A Practical and Empirical Comparison of Three Topic Modeling Methods using a COVID-19 Corpus: LSA, LDA, and Top2Vec

Ferhat D. Zengul
The University of Alabama at Birmingham
ferhat@uab.edu

Aysegul Bulut
The University of Alabama at Birmingham
abulut@uab.edu

Nurettin Oner
The University of Alabama at Birmingham
oner@uab.edu

Abdulaziz Ahmed
The University of Alabama at Birmingham
aahmed2@uab.edu

Manju Yadav
The University of Alabama at Birmingham
manju@uab.edu

Hope Gray
The University of Alabama at Birmingham
hopegray@uab.edu

Bunyamin Ozaydin
The University of Alabama at Birmingham
bozaydin@uab.edu

Abstract

This study was prepared as a practical guide for researchers interested in using topic modeling methodologies. This study is specially designed for those with difficulty determining which methodology to use. Many topic modeling methods have been developed since the 1980s namely, latent semantic indexing or analysis (LSI/LSA), , probabilistic LSI/LSA (pLSI/pLSA), naïve Bayes, the Author-Recipient-Topic (ART), Latent Dirichlet Allocation (LDA), Topic Over Time (TOT), Dynamic Topic Models (DTM), Word2Vec, Top2Vec and \variation and combination of these techniques. For researchers from disciplines other than computer science may find it challenging to select a topic modeling methodology. We compared a recently developed topic modeling algorithm– Top2Vec– with two of the most conventional and frequently-used methodologies–LSA and LDA. As a study sample, we used a corpus of 65,292 COVID-19-focused abstracts. Among the 11 topics we identified in each methodology, we found high levels of correlation between LDA and Top2Vec results, followed by LSA and LDA and Top2Vec and LSA. We also provided information on computational resources we used to perform the analyses and provided practical guidelines and recommendations for researchers.

Keywords: Topic modeling, LSA, LDA, Top2Vec, COVID-19

1. Introduction

Like the other data/text mining approaches, topic modeling methods have been experiencing accelerated and speedy changes and advancements in recent years. Topic modeling is a technique in machine learning that allows the researcher to mine through large volumes of unstructured text to detect word and phrase patterns (Miner et al., 2012). The advancements in computational resources and machine learning have led to the development of many topic modeling methods, some of which exhibit an incremental or slight variation of other topic modeling methods (Chen et al., 2021; Kherwa & Bansal, 2020). Almost each new topic modeling method has been disseminated to the research community through an archived or peer-reviewed publication along with publicly available codes. These publications tap into publicly available and frequently-used textual datasets to highlight the contribution of the new topic modeling method using fairly technical language (e.g., Angelov, 2020). However, researchers in disciplines other than computer science who are not familiar with the technical language yet interested in using topic modeling to extract patterns in textual data in their respective disciplines have difficulty determining which methods to use.

Several systematic reviews provide an extensive list and evaluation of the topic modeling methodologies in regards to their popularity, ease of use, and similarities and differences with other methods (Alghamdi & Alfalqi, 2015; Chauhan & Shah, 2021; Goldberg & Levy, 2014; Kherwa &

URI: https://hdl.handle.net/10125/102745
978-0-9981331-6-4

Bansal, 2020; Vayansky & Kumar, 2020). However, one method seems to dominate, namely Latent Dirichlet Allocation (LDA), despite the increasing availability of newly developed algorithms such as Top2Vec, NMF, CorEx, and BERTopic (Angelov, 2020; Obadimu et al., 2019; Sánchez-Franco & Rey-Moreno, 2022). While conducting their topic modeling application in their respective disciplines, researchers tend to use three established and commonly used topic modeling methods, including LDA, Latent Semantic Analysis (LSA), and Probabilistic LSA (Albalawi et al., 2020). However, few studies utilize multiple topic modeling methods on a particular dataset and provide side-by-side comparisons of these methods, especially the recently developed ones (Egger & Yu, 2022). Instead, the comparative studies primarily focus on two conventional methods including LSA and LDA (Anaya, 2011; Bergamaschi & Po, 2014; Cvitanic et al., 2016; Kalepalli et al., 2020; Mohammed & Al-augby, 2020). Moreover, from a practical and empirical perspective, there is a lack of attention to comparing multiple topic modeling methods on current topics of high interest such as COVID-19 (Egger & Yu, 2022).

Other than a couple of exceptions (Egger & Yu, 2022; Gutierrez et al., 2020), COVID-19-focused topic modeling studies to date have mainly utilized conventional methods such as LSA and LDA. For example, Zengul et al. (2021) used LSA to analyze 65 thousand abstracts and titles from the National Institutes of Health (NIH)'s COVID-19 Portfolio through November 2021. In another study, Älgå et al. used the Latent Dirichlet Allocation (LDA) method to analyze the articles early in the pandemic between February and June 2020 (Älgå et al., 2020). Others have used topic modeling to analyze COVID-19 on Twitter (Doogan et al., 2020; Ordun et al., 2020; Sha et al., 2020). Gutiérrez et al. analyzed multi-document classification models on 23,000 documents in the LitCovid dataset, which contains the most up-to-date scientific literature on COVID-19 (Chen et al., 2020a, 2020b; Gutierrez et al., 2020). Their analysis compared models such as linear regression and linear support vector machine (SVM) for multi-label classification, conventional neural models, i.e., KimCNN and XML-CNN, and Pre-Trained Language Models, i.e., BERT and BioBERT(Gutierrez et al., 2020). However, no study to date provides an empirical and practical comparison of conventional and recently developed topic modeling methods.

This study addresses this niche by comparing two conventional and most frequently used topic modeling methods, including LSA and LDA, and a recently developed method Top2Vec. In addition, our analysis is based on a large corpus of abstracts (n=65,292) from NIH"s COVID-19 portfolio used by Zengul et al. (2021). In this study, they identified eleven significant areas or topics of COVID-19 research. Using the same corpus of articles would allow us to make a direct comparison with this earlier published study.

2. Literature Review

Liu et al. describe topic modeling as a probabilistic generative model used for text mining and information retrieval in the computer science field (Liu et al., 2016). In 1990, Deerwester et al. introduced a new automatic indexing and retrieval method called *latent semantic indexing (LSI)* (Deerwester et al., 1990; Liu et al., 2016). However, LSI was not designed as a probabilistic model (Liu et al., 2016). Deerwester et al. noticed deficiencies in term-matching retrieval where words have been indexed, yet when a word search is performed, the search term does not match the indexed term (Deerwester et al., 1990). So, they designed LSI using a singular-value decomposition (Deerwester et al., 1990). Then in 1999, Hofmann proposed a novel statistical technique called *Probabilistic Latent Semantic Analysis (PLSA)* at the Fifteenth Conference on Uncertainty in Artificial Intelligence (Hofmann, 1999). PLSA was developed to analyze two-mode and co-occurrence data using a mixture decomposition derived from a latent model class (Hofmann, 1999). Then in 2003, David Blei and his team built the algorithm called *Latent Drichlet Allocation (LDA)*, which was proposed to improve the prior topic model techniques (Blei et al., 2002). LDA has become one of the most widely used methods for topic modeling (Alghamdi & Alfalqi, 2015). LDA is based on the Bayesian statistical topic models (Alghamdi & Alfalqi, 2015). Some LDA-based models include temporal text mining, author-topic analysis, supervised topic models, latent Dirichlet co-clustering, and LDA-based bio-informatics (Alghamdi & Alfalqi, 2015). LDA requires the programmer to remove stop-words manually. One of the drawbacks of LDA is that it cannot model the relations between the topics. Yet, other models are developed based on the LDA method, such as The Pairwise-Link-LDA model, which contains the LDA and Mixed Membership Stochastic Block model (MMSB), the Author-Recipient-Topic (ART) model, which combines LDA with the Author Topic model, and an extension of LDA called latent Class-Topic Model (CLTOM) (Alghamdi & Alfalqi, 2015). In addition, there is another form of topic modeling called Correlated Topic Modeling (CTM), characterized by using normal logistic distribution to create the relationships among the topics (Alghamdi &

Alfalqi, 2015). However, a couple of the limitations of CTM are that it requires a significant number of calculations and there can be a lot of general words inside the topics (Alghamdi & Alfalqi, 2015). On the other hand, topic evolution models model topics by considering time, such as Topic Over Time (TOT), Dynamic Topic Models (DTM), Multiscale Topic Tomography, and Dynamic Topic Correlation Detection. Overall, modeling topics over the results would reveal more precise topics from the corpus (Alghamdi & Alfalqi, 2015).

When the LDA was created, the fundamental tenet was that documents are represented as random mixtures of hidden subjects and each topic is described by a scattering of words (Cvitanic et al.,2016). As a result of being a probabilistic model, LDA can be utilized with ease for more intricate goals with more complicated models (Blei, 2012). Thus, LDA has more variants than LSA. On the other hand, in terms of supervised topic modeling, LSA has drawbacks in estimating observed data because it is based on similarities between words while LDA has comparable results with variants of Primary Component Analysis (PCA) and other least-squares regression models (Cvitanic et al.,2016).

However, state-of-the-art text mining includes both topic modeling and deep learning approaches (Chai & Li, 2019). Deep (having many layers) learning builds upon unsupervised representation learning where data are automatically extracted if they are useful representations using deep neural networks, such as convolutional neural networks (CNN) and recurrent neural networks (RNN) (Chai & Li, 2019, Jurafsky & Martin, 2022). Hence, the aim of deep learning for text mining is to find distributed representations that capture word semantics while performing language modeling (Chai & Li, 2019).

The previous techniques that Blei et al. reference included naïve Bayes, unigram, mixture of unigram, and "pLSI" (referring to Hofman's PLSA as cited above) (Blei et al., 2002). Naïve Bayes is an algorithm based on Bayes' theorem, a classification method (Rish, 2001). A unigram model assumes that "each word occurs independently, and consequently, the probability of a word sequence becomes the product of the probabilities of the individual words (Song & Croft, 1999). Blei et al. describe LDA as a mixture model compared to unigram models with mixture components and a mixture weights (Blei et al., 2002). They also emphasize that the LDA distribution has infinite continuous-varying mixture components (Blei et al., 2002). It must also be noted that natural language processing includes stemming, lemmatization, stop words, and Bag of Words (Balakrishnan & Lloyd-Yemoh, 2014; Manning et al., 2020). A stemming algorithm will remove the prefix or suffix to reduce words having the same stem (Manning et al., 2020). In contrast, a lemmatization algorithm removes inflectional endings and returns a word's root form.

On the other hand, stop words, such as 'as, an, of, the, etc.' are extremely common (Manning et al., 2020). These words are placed on a stop word list by the programmer/researcher. Deerwester identified deficiencies in current automatic indexing and retrieval methods, which are synonymy and polysemy (Deerwester et al., 1990). Synonymy means many ways to denote the same object. The programmer may not index information using the same word or term word searchers use, resulting in a diminished "recall performance for the retrieval system" (Deerwester et al., 1990). On the other hand, polysemy means that words can have more than one definitive meaning. Examples include words such as 'read' or 'chip' which can have different meanings based on the context of a sentence (Deerwester et al., 1990). Deerwester identifies that polysemy can reduce precision measures for the performance of the retrieval system (Deerwester et al., 1990). As Hoffman analyzed information retrieval, he found that one of the primary issues with information retrieval was automatic indexing which is primarily applied in query-based retrieval (Hofmann, 1999). He notes that the Vector Space Model technique was the most popular technique for information retrieval of documents (Hofmann, 1999).

In 2013, Mikilov et al. introduced an efficient estimation of word representations in vector space called the *word2vec* tool (Goldberg & Levy, 2014; *word2vec*, 2013). The tool utilized continuous bag-of-words and skip-gram models to compute the representation of the words into a vector form (Angelov, 2020; *word2vec*, 2013). Similarly, *doc2vec* is a tool that learns "jointly embedded document and word vectors (Angelov, 2020)." Paragraph Vector with Distributed Memory (DM) and Distributed Bag of Words are the two versions of the *doc2vec* model (Angelov, 2020). In 2020, Angelov introduced *top2vec* to leverage joint document and word semantic embedding to find topic vectors that do not require the researcher to enter stop words, stemming, or lemmatization. In this article, we chose to explore this recent method since the creator emphasized that this method would be more informative and representative of the corpus than the other topic modeling techniques (Angelov, 2020). Moreover, given these highlighted advantages of Top2Vec, we decided to compare Top2Vec with LSA and LDA–the two most established and frequently-used conventional methods. As can also be observed in the literature

review timeline, LSA was one of the earliest topic modeling methods. LDA was developed later and became the most commonly used topic modeling method. In Table 1 we provide a comparison of Top2Vec, LDA and LSA.

Table 1. Comparison of Top2Vec, LSA and LDA (Albalawi, et al., 2020; Egger & Yu 2022)

	Advantages	Disadvantages
Top2Vec	• Support multilingual analysis. • The number of optimal topics is not defined by a user. • Support very large dataset. • Preprocessing is not needed since it uses word embedding.	• It may result too many topics. • Does not work well with small dataset. • Results in outliers. • Evolution metric is missing. • It does not capture the relationship between several topics.
LSA	• Can catch words synonyms. • Deals well with data sparsity. • A solid understanding of probability theory and statistics are not necessary.	• Topic labeling is not easy • depends on human judgment for labeling topic.
LDA	• Can deal with small data. • Generates smaller number of topics comparing with word-embedding-based approaches. • Easy to interpret. • Domain knowledge is not extremely important. • Can catch noun and adjective within topics.	• Requires many experiments to fine-tune parameters. • May result in overlapped topics. • Depends on the frequency of common words and assumes topic independence. • May result in incoherent topics. • The number of topics is a user-defined parameter.

3. Methods

Lin et al. found that utilizing machine learning for literature analysis can be helpful in summarizing key research themes and trends (Lin et al., 2020). In addition, topic modeling can be a very useful tool in analyzing and classifying large amounts of textual data, especially for critical and timely issues. That is why there have been several attempts to utilize topic modeling on COVID-19 literature and we wanted to compare three different topic modeling methods using COVID-19 textual data.

3.1. Data

The textual data for this study includes a corpus of 65,292 abstracts (as of November 2021), which was curated from the reference of studies listed in NIH's COVID-19 Portfolio by Zengul et al. (2021) and made available by the authors for future research. NIH generated this portfolio as a repository of archived, preprint, and peer-reviewed publications to enhance COVID-19-related research activities (NIH-COVID-19-Portfolio, 2020).

3.2. Topic Modeling Methods: LSA, LDA and Top2Vec and their applications

Latent Semantic Analysis (LSA) extracts word usage patterns into a document-term matrix (DTM) and then reduces the dimensionality by applying a singular value decomposition (SVD) (Debortoli et al., 2016). Debortoli et al. cite Landauer by stating that "the resulting latent semantic factors, which share many similarities with the outputs of factor analysis or principal components analysis, are often interpreted as topics" (Debortoli et al., 2016). In our study, we utilized a personal laptop and JMP 16 pro for LSA analyses. For the details of the application of LSA on corpus of 65,292 abstracts, we refer the readers to Zengul et al. (2021) study.

Latent Dirichlet Allocation (LDA) is based on the Bayesian statistical topic modeling method (Alghamdi & Alfalqi, 2015). It extends LSA to provide clarity for interpretability issues around the computed factor loadings. The associations between documents and topics as well as between topics and words are represented as probability distributions that can be

used for further statistical analyses (Debortoli et al., 2016; Kao & Poteet, 2007). We performed LDA analysis in Python. The program was run on Kaggle using GPU Accelerator. It took almost 5 hours for the whole process to run, including 3 hours of preprocessing the data and 2 hours of model training and developing the results. Python packages used: NumPy: used for scientific computing in Python, Pandas: for easy-to-use data structure and data analysis tools, NLTK: used for building python programs, Gensim: used for the implementation of LDA Mallet, Spacy: en model for text preprocessing, and matplotlib package: used for building word clouds and charts.

LDA is a form of unsupervised learning that views documents as bags of words. LDA works by first making a key assumption: the way a document was generated was by picking a set of topics and picking a set of words for each topic. Once we provide the algorithm with the number of topics, it rearranges the topics distribution within the documents and keywords distribution within the topics to obtain a good composition of the topic-keywords distribution.

The data preprocessing involved a standard workflow of text analysis. The first step in data preprocessing was to tokenize the free text abstract, transforming the data to one line per word ("token"). We need to break down each sentence into a list of words through tokenization, removing punctuations and unnecessary characters altogether. We used Gensim's simple preprocess by setting the set 'deacc=True' to remove the punctuations. After tokenization, stop words were removed. Standard English "stop words" from NLTK and Spacy's en model, including some extended stop words, were used to eliminate tokens that do not represent any significant aspects of language parts. We used extended stop words consisting of 1149 unique words derived from the dataset, which are deemed undesirable for meaningful analysis. Next, we created bigrams and five grams models using Gensim's Phrases model to build and implement the bigrams and five grams. Subsequently, words were lemmatized — words in the third person are changed to the first person, and verbs in the past and future tenses are changed into the present. We also created a unique id for each word in the document using Gensim as a dictionary (id2word) and the corpus.

We used LDA Mallet's version of the LDA algorithm as it often gives a better quality set of topics. Gensim provides a wrapper to implement Mallet's LDA from within Gensim itself. The input of the LDAMallet algorithm is a set of documents, and the output of the LDA algorithm is a set of topics. Dictionary (id2word) and the corpus is the two main inputs to the LDA topic model. In addition to the corpus and dictionary, other LDA requirements are the path to LDAMallet and providing the number of topics. The number of topic value was provided as 11.

Topic coherence scores were used to evaluate the performance of the model. The coherence score is used to decide the optimum number of topics to be extracted using LDA. It is used to measure how well the topics are extracted. The coherence score of 0.50106 was obtained.

$$\text{Coherence} = \sum_{i<j} \text{score}(w_i, w_j). \quad (1)$$

where wi, wj are the top words of the topic.

Lastly, Top2Vec is a form of topic modeling that finds topic vectors (Angelov, 2020). According to Angelov, Top2Vec "does not require stop-word lists, stemming or lemmatization, and it automatically finds the number of topics" (Angelov, 2020). The other topic modeling methods require the researcher to know the number of topics there should be for the corpus being studied. Angelov asserts that this model is more informative compared to the standard probabilistic generative topic model methodologies (Angelov, 2020). One of the major benefits for using top2vec is that we did not have to do as much preliminary work, such as stemming and creating the stop word list (Angelov, 2020).

We used Python language for analyses. However, we were not able to use an average laptop or a desktop to run the analyses since it was taking too much time for a large corpus like ours. Therefore we used a university-provided research computing center and a super computer to run our analyses, which still took about an hour or two.. In its essence, Top2Vec is a class that calls other classes. The provided parameters need to be examined and changed depending on the project at hand. To achieve the desired results, one needs to consider the other classes called by Top2Vec and their parameters.

To achieve the provided Top2Vec results, we considered Uniform Manifold Approximation and Projection (UMAP) class which provides dimension reduction, and the Hierarchical Density-Based Spatial Clustering of Applications with Noise (HDBSCAN) class which is used for finding dense clusters of documents. All the parameters in these two classes are optimized for topic modeling except n_neigbhors for UMAP and min_cluster_size for HDBSCAN. These need to be considered according to the data at hand and the goal. In our study, we found that 35 gives the best outcome for our purpose. There is also min_count metrics provided by the Top2Vec model that is particularly important since any frequency of words less than the supplied is not considered by the

Word2Vec class. The default value is 50 for this parameter which a suitable number, especially if you have a generous size of data like ours. However, since we changed the UMAP's n_neigbhors and the HDBSCAN's min_cluster_size parameters to 35 causing bigger clusters of documents, we found that 50 is not a suitable size for vocabulary, and we had to bring it down to 30. Changing these three parameters allowed us to approach the optimized outcome that would allow comparison. Another important parameter we needed to consider was Vector Size in Doc2Vec class. The default size (300) was the right size for our NIH_Covid19 data.

We used two methods for the evaluation of the results. Cosine similarity and the human interpretation of the topic words. To allow our results to be comparable with previously published LSA study (Zengul et al., 2021) we decided to limit the topic number to 11 in both LDA and Top2Vec. In normal circumstances Top2Vec finds optimum numbers of topics automatically. When we introduced this topic number of limitation of 11 topics, we realized that cosine similarity is changing with the topic reduction on the data. This caused us to find an n_neigbhors and min_cluster_size parameters that are optimum for the original number of topics as well as when the number of topics is reduced. The main criteria for the human interpretation were to make sure there were no stop words in the topic words (i.e., the, a, an, some, etc.) and all the words in a topic were making sense toward a specific topic.

We run LSA, LDA, and Top2Vec separately to classify each the 65,292 abstracts into one of the eleven topics. Then we combined our results from LSA, LDA, and Top2Vec to run further statistical analyses. We generated binary values for total of 33 topics from LSA, LDA and Top2Vec. Then, we utilized Spearman correlation to examine the classification similarity of abstracts among the three topic modeling methods. We also generated word clouds and top 10, 20, and 30 terms from each topic modeling method to visually examine the similarity of topics, especially for those similarities that achieved more than 40% correlation.

4. Results

Table 2 displays the results for comparison of LDA and Top2Vec topic modeling methods by displaying both the top 10 terms for each method and the Spearman correlation (ρ) results between LSA and Top2Vec topics that achieved 0.4 or more correlation. As one can observe from the table, the highest correlation (0.8) was between LDA-T6 and Top2Vec-T10. Examining the common top 10 terms from these two topics indicates that there was a steam of COVID-19 literature focusing on the implications of the pandemic on mental health. There was a total of eight comparisons that achieved more than 0.4 or more correlation.

Table 2. The Comparison of Top 10 Terms and Overall Spearman Correlation Results between LDA and Top2Vec (n=65,262)

Topic Comparison	**Top 10 terms**	**ρ <.001**
LDA-T6 **vs** Top2Vec-T10	health, pandemic, covid, participant, mental, impact, people, survey, anxiety, measure	0.80
	perceived, anxiety, psychological, coping, mental, depression, feelings, emotional, behaviors, loneliness	
LDA-T9 **vs** Top2Vec-T3	virus, drug, viral, human, vaccine, protein, target, potential, cell, host	0.77
	affinity, mutagenesis, mpro, residues, residue, dimeric, atom, binding, intermolecular, spike	
LDA-T7 **vs** Top2Vec-T11	test, covid, sample, testing, infection, negative, case, diagnosis, detect, symptom	0.67
	assays, assay, samples, qpcr, serological, lod, elisas, kit, immunoassays, dilutions	
LDA-T2 **vs** Top2Vec-T2	case, country, spread, measure, transmission, rate, epidemic, population, control, estimate	0.64
	reproduction, seir, mathematical, stochastic, sird, compartmental, relaxing, deterministic, exponential, forecast	
LDA-T1 **vs** Top2Vec-T9	care, risk, pandemic, hospital, medical, covid, staff, patient, procedure, management	0.52
	elective, surgeries, surgery, surgical, operative, oncologic, postponed, oncological, postoperative, preoperative	
LDA-T4 **vs** Top2Vec-T4	severe, covid, respiratory, cell, acute, disease, increase, infection, level, response	0.50
	dysregulated, cytokines, proinflammatory, downregulation, macrophage, macrophages, modulating, transcriptional, signaling, cytokine	
LDA-T10 **vs** Top2Vec-T1	health, pandemic, public, care, system, community, challenge, response, crisis, covid	0.43
	curriculum, aorn, chiropractic, educators, curricula, struggles, tutorials, unprecedented, tasked,	
LDA-T8 **vs** Top2Vec-T5	patient, clinical, treatment, covid, hospital, severe, therapy, trial, admission, outcome	0.41

	randomized, rcts, mechanical, tocilizumab, auxora, imv, azithromycin, torsades, milliseconds, ventilation	

Table 3 displays the top 10 terms and Spearman correlation results between LDA and LSA methods. Again the mental health issues achieved the highest level of correlation (0.74) between LDA and LSA. There were a total of five comparisons that achieved 0.4 or more correlation.

Table 3. The Comparison of Top 10 Terms and Overall Spearman Correlation Results between LDA and LSA (n=65,262)

Topic Comparison	Top 10 terms	ρ <.001
LDA-T6 **vs** LSA-T9	health, pandemic, covid, participant, mental, impact, people, survey, anxiety, measure	0.74
	anxieti·, particip·, stress·, mental· health·, survey·, psycholog·, score·, respond·, questionnair·	
LDA-T7 **vs** LSA-T10	test, covid, sample, testing, infection, negative, case, diagnosis, detect, symptom	0.52
	test·, sars-cov-2·, assay·, sampl·, detect·, antibodi·, igg·, sensit·, sequenc·, specimen·	
LDA-T1 **vs** LSA-T8	care, risk, pandemic, hospital, medical, covid, staff, patient, procedure, management	0.45
	aerosol·, mask·, ppe·, procedur·, droplet·, patient·, air·, ventil·, surgeri·, surgic·	
LDA-T10 **vs** LSA-T1	health, pandemic, public, care, system, community, challenge, response, crisis, covid	0.44
	care·, health·, countri·, pandem·, peopl·, communiti·, patient·, servic·, public·, social·	
LDA-T4 **vs** LSA-T5	severe, covid, respiratory, cell, acute, disease, increase, infection, level, response	0.42
	cell·, ace2·, express·, protein·, sars-cov-2·, activ·, receptor·, immun·, gene·, lung·	

Table 4 exhibits the top ten terms and Spearman correlation between Top2Vec and LSA topics. Again, the highest correlation of 0.76 between Top2Vec and LSA was achieved on the mental health topic. Only four comparisons of Top2Vec and LSA achieved 0.4 or more correlation.

Table 4. The Comparison of Top 10 Terms and Overall Spearman Correlation Results between Top2Vec and LSA (n=65,262)

Topic Comparison	Top 10 terms	ρ <.001
Top2Vec-T10 **vs** LSA-T9	perceived, anxiety, psychological, coping, mental, depression, feelings, emotional, behaviors, loneliness	0.76
	anxieti·, particip·, stress·, mental· health·, survey·, psycholog·, score·, respond·, questionnair·	
Top2Vec-T4 **vs** LSA-T5	dysregulated, cytokines, proinflammatory, downregulation, macrophage, macrophages, modulating, transcriptional, signaling, cytokine	0.52
	cell·, ace2·, express·, protein·, sars-cov-2·, activ·, receptor·, immun·, gene·, lung·	
Top2Vec-T11 **vs** LSA-T10	assays, assay, samples, qpcr, serological, lod, elisas, kit, immunoassays, dilutions	0.45
	test·, sars-cov-2·, assay·, sampl·, detect·, antibodi·, igg·, sensit·, sequenc·, specimen·	
Top2Vec-T6 **vs** LSA-T4	prothrombotic, infarcts, thrombotic, hypercoagulable, dysautonomia, infarct, thrombosis, neurologic, postinfectious, coagulopathy	0.44
	patient·, imag·, stroke·, lung·, acut·, pulmonary, case·, arteri·, lesion·, cardiac	
Top2Vec-T8 **Vs** LSA-T2	ggo, intralobular, hypersensitive, subpleural, opacity, lobes, hbdh, bronchogram, consolidation, nomogram	0.42
	patient·, group·, hospit·, sever·, age·, mortal·, associ·, admiss·, covid-19· patient·	

5. Discussion

This study was prepared as a practical guide for researchers who are interested in using topic modeling methodologies and have difficulty determining which methodology to use. We compared two of the most conventional and frequently used methodologies, LSA and LDA, with a recently developed one, Top2Vec, by using a COVID-19-focused corpus of 65,292 abstracts. Our results have several practical implications.

First, while classifying documents into topics, all three topic modeling methods exhibited varying degrees of similarities. The most similar topic

modeling results were achieved between LDA and Top2Vec. This suggests that to a certain extent these two methods can be used interchangeably. When choosing one of these two methods, a researcher may need to examine other factors such as the ease of use, the needed computational resources, and the availability of analytical tools. In our case, both LDA and Top2Vec required us to use research computing since a traditional desktop or laptop was not sufficient. Even though LDA is more widely available in several statistical tools, it is generally difficult to run a large corpus, like in our case of 65,292 abstracts within a statistical tool. Therefore, we recommend Python and research computing resources for LDA and Top2Vec projects.

Our second finding and recommendation pertains to the comparison between LSA and LDA. The second most similar results were achieved between LSA and LDA, with five comparisons achieving 0.4 or more correlation. Even though LDA is the most commonly used algorithm, we found LSA can be attractive for some researchers for several reasons. First, both algorithms require time-consuming investment into data cleanup and NLP processes such as generation of the bag of words, term list curation, stop word list, and lemmatization. Therefore they both require very similar time for data cleanup processes. However, LSA does not require high computational resources compared to LDA. LSA can be run on an average desktop or laptop. Running the LSA algorithm on a statistical tool such as JMP takes merely minutes for a corpus of 65K abstracts, whereas the same job requires research computing resources for LDA. When more time is invested in data preprocessing and cleanup, both LSA and LDA have a great potential to generate clear topics. Therefore, we recommend researchers evaluate the computational resources while deciding between LSA and LDA.

Our third recommendation focuses on the third finding, the comparison between Top2Vec and LSA, which comes third regarding the overall level of correlations achieved. However, the highest level of correlation between Top2Vec and LSA (0.76) was slightly higher than the correlation between LDA and LSA (0.74). Again the medium to the high level of correlation that was achieved in four out of eleven topics suggests that depending on the project at hand, researchers may opt for using either of these two algorithms. However, researchers should pay attention to the less preprocessing requirements of Top2Vec compared to LSA. If a researcher is familiar with Python and has access to computational resources, Top2Vec would be a great choice. On the other hand, LSA can be opted for by some researchers due to its availability in major statistical tools like JMP and less dependence on computational resources.

6. Conclusion

Our study makes a valuable contribution to the topic modeling literature by providing a practical guide for researchers in various domains who are interested in utilizing topic modeling methods. We compared two conventional methods−LSA and LDA−with Top2Vec, a recently developed method. Our study findings indicate medium to high levels of similarities among these topic modeling methods while classifying documents into respective topics. Even though we picked two of the most frequently used and best-performing methodologies and compared them with a recently developed one, more than a dozen other topic modeling methods are available. Therefore, future studies may expand this current study by including more topic modeling methods. Another limitation of this study arises from the inherent differences between topic modeling methods. In LSA and LDA, researchers enter the number of topics to explore before running the algorithm and iteratively determine optimum numbers of topics by examining results. However, Top2Vec does not have such a requirement, and it automatically determines the number of topics. On the other hand, to be able to compare our findings with LSA and LDA, we had to limit the number of topics in Top2Vec to eleven. When we did not introduce this limitation, the Top2Vec generated hundreds of topics for the 65,392 abstracts of COVID-19 research articles. Obviously navigating such high numbers of topics can be difficult, and unnecessary, especially if the research was aiming to identify major themes. However, if the goal of the research is to generate both major and minor themes, Top2Vec can be considered a superior algorithm due to its less dependency on human input and preprocessing efforts.

7. References

Albalawi, R., Yeap, T. H., & Benyoucef, M. (2020). Using topic modeling methods for short-text data: A comparative analysis. *Frontiers in Artificial Intelligence, 3*, 42.

Älgå, A., Eriksson, O., & Nordberg, M. (2020). Analysis of Scientific Publications During the Early Phase of the COVID-19 Pandemic: Topic Modeling Study [Original Paper]. *J Med Internet Res, 22*(11), e21559. https://doi.org/10.2196/21559

Alghamdi, R., & Alfalqi, K. (2015, January 6, 2021). A survey of topic modeling in text mining. *Int. J. Adv. Comput. Sci. Appl.(IJACSA), 6*(1), 147-153.

Anaya, L. H. (2011). *Comparing Latent Dirichlet Allocation and Latent Semantic Analysis as Classifiers*. ERIC.

Angelov, D. (2020, January 7,). Top2Vec: Distributed Representations of Topics. *arXiv preprint arXiv:2008.09470*.

Balakrishnan, V., & Lloyd-Yemoh, E. (2014). Stemming and lemmatization: a comparison of retrieval performances.

Bergamaschi, S., & Po, L. (2014). Comparing LDA and LSA topic models for content-based movie recommendation systems. International conference on web information systems and technologies.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2002). Latent dirichlet allocation. Advances in neural information processing systems.

Blei, D. M. (2012). Probabilistic topic models. Communications of the ACM, 55(4), 77-84.

Chai, Y., & Li, W. (2019). Towards deep learning interpretability: A topic modeling approach.

Chauhan, U., & Shah, A. (2021). Topic modeling using latent Dirichlet allocation: A survey. *ACM Computing Surveys (CSUR), 54*(7), 1-35.

Chen, Q., Allot, A., & Lu, Z. (2020a). Keep up with the latest coronavirus research. *Natur, 579*(7798), 193-193.

Chen, Q., Allot, A., & Lu, Z. (2020b). LitCovid: an open database of COVID-19 literature. *Nucleic Acids Research*.

Chen, X., Zou, D., & Su, F. (2021). Twenty-five years of computer-assisted language learning: A topic modeling analysis. *Language Learning & Technology, 25*(3), 151-185.

Cvitanic, T., Lee, B., Song, H. I., Fu, K., & Rosen, D. (2016). LDA v. LSA: A comparison of two computational text analysis tools for the functional categorization of patents. International Conference on Case-Based Reasoning.

Debortoli, S., Müller, O., Junglas, I., & vom Brocke, J. (2016). Text mining for information systems researchers: An annotated topic modeling tutorial. *Communications of the Association for Information Systems, 39*(1), 7.

Deerwester, S., Dumais, S., Landauer, T., Furnas, G., & Harshman, R. (1990). Indexing by Latent Semantic Analysis. *J. Am. Soc. Inf. Sci., 41*, 391-407.

Doogan, C., Buntine, W., Linger, H., & Brunt, S. (2020). Public perceptions and attitudes toward COVID-19 nonpharmaceutical interventions across six countries: A topic modeling analysis of Twitter data. *Journal of Medical Internet Research, 22*(9), e21419.

Egger, R., & Yu, J. (2022). A Topic Modeling Comparison Between LDA, NMF, Top2Vec, and BERTopic to Demystify Twitter Posts. *Frontiers in sociology, 7*, 886498-886498. https://doi.org/10.3389/fsoc.2022.886498

Goldberg, Y., & Levy, O. (2014). word2vec Explained: deriving Mikolov et al.'s negative-sampling word-embedding method. *arXiv preprint arXiv:1402.3722*.

Gutierrez, B. J., Zeng, J., Zhang, D., Zhang, P., & Su, Y. (2020). Document classification for covid-19 literature. *arXiv preprint arXiv:2006.13816*.

Hofmann, T. (1999). *Probabilistic latent semantic analysis* Proceedings of the Fifteenth conference on Uncertainty in artificial intelligence, Stockholm, Sweden.

Jurafsky, D., & Martin, J. H. (2022). Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition. In (Third Draft ed.)

Kalepalli, Y., Tasneem, S., Teja, P. D. P., & Manne, S. (2020). Effective comparison of lda with lsa for topic modelling. 2020 4th International Conference on Intelligent Computing and Control Systems (ICICCS),

Kao, A., & Poteet, S. R. (2007). *Natural language processing and text mining*. Springer Science & Business Media.

Kherwa, P., & Bansal, P. (2020). Topic modeling: a comprehensive review. *EAI Endorsed transactions on scalable information systems, 7*(24).

Lin, H.-J., Sheu, P. C. Y., Tsai, J. J. P., Wang, C. C. N., & Chou, C.-Y. (2020). Text mining in a literature review of urothelial cancer using topic model. *BMC Cancer, 20*(1), 1-7. https://doi.org/10.1186/s12885-020-06931-0

Liu, L., Tang, L., Dong, W., Yao, S., & Zhou, W. (2016). An overview of topic modeling and its current

applications in bioinformatics. *Springerplus, 5*(1), 1608. https://doi.org/10.1186/s40064-016-3252-8

Manning, C., Raghavan, P., & Schütze, H. (2020). Dropping common terms: stop words. In *Introduction to Information Retrieval* (Vol. 2020, pp. 27). Cambridge University Press. https://nlp.stanford.edu/IR-book/html/htmledition/dropping-common-terms-stop-words-1.html

Miner, G., Elder, J., Fast, A., Hill, T., & Delen, D. (2012). *Practical Text Mining and Statistical Analysis for Non-structured Text Data Applications*. Academic Press. https://books.google.com/books?id=-B6amxqygTMC

Mohammed, S. H., & Al-augby, S. (2020). Lsa & lda topic modeling classification: Comparison study on e-books. *Indonesian Journal of Electrical Engineering and Computer Science, 19*(1), 353-362.

NIH-COVID-19-Portfolio. (2020). *NIH COVID-19 Portfolio*. NIH. Retrieved July 25 from https://icite.od.nih.gov/covid19/search/

Obadimu, A., Mead, E., & Agarwal, N. (2019). Identifying latent toxic features on YouTube using non-negative matrix factorization. The Ninth International Conference on Social Media Technologies, Communication, and Informatics, IEEE,

Ordun, C., Purushotham, S., & Raff, E. (2020). Exploratory analysis of covid-19 tweets using topic modeling, umap, and digraphs. *arXiv preprint arXiv:2005.03082*.

Rish, I. (2001). An empirical study of the naive Bayes classifier. IJCAI 2001 workshop on empirical methods in artificial intelligence,

Sánchez-Franco, M. J., & Rey-Moreno, M. (2022). Do travelers' reviews depend on the destination? An analysis in coastal and urban peer-to-peer lodgings. *Psychology & Marketing, 39*(2), 441-459.

Sha, H., Hasan, M. A., Mohler, G., & Brantingham, P. J. (2020). Dynamic topic modeling of the COVID-19 Twitter narrative among US governors and cabinet executives. *arXiv preprint arXiv:2004.11692*.

Song, F., & Croft, W. B. (1999). A general language model for information retrieval. Proceedings of the eighth international conference on information and knowledge management,

Vayansky, I., & Kumar, S. A. (2020). A review of topic modeling methods. *Information Systems, 94*, 101582.

word2vec. (2013, July 29, 2013). Google Code Archive. Retrieved January 24, from https://code.google.com/archive/p/word2vec/

Zengul, F. D., Zengul, A. G., Mugavero, M. J., Oner, N., Ozaydin, B., Delen, D., Willig, J. H., Kennedy, K. C., & Cimino, J. (2021, 2021/01/01/). A critical analysis of COVID-19 research literature: Text mining approach. *Intelligence-Based Medicine, 5*, 100036. https://doi.org/10.1016/j.ibmed.2021.100036