their distributions are $\varphi_1, \ldots ,_k$ are derived from Dirichlet (β) that are shared among all documents. Each document in the corpus D is considered as a mixture over these topics, indicated by θj. Then we generate the words for document dj by first sampling a topic assignment $z_{j,t}$ from the topic proportions θj, and then sampling a word from the corresponding topic $\varphi_{zj,t}$. $z_{j,t}$ is a variable it denotes which topic from 1, …., k was selected for the t-th word in document dj.

It is essential to identify some critical assumptions with this model. First, we assume that the number of topics K is a fixed quantity known beforehand, the number of distinct words in the dictionary is fixed and known ahead of time, and each φ_k is a fixed quantity to be approximated. Each word within a document and topic proportion θj is not dependent.

In this formulation, we can see that the joint distribution of the topic mixture Θ, the set of topic assignments Z, the words of the corpus W, and the topics Φ by

$$P(W,Z,\Theta,\Phi \mid \alpha,\beta) = \prod_{i=1}^{k} P(\Phi_{\mathrm{k}} \mid \beta) \prod_{j=1}^{M} P(\theta_j \mid \alpha) \prod_{t=1}^{N_j} P(Z_{j,t} \mid \theta_j) P\left(W_{j,t} \mid \phi_{z_{j,t}}\right)$$

2.3 Experimental step of the analysis

The experimental model diagram illustrated as in Figure 1 represents the steps of analysis in this research study from the input (Raw bible data) followed by preprocessing to remove the noise in the data and further removal of stop words and to find the root word of the given word by Lemmatization process to further assess the text to vector conversion and comparison of two topic modeling methods (LSA and LDA) in identifying the document similarity within-corpus and with the unseen document to categorize the word associations and coherence score as a measure for topic comparison and goodness of the topic model.

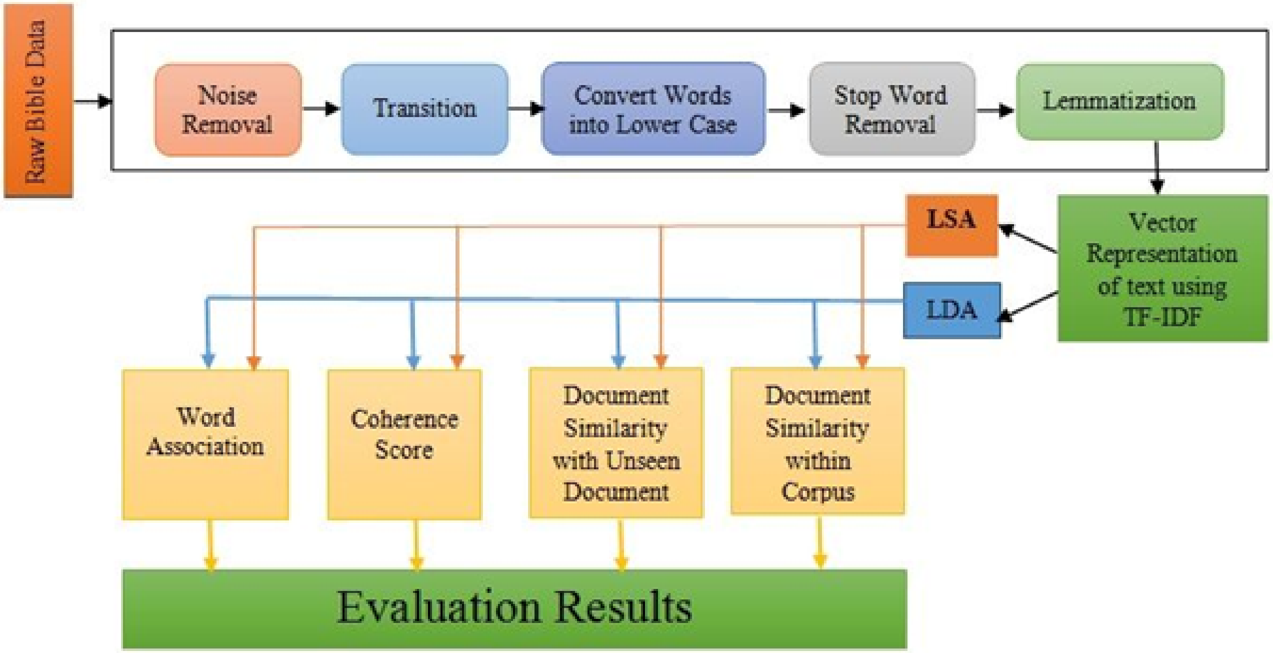

Raw Bible Data
Noise Removal
Transition
Convert Words into Lower Case
Stop Word Removal
Lemmatization
LSA
LDA
Vector Representation of text using TF-IDF
Word Association
Coherence Score
Document Similarity with Unseen Document
Document Similarity within Corpus
Evaluation Results

Fig 1. Comparative analysis of LSA and LDA

2.4 Data selection and preprocessing

The book of genesis was selected from the bible and divided into subtopics; each sub-topic is considered as a document in this corpus. Before the document term matrix was created, some preprocessing was also done on the data. Foremost, a program was created to read the documents from the concerned files. The second program removes punctuations and other symbols that are not useful to text analysis, and then documents were split into words. Then the program searches each word and then retains the words which are not in the ”stop words” list. However, the words contained in a list of ”stop words” were removed; these words are deemed to have no significance in describing the mechanical qualities of a data under study. The remaining words were converted into their base form using the lemmatization process. It helps to reduce the scope of the data for document matching, to get more consistent results from both the LDA and LSA methods. Now the corpus is ready for use.

Both LDA and LSA take a document- term matrix as an input. Each row represents a document from the entire dataset, and each column represents a word. Each location in the matrix has a number that corresponds to the number of times the word designated by the column appeared in the document designated by the row.

2.5 Analysis and comparison of the topic modeling methods

We used the GENSIM (Topic modeling and preprocessing), NLTK (Natural Language Processing), and SCIPY (Document comparison) and MATPLOTLIB (Visualization) Libraries in python that searches through a combination of the parameters. LSA gives a direct output of document similarities in the form of a cosine similarity matrix and coherence score based on the matter related to the book of genesis in this study(21–23). The text relevance is calculated where the values range from -1 to 1, where one is considered an exact match, and -1 represents two documents that are complete opposites. This output is enough to create a matrix related to the information whose columns each represent a document and whose rows contain documents in their order of similarity to the document associated with the column they are in.

Unlike LSA, LDA does not directly output document similarities. Instead, LDA outputs a matrix, whose rows represent all the documents in the dataset, and columns represent all the topics. Each value represents a particular topic’s weight in a document. The user specifies the total number of topics that the words are sorted into, and columns in the matrix range between 0 and the user-defined number of topics. LDA was run with different numbers of topics until a good topic range was found for the dataset.

The final step is to compare the document similarity matrices output by LDA and LSA. If only minor differences can be found between them, it can be inferred that LSA and LDA are more or less equal in their ability to sort the mechanics of the data. Nevertheless, if the two results differ significantly, the more efficient algorithm is determined by comparing one document with the other documents. In each column of the matrix, it counts the number of documents with the same core functions, and the central functions of mechanical data must be individually determined.

3 Results

3.1 Performance evaluation

We compared the performance of LSA and LDA models with two baselines, cosine similarity and coherence score as the primary evaluation metrics. In the following subsections, we illustrated and summed up the methods mentioned above. Because of document similarity within the corpus, entire documents were classified into four categories that are 0% to 25%, 26% to 50%, 51% to 75% and 76% to 100% similarity groups and chosen the documents from these groups and their most similar documents in similarity descending order and the same document were taken from the other method results and analyzed why the differences are shown between the results of two methods.

As per the results obtained from two methods, Figure 2 shows algorithms outperform significantly and almost with the same results at 76 to 100% similarity group when compared against the remaining three groups (0 to 75%). This result is an essential finding in the understanding of the similarities between documents, and this suggests and demonstrates that these methods can predict considerably better.

Further, the similarity results have been studied that in downstream to find which method giving the relevant results. Table 1 shows LSA results for the top ten similar documents with reference documents, and Table 2 shows the top ten similar documents from LDA with reference documents under study. The first column of the first row in the tables occupied by the reference document and their corresponding top five topics were occupied in the next columns. The second row onwards tables were filled with its top ten most similar documents and their top five topics in descending order of their similarity.

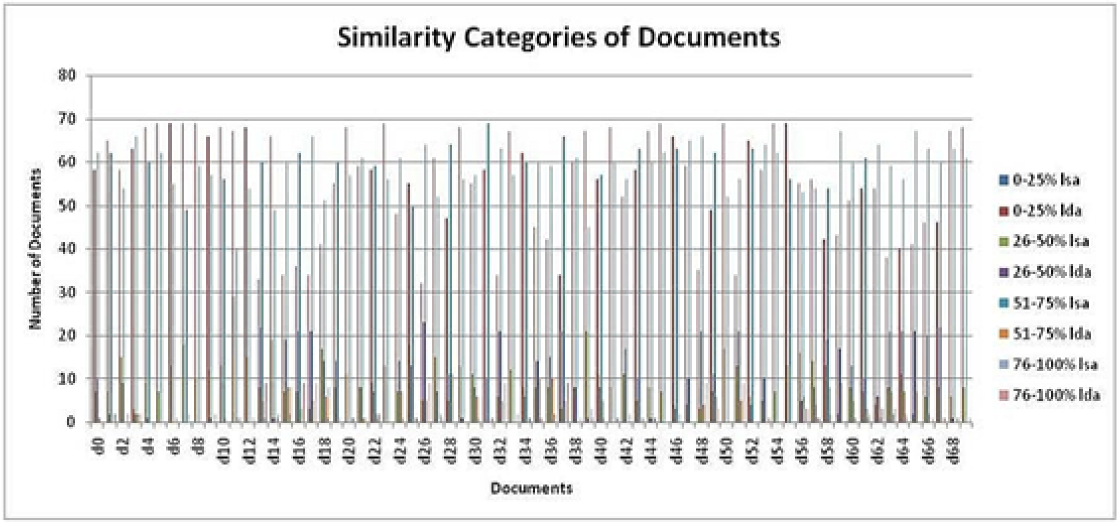


Fig 2. Document categorization based on similarity group

Table 1. Top ten similar documents from LSA of reference document

Document	Top 5 topics									
	Topic number	Probability	Topic number	Probability	Topic number	Probability	Topic number	Probability	Topic number	Probability
1	4	0.42302	54	0.310902	19	0.201036	55	0.188179	52	0.157631
2	4	0.284246	19	0.174393	32	0.163204	8	0.142758	50	0.133671
6	4	0.609205	9	0.155245	46	0.150951	29	0.143395	6	0.141816
0	4	0.454651	30	0.307792	31	0.226039	38	0.216571	44	0.151804
7	4	0.542851	11	0.352796	6	0.216504	39	0.145193	12	0.144648
37	14	0.390687	42	0.3	48	0.288584	17	0.253037	8	0.240016
15	8	0.307676	1	0.196223	13	0.13671	26	0.097108	50	0.094111
5	4	0.347579	35	0.345963	23	0.284666	38	0.171876	27	0.16461
61	18	0.162277	50	0.154305	47	0.124013	52	0.110536	54	0.102559
17	38	0.214936	49	0.18975	45	0.188285	47	0.172109	23	0.164105
9	9	0.524162	38	0.285517	43	0.219786	19	0.201452	4	0.191379

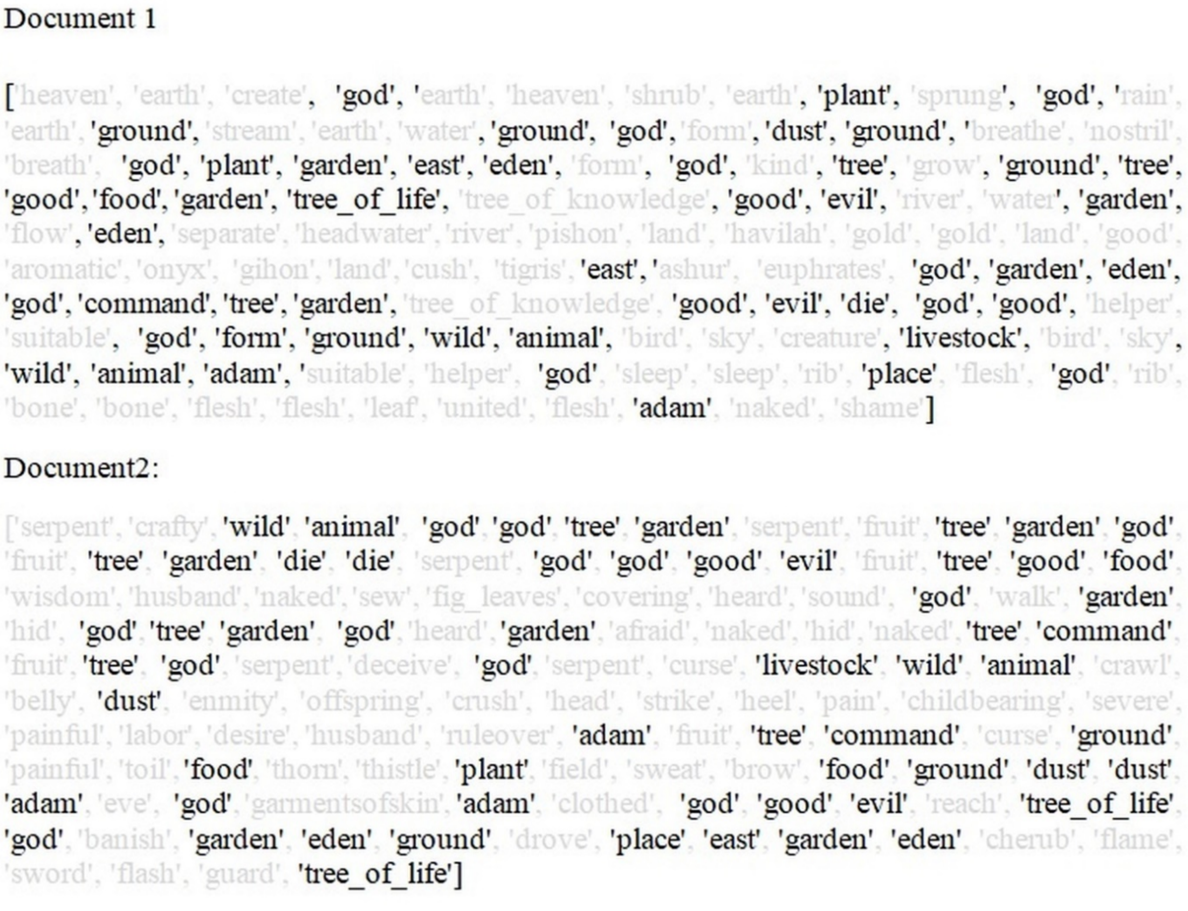
Table 2. Top ten Similar Documents from LDA of reference document.

Document	Top 5 topics									
	Topic number	Probability	Topic number	Probability	Topic number	Probability	Topic number	Probability	Topic number	Probability
1	39	0.665457	24	0.153983	26	0.059576	-	-	-	-
2	39	0.631368	24	0.179468	-	-	-	-	-	-
33	39	0.729882	24	0.123574	-	-	-	-	-	-
0	26	0.558888	24	0.175142	39	0.10438	-	-	-	-
15	24	0.462719	6	0.178772	16	0.155682	39	0.027457	-	-
32	24	0.860819	-	-	-	-	-	-	-	-
51	24	0.788457	-	-	-	-	-	-	-	-
37	24	0.875697	56	0.012098	-	-	-	-	-	-
26	24	0.824723	56	0.023041	-	-	-	-	-	-
17	24	0.759601	44	0.037791	-	-	-	-	-	-
13	24	0.799389	38	0.041327	-	-	-	-	-	-

First, from the LSA output, as shown in Table 3 , the reference document is placed in the first column of the first row and its top five topics were occupied in their respective columns. In the second row onwards, the table was tabulated with its top ten most similar documents and their top five topics by descending order, respectively. Comparatively, the LDA results in Table 2 identified some sort of similarity, like document 2 was showing the most similar document for the reference document 1; But, LSA output gives 51 to 75% similarity between the documents. However, LDA gives only 0 to 25% of similarity.

This result highlights that little is known about the correlation between the topics of reference documents and their top ten most similar documents (Table 1). Whereas in LDA, the correlation showed best among the topics of reference and other documents even they have 0-25% similarity with reference document 1. Owing to these results, a more in-depth downstream analysis was performed at the document level to check how many words are in common between the reference and most similar documents as both LSA and LDA results have been showing a clear difference. The complete difference has been reported in document 1 with both methods understudy; the following analysis was performed further with document 1 to understand the difference.

Figure 3 represents that document 1 and the topmost similar document, document 2. The common represented words are in thick colour, whereas, non-common words in both the documents are in light colour respectively. Document 1 explains the Garden of Eden, the creation of a woman from Adam, and his command. In comparison, document2 represents the entry of sin into humanity. However, these words do not convey the actual content of documents. LSA results showed that there is 56 to 75% similarity even there is no that much similarity among the documents, whereas LDA result representing there is 0 to 25% similarity among the document1 and document2. So, from the result is understandable, the LDA results are more appropriate than LSA.



Document 1

['heaven', 'earth', 'create', 'god', 'earth', 'heaven', 'shrub', 'earth', 'plant', 'sprung', 'god', 'rain', 'earth', 'ground', 'stream', 'earth', 'water', 'ground', 'god', 'form', 'dust', 'ground', 'breathe', 'nostril', 'breath', 'god', 'plant', 'garden', 'east', 'eden', 'form', 'god', 'kind', 'tree', 'grow', 'ground', 'tree', 'good', 'food', 'garden', 'tree_of_life', 'tree_of_knowledge', 'good', 'evil', 'river', 'water', 'garden', 'flow', 'eden', 'separate', 'headwater', 'river', 'pishon', 'land', 'havilah', 'gold', 'gold', 'land', 'good', 'aromatic', 'onyx', 'gihon', 'land', 'cush', 'tigris', 'east', 'ashur', 'euphrates', 'god', 'garden', 'eden', 'god', 'command', 'tree', 'garden', 'tree_of_knowledge', 'good', 'evil', 'die', 'god', 'good', 'helper', 'suitable', 'god', 'form', 'ground', 'wild', 'animal', 'bird', 'sky', 'creature', 'livestock', 'bird', 'sky', 'wild', 'animal', 'adam', 'suitable', 'helper', 'god', 'sleep', 'sleep', 'rib', 'place', 'flesh', 'god', 'rib', 'bone', 'bone', 'flesh', 'flesh', 'leaf', 'united', 'flesh', 'adam', 'naked', 'shame']

Document2:

['serpent', 'crafty', 'wild', 'animal', 'god', 'god', 'tree', 'garden', 'serpent', 'fruit', 'tree', 'garden', 'god', 'fruit', 'tree', 'garden', 'die', 'die', 'serpent', 'god', 'god', 'good', 'evil', 'fruit', 'tree', 'good', 'food', 'wisdom', 'husband', 'naked', 'sew', 'fig_leaves', 'covering', 'heard', 'sound', 'god', 'walk', 'garden', 'hid', 'god', 'tree', 'garden', 'god', 'heard', 'garden', 'afraid', 'naked', 'hid', 'naked', 'tree', 'command', 'fruit', 'tree', 'god', 'serpent', 'deceive', 'god', 'serpent', 'curse', 'livestock', 'wild', 'animal', 'crawl', 'belly', 'dust', 'enmity', 'offspring', 'crush', 'head', 'strike', 'heel', 'pain', 'childbearing', 'severe', 'painful', 'labor', 'desire', 'husband', 'ruleover', 'adam', 'fruit', 'tree', 'command', 'curse', 'ground', 'painful', 'toil', 'food', 'thorn', 'thistle', 'plant', 'field', 'sweat', 'brow', 'food', 'ground', 'dust', 'dust', 'adam', 'eve', 'god', 'garmentsofskin', 'adam', 'clothed', 'god', 'good', 'evil', 'reach', 'tree_of_life', 'god', 'banish', 'garden', 'eden', 'ground', 'drove', 'place', 'east', 'garden', 'eden', 'cherub', 'flame', 'sword', 'flash', 'guard', 'tree_of_life']

Fig 3. Common and Non-Common words from document1 and document2 understudy

3.2 Comparison based on similarity with the unknown document:

The supplementary Table 3 lists the top ten most similar documents with the unseen document. The first row of the table explaining that both LSA and LDA provided document 0 and document 1 as most similar to unseen documents, but from the third row onwards, some discrepancy can be observed. Mainly, document 2 is listed as the third most similar in the results of LDA, but in LSA, it is in the eighth position of document similarity in descending order. To understand the difference, observe the next Tables, in supplementary Table 4, and Table 5 which contains unseen document and its corresponding top five topic proportions in the first row of the table and top ten most similar documents to unseen document with its corresponding top five topics has occupied the second row onwards.

From the results shown here, it is easily understood that topics of document 2 in LDA are more correlated with topics of unseen document than in LSA. So, we can understand that the LDA results are more appropriate than LSA in finding similar documents to unseen documents.

3.3 Coherence score

Coherence is a state in which a set of topics or concepts supports each other, and it computes the relative distance between terms in topics. Topic coherence is a measure used to assess the goodness of the topic models. These measurements help in distinguish between semantically comprehensible topics. Here, the c_v coherence measure is used to calculate the score. c_v measure is based on a Boolean sliding window calculation, one-set segmentation of the top words, normalized point wise mutual information(NPMI) for agreement between individual words and cosine similarity. The following two figures showing the coherence score of LSA and LDA.

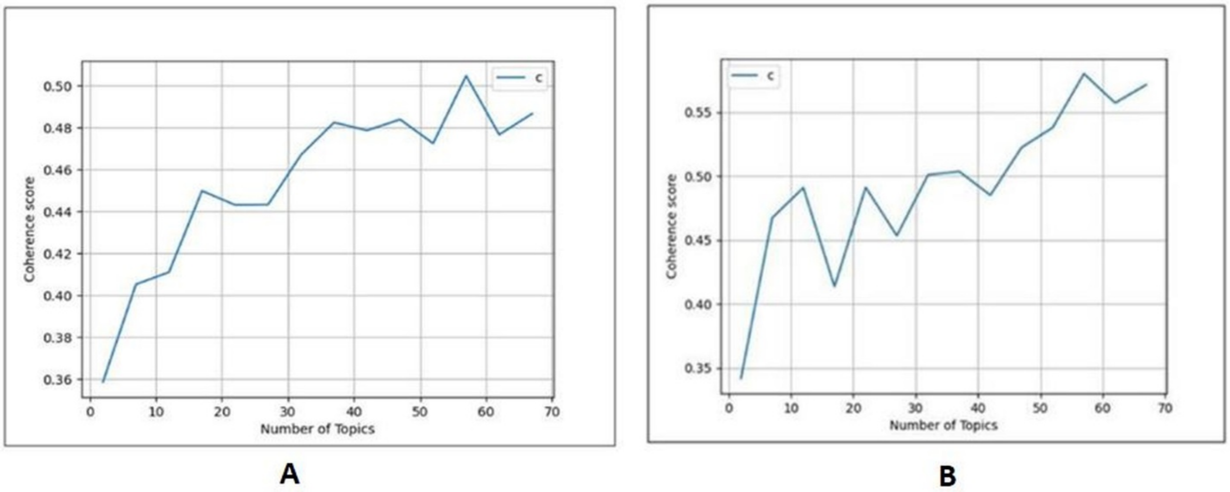


Fig 4. A) Coherence Score of LSA. B) Coherence Score of LDA

From the Figure 4, it is observed that the coherence score given by LSA is 0.50395, whereas LDA showed as 0.58018. So, LDA results are better than LSA results for the bible text.

3.4 Word association

The resemblance between the two words can be estimated by whether they share a common topic. Here, we found some word association between two words by calculating the cosine similarity between their topic proportions.

From the Table 3 , it is established that the semantic relationship between the words given by LDA results is better than the LSA results.

Table 3. Observed response distributions for the word bear from LSA and LDA

LSA		LDA	
Circumcise	0.711850447	Circumcise	0.978302519
Facedown	0.711850447	Punishment	0.897800252
Covenant	0.608353308	covenant	0.698462842
Blame	0.574658728	Birth	0.088462842
Steward	0.495564883	Facedown	0.088411842
Pardon	0.495564883	Misery	0.068462842
Misery	0.433461986	Female	0.051284628
Generation	0.390131378	Steward	0.050086417
Money	0.379946272	Money	0.004486417
Amount	0.360836249	Amount	0.004466417
Mistake	0.360836249	Blame	0.004365917

4 Discussion

In this study, we discussed some results and emerging trends and how they can be understandable from the perspective of earlier studies, including our comparisons. The difference between the two methods using bible text as corpora; the results give some indication about how evenly the distribution of words is between the documents[24]. The analysis shows that both techniques find the most significant percentage of instances and assessments of the context in which the words appear that contain words related to God's creation and his mandate for humanity.

Generally, human word associations, high-frequency words are more probable to be used as response words than low-frequency words. For example, in the studies of Griffiths and steyvers[25] compared the topic model with LSA in predicting word associations, finding the balance between the influence of word frequency and semantic relatedness found by the topic model can result in better performance than LSA on this task. In our study, the main questions related to the extraction of word meaning in natural language processing, but also for the extraction of its meaningful associations, have been observed. For example, the word 'bear' in the book of genesis in our dataset, implies in contexts like punishment and birth. For instance, in stock markets, the bear represents that the market is diminishing. Comparatively, the word bear that associates or correlates with LDA than LSA, respectively.

However, the studies conducted by Siti Qomariyaha et al. in 2019[26] by using Twitter data as text data were corroborated with our results in this study as they concluded that LDA considers the relationship between documents in the corpus with the best topic coherence than LSA. Also, in comparative studies using different text mining methods as applied to short text data, LDA showed more meaningful extracted topics and obtained good results with topic coherence as an evaluation metric for creating the content of a document collection[6,27].

The overall results showed clearly how the book of genesis is defined by the two text mining methods that complement each other. LSA and LDA agree with many of the texts and topics, yet they each generated some topics that the other method did not identify. This result indicates that using more than one text mining technique that uses different mechanisms to identify topics can result in more meaningful analysis and better identification of semantic structure from the text. Furthermore, we recommend using LDA due to its superior performance, and employing the LDA also provides the system with a more significant explanation as LDA is a probabilistic model in arriving at the conclusions. These insights can help in understanding the natural patterns in the data, when necessary.

5 Conclusion

Based on the result, the LDA showed the best topic coherence 0.58018 than the coherence score given by LSA (0.50395). Therefore, this study shows that LDA achieves superior performance when compared to LSA. The performance achieved by LDA using document similarity within-corpus, document similarity with the unseen document, and word associations also delivered maximum meaningful topics and implicitly, contextual word meaning from bible text corpora. Thus, the work presented in this comparative study can be a computationally efficient and vital reference for researchers on topic modeling.

References

1) Rüdiger M, Antons D, Salge TO. From text to data: On the role and effect of text pre-processing in text mining research. *Academy of Management Proceedings*. 2017;2017(1). Available from: https://dx.doi.org/10.5465/ambpp.2017.16353abstract.
2) Antons D, Joshi AM, Salge TO. Content, contribution, and knowledge consumption: Uncovering hidden topic structure and rhetorical signals in scientific texts. *Journal of Management*. 2019;45(7):3035–3076. Available from: https://dx.doi.org/10.1177/0149206318774619.
3) Antons D, Grünwald E, Cichy P, Salge TO. The application of text mining methods in innovation research: current state, evolution patterns, and development priorities. *R&D Management*. 2020;50:329–351. Available from: https://dx.doi.org/10.1111/radm.12408.
4) Hammed J, Wang Y, Yuan C. LDA and Topic Modeling- Models Applications, A Survey. *Multi Tools Appl*. 2019;78:15169–15211. Available from: https://doi.org/10.1007/s11042-018-6894-4.
5) Cho HW. Topic Modeling. *Osong Public Health and Research Perspectives*. 2019;10:115–116. Available from: https://dx.doi.org/10.24171/j.phrp.2019.10.3.01.
6) Albalawi R, Yeap TH, Benyoucef M. Using Topic Modeling Methods for Short-Text Data: A Comparative Analysis. *Frontiers in Artificial Intelligence*. 2020;3. Available from: https://dx.doi.org/10.3389/frai.2020.00042.
7) Kwantes PJ, Derbentseva N, Lam Q, Vartanian O, Marmurek HHC. Assessing the Big Five personality traits with latent semantic analysis. *Personality and Individual Differences*. 2016;102:229–233. Available from: https://dx.doi.org/10.1016/j.paid.2016.07.010.
8) Kherwa P, Bansal P. Latent Semantic Analysis: An Approach to Understand Semantic of Text. *IntConfCurTreCompuElecElectComm*. 2017. Available from: https://doi.org/10.1109/CTCEEC.2017.8455018.
9) Valdez D, Pickett AC, Goodson P. Topic Modeling: Latent Semantic Analysis for the Social Sciences. *Social Science Quarterly*. 2018;99(5):1665–1679. Available from: https://dx.doi.org/10.1111/ssqu.12528.
10) Blei DM, Ng AY, Jordan MI, et al. Latent Dirichlet Allocation. *JMLR*. 2003;3(4):993–1022. Available from: https://www.jmlr.org/papers/volume3/blei03a/blei03a.pdf.
11) Wang W, Feng Y, Dai W. Topic Analysis of Online Reviews for Two Competitive Products using Latent Dirichlet Allocation. *Electron Commer Res Appl*. 2018;p. 1–30. Available from: https://doi.org/10.1016/j.elerap.2018.04.003.
12) Xu A, Qi T, X D. Analysis of the douban online review of the mcu: based on LDA topic model. In: and others, editor. 2nd International Symposium on Big Data and Applied Statistics. J. Phys: Conf. Seri.;vol. 2019. Dalian. ;p. 1437–1437. Available from: https://doi.org/10.1088/1742-6596/1437/1/012102.
13) Korshunova I, Xiong H, Fedoryszak M, Theis L. Discriminative Topic Modeling with Logistic LDA. In: 33rd Conference on Neural Information Processing Systems. 2019. Available from: https://papers.nips.cc/paper/8902-discriminative-topic-modeling-with-logistic-lda.pdf.
14) Alghamdi R, Alfalqi K. A Survey of Topic Modeling in Text Mining. *International Journal of Advanced Computer Science and Applications*. 2015;6(1):147–153. Available from: https://dx.doi.org/10.14569/ijacsa.2015.060121.
15) Shi L, Liu L, Wu L, Jiang L. Event detection and user interest discovering in social media data streams. *IEEE Access*. 2017;5:20953–20964. Available from: https://doi.org/10.1109/ACCESS.2017.2675839.
16) Ray SK, Ahmad A, Kumar CA. Review and Implementation of Topic Modeling in Hindi. *Applied Artificial Intelligence*. 2019;33(11):979–1007. Available from: https://dx.doi.org/10.1080/08839514.2019.1661576.
17) Pilato G, Vassallo G. TSVD as a Statistical Estimator in the Latent Semantic Analysis Paradigm. *IEEE Transactions on Emerging Topics in Computing*. 2015;3(2):185–192. Available from: https://dx.doi.org/10.1109/tetc.2014.2385594.
18) Kaur R, Kaur M. Latent Semantic Analysis: Searching Technique for Text Documents. *Int J Eng Dev Res*. 2015;3(2):803–806. Available from: https://www.ijedr.org/papers/IJEDR1502143.pdf.
19) Tamizharasan M, Shahana RS, Subathra P. Topic modeling-based approach for word prediction using automata. *J Crit Rev*. 2020;7(7):744–749. Available from: https://doi.org/10.31838/jcr.07.07.135.
20) Safiie MA, et al. Latent Dirichlet Allocation (LDA) Model and kNN Algorithm to Classify Research Project Selection. *IOP Conf Ser: Mater Sci Eng*. 2018;333. Available from: https://doi.org/10.1088/1757-899X/333/1/012110.
21) Gunawan D, Sembiring CA, Budiman MA. The Implementation of Cosine Similarity to Calculate Text Relevance between Two Documents. *Journal of Physics: Conference Series*. 2018;978(1). Available from: https://dx.doi.org/10.1088/1742-6596/978/1/012120.
22) Stevens K, Kegelmeyer P, Andrzejewski D. Exploring topic coherence over many models and many topics. In: and others, editor. EMNLP. ;p. 952–961. Available from: https://www.aclweb.org/anthology/D12-1087.pdf.
23) Koltcov S, Ignatenko V, Koltsova O. Estimating Topic Modeling Performance with Sharma–Mittal Entropy. *Entropy*. 2019;21(7). Available from: https://dx.doi.org/10.3390/e21070660.
24) Lund J, Armstrong P, Fearn W, Cowley S, Hales E, Seppi K. Cross-referencing using Fine-grained Topic Modeling. In: and others, editor. Proceedings of NAACL-HLT. 2019;p. 3978–3987. Available from: https://www.aclweb.org/anthology/N19-1399.pdf.
25) Griffiths TL, Steyvers M. A probabilistic approach to semantic representation. *Proce of the 24th Annual Conference of the Cognitive Science Society*. 2002. Available from: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.9.5961&rep=rep1&type=pdf.
26) Qomariyaha S, Iriawanb N, Fithriasaric K. Topic modeling Twitter data using Latent Dirichlet Allocation and Latent Semantic Analysis. *AIPConfProc*. 2019. Available from: https://doi.org/10.1063/1.5139825.
27) Chakkarwar V, Sc T. Quick insight of research literature using topic modeling. In: Zhang YD, Mandal J, So-In C, Thakur N, editors. Smart Trends in Computing and Communications. Smart Innovation, Systems and Technologies, 2020;vol. 165. Springer. ;p. 189–197. Available from: https://doi.org/10.1007/978-981-15-0077-0_20.