Journal of Information Systems Engineering and Management
2025, 10(21s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

eLDA: Augmenting Topic Modeling with Word Embeddings for Enhanced Coherence and Interpretability

[1] Gobind Kumar Das, Panthadeep Bhattacharjee[2]

[1]*Department of Computer Science & Engineering, National Institute of Technology, Rourkela, India. gobinddas1999@gmail.com*

[2]*Department of Computer Science & Engineering, National Institute of Technology, Rourkela, India.* panthadeep.edu@gmail.com

ARTICLE INFO

Received: 25 Dec 2024

Revised: 30 Jan 2025

Accepted: 20 Feb 2025

ABSTRACT

Traditional topic modeling methods like Latent Dirichlet Allocation suffer from several challenges, especially concerning appropriate topic coherence, logical and consistent word groups that follow some semantic relationship, and interpretability. In this work, we propose an enhanced version of LDA, called eLDA, which incorporates Word2Vec embeddings (W2Ve) into LDA. This approach is adopted in order to improve the coherence of individual topics and improve the general topic interpretability by using established metrics such as the coherence score. Traditional LDA and eLDA coherence scores are compared to validate the results. In contrast to the former, we observe that eLDA provides much better interpretability with higher coherence scores, stronger semantic relationships, and improved visualization of topics.

Keywords: Latent Dirichlet Allocation, Corpora, Word2Vec Embedding, Topic Modeling.

INTRODUCTION

The expanding growth of digital text data requires the need for an advanced text analysis technique that can extract meaningful knowledge from a large data pool that is unstructured. This challenge may be addressed through the use of LDA [5], a robust method for topic modeling (TM). However, LDA has notable limitations, particularly in identifying the topic coherence (logical, consistent, and semantic connection between the words) along with its interpretability [20]. Interpretability means how easily a human understands and makes sense of the topics. In this paper, whenever we say interpretability, it will be in terms of (a) qualitative analysis or semantic relationships of words into topics, (b) quantitative evidence or coherence score, and (c) visualization of document mapping of the topic. This is important for ensuring that meaningful topics are identified and can be used in various practical applications.

While addressing traditional LDA challenges, recent developments in NLP (Natural Language Processing) have explored the utilities of models such as W2Ve [16]. This enables the extraction of semantic similarities between words based on their contextual usage in large corpora[11]. In this research, we therefore make the following contributions:

1. The proposed hybrid approach combines the utilities of LDA with W2Ve to produce eLDA for enhancing topic coherence (coherence score).

2. The proposed model also enhances the semantic relationship and thereby improves the topic document mapping visualization.

LITERATURE REVIEW

Topic modeling [10,5] is a technique used to discover hidden topics within data collection. The primary goal of TM is to group documents into topics such that the documents within the same topic are more akin to each other than to those in different topics.

Latent Semantic Analysis (LSA) [7] was first introduced in 1990. LSA employs Singular Value Decomposition (SVD) [22] for the purpose of reducing the dimensionality of term-document matrices [15]. Probabilistic LSA (pLSA) is an extension of LSA based on a probabilistic framework [21]. It models documents as a blend of topics, where all topics are distributed over words [9].

Latent Dirichlet Allocation (LDA) [2,6,24] was introduced in 2001 as an improvement over earlier models by explicitly modeling the allotment of topics across the documents, and the words within those topics. In general, probabilistic TM models, including the LDA, treat topics as distributions which are probabilistic in nature over words and documents [3]. LDA selects a topic distribution over all documents, and all words are distributed over topics [4].

Recently, the impact of word embeddings on text analysis has also been explored [1]. LDA has been applied to various fields, as discussed in works such as [25,12,14].

METHODOLOGY

The enhanced LDA or eLDA model extends the traditional version of LDA by incorporating an extra layer of feature extraction through embedding layers. It also involves a more complex preprocessing of the deep semantic relationship between the words. That is to say, though it transforms the text into a bag-of-words (BOW), similar to LDA, it then provides higher accuracy with better parameter tuning and advanced topic identification. In order to overcome the weakness of the LDA model, eLDA breaks away from the BOW approach, making it more robust for application to complex datasets.

Multiple documents are passed to the eLDA model as input, and the model pre-processes the text data (Refer Figure 1). The text is subjected to tokenization followed by the removal of stopwords. This preprocessing stage is strategically conducted in order to reduce the noise element from the input that is to be provided to the eLDA model.

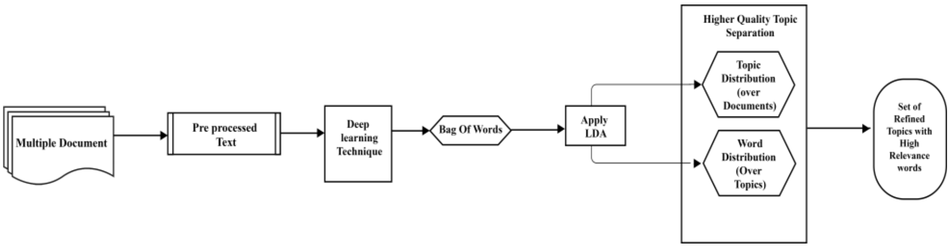

Multiple Document
Pre processed Text
Deep learning Technique
Bag Of Words
Apply LDA
Higher Quality Topic Separation
Topic Distribution (over Documents)
Word Distribution (Over Topics)
Set of Refined Topics with High Relevance words

Fig. 1: Proposed framework of the eLDA.

Following the pre-processing step, we integrate W2Ve into the LDA model to make the words more semantically related. W2Ve [16] was introduced in 2013 by Google [16]. The embedding architecture that we adopted for developing eLDA uses the Continuous Bag Of Word (CBOW) instead of relying on the conventional skip-gram [19] model. The CBOW executes in a sequenced manner. CBOW aims to maximize the probability of predicting the center word (*wcon*) around surrounding words $wsur_1, wsur_2, \dots, wsur_n$ from the corpus. Next, we state the objective function required for the purpose:

$$\boxed{f_{obj} = max\ \sum_{wcon \in corpus} log\ P\left(wcon|wsur_{1,}wsur_2, \dots, wsur_n\right)} \qquad (1)$$

The above objective function f_{obj} (Equation 1) generates a word vector for every word in the vocabulary. After this, an average context vector $\mathbf{vec}_{context}$ is computed using the formula given below (Equation 2) for each context word around a center word.

$$\boxed{Vec_{context} = \frac{1}{n}\sum_{i=1}^{n} Vec_{wsur_i}} \qquad (2)$$

Here, $wsur_i$ indicates words that surround the center word. After this, the softmax function (Equation 3) is used to transform the dot product of the context vector and each center word into a probability distribution over possible center words. This can be expressed as:

$$\boxed{P\left(wcon|wsur_{1,}wsur_2 \dots., wsur_n\right) = \frac{exp(Vec_{wcon}Vec_{context})}{\sum_{wcon' \in vocab} exp(Vec_{wcon'}Vec_{context})}} \qquad (3)$$

The softmax function is used to ensure that the model predicts a probability for each word in the vocabulary, while the gradient descent contributes towards maximizing the overall objective function.

Center word probability: The probability of the center word '*wcon*' is presented as a product of the conditional probabilities (Equation 4) of context words.

$$\boxed{P = p(wcon|wsur_1) \cdot p(wcon|wsur_2) \cdots p(wcon|wsur_n)} \qquad (4)$$

This product returns the independent assumption that each context word contributes individually to predict the center word [23]. This approach to topic modeling addresses the limitations of traditional LDA. An individual topic's coherence score is calculated, followed by the overall coherence [26].

We had also improved the topic-document mapping visualization [18, 13]. Newly discovered topics obtained through eLDA were visualized by using pyLDAvis[1]

(Figure 3). The same was done for traditional LDA in (Figure 2). After reducing the dimensions to two components, topics were mapped with documents for comparison (Figure 4). The number of documents associated with topics for traditional LDA and eLDA is given in Table 3

RESULTS AND ANALYSIS

For our experimental purpose, we acquired relevant datasets from Kaggle [2]. These datasets are: Amazon review dataset, BBC news articles, Indian airlines customer reviews, Medical transcription, and Stock market news data (Table 1).

Table 1: Dataset's description

Dataset	**Size**	**Description**	**Link**
D1	21214	Amazon review dataset	**Click Here**
D2	35860	BBC news articles	**Click Here**
D3	2210	Indian airlines customer reviews	**Click Here**
D4	4999	Medical transcription	**Click Here**
D5	4845	Stock market news data	**Click Here**

We evaluated our proposed model on the datasets as stated in Table 1. The output of our model showed improved interpretability in terms of coherence score (Table 2) and topic-document visualization (Figure 4). The semantic relations were also extracted from topics as a group of words with the use of W2Ve and topic-document mapping (Table 3). We used cosine similarity to measure the proximity between topic vectors and document vectors. Each document vector (a row) is a collection of

[1] Source of pyLDAvis: https://pypi.org/project/pyLDAvis/

[2] https://www.kaggle.com/datasets

Table 2: Calculated Coherence Scores

Topics	Topic 1		Topic 2		Topic 3		Topic 4		Topic 5		Overall	
Model	**LDA**	**eLDA**	**LDA**	**eLDA**	**LDA**	**eLDA**	**LDA**	**eLDA**	**LDA**	**eLDA**	**LDA**	**eLDA**
D1	-1.4744	0.4161	-2.0223	0.5059	-1.4834	0.5652	-1.7628	0.5281	-1.9196	0.4760	-1.7325	0.4983
D2	-0.6434	0.6514	-5.9017	0.7673	-5.9127	0.7836	-3.3974	0.8064	-5.5255	0.8281	-5.4343	0.7661
D3	-1.684	0.9539	-1.2329	0.9679	-1.7706	0.9288	-1.9763	0.9405	-1.8007	0.9649	-1.7770	0.9512
D4	-0.8317	0.3434	-1.1388	0.4123	-1.5330	0.3173	-0.7011	0.3710	-1.2923	0.3422	-1.1025	0.3572
D5	-4.3756	0.9840	-7.0795	0.9866	-2.6661	0.9752	-3.6399	0.9823	-5.5532	0.9021	-4.6629	0.9660

Table 3: Topic-Document mapping

Topics	Topic 1		Topic 2		Topic 3		Topic 4		Topic 5	
Number of Document associated	**in LDA**	**in eLDA**	**in LDA**	**in eLDA**	**in LDA**	**in eLDA**	**in LDA**	**in eLDA**	**in LDA**	**in eLDA**
D1	4473	2912	7576	3513	2887	6074	3537	5479	2741	3070
D2	6395	6048	5499	2517	9474	3972	6944	6678	7548	16645
D3	723	682	376	485	212	26	246	97	653	920

D4	1766	818	524	2377	652	1135	1515	156	542	480
D5	1075	373	1158	610	852	387	633	3066	1127	40

coherent words sourced from their respective documents. Multiple word vectors combine together to form a document vector. Our approach ensures that the evaluation considers the context and semantic relationships within the text [17].

Results displayed in Figures 2 and 3 highlight the comparison between the modeled topics for both the methods: LDA and eLDA respectively. The main difference in outcome lies in the fact that the semantic relationships that are captured by W2Ve in eLDA were otherwise ignored by the traditional LDA. Words involved in Figure 3 for respective topics focus on their semantic relation instead of only the term frequency as in LDA.

We have used the PCA [8] plots to demonstrate our obtained results. In Figures 4a and 4b, each point in the image represents a document, colored by its enhanced topic, while the black markers indicate the topic centroids.

1.1 Visualization

The pyLDAvis visualization for LDA with respect to Topic 1 from the BBC news articles is shown in Figure 2, while Figure 3 represents the same for our proposed eLDA model. The bar in light blue color depicts the overall term (word) frequency, while the red colored bar denotes the estimated word frequency within the selected document.

The output obtained by using traditional LDA for Topic 1 consumed 23.1% of tokens. In the right part of Figure 2, the top-30 most relevant words (terms) that occurred from top to bottom were:*‘bbc’, ‘says’, ‘people’, ‘party’, ‘euro’, ‘uk’, ‘say’, ‘isreal’, ‘minister’, ‘gaza’, ‘two’, ‘police’, ‘labour’, ‘years’, ‘one’, ‘said’, ‘man’, ‘election’, ‘king’, ‘could’, ‘first’, ‘prime’, ‘us’, ‘leader’, ‘former’, ‘number’, ‘government’, ‘ ’, ‘family’, ‘new’.*

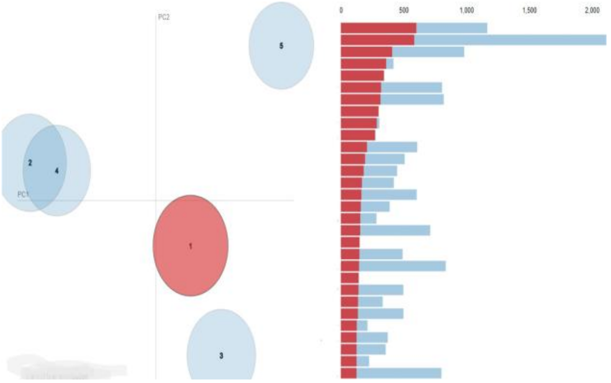


Fig. 2: Topic 1 pyLDAvis visualization for BBC news articles (traditional LDA).

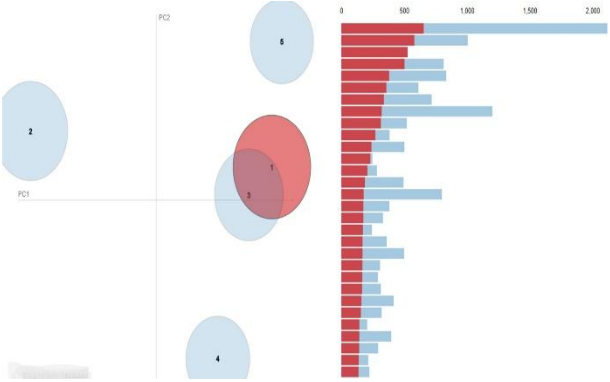
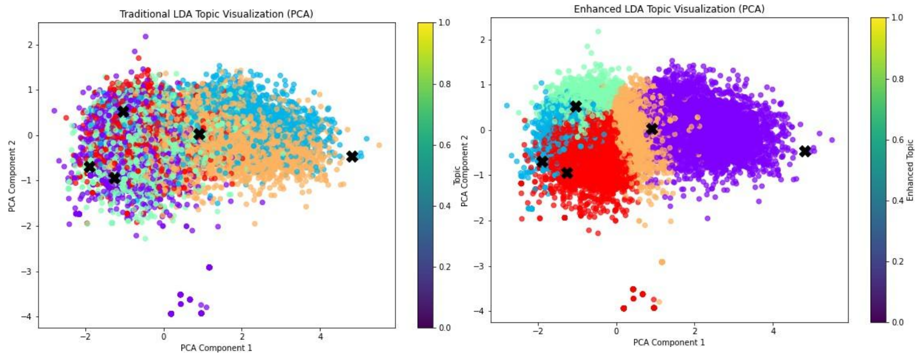


Fig. 3: Topic 1 pyLDAvis visualization for BBC news articles (eLDA).

The output obtained from eLDA for Topic 1 consumed 24.8% of tokens. In the right part of Figure 3, top-30 most relevant words (terms) occurred from top to bottom: *‘says’, ‘people’, ‘2024’, ‘uk’, ‘say’, ‘two’, ‘election’, ‘bbc’, ‘police’, ‘general’, ‘us’, ‘kill’, ‘died’, ‘could’, ‘new’, ‘government’, ‘president’, ‘attack’, ‘last’,‘former’, ‘found’, ‘gaza’, ‘london’, ‘year’, ‘three’, ‘told’, ‘said’, ‘man’, ‘death’, ‘ukraine’.*

While in Figure 4a, the plot visualizes the document-topic distribution for the traditional LDA, in Figure 4b, the same is represented for the eLDA model while inducing W2Ve. Due to space constraints, the remaining plots for rest of the datasets are available at: **results**. The code and the datasets, along with other supplementary materials, are available at **eLDA resources.**



(a) Traditional LDA (b) eLDA

Fig. 4: Topic-Document Visualization

DISCUSSION AND CONCLUSION

In this work, we combined the utilities of W2Ve with traditional LDA to generate an enhanced coherence score and an improved document-topic visualization. Through eLDA, we obtained a higher coherence score than the traditional LDA, as W2Ve generates semantic relationships between words. It may also be ob- served that the document-topic mapping generated by eLDA is well clustered (Figure 4), and the associative words in the generated topics are semantically meaningful. However, the model effectively works for only smaller datasets due to the involvement of CBOW. As the size of the dataset increases, the model's efficiency in generating semantic relationships between words may be reduced.

This research therefore concludes that the enhancement of traditional LDA towards eLDA elevates the coherence score of topics and their interpretability in terms of visualization. Our proposed scheme also has the potential to offer a powerful tool for text analysis in various applications. In the future, we would like to apply the same on more datasets for a better analysis and understanding of this domain.

REFERENCES

[1] Deepak Suresh Asudani, Naresh Kumar Nagwani, and Pradeep Singh. Impact of word embedding models on text analytics in deep learning environment: a review. *Artificial intelligence review*, 56(9):10345–10425, 2023.

[2] David Blei, Andrew Ng, and Michael Jordan. Latent dirichlet allocation. *Advances in neural information processing systems*, 14, 2001.

[3] David M Blei. Probabilistic topic models. *Communications of the ACM*, 55(4):77– 84, 2012.

[4] David M. Blei. Probabilistic topic models. *Commun. ACM*, 55(4):77–84, apr 2012.

[5] David M Blei, Andrew Y Ng, and Michael I Jordan. Latent dirichlet allocation.

[6] Journal of machine Learning research, 3(Jan):993–1022, 2003.

[7] Uttam Chauhan and Apurva Shah. Topic modeling using latent dirichlet allocation: A survey. *ACM Computing Surveys (CSUR)*, 54(7):1–35, 2021.

[8] Scott Deerwester, Susan T Dumais, George W Furnas, Thomas K Landauer, and Richard Harshman. Indexing by latent semantic analysis. *Journal of the American society for information science*, 41(6):391–407, 1990.

[9] Basna Mohammed Salih Hasan and Adnan Mohsin Abdulazeez. A review of prin- cipal component analysis algorithm for dimensionality reduction. *Journal of Soft Computing and Data Mining*, 2(1):20–30, 2021.

[10] Thomas Hofmann et al. Probabilistic latent semantic analysis. In *UAI*, volume 99, pages 289–296, 1999.

[11] Hamed Jelodar, Yongli Wang, Chi Yuan, Xia Feng, Xiahui Jiang, Yanchao Li, and Liang Zhao. Latent dirichlet allocation (lda) and topic modeling: models, applications, a survey. *Multimedia tools and applications*, 78:15169–15211, 2019.

[12] Ibrahim Kaibi, El Habib Nfaoui, and Hassan Satori. Sentiment analysis approach based on combination of word embedding techniques. In *Embedded Systems and Artificial Intelligence: Proceedings of ESAI 2019, Fez, Morocco*, pages 805–813. Springer, 2020.

[13] Na Li, Tao Lv, Xingyu Wang, Xiangyun Meng, Jie Xu, and Yuxia Guo. Research progress and hot topics of distributed photovoltaic: Bibliometric analysis and latent dirichlet allocation model. *Energy and Buildings*, page 115056, 2024.

[14] Moses, M. B., Nithya, S. E. & Parameswari, M. (2022). Internet of Things and Geographical Information System based Monitoring and Mapping of Real Time Water Quality System. International Journal of Environmental Sciences, 8(1), 27-36. https://www.theaspd.com/resources/3.%20Water%20Quality%20Monitoring%20Paper.pdf

[15] Peter Madzík, Lukáš Falát, and Dominik Zimon. Supply chain research overview from the early eighties to covid era–big data approach based on latent dirichlet allocation. *Computers & Industrial Engineering*, page 109520, 2023.

[16] Yishu Miao, Lei Yu, and Phil Blunsom. Neural variational inference for text pro- cessing. In *International conference on machine learning*, pages 1727–1736. PMLR, 2016.

[17] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013.

[18] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Dis- tributed representations of words and phrases and their compositionality. *Advances in neural information processing systems*, 26, 2013.

[19] Jeffrey Pennington, Richard Socher, and Christopher D Manning. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543, 2014.

[20] P Preethi Krishna and A Sharada. Word embeddings-skip gram model. In ICI- CCT 2019–System Reliability, Quality Control, Safety, Maintenance and Manage- ment: Applications to Electrical, Electronics and Computer Science and Engineer- ing, pages 133–139. Springer, 2020.

[21] Nikhil Rasiwasia and Nuno Vasconcelos. Latent dirichlet allocation models for im- age classification. *IEEE transactions on pattern analysis and machine intelligence*, 35(11):2665–2679, 2013.

[22] Hofmann Thomas. Probabilistic latent semantic analysis. *Uncertainity in Artificial Intelligence*, 1999.

[23] Michael E Wall, Andreas Rechtsteiner, and Luis M Rocha. Singular value decom- position and principal component analysis. In *A practical approach to microarray data analysis*, pages 91–109. Springer, 2003.

[24] Haowen Xia. Continuous-bag-of-words and skip-gram for word vector training and text classification. In *Journal of Physics: Conference Series*, volume 2634, page 012052. IOP Publishing, 2023.

[25] Lotfi A Zadeh. Fuzzy sets. *Information and control*, 8(3):338–353, 1965.

[26] Sulong Zhou, Pengyu Kan, Qunying Huang, and Janet Silbernagel. A guided latent dirichlet allocation approach to investigate real-time latent topics of twitter data during hurricane laura. *Journal of Information Science*, 49(2):465–479, 2023.

[27] Ben Mabey. pyldavis documentation, 2018.